Pontifícia Universidade Católica
do Rio de Janeiro

Ariane Moraes Bueno Rodrigues

Uncovering factors that influence how data visualizations are interpreted by non-experts

Tese de Doutorado

Thesis presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Ciências – Informática.

Advisor: Profa. Simone Diniz Junqueira Barbosa

Rio de Janeiro
March 2022

Pontifícia Universidade Católica do Rio de Janeiro

Ariane Moraes Bueno Rodrigues

Uncovering factors that influence how data visualizations are interpreted by non-experts

Thesis presented to the Programa de Pós–graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Ciências – Informática. Approved by the Examination Committee:

Profa. Simone Diniz Junqueira Barbosa
Advisor
Departamento de Informática – PUC-Rio

Profa. Milene Selbach Silveira
PUCRS

Prof. Marcos de Oliveira Lage Ferreira
UFF

Prof. Bruno Feijó
Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio

Prof. Hélio Côrtes Vieira Lopes
Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio

Rio de Janeiro, March 22nd, 2022

Ariane Moraes Bueno Rodrigues

graduated in Computer Engineering from the Pontifical Catholic University of Rio de Janeiro. She has a master's degree in the Department of Informatics at the same University, specializing in Human-Computer Interaction. She worked as an HCI Consultant at Petrobras, responsible for the usability and evaluation of systems developed in the E&P management.

Bibliographic Data

Moraes Bueno Rodrigues, Ariane

Uncovering factors that influence how data visualizations are interpreted by non-experts / Ariane Moraes Bueno Rodrigues; advisor: Simone Diniz Junqueira Barbosa. – 2022.

199 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2022.

Inclui bibliografia

1. keywordpre – Teses. 2. keywordpre – Teses. 3. Alfabetização visual. 4. Visualização de dados. 5. Análise exploratória de dados. I. Diniz Junqueira Barbosa, Simone. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

To my dear sister Aline (in memoriam),
who remains my greatest inspiration in life.
To God, who strengthens me every day
and gives me much more than I deserve.
To my beloved husband, precious children,
and family for supporting and understanding
the countless times I couldn't be with them.

Acknowledgments

My sincere thanks to:

My advisor, Simone D.J. Barbosa, for offering guidance and encouragement, always blending experience and humor. I'm proud of and grateful for my time working with her three times in my academic life.

Prof. Hélio V. Lopes, who welcomed me from the first term, sharing his teachings, projects, and motivational words.

My co-authors, who actively collaborated on the papers to be concluded and accepted.

All my colleagues and other anonymous, who participated in the empirical studies, donating their time and experiences and, in a way, are also part of this work.

To my friends from the Ideias and DasLab laboratories, who made difficult times much lighter.

To my parents, for their great encouragement to study, for having financed in my education, opening doors to everything I have achieved so far.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

And also to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), for partially financing this research under grant #146438/2018-3.

Abstract

Moraes Bueno Rodrigues, Ariane; Diniz Junqueira Barbosa, Simone (Advisor). **Uncovering factors that influence how data visualizations are interpreted by non-experts**. Rio de Janeiro, 2022. 199p. Tese de Doutorado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Data visualizations are increasingly common in traditional media and social networks. However, the visualization literacy of the population did not follow this growing popularity. It is necessary for those who create the charts to assemble a visual communication that contains the necessary information in an attractive and easy-to-understand way. By contrast, it is necessary for those who consume them to capture information represented by the charts and extract the analyses of what they see. The importance of visual literacy is the ability to "read" a chart, *i.e.*, look at a chart and identify relevant information, trends, and outliers in a given scenario. In this work, we conducted four studies to explore factors related to the success of visual data analysis. We identified issues ranging from data distribution to formulating good questions to enrich exploration. The first study discovered how people try to make sense of specific data visualizations through questions they ask when they first encounter a visualization. In the second study, we explored how data distributions can affect the effectiveness and efficiency of data visualizations. In the third study, we investigated when non-experts identify that particular visualization is not adequate to answer a specific analysis question, when they make good suggestions for changes to make these visualizations adequate, and when they evaluated well the adequacy of some suggestions offered to them. In the fourth study, we created a test to assess people's understanding of the applied (answering analysis questions supported by a visualization) and conceptual (questions about function and structure) aspects of data visualization. Our results provide resources for developing of educational material and tools for recommending data visualizations to answer specific data-relation questions. An additional contribution of this work to the results of the studies was the structuring of a unified list of different visualization tasks that we found in the literature.

Keywords

Visualization literacy; Data Visualization; Exploratory data analysis.

Resumo

Moraes Bueno Rodrigues, Ariane; Diniz Junqueira Barbosa, Simone. **Explorando fatores que influenciam como as visualizações de dados são interpretadas por não especialistas**. Rio de Janeiro, 2022. 199p. Tese de Doutorado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

As visualizações de dados são cada vez mais comuns na mídia tradicional e nas redes sociais. No entanto, a alfabetização visual da população não acompanhou essa crescente popularidade. É necessário para quem cria os gráficos montar uma comunicação visual que contenha as informações necessárias de forma atrativa e de fácil compreensão. Em contrapartida, é necessário para quem os consome, captar as informações representadas pelos gráficos e extrair as análises do que vê. A importância da alfabetização visual é a capacidade de "ler" um gráfico, ou seja, olhar para um gráfico e identificar informações relevantes, tendências e discrepâncias em um determinado cenário. Neste trabalho, realizamos quatro estudos para explorar os fatores que influenciam o sucesso da análise de dados visuais. No primeiro estudo descobrimos como as pessoas tentam dar sentido a visualizações de dados específicas, através de perguntas que elas fazem ao encontrar uma visualização pela primeira vez. No segundo estudo exploramos como as distribuições de dados podem afetar a eficácia e eficiência das visualizações de dados. No terceiro estudo investigamos quando não especialistas identificam que uma visualização não é adequada para responder uma pergunta de análise específica, quando eles fazem boas sugestões de alteração para tornar essas visualizações adequadas e quando avaliam bem a adequação de algumas sugestões oferecidas a eles. No quarto estudo, criamos um teste para avaliar a compreensão das pessoas sobre os aspectos aplicados (responder perguntas de análise com o apoio de uma visualização) e conceituais (questões sobre a função e estrutura) da visualização de dados. Nossos resultados fornecem recursos para o desenvolvimento de material didático e ferramentas para recomendação de visualizações de dados relacionadas a perguntas que se visa responder. Uma contribuição adicional deste trabalho aos resultados dos estudos foi a estruturação de uma lista unificada de diferentes tarefas de visualização que encontramos na literatura.

Palavras-chave

Alfabetização visual; Visualização de dados; Análise exploratória de dados.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

EngSem – Engenharia Semiótica

IA – Inteligência Artificial

IHC – Interação Humano-Computador

InfoVis – *Information Visualization* (Visualização de Informação)

NLP – *Natural Language Processing* (Processamento de Linguagem Natural)

1
Introduction

In an increasingly data-driven world, we use visualizations to understand our personal lives, work activities, and society. We are already familiar with using matrices to represent appointments on a calendar, maps for city subway location, or gauges to measure car speed. These are visualization examples that can communicate different information (Bertin, 1983): they transform raw data into visual information that can reveal to the reader relationships and other pieces of information that could not be clear if they looked at several numbers and attributes. This communication also happens in a visual analytics environment when dealing with complex data.

In this scenario, visualizations can either amplify or hamper human cognitive capabilities. Some problems occur because the analyst does not know what they are seeking (Cox et al., 2001) or they have a lack of visual analytics expertise. They may also be due to the complexity of the visualization itself (Huang et al., 2009) or because the visualization does not reflect the raw data's message in its totality and may inform less than the analyst wants to know (Dasgupta and Kosara, 2010).

Data Visualization Literacy involves recognizing a given chart, reading it correctly, and extracting information from it (Lee et al., 2017). Familiarity with a specific type of visualization does not imply that the person can read or interpret it correctly (Boy et al., 2014). Data visualization researchers have attempted to explore and promote solutions to support data visualization comprehension activities. Studies address developing and applying tests to assess literacy (Lee et al., 2017; Börner et al., 2019), understanding how analysts interpret visualizations (Lee et al., 2016; Boy et al., 2014), how unfamiliar visualizations are taught (Alper et al., 2017; Bishop et al., 2020), and even identifying the cognitive activities involved in creating visualizations (Grammel et al., 2010; Huron et al., 2014). However, we have not found more comprehensive studies, which relate these different aspects, which we believe is important for teaching and learning about data visualizations and data analysis tasks.

This work advances Data Visualization education research by introducing a set of studies that identify some understanding gaps that can influence how non-experts interpret data visualizations.

1.1 Motivation

Any person who reads an image may misinterpret it (Tufte and Graves-Morris, 1986). The interpretation of visualization may depend on the observer's familiarity and previous experience with it (Bresciani, 2009). Our greatest motivation with this work is that, as people make mistakes when interpreting a visualization, our understanding of those mistakes may contribute to data visualization education and help increase data visualization literacy.

1.2 Problem definition

As the availability of data reaches unprecedented volumes, attention has shifted from data *acquisition* (when there were poor datasets) to data *analysis* (what to do with the recently available rich datasets) (Key et al., 2012). Human attention and capacity to process that data are now the limited resources. To extract information and gain insights from those data, several tools to support data analysis and visualization techniques have been developed (Keim et al., 2008). Analysts may need to create many visualizations to make sense of these growing data and communicate the insights obtained from them. A direct consequence of this is the proliferation of misunderstandings: the reader cannot interpret the visualization and the creator does not know the best way to communicate insights visually.

There are several studies that define and/or apply visualization literacy assessments (*e.g.*, Lee et al., 2017). They differ in terms of:

- the visualization activity:
  - interpreting or reading: (*e.g.*, Börner et al., 2016) or
  - creating: (*e.g.*, Lee et al., 2016)),
- the types of charts analyzed: (*e.g.*, Boy et al., 2014); and
- the analyst's expertise level:
  - children: (*e.g.*, Alper et al., 2017),
  - students: (*e.g.*, Chevalier et al., 2018),
  - novice/experienced: (*e.g.*, Maltese et al., 2015), and
  - teaching or learning.

In this work, we focus on investigating how data visualizations are interpreted by non-experts, as a first step towards devising approaches to increase data visualization literacy. Our primary research question is the following:

RQ: What factors play a role in how novices interpret data visualizations?

We can unfold this research question in the following subquestions:

SQ1: What are the common novices' misinterpretations when trying to make sense of data visualizations?

SQ2: Does the data distribution have a role in the interpretation of data visualizations?

SQ3: How suitable do non-experts find certain data visualizations for a given analysis question?

SQ4: How can we assess a particular individual's data visualization literacy in detail, so as understand how to improve it?

1.3 Methodology and Contributions

To answer the research questions in this work, we adopted the following procedure. First, we reviewed the literature on data visualization literacy (section 2.4). Then, we conducted a series of empirical studies, to address each research subquestion.

We carried out the studies with the approval of the Pontifical Catholic University of Rio de Janeiro's Institutional Review Board (Câmara de Ética em Pesquisa da PUC-Rio), PUC-Rio 063/2020 - Protocol 97/2020. We conducted them through an online questionnaire. The first two had time constraints because most of the participants were students of data visualization classes who volunteered to take part in the studies, and we would need to take advantage of the available window of the course. After that, there could be a possible loss of interest in participating in the research. The last two were carried out during the pandemic, when much of the population was in confinement. In this way, we gathered a sufficient number of participants without compromising the deadline for completing this work. To proceed with the studies, each participant received an informed consent form, where we explained the purpose of the study, the risks and benefits of participation, confidentiality and ethical issues, and information for contacting us (appendix A shows an example).

To address SQ1, we asked participants to pose questions about a set of data visualizations and analyzed both their misunderstandings and the question patterns that emerged for each visualization type (chapter 3). The compilation of misunderstandings can help us map specific limitations of data visualization education and the question patterns may be used as a resource to provide more refined recommendations for creating visualizations to answer certain analysis questions.

To address SQ2, we investigated how visualization efficiency and effectiveness vary according to the data distribution (chapter 4). Typical data visualization guidelines and recommendations give little consideration to how diverse data distributions should be handled when creating data visualizations. Our study sheds some light on this issue, reveals limitations of general recommendations, and points to the need for more specific recommendations that take into account the data distributions.

To address SQ3, we studied how participants assess the suitability of certain data visualizations for answering specific analysis questions, before and after being exposed to related guidelines (chapter 5). Our results provide some evidence that guidelines may not help novices to effectively relate analysis questions to specific chart properties.

Finally, to address SQ4, based on existing visualization literacy tests, we devised a test for assessing people's understanding of both applied and conceptual aspects of data visualization (chapter 6), which provides a fine-grained evaluation of knowledge gaps, which in turn may point to personalized educational opportunities.

When studying the visualization tasks taxonomies for creating the studies, we identified several overlaps, inconsistencies, and ambiguities. This motivated us to create a unified list of visualization tasks in a structured format, resulting in an additional contribution of this work (chapter 7). Figure 1.1 illustrates the sequence of activities carried out in this work.

The structure of this document follows the studies conducted, and Chapter 8 concludes the thesis and points to future work.

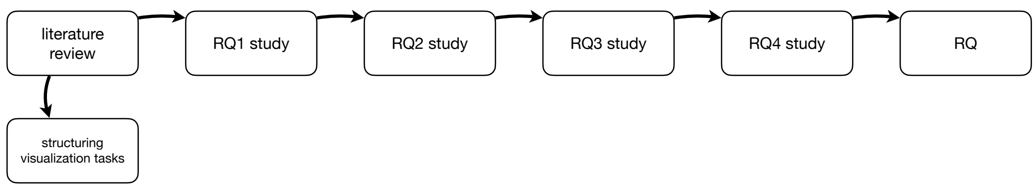


Figure 1.1: Activities carried out in this work.

2
Related Work

In this chapter, we present existing work related to each research subquestion and corresponding study. We also included a section about visualization task taxonomies, which underlie most of the studies we conducted.

2.1
Making sense of visualizations

Data visualization can help interpret and understand a large volume of data. Some factors can make a particular visualization challenging to understand: the chart type may not be the most appropriate for the data type, there is an excess or lack of information, the visual channels are not correctly mapped, the reader or analyst does not know the schema to interpret it, to name a few.

Familiarity with a particular visualization type is crucial in ensuring that one can extract correct information about it. Börner et al. (2016), informally investigated the familiarity of young and adult museum visitors with different visualization types. They seek to know whether and where they have seen them before, how they read and call them, and what types of data or information they would visualize similarly. Results showed that charts are easiest to read, followed by maps, graphs, and Network layouts. However, a very high proportion of the studied population could not name or interpret visual representations beyond fundamental charts.

Lee et al. (2016) investigated how people make sense of unfamiliar information visualizations such as parallel-coordinates plot, chord diagram, and treemap. They experimented with three sessions, one for each visualization type. For each session, they asked participants to verbalize their thoughts and behavior while making sense of the visualization, seeking to extract the participant's visual knowledge and understanding. The participant has only the possible prior experience of similar visualizations to collect the information. As a result, they established a model for novice visualization sensemaking: the NOVIS. This model consists of five activities and a miscellaneous one, as follows:

1. Encountering visualization: a cognitive activity that the reader faces and looks at visualization as a whole image and can build their first

impression of it.

2. Constructing a frame: a cognitive activity that the reader attempts to construct a frame (*i.e.*, "an explanatory internal structure that accounts for visual objects" Klein et al. (2007)) to make sense of a given visualization.

3. Exploring visualization: the cognitive activity that the reader interacts with visualization to discover facts and insights based on the constructed frames.

4. Questioning the frame: the cognitive activity the reader tries to confirm that the constructed frame reasonably explains the visual object. They can use the frame to explore the visualization.

5. Floundering on visualization: the cognitive activity that the reader does not know what to do with a visualization because they failed to construct any reasonable frames.

Understanding a person's cognitive process when creating a visualization can reveal how much they understand and express the visual mappings and how this can impact the final visualization. Grammel et al. (2010) discovered, through an empirical study, how information visualization novices construct visualizations. Participants asked questions about the data to be analyzed in all visualization construction cycles, mapping these questions to visualization construction; however, the formulation of questions was not the main point of the study. Participants received a task sheet with data attributes, visual properties to map, task operations, and the task description with a scenario. They were allowed to construct any visualization they wanted and were encouraged to use various visualization types, different from the common ones (bar, line, and pie charts). They faced several barriers, and the first one was selecting the right data attributes for their high-level questions.

With a similar goal, Huron et al. (2014) deconstructed the visual representation process by observing people creating representations using tangible tokens (mapped as a data unit). Participants created a visualization based on the available data. After crafting the visualization, they explained it as if to a friend. Although researchers did not define questions for study tasks, they noticed that some participants created particular questions about the scenario. They used these questions to customize visual mappings made in initial designs. They identified 11 logical tasks requiring several mental and physical actions to perform different combinations and execution orders (Table 2.1).

Table 2.1: Identified gols, tasks and actions by Huron et al. (2014)

	Logical task	Mental and Physical actions
1	Load data	Read, compute, select color, grasp, create
2	Build constructs	Organize, move
3	Combine constructs	Arrange, align
4	Extend	Read, compute, select, color, grasp, create, organize, mode, arrange, align
5	Correct	Increase, decrease, remove
6	Categorize	Select color, arrange, merge, split
7	Aggregate	Move, merge
8	Compute new value	Split, compute + load
9	Unitize	Organize, arrange, split, merge
10	Highlighting	Split (temporarily)
11	Marking	Create, select color

In an attempt to make analytics accessible to a broad audience, Setlur et al. (2016) developed Eviza, a natural language interface for visual analysis. They conducted a user study to gather a repository of questions that people would naturally ask, given different types of charts to guide this development. The chart types used were: maps, bar charts, scatterplots, time-series line plots, bubble charts, treemaps, heatmaps, and dashboards. They asked participants to provide three to five statements or questions about five random visualizations. They categorized the responses into 12 categories. They used the questions as a basis for the design and conception of Eviza. Unfortunately, they did not describe the question repository nor the relationships between them and the types of charts analyzed.

Recently, Kim et al. (2020) conducted a study to identify how people usually ask questions when they are analyzing a chart. It was a formative study to start broader research, whose ultimate goal was to develop a pipeline of automatic chart answers with visual explanations of how these answers were obtained. For a given chart, they asked crowdworkers to write questions, answers, and explanations for their answers. They limited the study to bar (simple, grouped, and staked) and line charts. They reviewed the responses and, after removing questions that were not valid, they classified them as lookup questions or visual versus non-visual questions and explanations. They found that people regularly ask visual questions, and those visual explanations are both common and effective.

The mentioned studies reveal that the participants' path to extract information from the visualization is to analyze the charts based on a question. When not provided by the study's methodology, participants formulated their analysis questions to analyze the visualization. However, the aim of the studies was not to identify the questions but the cognitive process resulting from them. We identified a value loss in dismissing these questions, so we propose a study to collect these questions and analyze the emerging patterns and misunderstandings. The studies we analyzed limited the diversity of visualizations to deepen the participants' analytical procedure. Asking, answering, and explaining are cognitive tasks that encompass more than we want to investigate. So, to address RQ1 and support the related work gaps, we simplified an empirical study task to analyze only the data-related questions that the participants would ask and diversify the set of visualizations.

2.2
Evaluating the effectiveness of different types of visualization

Many empirical studies have evaluated the effectiveness of different types of visualization, assessing whether the combination of data and visualization allows the success of the analysis task (Ondov et al., 2019; Saket et al., 2019; Ware, 2019; Bertin, 1983; Cleveland and McGill, 1984; Heer and Bostock, 2010; Mackinlay et al., 2007; Wongsuphasawat et al., 2016; Lee et al., 2017; Kim and Heer, 2018; de Santana et al., 2015, 2017). Some are specific to the chart type: bar charts (Srinivasan and Stasko, 2017; Skau et al., 2015), scatterplots (Sarikaya and Gleicher, 2017; Kim and Heer, 2018) and time series (Albers et al., 2014; Heer et al., 2009), for example. Others compare two types of visualization: bar *vs.* line charts (Siegrist, 1996), tables *vs.* pie charts (Spence and Lewandowsky, 1991) and bar *vs.* radar charts (Toker et al., 2012). These studies were conducted under different combinations of data sets and tasks.

Saket et al. (2019) used crowdsourcing to evaluate the effectiveness (*i.e.*, proportion of correct answers) of five types of bi-dimensional visualizations on a small scale (5-34 data points) for two data sets (cars and films). They evaluated tables and line charts, bar charts, scatter plots, and pie charts. They chose few data points because they would face two challenges for more than 50 data points: difficulty in task completion and task duration over 2 minutes. However, some of their conclusions cannot be generalized to a larger dataset with more categories. For example, the pie chart was one of the most effective charts for finding an extreme value (minimum or maximum value). For a data set with many categories, this same chart would have too many slices, perhaps some very similar ones, decreasing their effectiveness for this task.

Visualization tasks may be defined as analysts' goals when visualizing data. Since the 90s, there has been interest in identifying and classifying such tasks (Wehrend and Lewis, 1990; Roth and Mattis, 1990), mapping them onto efficient visualizations. Later, there has been greater interest in formalizing these classifications in taxonomies for diverse ends within the data visualization field. We will further specify related work on defining visualization tasks in the section 2.5. Our intention is only to introduce this subject to link it to works on the effectiveness of visualizations.

Kim and Heer (2018) conducted a study to evaluate performance at different task types (comparison of individual values *vs.* aggregate values) and data distribution (cardinality and entropies). They used four analysis tasks and five variations of visual codings (alternating the analysis variable in x, y, color, size, and position). The data sets had at most 30 points, and only three analysis variables were used (1 categorical and two quantitative). All of the evaluated visualizations were chart variations at the position of points (in a scatterplot) or mapping some variable onto the point size (bubble). The questions had only two available answers, reducing the opportunities for errors. Each person answered the same task and distribution despite balancing the number of people. There were eight different questions for each coding.

Despite some related work discussing visualization effectiveness and efficiency, they do not consider the data distribution, especially when there are very close similarities or significant discrepancies within the values. Our work extends related work in various aspects, which we present in detail in section 4. So, to address RQ2, we conducted an empirical study seeking to identify, for each task, the best visualization type according to data distribution, in terms of effectiveness, time on task, and adequacy to the task.

2.3 Recommending visualizations

In recent years, researchers developed many visualization systems to help users see, explore, and analyze information. The features supported by these systems vary widely, from supporting casual visual collaborations (*e.g.*, ManyEyes - (Viegas et al., 2007)) to commercial visual analyses (*e.g.*, Spotfire - (Ahlberg, 1996), and Tableau - (Gotz et al., 2010)). They can assist users in creating visualizations to facilitate data information analysis. The user then has two goals: to interact with the system to generate the charts and to visually analyze them.

These systems require additional programming to create visualizations and construct a highly customized chart. These systems' learning curve is high, and data analysts are the frequent (maybe only) users. By contrast, several toolkits and systems support lay users in the visual analysis process, suggesting visualizations accordingly to the data.

Vartak et al. (2017) described their vision of the essential requirements and design considerations for recommending visualizations. They defined recommendations based on five relevant axes: data characteristics, task, domain, ease of understanding, and user preference.

Systems that rely on **data characteristics** to recommend visualizations can be based on:

(i) certain properties of the data set, such as its attributes (Gnanamgari, 1981) and its dimensions (Shneiderman, 1996);

(ii) rules or techniques (Viegas et al., 2007);

(iii) its own formalized language (Hanrahan, 2006; Satyanarayan et al., 2017);

(iv) statistics (Wongsuphasawat et al., 2016; Key et al., 2012; Vartak et al., 2017); or

(v) context (the data set, the type of task, and the perspective on the data: general or detailed) (Ribeiro et al., 2014). Visualizations, according to these characteristics, allow the analyst to have an overview of the data distribution, identify related attributes, and contextualize trends.

Other recommendation systems are **task-oriented**, based on the analysis intention. Gotz and Wen's (2009) approach uses inference of intention through user actions. The "observation" of the interaction with a visualization suggests new analyses with other visualizations. Srinivasan et al. (2019) use a similar but fact-based approach. Voder, a system they developed for this purpose, uses a set of predefined heuristics to generate data facts associated with a specific visualization. They defined these heuristics according to the visualization task, defined in Amar et al.'s (2005) work. Each fact has its visualization, and, thus, a set of recommended visualizations are shown to the analyst. Task-oriented recommendations can also be related to the analysis style: exploratory, comparative, predictive, or with a clear goal (Vartak et al., 2017). In the latter style, we can mention VizAssist (Bouali et al., 2016), which allows the user to define which analysis tasks they want to perform and then recommend appropriate visualizations.

Concerning the **domain**, some recommendation systems may use information about the typical behavior of the attributes or data set, relationships between groups of attributes, and even factors external to the database (Vartak et al., 2017). Munzner (2009) proposed a model that first characterizes the task and the data in the vocabulary of the problem domain so that visualization can meet the requirements of users in any specific target domain.

Voigt et al. (2012) proposed an approach based on knowledge obtained from the domain ontology to provide visualization components that support both the user's task and the data.

The **ease of understandin**g axis encompasses recommendations that consider whether the visualization displays the data intuitively. Some suggestions may involve aspects of appropriate mapping of visual codes (Zhou and Feiner, 1998), color scales (Keller et al., 2006; Szafir, 2018), and data dimensionality (Liu et al., 2014).

The last axis encompasses recommendations oriented by **users' preferences**. Preferences for particular visualizations may differ depending on the analysis stage, or even the characteristics of the investigation. The study by Toker et al. (2012) shows the impact of four personal components (perceptual speed, verbal memory, visual memory, and experience) on the effectiveness of two visualization techniques (bar and radar graphics).

In addition to these axes, we have identified the **analysis question** axis, perhaps more associated with the task-based axis. The analysis question is a more refined form of the analysis task. Different questions can be associated with the same task. A visualization that allows analyzing a specific question will not necessarily allow analyzing another, even though they participate in the same analysis task.

Some systems present visualizations in response to analysis questions. Eviza (Setlur et al., 2016) enables users to have a conversation with their data using natural language. Users can post a query about some data, and the system provides graphical answers. ViSC (de Sousa and Barbosa, 2014) is a visualization recommender tool that generates visualizations according to the mapping between data variables and visual channels and also maps the task onto questions about the data. This mapping and the current visualization populate a panel with related questions and suggest new visualizations. VisMaker (de Araújo Lima and Barbosa, 2020) is a recommender visualization tool that uses a set of rules to determine appropriate visualizations for a specific selection of variables and make suggestions through questions based on the user's selected variables and the domain ontology.

In addition to recommendation systems, we can also mention the visualization catalogs available online or in textbooks, which suggest the most appropriate visualization types for each situation[1]. The visualization types are generally grouped into navigable sets: function (comparison), data type (numeric), and variables number. They present a more detailed description and

[1] https://datavizcatalogue.com/, https://datavizproject.com/, https://www.data-to-viz.com/ - last visited, February 2022

usage examples upon reaching the specific visualization in the set.

Analyzing these different axes, we conclude that the best visual representation choice depends on different factors. An analyst who uses these systems or catalogs needs to know how to identify whether the suggested recommendation is sufficient to cover the goals of their analysis. So, to address RQ3, we analyzed, through an empirical study, how participants assess the suitability of certain data visualizations for answering specific analysis questions before and after being exposed to related guidelines (chapter 5).

2.4 Data visualization literacy

Data visualization literacy is "the ability and skill to read and interpret visually represented data in and extract information from data visualizations" (Lee et al., 2017). It can occur in three steps (Börner et al., 2016):

- *external identification*, where the reader recognizes the frame of visual encodings;
- *internal identification*, where the reader identifies visual characteristics or patterns; and
- *perception of correspondence stage*, where the reader makes analysis and extracts the messages from the visualizations.

How people understand, create, and extract information from visualizations concerns their visualization literacy. Existing work in visualization literacy spans different research communities. Some of them focus on the accurate assessment and representation of visualization proficiency (Bishop et al., 2020; Ruchikachorn and Mueller, 2015; Huron et al., 2014). Others are specific to teaching and learning about visualizations (Wang et al., 2020; Alper et al., 2017; Börner et al., 2019).

Concerning the assessment of visualization literacy, some researchers conducted empirical studies with analysts and reported their results to contribute to the area (Krekhov et al., 2019; Koedinger et al., 2001; Kodagoda et al., 2012; Lee et al., 2019). Others followed a different path, defining a test model as a contribution to apply in these studies (Lee et al., 2017; Boy et al., 2014; Maltese et al., 2015).

There are several variables to analyze in data visualization literacy assessments:

- visualization types: *e.g.*, bar charts, scatterplot, boxplots;
- analysis tasks: *e.g.*, find correlations, identify outliers;

– the goal: *e.g.*, reading, creating, interpreting; and

– the expertise degree: *e.g.*, novices, experts, children.

Some studies look for approaches to pedagogically improve people's ability to solve problems and obtain information through visualizations. Alper et al. (2017) analyzed current practices and challenges in teaching and learning data visualization in early education. They concluded that concrete examples should guide students to abstract knowledge about the visualizations. Maltese et al. (2015) created an assessment tool to measure differences between novices and experts when reading and interpreting visualizations. They wanted to identify when and how students developed proficiency in reading and interpreting charts. They reported a relatively small difference in ability level across participants. Even students with advanced coursework in science and mathematics had difficulty in the basic interpretation of common data visualizations. Chevalier et al. (2018) explored how basic visualization principles and skills are taught and learned at an elementary school in the United States. They concluded that visualizations are omnipresent in grades K–4, teachers believe visualizations are intuitive, and elementary students learn to read and create visualizations in early grades.

2.4.1 Assessing Visualization Literary Through Tests

Typically, studies that assess how well people understand visualizations aim to investigate the ability of a user to extract information from a graphical representation (Boy et al., 2014; Lee et al., 2017; Livingston et al., 2019). Boy et al. (2014) proposed a method for assessing the visualization literacy of a user inspecting the ability scores, derived from the item response theory (IRT) models (Cohen et al., 1996). Their method tests visualization literacy for line charts, bar charts, and scatterplots, by asking participants to read and answer questions about these charts. Each chart type has a separate test, and each test has a set of 12 items using different stimulus parameters and six tasks (minimum, maximum, variation, intersection, average, comparison). They created two line graph tests with slightly different designs. Both line and bar chart tests are useful for differentiating examinees with relatively low abilities from ones with average skills. In contrast, the scatterplot test is adequate for testing examinees with relatively low skills.

Proceeding very much in a similar way, Lee et al. (2017) developed VLAT, a visualization literacy test that associates tasks, chart types, and questions to assess user visualization understanding. It proposes metrics for difficulty indices and discrimination in evaluations. In one of the steps to create the systematic test, they asked the participants to formulate a sentence with information gained from the chart. The researchers analyzed the sentences collected and defined some potential test items, which are transcriptions of the sentences into analytical questions. The final set of test items has 53 questions for 12 different chart types and eight tasks.

To evaluate graph comprehension capability, Livingston et al. (2019) developed an algorithmic method for generating queries, based on the Sentence Verification Technique (SVT). Instead of defining transformations of prose sentences into query probes (from SVT), they made changes to graph information statements or assertions to graph queries: content (*i.e.*, representation of the data like points, lines or bars), labels (*i.e.*, variables names, title, axes names, and legend), and framework (*i.e.*, axes). First, they showed a source graph, some diversionary images, and a source prose. Then, they showed a graph query (bar or line chart) and asked participants whether the information in the graph query was "stated" or "not stated" in the source graph. Next, they did the same with a prose query, asking participants whether the information in this query was "stated" or "not stated" in the source prose shown before. Their goal was to assess the participants' understanding of each graph. They noticed a slight tendency for participants to be more accurate as queries showed more data values. Although the authors concluded that the participants understood the tasks and objectively demonstrated understanding in the graphs, this test is too simplified for our purposes.

Maltese et al. (2015) created an assessment tool to measure differences between novices and experts when reading and interpreting visualizations. They wanted to identify when and how students developed proficiency in reading and interpreting charts, using 19 visualizations commonly found in textbooks and school curricula and an additional scatterplot with a two-variable relationship. They conducted an item analysis to measure the items' psychometric qualities in the assessment test: item difficulty, item discrimination, and distractor. After the test, each response received a score: correct, incorrect, and missing. They reported a relatively small difference in ability level across participants. Even students with advanced coursework in science and mathematics had difficulty in the basic interpretation of common data visualizations.

After analyzing the literature studies, we identified that they seek to assess visualization literacy through suitable analysis questions. These questions always have a possible answer. We noticed that knowledge about visualizations goes beyond the visual search for answers about data. We, therefore, decided to assess whether conceptual questions and unsuitable analysis questions contribute to a comprehensive assessment of data visualization literacy. So, to address RQ4, we analyzed through an empirical study a visualization literacy assessment test for both applied and conceptual aspects of data visualization.

2.5 Visualization Task Taxonomies

Several approaches attempt to classify the different intentions (the visualization tasks) that an analyst may have when visualizing the data. These approaches are split in two fundamental ways in the literature: high-level tasks, which define conceptual tasks (Brehmer and Munzner, 2013; Schulz et al., 2013) or general goals (Shneiderman, 1996; Keller et al., 1994); and low-level tasks, which identify specific goals (Roth and Mattis, 1990; Amar et al., 2005) or analytical actions (Fujishiro et al., 2000; Wehrend and Lewis, 1990). Although the researchers agree to name these classifications as visualization task taxonomies, they differ in some of their definitions. For example, while Wehrend and Lewis (1990) call 'Identify' an *operator*, Sarikaya and Gleicher (2017) call it an *analysis task*.

Visualization tasks have been an object of study since the 1990s. Wehrend and Lewis (1990) defined a classification scheme that maps objects (data attributes) and "operations" (representation objectives) to find an appropriate visualization technique for a given problem – the user's goal in analyzing the representation. Roth and Mattis (1990) classified visualization problems and their solutions independently of the domain and proposed a taxonomy of information characteristics that provides a list of different user objectives in seeing a visual representation. Their proposed classification is similar to Wehrend and Lewis', albeit more succinct and focused on the automatic generation of a representation. These taxonomies are considered low-level and user-focused.

Shneiderman (1996) proposed TTT (Task by data type), a high-level, system-focused taxonomy based on data types and on the problem the user seeks to solve. He wanted to guide graphical user interface design for data visualization analysis. Amar et al. (2005) defined ten low-level analysis tasks that a person may perform when working with data. They defined "aggregate functions", which create a numeric representation for a set of entities in the data set. They claim that high-level tasks do not express a specific objective or task but require an answer for a more direct question, which is usually derived by using one or more low-level analytic operations (Srinivasan and Stasko, 2017). Some of these questions may be efficiently answered by text; others require visualizations for an efficient answer. However, even when a textual representation is sufficient to answer a specific question, visualization may amplify the understanding of an answer and its context.

Later studies became more specific, as is the case of Lee et al. (2006). They defined the list of chart visualization tasks with enough detail to be useful both for designers who seek to improve their systems and for evaluators who seek to compare chart visualization systems. In contrast, all tasks were composed of tasks created by the primitive tasks described by Amar et al. (2005), as well as two generic tasks and one chart-specific task.

Chen et al. (2009) explored tasks related to “data, visualization and objective”, and defined a taxonomy to categorize facts that may be extracted from multidimensional data in a visual data analysis task. Facts are patterns, relationships, or anomalies extracted from data through analysis (Chen et al., 2009). More recently, Brehmer and Munzner (2013) asserted that visualization tasks ought to be described abstractly, through different levels: *why* the task is conducted, *how* the task is conducted, and *what* are the task inputs and outputs.

The low-level taxonomies proposed previously only answer *how*, while the high-level ones only answer *why*. For this reason, some researchers consider the use of more than one taxonomy, as is the case of VLAT (Lee et al., 2017). To develop VLAT, a visualization literacy test, Lee et al. (2017) associated tasks from three different taxonomies. First, they combined the low-level taxonomy (Amar et al., 2005) with the facts-based one (Chen et al., 2009) and afterward discarded some of the tasks that were included in *how* and *why* from (Brehmer and Munzner, 2013) – the discarded tasks were related to manipulation and generation of new elements, and not reading and interpretation of visual representations of data.

Visualization tasks are fundamental tools for this work. We used them in all the empirical studies we performed. However, as taxonomies present in the literature use slightly different definitions of visualization tasks, we need to look deeper into the definitions and not rely on the labels. They are full of overlaps, inconsistencies, and ambiguities that motivated us to create a precise specification of visualization tasks (chapter 7).

2.6 Concluding Remarks

This chapter presented existing works that contribute to research on data visualization. In each subsection, we presented the main concepts on each topic related to each research subquestion. We also present uncovered concepts by the literature that became important research points to help address the primary research question.

3
Understanding novices' attempts to make sense of data visualizations

Researchers have been striving to develop new tools and visualization techniques to support analysts in extracting information and gaining insights from the data. It is still challenging to ensure that both the producers and consumers of data visualization can understand them. In this chapter, we investigate *what* are the common novices' misinterpretations when trying to make sense of data visualizations.

We assume that people's questions about the represented data allow us to recognize gaps in their understanding of data visualization concepts. This understanding contributes to data visualization education and the design of visualization authoring tools and other tools that make use of visualizations, such as question-answering systems.

3.1
Goal

To understand how people make sense of data visualizations, we set out to learn data-related questions produced by people with minimal knowledge of data visualization, when exposed to different kinds of visualizations.

3.2
Study Design

We created twenty visualizations: bar (ordered by category); bar (ordered by frequency); bar (clustered); bar (stacked); boxplot; heatmap; chord; Sankey; network; line; line (multiple); ridge; histogram; scatterplot; scatterplot (+ color); bubblechart; bubblechart (+ color); map (cartogram); map (choropleth); and table. The visualizations are available in appendix B.

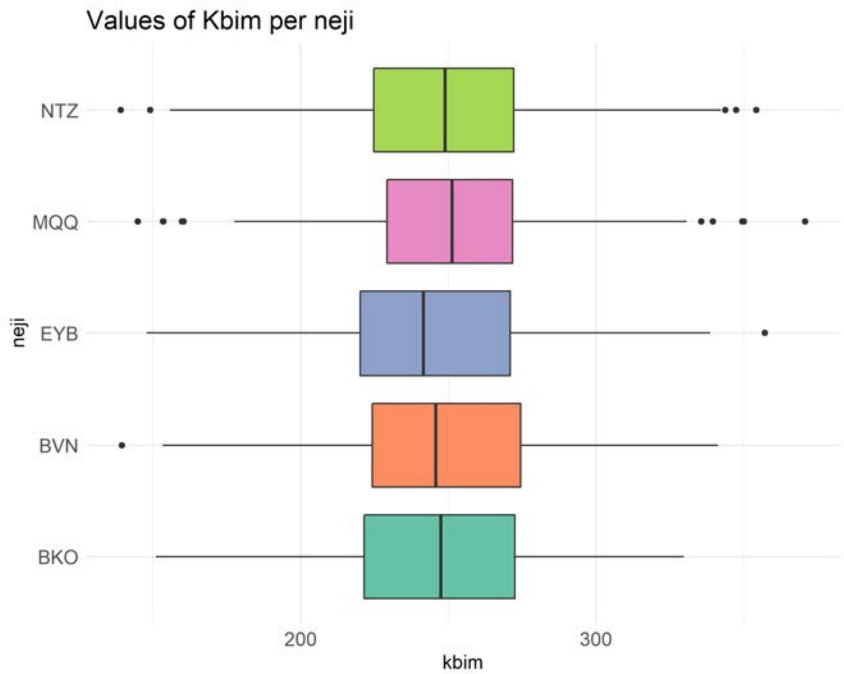
The dataset used to create the visualizations had 15 variables of different types, with randomly generated values. The variables had meaningless names in either Portuguese or English. For example, *klod*, *nili*, and *neji*. We used dummy variables to prevent a participant's domain knowledge from influencing the literacy assessment. They may rely on their domain knowledge to fill the gaps in their interpretation of the visual representation.

3.3 Procedure and Participants

We created an anonymous online questionnaire and presented all twenty visualizations in random order, one at a time. For each visualization, each participant should generate up to five questions about the underlying data, questions they believed that could be answered by examining the visualization. They also indicated the level of effort required to generate each question on a seven-point scale, with 1 meaning "no effort", and 7 meaning "excessive effort". Table 3.1 exemplifies the task. The remaining visualizations are in appendix B.

Table 3.1: Study 2 task example

Task: Analyze the visualization below and create up to 5 questions you consider you can answer using the visualization. Then, select the effort level to create each one.



	Effort level						
Question	1 - no effort	2	3	4	5	6	7 - excessive effort
1.	( )	( )	( )	( )	( )	( )	( )
2.	( )	( )	( )	( )	( )	( )	( )
3.	( )	( )	( )	( )	( )	( )	( )
4.	( )	( )	( )	( )	( )	( )	( )
5.	( )	( )	( )	( )	( )	( )	( )

In addition to a qualitative analysis, we also standardized the answers using an open coding approach. We derived templates for the recurring types of questions that emerged as information-seeking patterns. Next, we classified the various kinds of error classification participants made when creating the questions. Figure 3.1 shows the procedure schema.

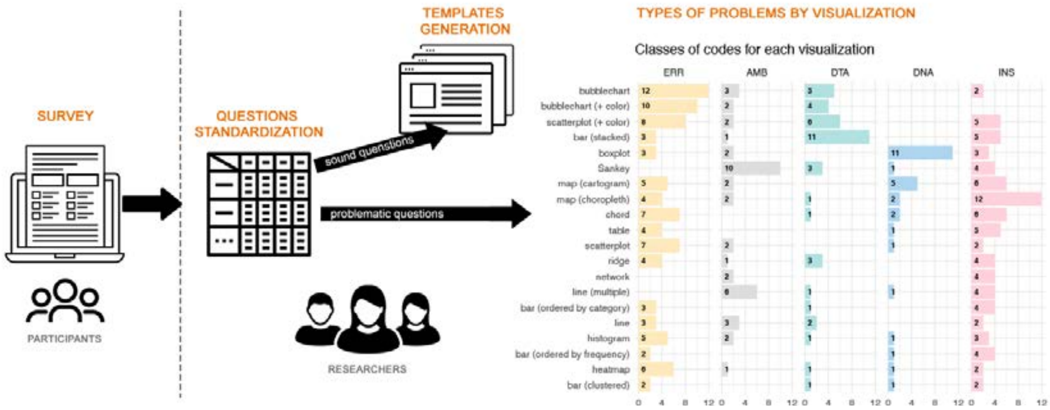


Figure 3.1: Study procedure

The volunteer participants we invited for our study were graduate and undergraduate students in Computer Science, Design, Engineering, and Social Sciences, with little to no data visualization knowledge. Although we have not systematically tested their level of visualization literacy, we asked them to self-assess their previous experience and knowledge with data visualization on a series of a 7-point Likert scale statements (1 meaning no knowledge or experience, and 7 meaning specialist). We asked whether they had already:

- taken a course (median M = 1, interquartile range IQR = 3);
- read any textbook material or blogs on data visualization (M = 1, IQR = 1), selected a type of chart for a visualization (M = 3, IQR = 2);
- adjusted the mapping of visual variables in a chart (M = 2, IQR = 2); and
- evaluated a data visualization (M = 2, IQR = 2).

Their participation was completely voluntary and the data collected was anonymous. Twenty-two people participated in the study.

3.3.1 Analysis and Results

The questionnaire collected a total of 1058 answers. Three researchers (including the author of this thesis), with international publications in conferences and journals on data visualization, examined the data collected and discarded eight non-questions. Each researcher individually and independently created standardized versions of the questions. Then we identified the differences, discussing them one by one until we reached a consensus. The guiding principle to standardize was to rephrase each question following basic English grammar, using a simple structure starting with an interrogative pronoun and using the variable types as nouns. From the 1050 questions, we deemed 800 as sound, *i.e.*, we could clearly answer them.

We also flagged 250 questions as problematic, which means that, by inspecting the visualizations, we could not answer those question instances clearly. When examining the 800 standardized questions each researcher had written, we found that we had fully agreed on 641 of them (80.1%). In 150 cases (18.8%), one of us had created a different standardized question, and in 9 cases (1.1%) we had all created different standardized questions. Most of the discrepancies were caused by distraction and easily resolved. In the cases where there were equivalent ways of posing the same question, we opted for the simplest phrasing.

Figure 3.2 shows the number of questions created per type of visualization, according to our assessment: 800 clear and conceptually sound questions ("OK", in blue), and 250 problematic questions (in orange).

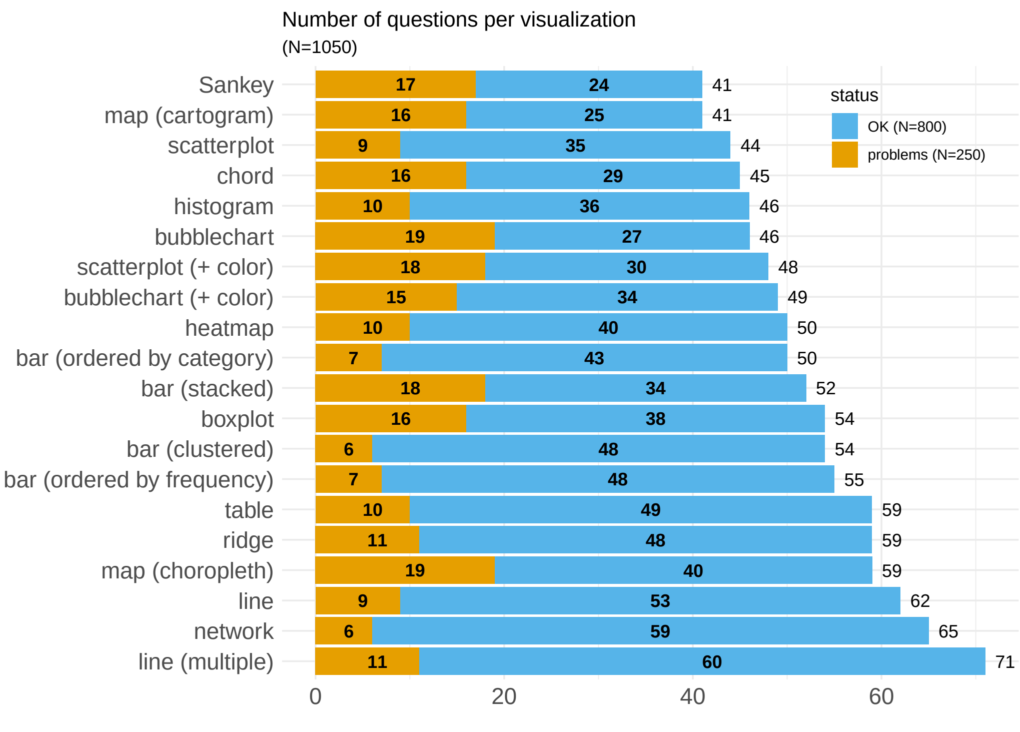


Figure 3.2: Number of questions created per type of visualization.

To assess the possible influence of the participants' previous knowledge of data visualizations on the results, we (i) calculated each participant's error rate (number of problematic questions / total number of questions generated by the participant); and (ii) calculated the Spearman correlation between the degree of self-reported knowledge and the error rate. The correlation coefficients were very low (ρ in [-0.28023163, 0.09411151]) and none of the correlations were significant (the lowest p-value was $p = 0.2185527$).

3.3.2 Levels of effort and question order

Regarding the level of effort to create each question on a seven-point scale and the question order (one to five), our hypotheses were:

H1: There is a significant difference in the perceived effort level to create a clear question and a problematic question.

We ran a Mann-Whitney non-parametric test on the effort level for OK *vs.* problematic questions, and found a significant difference ($U = 63326, p = 5.047 \times 10^{-16}$). Therefore, we accept H1, *i.e.*, participants perceived they expended more effort in creating the questions that were later assessed as having lower quality. In future work, we may consider the participants' assessment of their effort when we first select questions for further analysis. Figure 3.3 depicts the distribution of OK and problematic questions over the perceived (self-reported) effort level.

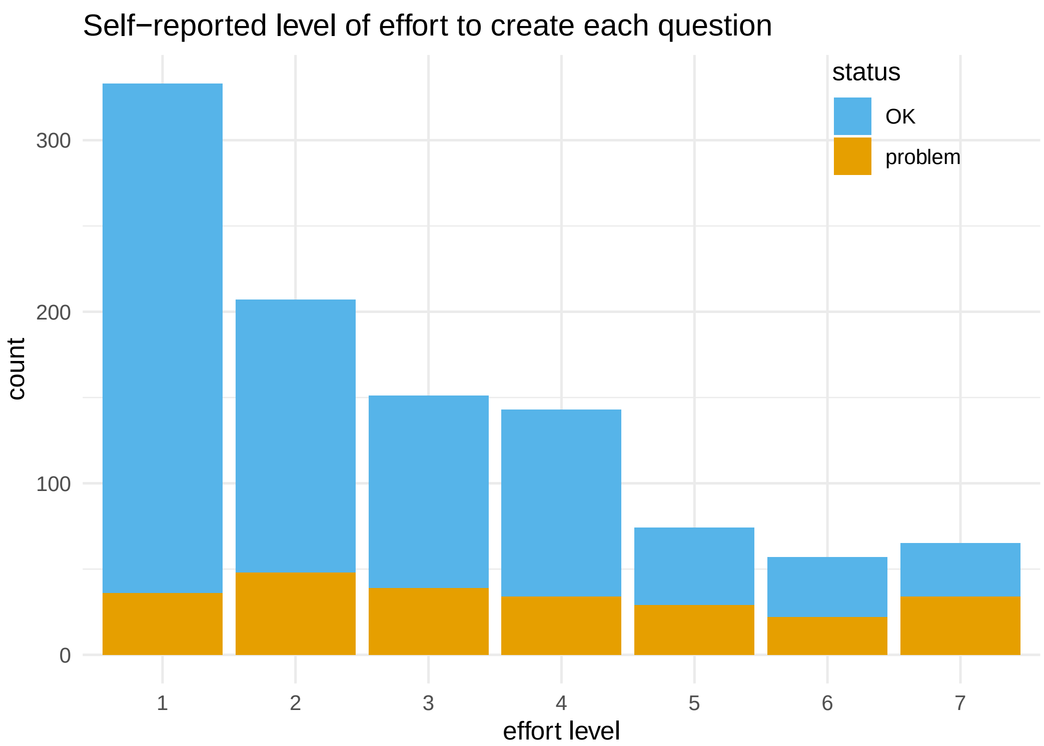


Figure 3.3: Self-reported effort level to create each question.

H2: There is a significant difference in the question order between clear and problematic questions.

We ran a Mann-Whitney non-parametric test on the question order for OK *vs.* problematic questions, but found no significant difference ($U = 94548, p = 0.1722$). Therefore, we reject H2, *i.e.*, there was no difference in the quality of questions created earlier or later for each visualization. We still do not know up to how many questions we might ask of participants before getting lower-quality results. Figure 3.4 depicts the distribution of OK and problematic questions over the question order.

Figure 3.4: Question order and the initial quality assessment of each question.

H3: The effort levels to create a question and the question order are correlated.

We calculated the Spearman rank-correlation coefficient between effort level (on a 7-point scale) and the question order (1 to 5). We found a weak correlation ($\rho = 0.349$ ($p \leq 2.2 \times 10^{-16}$)), so we may reject H3. This result may be due to the low number of questions created in the fourth and fifth order, given that the participants were not obliged to provide all five responses for each visualization. We expected that the perceived effort level would increase with the number of questions asked (up to the maximum number of five questions per chart). However, this did not happen. We found no correlation, so, again, the study did not reveal an appropriate threshold for the number of questions we might ask participants to create. Figure 3.5 depicts the relation between the effort level and the question order.

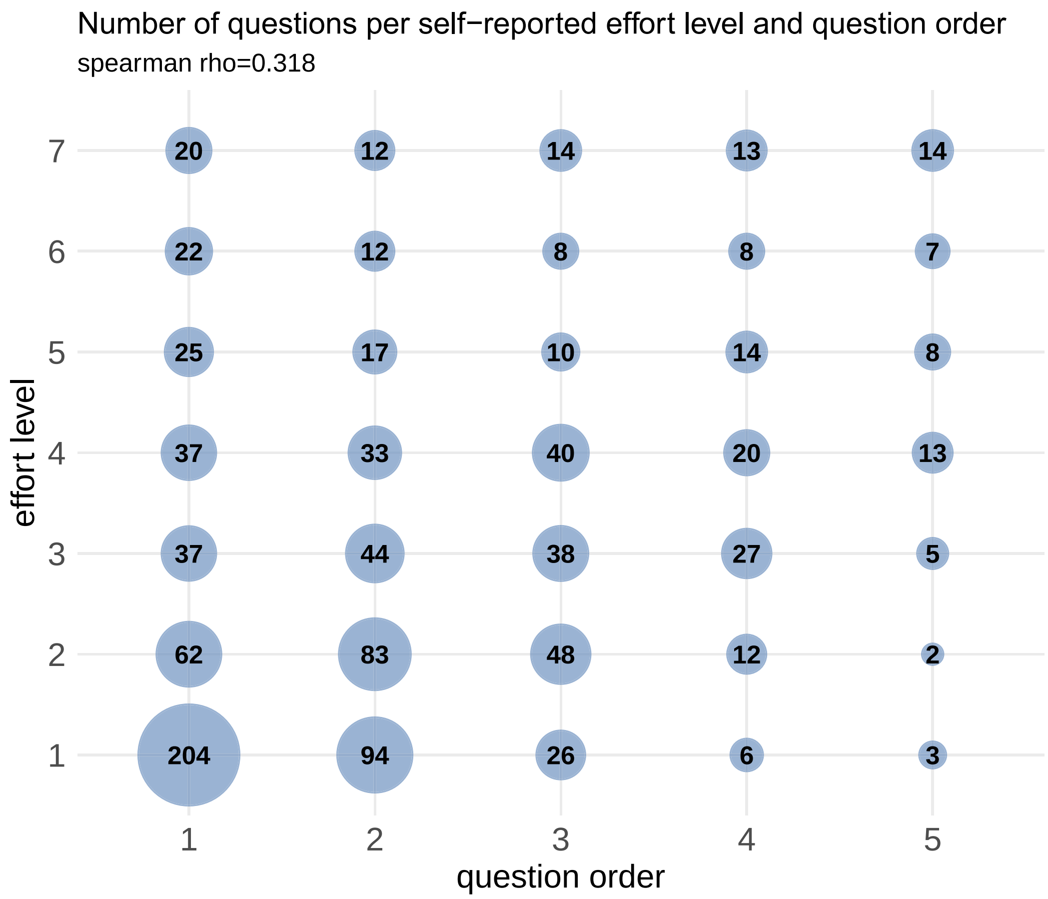


Figure 3.5: Number of questions per self-reported effort level and question order.

3.3.3
Clear, conceptually sound questions

We rewrote the 800 clear, conceptually sound questions to analyze what kind of questions people would ask about each chart. We generated a consolidated list of 249 unique question-templates (265 ⟨visualization, template⟩ pairs), each one associated to at least five question-instances (table 3.2). The templates replaced the variable names with their corresponding types: N for nominal variables, Q for continuous numeric, T for temporal, S for spatial, and Objs for objects. We also parameterized question variations. For example, The questions: "In which T did Q reach its largest value?" And "In which T did Q reach its smallest value?" were subsumed under "In which T did Q reach its [largest | smallest] value?").

Table 3.2 illustrates the question-templates. If two or more variables of the same type are used (for instance, three continuous numeric variables in a bubblechart), they were indexed with numbers, as in Q1, Q2, Q3. Moreover, when referring to a specific value, we use a dot notation, such as N1.A, meaning a specific value of one of the nominal variables. Expressions within brackets indicate they may or may not be present in the template instances. For instance, for the clustered bar chart, “What is the value of klod per nili?” and “What is the value of klod in nili BKL and neji MQQ?” were both instances of the template “What is the value of Q [per N1 | per N1 and N2 | in N1.A | in N1.A and N2.B]?”

3.3.4 Problematic questions

We analyzed the 250 problematic questions and coded them using an open coding approach to identify what kinds of mistakes people made when asking questions, resulting in a unified set of 20 codes, categorized into five major classes.

- *ERR-** questions containing conceptual errors (88 occurrences)
- *AMB-** questions that contain some ambiguity (41)
- *DTA-** questions that are technically answerable, but are difficult to answer with the visualization, *i.e.*, questions for which the visualization was not appropriate (43)
- *DNA-** questions the visualization does not answer (28)
- *INS-** failures to follow the instructions when filling out the questionnaire (79)

Again, each researcher individually and independently coded the problematic questions. We had a round of discussion to analyze and standardize the codes created. We reclassified the questions individually and compared the results to ensure we agreed with the codes using the generated codes. When there was disagreement, we discussed and adjusted until we reached a consensus. Figure 3.6 shows how the code classes are distributed across the different visualizations.

We calculated the inter-rater agreement using the Fleiss kappa metric (Fleiss et al., 2003). We obtained $\kappa = .693$, which, according to Landis and Koch's proposed benchmark, is considered a substantial agreement (Landis and Koch, 1977). We then examined the questions presenting conflicts and decided on the final coding. Below, we describe each of the major problematic classes.

Conceptual. In 88 cases, there were comprehension errors about the variables represented in the chart. The higher incidence of error issues was in the bubble chart. It was related mostly to treating a continuous numeric variable as if it were discrete, an example of misunderstanding on how to

Table 3.2: Question templates with five or more occurrences.

Visualization	Question template	Count
bar (clustered)	What is the value of Q [per N1 | per N1 and N2 | in N1.A | in N1.A and N2.B]?	9
	Which N1 has the [largest | smallest] Q in N2?	9
bar (ordered by category)	Which N has the [largest | smallest] number of Objs?	18
	How many Objs are there per N?	6
	What is the [average | median | mode] number of Objs?	5
bar (ordered by frequency)	Which N has the [largest | smallest] number of Objs?	19
	How many Objs are there [per N | in N.A]?	10
	What is the [average | median | mode] number of Objs?	5
	Which Ns have [more | fewer] than A Objs?	5
bar (stacked)	Which N has the [largest | smallest] number of Objs?	14
	Which N1 has the [largest | smallest] number of Objs [per N2 | in N2.A]?	12
boxplot	Which N has the [largest | smallest] variation of Q?	6
	What is the median of Q per N?	5
	Which N has the [most | fewest] outliers?	5
bubblechart	For what ranges of [Q1 | Q2] do we have the [most | fewest] Objs?	12
bubblechart (+ color)	For what ranges of [Q1 | Q2] do we have the [most | fewest] Objs (per N | in N.A)?	12
chord	Which N1 is the [most | least] associated with N2?	9
	What is the [most | least] frequent N?	7
heatmap	Which N1 and N2 has the [largest | smallest] Q?	12
histogram	Which range of Q has the [most | fewest] Objs?	16
	What is the distribution of Q?	5
line (single)	How has Q [behaved | increased | decreased] (in T.A)?	11
	In which T did Q reach its [largest | smallest] value?	11
	In which (period of) T did Q [increase | decrease] the [most | least]?	7
	In which (period of) T did Q [increase | decrease] monotonically?	5
line (multiple)	Which N had the [most | least] variation of Q (in (period of) T)?	16
	How has Q behaved [in each N | in N.A] (since T.A)?	7
	In which N did Q [increase | decrease] the [most | least] over T?	6
	Which N had the [largest | smallest] Q (in (period of) T)?	6
map (cartogram)	Which S (or set of Ss) has the [largest | smallest] values of Q?	16
map (choropleth)	Which (set of) Ss have the [largest | smallest] values of Q?	21
network	Which Vs have the [largest | smallest] degree (number of connections)?	11
	What is the [shortest | longest] path between V.A and V.B?	7
	How many cycles are there in this graph?	6
	Which Vs are (indirectly) connected (to vertex V.A)?	5
ridge	In which T(year) did Q have its [largest | smallest] value (per N | in N.A)?	10
	What are the values of Q per N (in T(year) | in T.A-T.B)?	5
	Which N had the [largest | smallest] Q (per T | in T.A)?	5
Sankey	Which N1 is associated with (the [most | least]) Objs [in each N2 | in N2.A]?	9
scatterplot	In which range of [Q1 | Q2] are there the [most | fewest] Objs?	13
	What is the relation between Q1 and Q2?	6
scatterplot (+ color)	What is the relation between Q1 and Q2 (in each N | in N.A)?	11
	Which N has the [most | least] Objs (in the range Q1.A-Q1.B, Q2.C-Q2.D)?	7
table	Which N has the [largest | smallest] number of Objs?	20
	What is the [average | median | standard deviation | variance] of the number of Objs?	8
	What is the number of Objs in each N?	6

Table 3.3: Codes resulting from the open coding process. A total of 277 code occurrences were associated with the 250 problematic questions.

Code	**Count**	**Definition**
Questions containing errors		
ERR-COUNT-OBJ-IO-CONT-VAR	37	Question called for a countable object, but mentioned a continuous variable instead
ERR-MISUNDERSTOOD-VAR	20	Participant seems to have misunderstood the variables encoded in the visualization
ERR-DISC-VAR-IO-CONT-VAR	18	Question called for a discrete variable, but mentioned a continuous variable instead
ERR-TREND-IN-CATEG-VAR	5	Participant asked about a trend along a categorical variable
ERR-POINT-IO-RANGE	5	Question called for a range of values, but mentioned a single value instead
ERR-OBJ-VAL-IO-OBJ	1	Question called for objects, but mentioned object values instead
ERR-RANGE-IO-POINT	1	Question called for a single value, but mentioned a range of values instead
ERR-STAT-IO-VAL	1	Question called for values, but mentioned a summary statistics or aggregate value instead
Ambiguous questions		
AMB-QUESTION	41	Question used ambiguous terminology that allows for multiple interpretations
Questions that were difficult to answer with the visualization		
DTA-ALL-VALS	23	Participant asked for all the values of one or more variables of all objects
DTA-VIS-TYPE	17	The question cannot be answered with the current visualization type
DTA-DATASET	3	Although the question could potentially be answered with the current type of visualization, the particular dataset made it very difficult to answer it.
Questions the visualization does not answer		
DNA-REQ-INDIV-OBJS	9	The question requires individual objects to answer, but only aggregates are represented in the visualization
DNA-OTHER	8	The question cannot be answered by this visualization (for some other reason than the ones specified in the other DNA-* codes)
DNA-REQ-ADDIT-VARS	4	The question requires additional variables that are not represented in the visualization
DNA-WHY	4	The question requires some statement of causal relationship that cannot be asserted based only on the visualization?
DNA-REQ-DATA-VALS	3	The question requires individual data values to answer, but they are not represented in the visualization
Failures to follow instructions		
INS-MAPPED-ONTO-DOMAIN	38	The participant phrased an analogous question about a familiar domain, ignoring the dummy variables
INS-ABOUT-VIS	37	Question about the chart type and its elements, and not about the underlying data.
INS-MULTIPLE-QUESTIONS	4	The participant provided multiple questions in a single text field

use a variable: *e.g.*, "What value of Q_x had the least frequency of Q_y?", in a bubblechart.

Ambiguity. In 41 cases, the questions were ambiguous, meaning we could assign different interpretations to it, each with a different answer: *e.g.*, "What is the intensity of this relation?" without specifying the values of interest (nor what they meant by "intensity"), in a Sankey diagram. This kind of question brings challenges when implementing a system to answer users' queries about specific visualizations.

Difficult to answer. In 43 cases, the questions were difficult to answer. The issues were more related to the inadequacy of the visualization to answer the stated question. For instance, the high incidence of difficult to answer issues was in the stacked bar chart. It was related mostly to questions that inquired about the size of specific segments, which are often not straightforward to answer with this type of visualization: *e.g.*, "What is the exact value of Q_y in each segment of the curve?", in a line chart.

Does not answer. In 28 questions, participants posed questions that the visualization could not answer. For instance, the high incidence of does not answer issues was in boxplots. It is related to questions that require observing individual objects or computing derived values from them, which were not represented in the visualization: *e.g.*, "How many countries are above average of Q_color ? (only ranges of values are provided, and the mean is not provided nor can it be calculated)", in a map (cartogram).

Failures in instructions. In 29 questions, participants failed to follow the instructions when filling out the questionnaire. We asked them to pose questions about the underlying data and not about the chart or any imagined domain. For instance, the highest incidence of instructions issues was in the choropleth map. It is mostly related to assumptions about the underlying domain represented in the map. For example, they were assuming that the numeric variable meant the population of each country: *e.g.*, "What are the most populous countries?".

3.3.5 Distribution of problems across participants

Finally, we analyzed how the problems were distributed across participants, to assess how common each type of problem was, and whether a problem was particular to only one or few participants. Table 3.4 summarizes the number of participants that introduced which kind of problem, and Figure 3.7 details this distribution. We omitted the *INS-** problems, as they are related only to not following the study instructions and are therefore unrelated to visualization

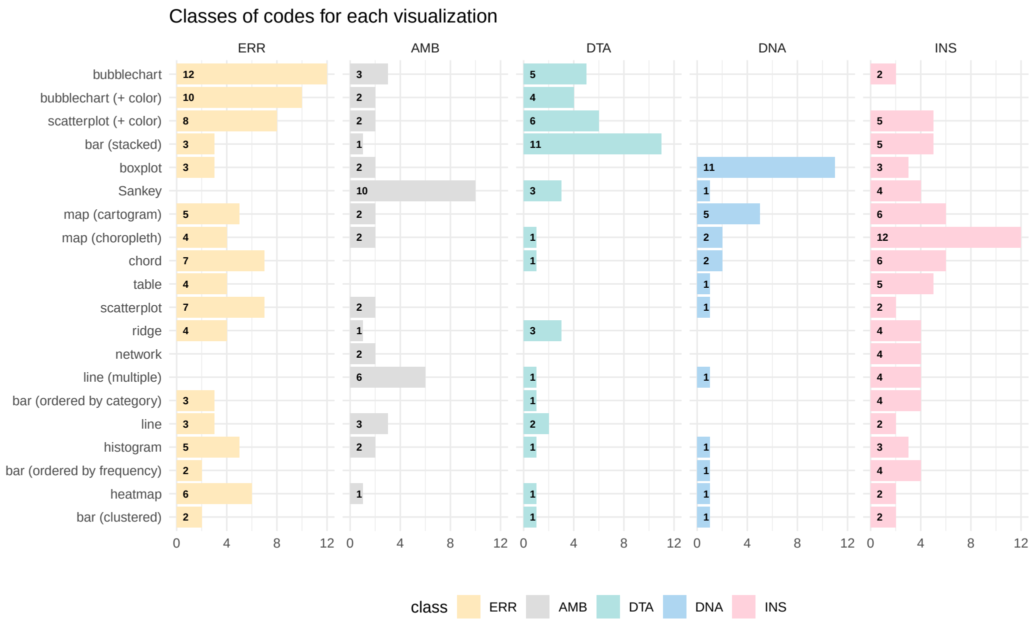


Figure 3.6: Distribution of code classes per visualization.

literacy.

Table 3.4: Number of participants who introduced a problem in their questions.

Code	Number of participants
AMB-QUESTION	16
ERR-COUNT-OBJ-IO-CONT-VAR	12
DTA-VIS-TYPE	11
DTA-ALL-VALS	7
ERR-DISC-VAR-IO-CONT-VAR	7
ERR-MISUNDERSTOOD-VAR	7
DNA-OTHER	6
DNA-REQ-INDIV-OBJS	4
DTA-DATASET	3
ERR-POINT-IO-RANGE	3
ERR-TREND-IN-CATEG-VAR	3
DNA-REQ-ADDIT-VARS	2
DNA-REQ-DATA-VALS	2
DNA-WHY	2
ERR-OBJ-VALUE-IO-OBJ	1
ERR-RANGE-IO-POINT	1
ERR-STAT-IO-VALUE	1

As we can see, most participants created ambiguous questions (*AMB-QUESTION*); misunderstood the types of variables represented in the visu-

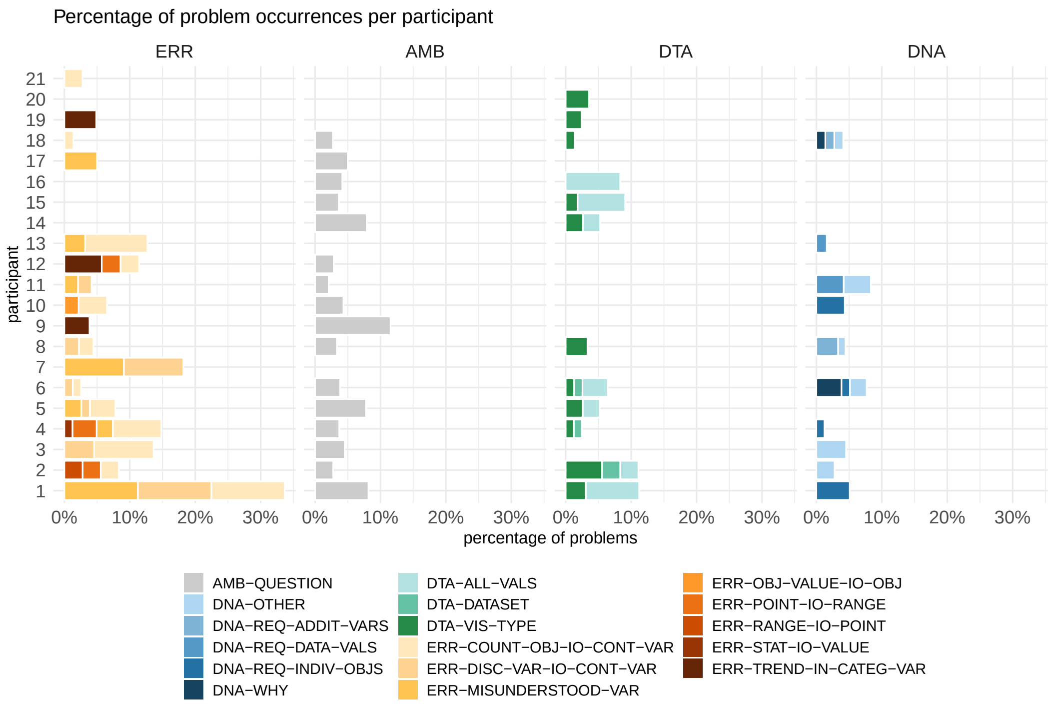


Figure 3.7: Distribution of codes per participant, as a percentage of each participant's errors.

alizations (*ERR-COUNT-OBJ-IO-CONT-VAR*, *ERR-DISC-VAR-IO-CONT-VAR*, and the more general *ERR-MISUNDERSTOOD-VAR*); or created questions that the corresponding visualization type could not answer effectively (*DTA-VIS-TYPE*), *i.e.*, questions that could be answered much better by other types of visualization (*e.g.*, attempting to compare or find the difference between bar segments in a stacked bar chart, instead of a clustered bar chart).

For the three kinds of errors that were made by only one participant (*ERR-OBJ-VALUE-IO-OBJ*, *ERR-RANGE-IO-POINT*, and *ERR-STAT-IO-VALUE*), we note that each one was made by a different participant.

3.4 Concluding Remarks

This chapter introduced our study about the questions people with minimal knowledge of data visualizations ask when exposed to different visualizations. We created 20 different visualizations and asked participants to formulate questions to answer using them. After having collected 1058 questions from 22 participants, we classified the questions into two groups: (i) clear and conceptually sound (800 questions) and (ii) problematic (250 questions). We derived 249 unique question templates from the first group; those templates describe the different types of questions our participants expected each visualization should help answer. For the second group, we applied an open coding technique to identify the types of problems found in the questions, yielding 20 types of problems subsumed under five classes of problems.

The paper "What questions reveal about novices' attempts to make sense of data visualizations: patterns and misconceptions", published at Computers & Graphics 2020 (Rodrigues et al., 2021), has the full description of this study.

4
Comparing the effectiveness of visualizations of different data distributions

When the analyst faces some data, they may have some predefined exploration goals. Their goals may be translated into analysis questions about the data. Analysts may then use visualizations to answer these questions. There has been a great effort in defining visualization recommending systems that suggest better visualizations based on specific aspects, such as data characteristics (Gnanamgari, 1981; Mackinlay, 1986; Hanrahan, 2006; Roth and Mattis, 1990; Shneiderman, 1996; Viegas et al., 2007; Wongsuphasawat et al., 2016; Satyanarayan et al., 2017; Key et al., 2012; Vartak et al., 2017; de Sousa and Barbosa, 2014). Likewise, much-related work has sought to evaluate the effectiveness of diverse visualizations (Ondov et al., 2019; Saket et al., 2019; Ware, 2019; Bertin, 1983; Cleveland and McGill, 1984; Heer and Bostock, 2010; Mackinlay et al., 2007; Lee et al., 2017; Kim and Heer, 2018; de Santana et al., 2015, 2017). However, we noticed that they do not focus on the specific effects of data distributions on how these visualizations perform in different situations. We, therefore, arrived at the more specific goal of exploring how data distributions can affect data visualization effectiveness and efficiency.

While looking at related work, we identified two groups of interest. The first one is about studies that define visualization task taxonomies. Amar et al. (2005) defined a set of low-level analysis tasks that people may perform when exploring data. Later studies based on their work became more specific, as is the case of Lee et al. (2017). Based on other taxonomies, they proposed a visualization literacy test, VLAT. It combines visualization type, task, and question, with metrics for task difficulty and discrimination in visualization evaluations.

The second group of interest is studies that evaluate different visualizations' efficiency and/or effectiveness. They have evaluated the effectiveness of different visualizations, but not in the same way. Some are specific to the chart type: bar charts Srinivasan and Stasko (2017); Skau et al. (2015), scatterplots Kim and Heer (2018); Sarikaya and Gleicher (2017), and time series Albers et al. (2014); Heer et al. (2009), for example. Others compare two types of visualization: bar vs. line charts Siegrist (1996), tables vs. pie charts Spence and

Lewandowsky (1991) and bar vs. radar charts Toker et al. (2012). They differ in the set of analyzed charts and tasks and in what they assess: effectiveness, efficiency, or both. We have found no comprehensive research that measured the user's perception of chart-task fit and confidence in answers in empirical evaluations.

Our work extends the literature in various aspects:

(i) we consider more than one visualization type;

(ii) the data set used in our study contains a significant number of data points (3,722), making it more realistic;

(iii) we consider a wider range of visualization tasks; and

(iv) we compare our results to other studies through predefined metrics.

4.1 Goal

Our primary goal in this work is to identify how well some common visualization types support an analyst in answering specific analysis questions given a data set.We seek to identify, for each task (define by an user question), the best visualization type, in terms of effectiveness, time on task, and adequacy to the task, for two types of data distribution.

4.2 Study Design

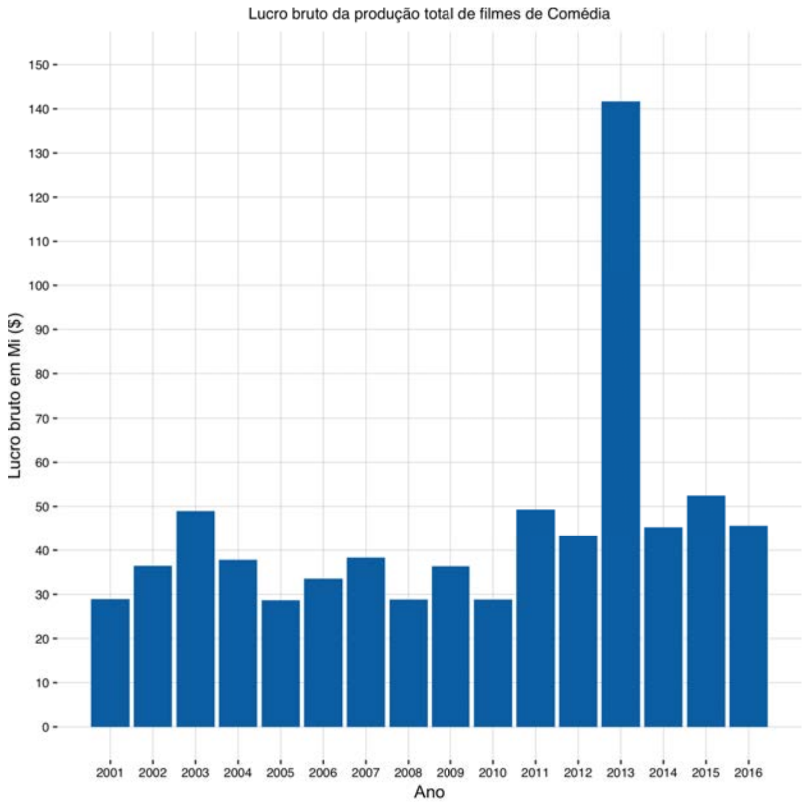
We conducted an empirical study through an online questionnaire. We used nine visualization charts and seven visualization tasks to measure efficiency and effectiveness for two different data sets: one clear of disturbances and a confusing one. The confusing one had distribution disturbances inserted on purpose.

We also measured the user's perception of chart-task fit and confidence in their answers. Participants investigated different chart-task pairings in both clear and confusing distributions in two subsets, so that the evaluation would be less exhausting. Each participant answered the questions, in random order, of one of the subsets. After answering all questions in the first subset, the participant decided whether he/she would like to answer an additional subset of questions, which were also randomized. Table 4.1 exemplifies the task. The remaining visualizations are in appendix C.

We created the charts using the IMDB data set, which comprises music, cinema, and TV series. We selected data from 16 years, almost 4000 records,

Table 4.1: Study 2 task example

Task: Analyze the chart below and answer the questions by ticking a single answer.



When Comedy gross profit was minimal?

( ) None
( ) 2001
( ) 2005
( ) 2008
( ) 2010
( ) 2013
( ) The chart does not allow me to answer

What is your confidence in your answer?

( ) 1 - None	( ) 2	( ) 3	( ) 4	( ) 5	( ) 6	( ) 7 - I'm sure it's right

How good is this visualization to help answer the question?

( ) Terrible	( ) Very good	( ) Bad	( ) Not so bad	( ) Good	( ) Very good	( ) Excellent

and some attributes (like age, genre, rating, for example). We based our choice of chart type and task pairings on VLAT (Lee et al., 2017), and included boxplot, as it has been considered a good visualization tool to analyze data distributions Benjamini (1988). Table 4.2 shows the relationship between the selected charts for each type of visualization task. We did not use georeferenced or hierarchical data, nor the corresponding charts, to keep the questionnaire length reasonable, .

We selected the questions from a previous study we conducted, inspired

Table 4.2: Relationship between selected tasks and visualizations (adapted from Lee et al. (2017))

Task types	**Chart types**
Return value (RV)	Bar, Line, Area, Pie, Stacked bar, Stacked area, Scatterplot, Bubble
Find extremum (FE)	Bar, Line, Area, Pie, Stacked bar, Stacked area, Scatterplot, Bubble
Make comparisons (MC)	Bar, Line, Area, Pie, Stacked bar, Stacked Area, Scatterplot, Bubble
Determining range (DR)	Bar, Line, Stacked area, Scatterplot, Bubble
Find correlation (FC)	Line, Area, Scatterplot, Bubble
Characterize distribution (CD)	Histogram, Boxplot
Find anomalies (FA)	Histogram, Boxplot, Scatterplot, Bubble

by the work of Amar et al. (2005), where we asked a group of people to write questions about the same data set. We obtained a total of 76 questions, and for each task, we chose a representative question from that pool of questions, which we organized according to the tasks shown in table 4.2.

Each visualization concerned its subset of data, which we modified by multiplying them by a random factor and or switching the values randomly between the years. This way, we maintained the distributions' shapes while changing the values. We created a confusing counterpart for each subset by inserting some disturbance in the data, either peaks, gaps, or anomalies:

– Peaks: We randomly chose a variable in the set and increased or decreased its value by 70% of the greatest value in the set, *e.g.*, in bar charts.

– Gaps: We randomly chose a value from the set and removed its n closest data points, *e.g.*, in histograms.

– Anomalies: We randomly added n points (*e.g.*, in histograms and boxplots) with values within [MIN, Q1 - 1.5*IQR] or [Q3 + 1.5*IQR, MAX], where: MIN is the smallest value, MAX is the largest value, Q1 is the value in the first quartile, Q3 is the value in the third quartile, and IQR is the interquartile range.

The relationships we found between the nine chart types we selected for each of the seven visualization tasks in the clear and confusing distributions resulted in seventy-eight question items. The entire selection of tasks, questions, and charts types and resulting visualization is in appendix C.

4.3 Procedure and Participants

We executed our survey through an online questionnaire. After introducing participants to the research and procedure, we asked some profile questions on 7-point Likert scales:

1. Their frequency in creating and analyzing charts

2. Their knowledge about different data concepts

3. And their familiarity with all the types of visualization charts used in our study

The central part of our questionnaire had the following structure. Participants received the resulting visualization and task-related question for each combination of chart type, visualization task, and data distribution and had to choose an answer for the question. For every question, we added a checkbox "The chart does not allow me to answer", to allow us to capture the participants' assessment of the inadequacy of the chart. The task-related questions were mandatory and had a range of possible answer formats: general text field, True/False multiple choice, and non-exclusive multiple choice (with an added option "None"). We also included two Likert statements concerning the confidence level in their answer (to have an approximate guesswork indication), motivated by existing work (Correll and Gleicher, 2014; Hullman et al., 2018), and their perception of the chart type's quality in answering that question. Also, we collected the response times for each question. By collecting these four items, we measured all of the desired aspects of the survey.

Our questionnaire was available to participants for fifteen days. We invited students and researchers from the Department of Informatics at PUC-Rio. We also invited practitioners with experience in chart analysis, using social networks like Facebook and Linkedin. We obtained 119 responses in this period, but only included the 50 complete ones, where participants answered all the questions of at least one of the groups. Participants took, on average, 35 min to answer one group of questions, and 56 min to answer both groups.

Most participants (84%) were between 21 and 30 years old. Twenty-six had Bachelor's degrees, 6 had some specialization, 16 had Master's degrees, and 4 had Doctorate degrees. Only 12% of participants had formal education in the Humanities; all others came from STEM fields, mostly IT or Engineering.

The sample showed that most participants create (56%) and analyze (62%) charts frequently, with few (8%) knowing little to nothing about numeric data distribution concepts. Participants were, in general, familiar with the majority of chart types we used. The best known were Bar, Line, Pie, and Histogram (median M=6, interquartile range IQR=1, in a 1-7 scale), followed by Stacked bars (M=6, IQR=2); Area, Bubble, and Stacked area (M=5, IQR=2); Scatterplot (M=5, IQR=3); and Boxplot (M=4; IQR=2). See fig. 4.1 for a complete analysis.

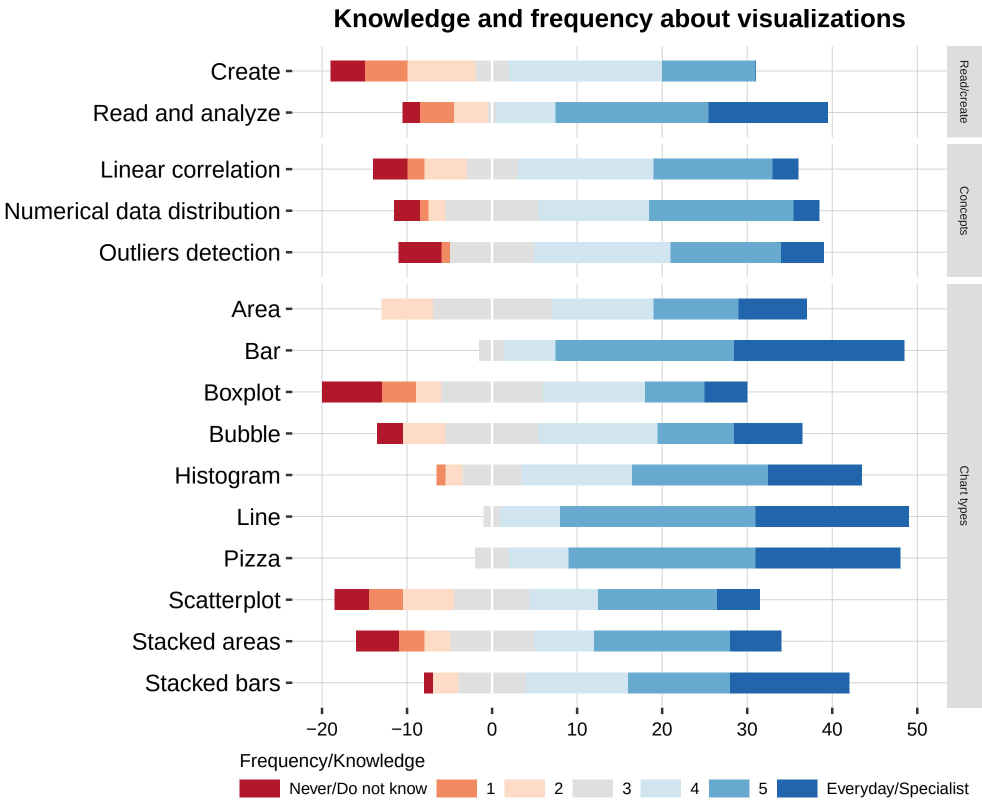


Figure 4.1: Participants' knowledge about visualization types and concepts, and frequency of reading and creating visualizations.

4.4 Analysis and Results

We formulated six hypotheses to aid in the analysis of the results.

H1: The data distribution would affect the charts' effectiveness in fulfilling the tasks.

To evaluate this hypothesis, we conducted, for each task to chart pair, a Fisher's exact test (FET) (Fisher, 1922), assuming an arbitrary threshold of sixty percent correct answers to consider a chart suitable for a specific task. Table 4.3 shows the effectiveness of each chart type for each task, for both clear and confusing distributions. It also shows the p-value resulting from the FET and indicates the test significance.

The Boxplot performed worse than the Histogram in the *Characterize the distribution* task. This was somewhat expected, given the participants'

Table 4.3: Effectiveness (% of correct answers)

Task/Chart type	Clear	Confusing	p-value	Sig
Characterize distribution (CD)				
Histogram	**66.67%**	12.12%	1.02e-05	**
Boxplot	51.52%	3.03%	1.15e-05	**
Determine range (DR)				
Stacked area	**93.94%**	12.12%	6.80e-12	**
Bar	**93.94%**	3.03%	4.98e-15	**
Line	**90.91%**	51.52%	8.11e-04	**
Scatterplot	**81.82%**	42.42%	2.02e-03	**
Bubble	**63.64%**	15.15%	1.12e-04	**
Find anomalies (FA)				
Histogram	**84.85%**	12.12%	2.84e-09	**
Boxplot	**75.76%**	42.42%	1.16e-02	*
Scatterplot	**63.64%**	51.52%	4.55e-01	
Bubble	36.36%	6.06%	5.35e-03	**
Find correlations (FC)				
Line	**89.19%**	0.00%	1.16e-16	**
Area	**72.97%**	18.92%	5.67e-06	**
Bubble	**72.97%**	18.92%	5.67e-06	**
Scatterplot	**72.97%**	5.41%	1.51e-09	**
Find extremum (FE)				
Area	**100.00%**	**75.68%**	2.25e-03	**
Bar	**100.00%**	16.22%	6.98e-15	**
Line	**97.30%**	**67.57%**	1.38e-03	**
Stacked area	**91.89%**	29.73%	4.40e-08	**
Scatterplot	**89.19%**	5.41%	5.64e-14	**
Bubble	**75.68%**	35.14%	9.21e-04	**
Stacked bar	**72.97%**	29.73%	4.07e-04	**
Make comparisons (MC)				
Bar	**81.82%**	0.00%	9.04e-13	**
Line	**81.82%**	0.00%	9.04e-13	**
Stacked bar	**75.76%**	3.03%	5.59e-10	**
Scatterplot	**60.61%**	9.09%	1.91e-05	**
Bubble	54.55%	6.06%	2.83e-05	**
Area	30.30%	0.00%	8.77e-04	**
Stacked area	27.27%	0.00%	2.08e-03	**
Retrieve value (RV)				
Bar	**91.89%**	18.92%	1.43e-10	**
Line	**86.49%**	27.03%	3.58e-07	**
Area	**86.49%**	2.70%	2.84e-14	**
Stacked area	**83.78%**	5.41%	2.75e-12	**
Scatterplot	**75.68%**	16.22%	4.47e-07	**
Bubble	**72.97%**	0.00%	5.93e-12	**
Stacked bar	**64.86%**	13.51%	1.08e-05	**

Fisher's exact test results: ** means $p < 0.01$ and * means $p < 0.05$.

self-reported knowledge levels (fig. 4.1) and given that histograms convey more information than boxplots. The Bubble chart also did not perform well for the task of *Finding anomalies (outliers)*, even in the clear distribution, and even though it had a similar structure as the Scatterplot, plus a third variable (unrelated to the task) mapped onto the size of the bubbles. We hypothesize that the inclusion of visual clutter from the different sizes of the bubbles may have caused the difference in performances between the Scatterplot and the Bubble chart, but this requires further studies. The task of *Making comparisons* also had some ineffective chart types: Bubble, Area, and Stacked Area. Analyzing the comments, we have identified that people confused the Area and Stacked area charts: "*I cannot even tell whether it is stacked or whether the vertical value starts from the horizontal axis.*"

Some chart types did not perform well, even with the clear distribution (fig. 4.2, *e.g.*, Bubble chart for *Make Comparisons*). We also found significant differences in all cases except one: the scatter plot for **Finding anomalies**.

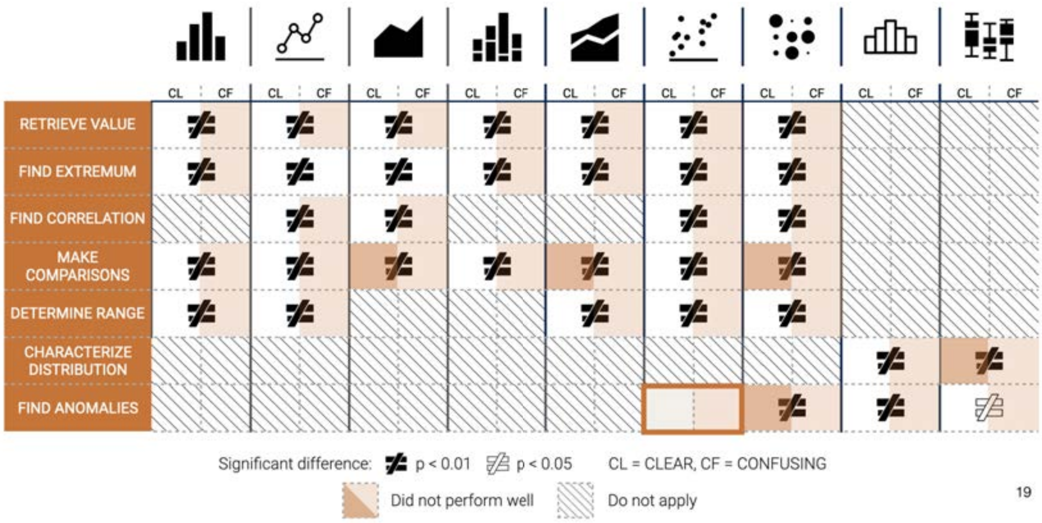


Figure 4.2: Chart effectiveness per task

H2: The task duration did not have a normal distribution.

To compare response time for correct answers across distributions, we used a Mann-Whitney test because the task duration was not normally distributed. The response times did not differ significantly. In three cases, participants took significantly longer to provide a correct answer with the confusing distribution: Scatterplot for *Determine range*, Area chart for *Find correlation*, and Bubble plot and Line chart for *Find extremum*:

- DR, Scatterplot: $M_{cl} = 29.5, M_{co} = 54, p = 4.42e^{-03}$

- FC, Area: $M_{cl} = 24, M_{co} = 46, p = 4.03e^{-02}$

- FE, Bubble: $M_{cl} = 34.5, M_{co} = 52, p = 2.02e^{-02}$
- FE, Line: $M_{cl} = 23, M_{co} = 36, p = 1.69e^{-02}$

H3: There is a significant difference in user preference between each pair for charts with similar effectiveness.

We ran Mann-Whitney tests to analyze users' preferences. We found significant differences in user ratings about how adequately each chart answered their related questions. For the *Determine range task*, the Stacked Area and Bar charts had the same percentage of correct answers, but Bar charts received higher adequacy ratings with $p = 0.002$. The same was true in the *Find correlations* task in regards to Area charts and Scatter plots, although they had the same percentage of correct answers. The area received higher ratings than Scatterplot, with $p = 0.040$. For *Make comparisons*, Bar received higher ratings than Pie, with $p = 0.015$.

H4: Prior knowledge is significant related to the response confidence rating.

We ran Mann-Whitney tests concerning the self-reported knowledge about each chart to each answer's rating and confidence level, as these measures have ordinal scales. Table 4.4 shows the results. In this table, y represents the correct answers, n the incorrect. Also, p is the $p\!-\!value$, and s is the significance.

The higher prior knowledge was related to getting the answer correct in only three cases, all with the confusing distribution: Boxplot and Scatterplot for *Finding anomalies* ($p = 0.018$ and $p = 0.008$) and Stacked area for *Finding extremum* ($p = 0.042$).

Table 4.4: Correctness x Rating x Confidence x Knowledge

		Knowledge				**Rating**				**Confidence level**			
Task/Chart type	**Distribution**	**Ky**	**Kn**	**Kp**	**Ks**	**Ry**	**Rn**	**Rp**	**Rs**	**Cy**	**Cn**	**Cp**	**Cs**
Characterize distribution													
Boxplot	cl	5.0	4.0	0.070		4.0	3.5	0.025	*	5.0	4.0	0.377	
Boxplot	cf	2.0	4.0	-		4.0	4.0	-		4.0	4.0	-	
Histogram	cl	5.5	5.0	1.000		5.0	4.0	0.117		5.0	4.0	0.096	
Histogram	cf	6.0	5.0	0.588		5.5	5.0	0.261		4.5	5.0	0.255	
Determine range													
Bar	cl	6.0	6.5	-		6.0	3.0	-		7.0	3.0	-	
Bar	cf	7.0	6.0	-		3.0	3.0	-		4.0	5.0	-	
Bubble	cl	5.0	5.0	0.773		5.0	1.0	1.23e-04	**	6.0	5.5	0.508	
Bubble	cf	6.0	5.0	0.115		2.0	2.0	0.854		4.0	4.0	0.370	
Line	cl	6.0	7.0	0.711		5.0	4.0	0.120		6.0	4.0	0.192	
Line	cf	6.0	6.0	0.684		3.0	3.0	0.839		4.0	4.5	0.794	
Scatterplot	cl	5.0	4.5	0.757		4.0	1.5	0.001	**	6.0	2.5	0.007	**
Scatterplot	cf	4.5	5.0	0.753		3.0	2.0	0.060		4.0	4.0	0.294	
Stacked area	cl	5.0	3.0	-		5.0	4.5	-		6.0	6.0	-	
Stacked area	cf	3.5	5.0	0.283		5.0	4.0	0.073		5.0	5.0	0.819	
Find anomalies													
Boxplot	cl	5.0	3.0	0.076		6.0	4.0	0.029	*	6.0	6.0	0.739	
Boxplot	cf	5.0	4.0	0.018	*	5.5	4.0	0.008	**	5.5	4.0	0.085	
Bubble	cl	5.0	5.0	0.803		4.5	4.0	0.023	*	5.0	4.0	0.065	
Bubble	cf	6.5	5.0	-		1.5	4.0	-		1.0	5.0	-	
Histogram	cl	5.5	5.0	0.938		5.0	1.0	0.001	**	6.0	4.0	0.100	
Histogram	cf	5.5	5.0	0.932		4.5	4.0	0.292		4.5	4.0	1.000	
Scatterplot	cl	5.0	4.0	0.261		5.0	4.5	0.127		6.0	4.5	0.018	*
Scatterplot	cf	5.0	3.5	0.008	**	5.0	3.5	0.167		5.0	5.0	0.393	

Table 4.4: Correctness x Rating x Confidence x Knowledge, continued

		Knowledge				Rating				Confidence level			
Task/Chart type	**Distribution**	**Ky**	**Kn**	**K*p***	**Ks**	**Ry**	**Rn**	**R*p***	**Rs**	**Cy**	**Cn**	**C*p***	**Cs**
Find correlations or trends													
Area	cl	5.0	4.5	0.370		5.0	3.5	0.001	**	6.0	4.0	0.005	**
Area	cf	6.0	5.0	0.264		5.0	4.0	0.444		5.0	5.0	0.769	
Bubble	cl	5.0	4.0	0.220		6.0	3.0	0.001	**	6.0	4.5	0.002	**
Bubble	cf	5.0	5.0	1.000		4.0	3.0	0.566		4.0	5.0	0.068	
Line	cl	6.0	5.5	0.250		5.0	3.5	0.012	*	6.0	4.5	0.014	*
Line	cf	-	6.0	-		-	5.0	-		-	5.0	-	
Scatterplot	cl	4.0	3.5	0.251		5.0	4.0	0.001	**	6.0	4.0	0.001	**
Scatterplot	cf	5.0	4.0	-		3.5	5.0	-		3.5	5.0	-	
Find extremum													
Area	cl	-	5.0	-		-	6.0	-		-	7.0	-	
Area	cf	5.0	5.0	0.344		4.0	3.0	0.010	**	5.0	6.0	0.262	
Bar	cl	-	6.0	-		-	6.0	-		-	7.0	-	
Bar	cf	6.0	6.0	0.877		4.0	5.0	0.068		4.0	7.0	0.002	**
Bubble	cl	5.0	4.0	0.677		4.0	4.0	0.520		6.0	5.0	0.035	*
Bubble	cf	4.0	5.0	0.055		3.0	2.5	0.987		4.0	4.5	0.176	
Line	cl	6.0	7.0	-		6.0	6.0	-		7.0	7.0	-	
Line	cf	6.0	6.0	0.876		4.0	4.0	0.729		5.0	4.5	0.481	
Pie	cl	6.0	7.0	0.238		5.0	6.0	0.474		7.0	7.0	0.882	
Pie	cf	5.0	6.0	-		3.0	2.0	-		4.5	6.0	-	
Scatterplot	cl	4.0	6.0	0.156		4.0	2.0	0.012	*	6.0	4.0	0.004	**
Scatterplot	cf	5.0	4.0	-		2.0	2.0	-		2.5	4.0	-	
Stacked area	cl	5.0	4.0	0.733		5.0	4.0	0.115		6.0	4.0	0.010	*
Stacked area	cf	6.0	4.0	0.042	*	4.0	3.0	0.384		4.0	5.0	0.103	
Stacked bar	cl	6.0	5.0	0.400		4.0	3.0	0.055		5.0	4.5	0.297	
Stacked bar	cf	6.0	5.0	1.000		3.0	3.0	0.865		4.0	4.0	0.433	
Make comparisons													
Area	cl	5.0	5.0	0.936		4.0	5.0	0.499		6.0	5.0	0.243	
Area	cf	-	5.0	-		-	5.0	-		-	5.0	-	
Bar	cl	6.0	5.5	0.094		6.0	6.0	0.498		7.0	7.0	0.597	
Bar	cf	-	6.0	-		-	6.0	-		-	7.0	-	
Bubble	cl	5.0	4.0	0.091		6.5	2.0	1.34e-04	**	7.0	4.0	1.56e-04	**
Bubble	cf	5.0	5.0	-		2.0	2.0	-		4.5	4.0	-	
Line	cl	6.0	6.0	0.248		7.0	5.0	0.062		7.0	6.5	0.481	
Line	cf	-	6.0	-		-	6.0	-		-	7.0	-	
Pie	cl	6.0	6.5	0.651		6.0	5.0	0.209		7.0	7.0	0.840	
Pie	cf	-	6.0	-		-	5.0	-		-	7.0	-	
Scatterplot	cl	5.0	4.0	0.103		6.0	3.0	0.001	**	7.0	5.0	0.002	**
Scatterplot	cf	6.0	4.0	0.104		3.0	2.0	0.895		4.0	4.0	0.329	
Stacked area	cl	5.0	5.0	0.605		4.0	5.0	0.635		6.0	6.5	0.635	
Stacked area	cf	-	5.0	-		-	5.0	-		-	5.0	-	
Stacked bar	cl	6.0	5.0	0.476		5.0	5.5	0.121		7.0	6.5	0.711	
Stacked bar	cf	5.0	5.5	-		3.0	4.0	-		3.0	6.0	-	
Retrieve value													
Area	cl	5.0	6.0	0.075		5.0	6.0	0.619		6.0	6.0	0.811	
Area	cf	5.0	5.0	-		3.0	2.5	-		5.0	5.0	-	
Bar	cl	6.0	6.0	0.145		5.0	5.0	0.842		7.0	6.0	0.269	
Bar	cf	7.0	6.0	0.221		4.0	3.0	0.014	*	5.0	5.0	0.593	
Bubble	cl	5.0	5.0	0.806		5.0	2.5	0.017	*	6.0	5.5	0.451	
Bubble	cf	-	5.0	-		-	2.0	-		-	4.0	-	
Line	cl	6.0	7.0	0.406		5.0	2.0	0.002	**	6.5	4.0	0.142	
Line	cf	6.0	6.0	0.315		4.0	4.0	0.179		5.0	6.0	0.345	
Pie	cl	6.0	6.5	0.600		5.0	6.0	0.966		7.0	7.0	0.089	
Pie	cf	6.0	6.0	0.518		3.0	3.0	0.243		6.0	6.0	0.668	
Scatterplot	cl	4.5	4.0	0.773		5.0	3.0	0.038	*	6.0	5.0	0.542	
Scatterplot	cf	5.5	4.0	0.068		3.0	3.0	0.800		5.0	5.0	0.501	
Stacked area	cl	5.0	4.5	0.515		4.0	3.0	0.140		6.0	6.5	0.392	
Stacked area	cf	5.0	5.0	-		2.5	2.0	-		5.0	4.0	-	
Stacked bar	cl	5.5	5.0	0.794		4.0	4.0	0.200		6.0	4.0	0.165	
Stacked bar	cf	5.0	5.5	0.699		2.0	2.5	0.436		3.0	5.0	0.288	

H5: The confidence level in the answer is directly related to its assertiveness.

Regarding the **participants' confidence level**, we analyzed the data from all distributions of three groups: participants who answered correctly (y), incorrectly (n), and who stated the chart did not answer the question (dna), see table 4.4. We found a significant difference with a Kruskal-Wallis test ($p = 4.06e - 63$), so we ran post-hoc Mann-Whitney tests with Bonferroni correction, and found a significant difference in all three cases:

- $y \times n$: $M_y = 6, M_n = 5, p = 3.98e^{-34}$
- $y \times dna$: $M_y = 6, M_{dna} = 4, p = 1.12e^{-31}$
- $n \times dna$: $M_n = 5, M_{dna} = 4, p = 2.88e^{-04}$

It is worth noting that the lowest confidence level occurred when the participants believed the chart did not answer the question, even lower than when they got the answer wrong. By contrast, when analyzing the confusing distribution alone, there are only significant differences between the groups (y,n), but not between (y,dna) nor (n,dna). In other words, the underlying data distribution affected the participants' confidence level in their answers.
H6: The rating in the answer is directly related to its assertiveness.

Regarding the **participants' rating** - a scale for the charts' suitability in helping to answer the question, we performed an analogous analysis, with similar results. When analyzing data from all distributions, we found a significant difference with a Kruskal-Wallis test ($p = 1.20e - 128$), so we ran post-hoc Mann-Whitney tests with Bonferroni correction, and found a significant difference in all three cases:

- $y \times n$: $M_y = 5, M_n = 4, p = 6.04e^{-12}$
- $y \times dna$: $M_y = 5, M_{dna} = 2, p = 3.92e^{-145}$
- $n \times dna$: $M_n = 4, M_{dna} = 2, p = 2.19e^{-94}$

As expected, when the participants believed the chart did not answer the question, they rated it as inadequate (median 2 in a 1-7 scale). Similar results were found when analyzing the clear and the confusing distributions separately, with significant differences in all three pairs of comparisons.

4.4.1 Chart Ranking According to Task

In an attempt to rank the charts in terms of effectiveness for the same task, we compared the percentage of correct answers across pairs of charts. Table 4.5 shows the result for significant cases for clear, confusing, and all distributions.

Analyzing the *clear* distribution, for *Determine range*, Bar, Line, and Stacked area were all better than Bubble; for *Find anomalies*, Boxplot and Histogram were better than Bubble; for *Find extremum*, Area, Bar, and Line were better than Bubble or Stacked bar; for *Make comparisons*, Bar, Line, and

Table 4.5: Comparing chart effectiveness

Better	**Worse**	**All**	**Clear**	**Confusing**
Determine range (DR)				
Bar	Bubble		**5.35e-03**	
Line	Bubble	**4.63e-04**	1.69e-02	
Stacked area	Bubble		**5.35e-03**	
Find anomalies (FA)				
Boxplot	Bubble		**2.92e-03**	
Histogram	Bubble		**1.58e-04**	
Find extremum (FE)				
Area	Bar	**1.02e-04**		**9.66e-07**
Area	Bubble	**2.75e-05**	**2.25e-03**	**1.06e-03**
Area	Scatter	**3.54e-07**		**3.59e-10**
Area	Stacked area	**3.50e-04**		**1.95e-04**
Area	Stacked bar	**3.38e-06**	**9.70e-04**	**1.95e-04**
Bar	Bubble		**2.25e-03**	
Bar	Stacked bar		**9.70e-04**	
Line	Bar	**2.23e-03**		**2.22e-05**
Line	Bubble	**7.40e-04**	1.38e-02	1.05e-02
Line	Scatter	**1.67e-05**		**2.16e-08**
Line	Stacked area	**6.24e-03**		**2.50e-03**
Line	Stacked bar	**1.22e-04**	**6.58e-03**	**2.50e-03**
Make comparisons (MC)				
Bar	Area		**7.24e-05**	
Bar	Stacked area		**2.64e-05**	
Line	Area		**7.24e-05**	
Line	Stacked area		**2.64e-05**	
Stacked bar	Area		**5.55e-04**	
Stacked bar	Stacked area		**2.20e-04**	
Retrieve value (RV)				
Bar	Stacked bar		**9.55e-03**	

Stacked Bar were better than Area or Stacked area; and for *Retrieve Value*, Bar was better than Bubble and Stacked bar.

When we analyze the cases independent of the distribution (column *all*), we see a different picture. This means that a chart that works for *clear* distributions may not work as well for any distribution. Figure 4.3 shows the result for significant cases of clear, confusing, or disregarding distribution. The checkpoints show whether charts in the columns were significantly better than those in the rows, also greyed out.

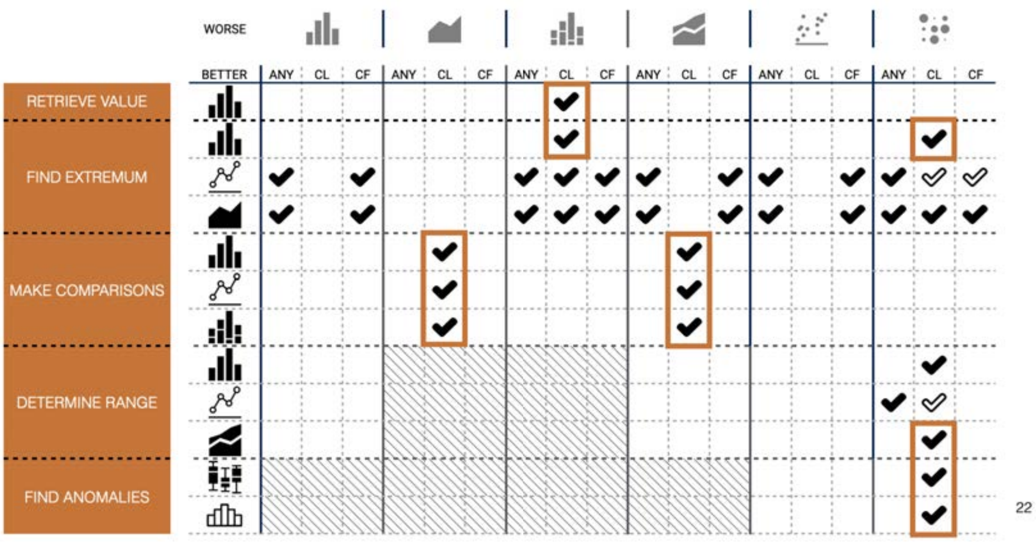


Figure 4.3: Pairwise comparison of chart effectiveness for each task

Table 4.6: Chart types ranking according to effectiveness

Task	**Consider using**	**Avoid**
Characterize distribution	Histogram	
Determine range	Stacked area, Bar, Line	Bubble
Find anomalies	Histogram, Boxplot, Scatterplot	Bubble
Find correlations	Line, Area, Bubble, Scatterplot	
Find extremum	Area, Bar, Line, Stacked area, Scatterplot	Bubble, Stacked bar
Make comparisons	Bar, Line,	Area, Stacked area, Bubble
Retrieve value	Bar, Line, Area, Stacked area, Scatterplot	Bubble, Stacked bar

After these analyses, we can recommend chart types for each visualization task, as shown in table 4.6. The charts in the *avoid* column fared significantly worse than their counterparts. It is important to note that here we have focused only on the data distributions. Other characteristics may influence the chart choice, as extensively discussed elsewhere (*e.g.*, (Chambers et al., 1983; Tufte, 2001; Few, 2009; Cairo, 2016)).

We could relate the prior knowledge in getting the answer correct in only three cases, all with the confusing distribution: Boxplot and Scatterplot for Find anomalies, and Stacked area chart for Find Extremum. We found a significant difference in participants' confidence levels and chart-task fit rating in all three pairwise combinations of an answer: correct, incorrect, and the chart does not answer. Also, the lowest confidence level occurred when the participants believed the chart did not answer the question. And, as expected, they rated it as inadequate.

4.5 Discussion

Saket et al. (2019) evaluated the effectiveness of five types of bi-dimensional visualizations for two different data sets and ten analysis tasks. They chose line charts, bar charts, scatter plots, and pie charts. The tasks were: find anomalies, find clusters, find correlation, compute derived value, characterize distribution, find extremum, filter, order, determine range, and retrieve value. They conducted an empirical study with 180 participants. Each participant was randomly assigned an analysis task. Participants answered 30 questions each (5 Visualizations $\times$ 2 Datasets $\times$ 3 Trials), plus two other questions to detect guessing. Based on the results of their study, they defined five guidelines to help choose which visualization type to use, based on time on task, accuracy (effectiveness) and user preferences (rating). They were:

G1: Use bar charts for finding clusters;

G2: Use line charts for finding correlations;

G3: Use scatterplots for finding anomalies;

G4: Avoid line charts for tasks that require readers to precisely identify the value of a specific data point;

G5: Avoid using tables and pie charts for correlation tasks.

In their study, Line chart performed better than Scatterplot in all measured variables (G2). However, we did not find significant differences between Line and Scatterplot, in any distribution, regarding either accuracy or user preference. Regarding time on task, our results go in the opposite direction:

Scatterplot performed significantly better than Line. This discrepancy suggests that additional studies need to be conducted to further explore these charts.

In contrast to G3, in our study Histograms and Boxplots were more effective than Scatterplots, although Scatterplots were more effective ($p = 4.19e^{05}$) and were rated higher ($p = 0.004$) than Bubble charts, regardless of the distribution. There was no significant difference on time on task between Scatterplots and the other charts for this task.

Despite G4, in our study, Line charts were highly effective for all the tasks in which they were tested (DR, FC, FE, MC, RV) with the clear distribution. Moreover, our pairwise comparison of effectiveness shows that Line is significantly better than many other charts type for *Find extremum* and *Make comparisons*, regardless the distribution. Regarding user preferences, Line received significant higher rating in several cases, for example, compared to Scatterplot ($p = 0.045$), Stacked area ($p = 0.02$), and Stacked bar ($p = 0.04$), for *Retrieve value*, regardless the distribution. Our study showed no significant difference in time on task involving Line charts.

Guidelines G1 and G5 lie outside the scope of our work, because we did not investigate the *Finding clusters* task, nor did we use Tables or Pie charts for *Finding correlation*.

4.6
Concluding Remarks

This chapter introduced our empirical study to assess the effectiveness (accuracy), efficiency (time on task) and user preference (rating) to identify which types of visualization better support specific visualization tasks. We used seven different tasks, ten chart types, and two variations of a data set (clear and confusing distributions). We set out to verify whether and how data distribution affects participants' answers for each <task, chart type, distribution>.

Comparing the results of the two types of distribution, we verified that there is a significant difference in effectiveness in all cases except one: Scatterplot for *Finding anomalies* which, although it had a good result with the clear distribution, the difference was not significant. For *Finding extremum*, although Area and Line charts had significantly different effectiveness across distributions, in both cases their effectiveness was deemed good ($\geq 60\%$).

With this study, we were able to identify some charts that perform better according to the task, regardless of the distribution (fig. 4.4). Our results show that Area charts and Scatterplots are good to *Find correlations*, but people prefer Area charts over Scatterplots for this task. For *Make comparisons*, Bar was the most effective chart.

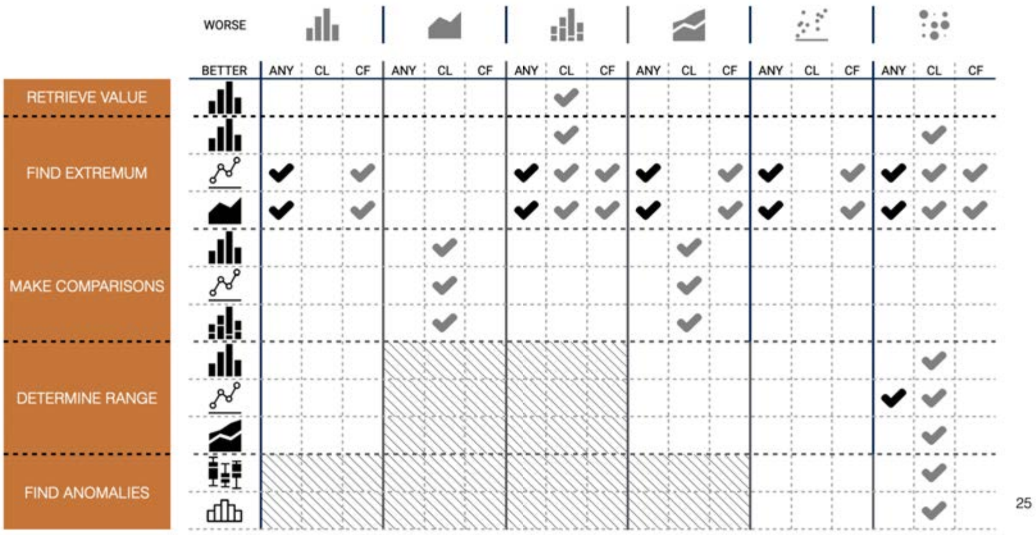


Figure 4.4: Charts performance: accordingly to the task, regardless of the data distribution

We also identified that the Bubble chart is not recommended for *Retrieve value* when the analysis variable is mapped onto the size of the bubble. Likewise, Stacked area is not recommended for *Make comparisons* when the analysis category is not close to the axis. Moreover, participants could not *Find extremum* using Scatterplots with non-proportional axes.

This work cannot assume that, since the questionnaire was online and responded without supervision, the participants correctly used the option to pause the study whenever necessary to interrupt the task without distorting the time data. Ideally, this type of data should be measured in supervised tests, whether online through videoconference or in person.

Most results pointed to a significant difference between effectiveness, confidence, and rating across distribution (clear *vs.* confusing). This calls for further comprehensive studies, as well as combining different disturbances in each pair <task, chart>, to derive more fine-grained recommendations.

The paper "Comparing the effectiveness of visualizations of different data distributions", published at SIBGRAPI 2019 (Rodrigues et al., 2019), describes this study in detail.

5 Uncovering whether people identify whether a data visualization is suitable for answering an analysis question

Choosing the type of chart to represent the data can go beyond the combination of data type and analysis task. Abela (2008) suggested a diagram for choosing charts based on the type of data (number of variables and type of variables) and the message to be conveyed (comparison, relationship, distribution, and composition), see fig. 5.1 for reference.

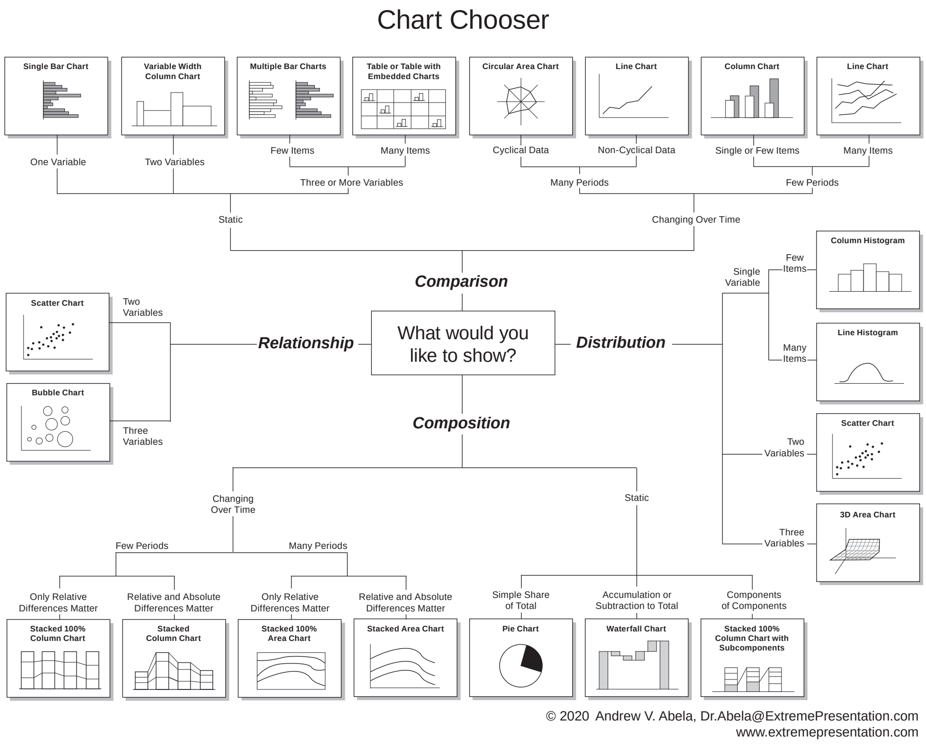


Figure 5.1: Chart Chooser proposed by Abela (2008).

However, it is possible to notice that this diagram can be very generalized and allow not-so-adequate choices. For example, the diagram suggests a multiple line chart to compare a few periods and many categories over time. But if the analyst's goal is to compare intersections between curves and the chart has many curves, this is not ideal. In this case, it might be better to use small multiples of the line chart: there would no longer be a tangle of lines, allowing for better comparisons (Tufte et al., 1990).

Abela's diagram is just one of many other diagrams and catalogs found in books and websites on data visualization. [1] The previous example falls short because it may not have considered the analyst's goal as expressed by the following question: "In which period was there the minimum sales drop in all sectors?" From this example, we see that visualization becomes more or less efficient depending on the analysis question.

5.1 Goal

We conducted a study to investigate whether people can identify when a visualization is suitable for answering a particular analysis question. More specifically, we investigated whether the analyst can:

- realize that the visualization is not suitable;
- suggest changes that make it suitable;
- assess whether or not some given suggestions make it more suitable.

We also evaluated whether suggestions provided in text only are assessed in the same way as suggestions coupled with the corresponding visualizations, *i.e.*, whether participants could foresee the results of applying a recommendation or they relied on the concrete image of the visualization to understand the recommendation.

5.2 Study Design

The survey had 31 pages: an introductory page contained ethical information about the research and the request for the participants to consent to the terms regarding the use of the data they would provide. The second page presented a set of demographic questions about the participant. The third page was read-only, with a description of the tasks that the participant should perform. Then, we presented the main study tasks and questions on the following pages: 4-29 (appendix D). Finally, the last page showed two more questions about the impact of the suggestions on answers' choice and an open text field for comments. The questionnaire was built in Portuguese, but in this text we present translations of the relevant texts to English.

We split the main study questions into three parts (fig. 5.2), the first two with similar tasks. In the first part, we showed a bar chart ordered by category

[1] https://datavizcatalogue.com/, https://datavizproject.com/, https://www.data-to-viz.com/ - last visited, February 2022

to answer an analysis question that would be more easily answered with a frequency-ordered bar chart. We then presented a series of questions related to this issue and a set of explanations about the problems and suggestions related to ways to improve or design a better visualization for answering the proposed question. For brevity, the set of short explanations and general recommendations will be henceforth called tutorial.

In the second part, we showed a grouped bar chart to answer an analysis question that would be more easily answered with a line chart. Once again, we present several questions related to this issue and a tutorial explaining the problems and suggestions for improvement for each case.

In the final part, we presented a bar chart ordered by frequency to answer an analysis question that would be more easily answered with a bar chart ordered by category. We also presented a line chart and an analysis question more easily answered with either a grouped bar chart or small multiples. We hypothesized that people could associate the problem and suggestion explored in the first two parts of the study with the final charts.

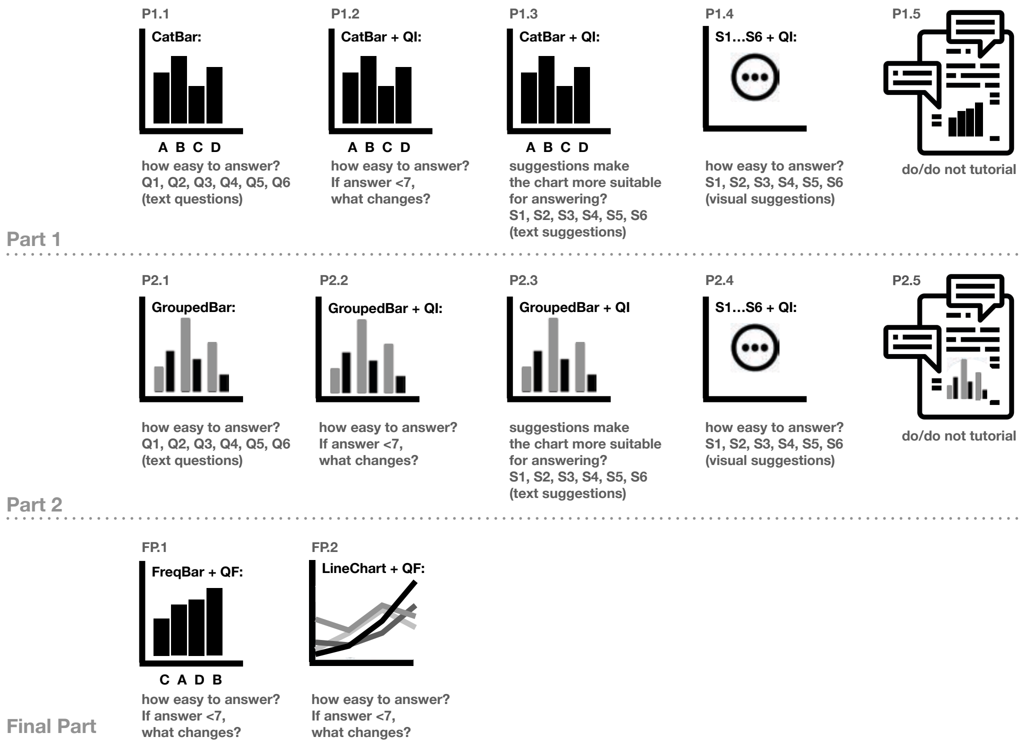


Figure 5.2: Study Design Overview.

We presented a set of questions about one chart (P1.*) and then the same set of questions (with reasonable modifications) about another chart (P2.*). We wanted to investigate the range of answers by participants when we ask them to:

- assess which questions can be answered well with a chart (P1.1 and P2.1),
- assess and make suggestions in free text (P1.2 and P2.2),
- evaluate suggestions presented in text form (P1.3 and P2.3), and
- evaluate suggestions presented in image form (P1.4 and P2.4).

Also, we investigated whether a tutorial on visualization suggestions can improve or hinder the proposition of suggestions (final part). We used a commercial analysis tool, Tableau (2003), to generate all charts in this study. It suggests visualizations when we load a database and choose analysis variables. In the tool's "Show me" panel, we choose some suggested visualizations. Tableau does not consider the analysis question, only the data and the variable types selected for the analysis. Its suggestions are compatible with the data but may not be good options for analyzing a certain question. Our aim is to investigate whether participants rate the suggestions as appropriate for the analysis questions we chose in our study.

5.2.1 Part 1

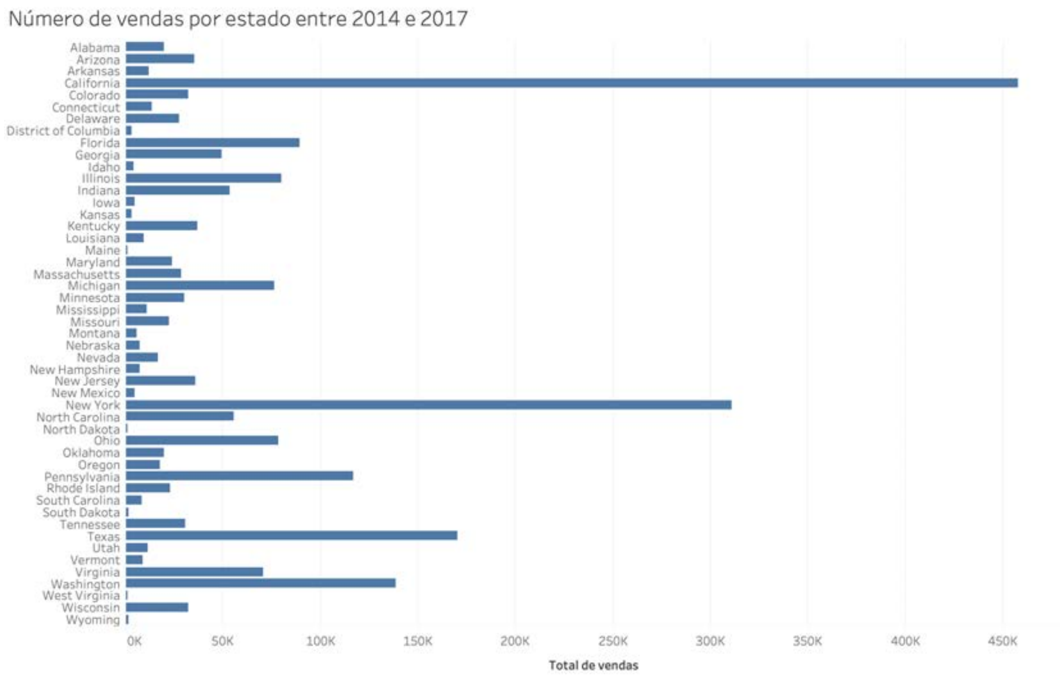
We created a bar chart ordered by category to answer an analysis question that would be more easily answered with a bar chart ordered by frequency.

P1.1: We showed the participant a data visualization and six data analysis questions to assess their agreement on how well they could answer them using the chart (see table 5.1). We chose five questions from Lee et al. (2017), specific for bar chart analysis(#1, 2, 4, 5, 6), and one question from Rodrigues et al. (2021) (#3) for bar charts ordered by frequency. We also used the same visualization in the following steps (P1.2, and P1.3). On a scale of 1 to 7 (impossible to trivial), we asked participants how easy would it be for them to answer each of the six data analysis questions with the suggested data visualization (P1.1.1-P1.1.6). This step aimed to identify whether the participants' assessment of which questions are easy (and how easy) to answer using data visualization are aligned with the data visualization literature.

P1.2: We used the same visualization as P1.1 and one of the analysis questions in this task (see table 5.2). This visualization, although adequate, is not suitable for the participant to answer the analysis question without much effort. In this step, we asked how adequate the participant considers the visualization for answering the question on a scale from 1 to 7 (totally inadequate to totally adequate). If they choose any option other than 7, we present an open-ended question for the participant to know what improvements they suggest to make the visualization suitable for answering the question. This

Table 5.1: Study 3 task example: P1.1

Task: A data visualization software suggested the following visualization after you selected two variables types: a nominal one on the Y-axis and a quantitative one on the X-axis.



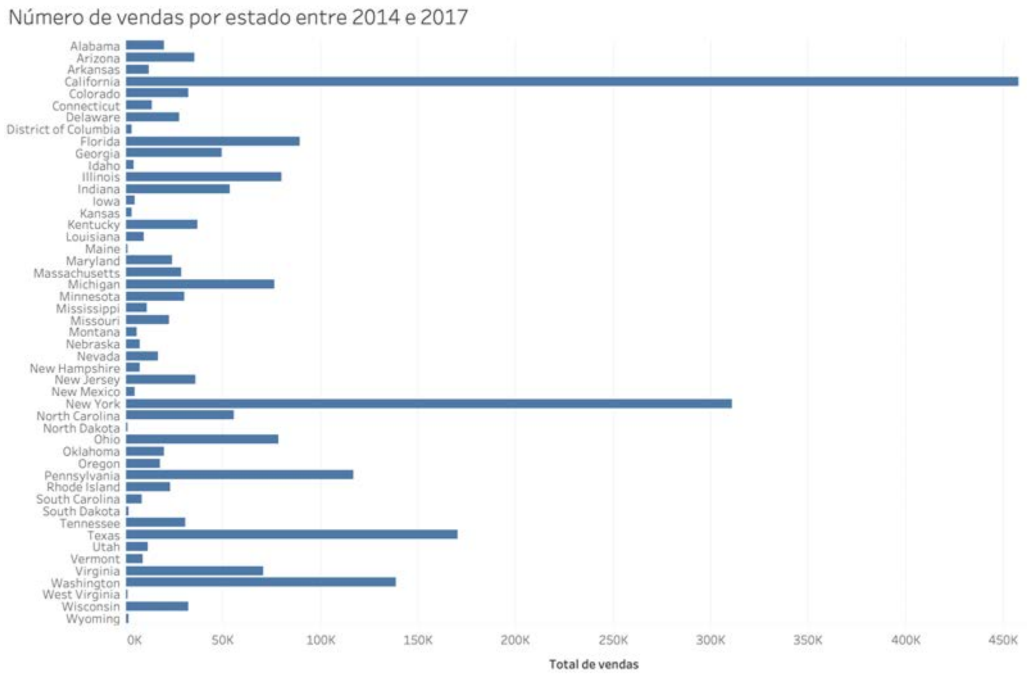
Rate how easy you find answering each question below using the visualization.	1 - impossible	2	3	4	5	6	7 - trivial
1. Did North Dakoka have more sales than Maine?	( )	( )	( )	( )	( )	( )	( )
2. Which states are below average sales?	( )	( )	( )	( )	( )	( )	( )
3. What are the top-10 states with the most sales?	( )	( )	( )	( )	( )	( )	( )
4. What is the number of sales range between 2014 and 2017?	( )	( )	( )	( )	( )	( )	( )
5. What is California's sales number?	( )	( )	( )	( )	( )	( )	( )
6. How many states have surpassed Illinois in sales?	( )	( )	( )	( )	( )	( )	( )

stage aims to identify whether the participant can identify flaws and propose improvements to solve them.

P1.3: Using the same data visualization and analysis question (*What are the top-10 states with the most sales?*), we presented participants with six suggestions of changes for improvement in random order (table 5.3) and, for each one, we asked how much the participant thought they made the chart

Table 5.2: Study 3 task example: P1.2

Task: Using the same visualization suggested by the data visualization software, answer the question by ticking a single answer:



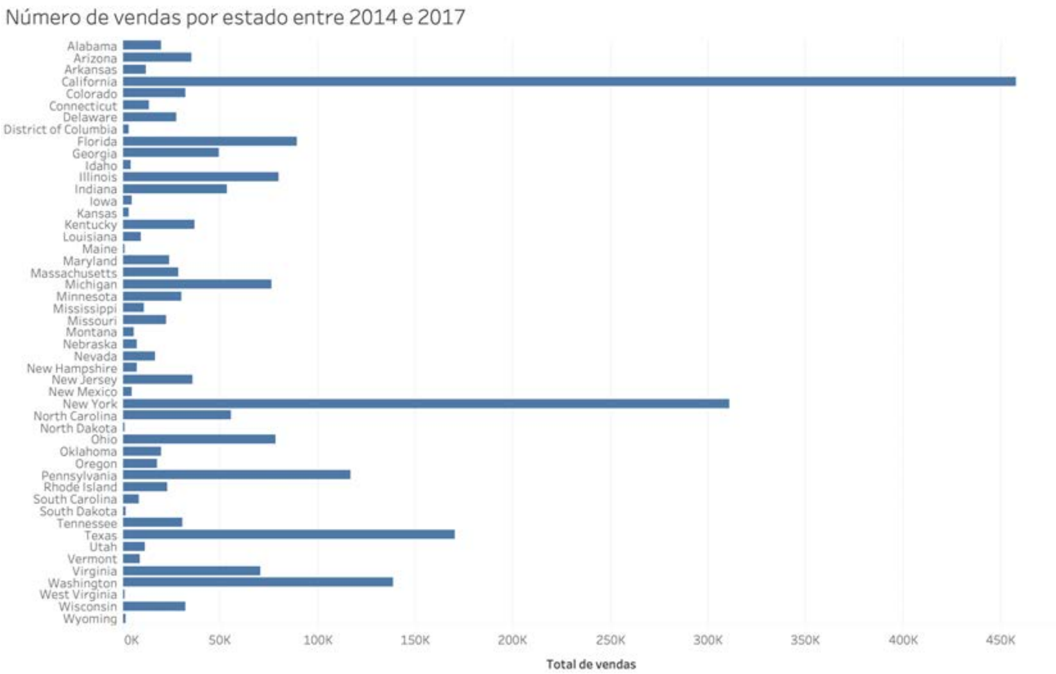
Since you would like to visually investigate the following question: "What are the top-10 states with the most sales?", how much do you consider this visualization suitable to answer it?

( ) 1 - totally inadequate ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - totally adequate

more suitable for answering the question, on a scale from 1 to 7 (totally disagree to totally agree)(P1.3.1-P1.3.6). We included three good options among them (marked with an "*"), one being much better than the others (bold) and two options that improved the situation but not as much. At this step, we presented the suggestions to the participant in a textual form only. This step aims to identify to what extent the participant can identify good and bad solutions presented in textual form.

Table 5.3: Study 3 task example: P1.3

Task: Analyze the question and the suggested visualization again.
What are the top-10 states with the most sales?



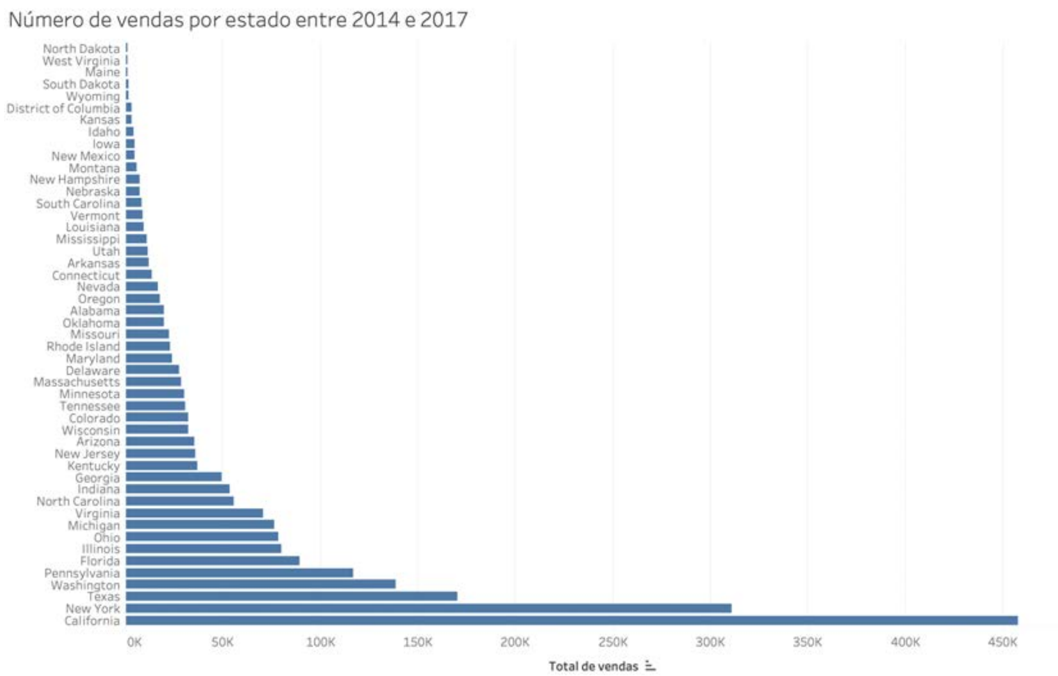
Rate your agreement for each of the following recommendations. To make the visualization more suitable for answering the question, I:

	1 - totally disagree	2	3	4	5	6	7 - totally agree
1.* I would keep the chart type (bar) but I would order the bars by the number of sales.	( )	( )	( )	( )	( )	( )	( )
2.* I would keep the chart type (bar), but I would highlight top-10 states bars with the highest number of sales.	( )	( )	( )	( )	( )	( )	( )
3.* I would keep the chart type (bars) but I would include the sales value in each bar.	( )	( )	( )	( )	( )	( )	( )
4. I would keep the chart type (bar) but I would use a different color scale (blue, yellow, ...) for each state.	( )	( )	( )	( )	( )	( )	( )
5. I would change the chart type to pie chart, where each slice would be a state, ordered by state name.	( )	( )	( )	( )	( )	( )	( )
6. I would change the chart to a table, ordered by the states name, coloring the cells with an intensity gradient according to the value of the total sales in each one.	( )	( )	( )	( )	( )	( )	( )

P1.4: We randomly presented each of the six suggestions from the previous step again, but this time in visual form (table 5.4 presents the first one, the others are in Appendix D, Part 1 - questions 4a-f). We asked participants how well each chart answers the analysis question on a scale from 1 to 7 (very poorly to very well)(P1.4.1-P1.4.6). This step aims to identify whether the participant can identify good and bad solutions when presented visually.

Table 5.4: Study 3 task example: P1.4

Task: Analyze the following visualization to answer the question: What are the top-10 states with the most sales?



This chart answers the question:
( ) 1 - very poorly ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - very well

P1.5: We provided a tutorial explaining how each visualization from step 3 or 4 helps or hinders answering the analysis question.

5.2.2
Part 2

We selected a grouped bar chart to answer an analysis question that would be more easily answered with a line chart. We repeated the same steps (1 to 5) from Part 1.

P2.1: We created the visualization represented in Figure 5.3 and the questions in Table 5.5 for this task (rate how easy would it be for them to answer each of the six data analysis questions with the suggested data visualization)(P2.1.1-P2.1.6).

P2.2: The question chosen for this task was Table 5.5 - 4, also from Rodrigues et al. (2021), and the visualization is the same as P2.1 (Figure 5.3). Again, we asked participants to rate how adequate they consider the visualization for answering the question.

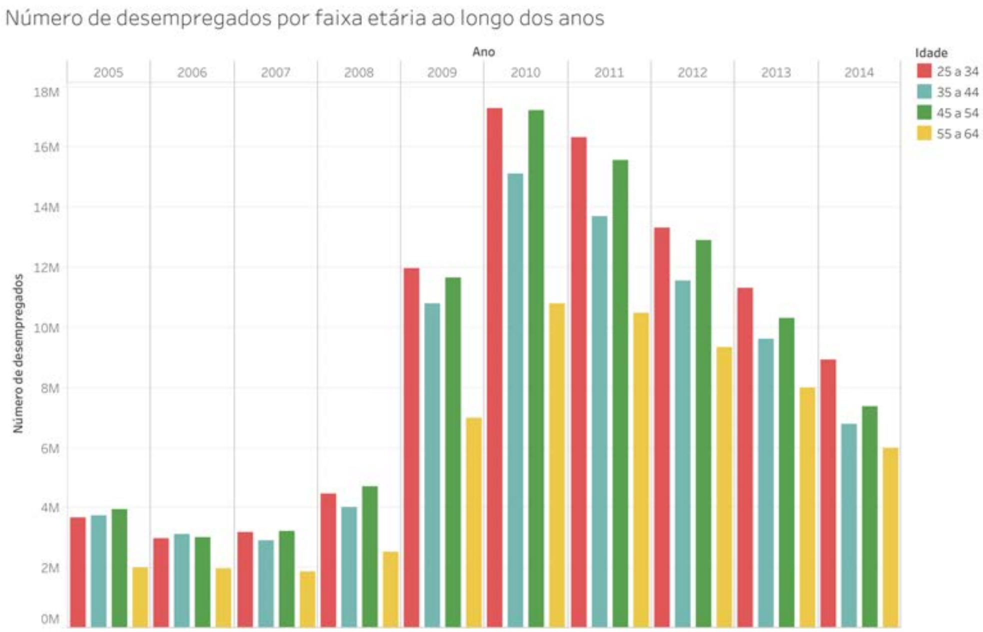


Figure 5.3: Study Chart P2.1.

Table 5.5: P2.1 questions

1	What is the unemployed number ratio between 25 to 34 years in 2008 to all unemployment for that year?
2	In 2006, unemployed number between 25 to 34 years was lower than that of people between 45 to 54.
3	What is the age group with the lowest unemployed unemployed?
4	What was the longest unemployed number decreasing period for all age groups?
5	In 2005 and 2006, was unemployed number between 55 to 64 years the same?
6	What is the unemployed number between 55 and 64 years in 2010?

P2.3: Using the exact data visualization (Figure 5.3) and analysis question (table 5.5 - 4), we introduced six suggestions for improvement, in random order (table 5.6). For each one, we asked the same as Part 1 (rate how much the participant thought the improvements made the chart more suitable for answering the question)(P2.3.1-P2.3.6). Good suggestions are marked with an "*", and the best suggestion is in boldface.

Table 5.6: P2.3 suggestions

1*	**I would change the chart type to line chart, but i would split it into small multiples, one chart for each age group.**
2*	I would change the chart type to line chart, one color for each age range.
3*	I would keep the chart type (bar), but I would group bars by the same age range, where each bar would be a year.
4	I would keep the chart type (bar), but I would stack the age ranges.
5	I would change the chart to a boxplot, distributing the number of unemployed by age group.
6	I would change to a table, sorting by year, coloring the cells with an intensity gradient according to the number of unemployment.

P1.4: We randomly presented each of the six suggestions from the previous step visually (Appendix D, Part 2 - questions 4a-f) and repeated the same task as in Part 1 (rate how well each chart answers the analysis question)(P2.4.1-P2.4.6).

P2.5: We introduced a tutorial explaining how each visualization from step 3 or 4 helps or hinders answering the analysis question.

5.2.3 Final Part

We repeated P1.2 and P2.2, modifying the analysis question, data, and the suggested visualization. With these modifications, the problem participants must identify differs from Part 1 and 2 but is still included in the context of the suggestions. This step aims to identify whether the tutorial suggestions help or hinder the participant in identifying the problem and proposing improvements to solve it.

FP.1: We presented a bar chart ordered by frequency (Figure 5.4) to answer an analysis question that would be more easily answered with a bar chart ordered by category. The question was: “Qual o número total de empresas de crescimento rápido em Charlotte?” (“What is the total number of companies with fast increase in Charlotte?”). We asked how much the participant considers the visualization adequate to answer the question on a scale from 1 to 7 (totally inadequate to totally adequate) and what improvements they suggest to make the visualization suitable for answering the question if they choose a grade below 7.

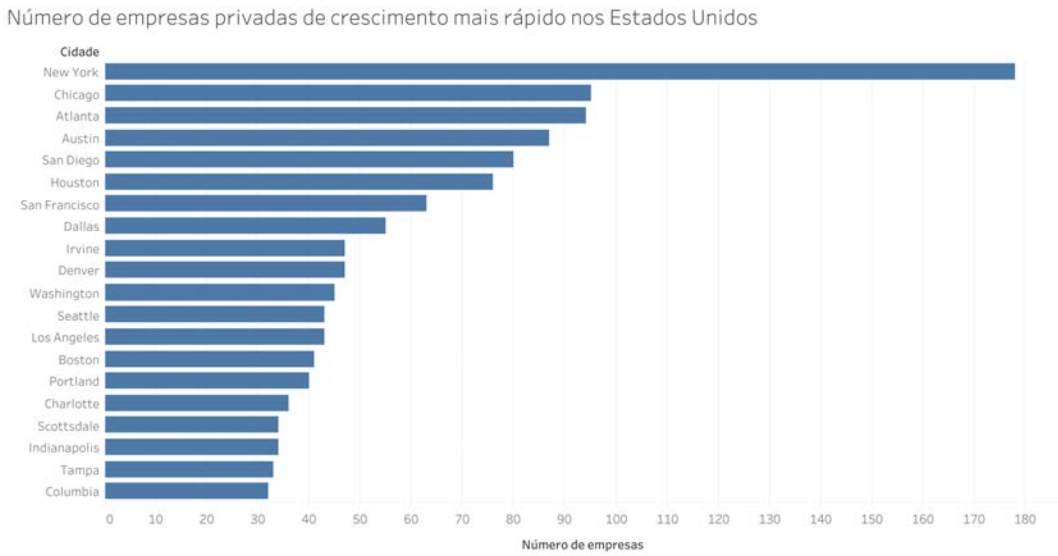


Figure 5.4: Study Chart FP.1.

FP.2: We presented a line chart (Figure 5.5), and an analysis question more easily answered with a grouped bar chart or small multiples. The question was: "Qual segmento se manteve mais constante (com menos picos) com relação aos valores das transações no ano de 2016?" ("Which segment kept constant (with fewer peaks) in regard to transaction values in 2016?"). We asked how much the participant considers the visualization adequate to answer the question on a scale from 1 to 7 (totally inadequate to totally adequate) and what improvements they would suggest to make the visualization suitable for answering the question if they choose a grade below 7.

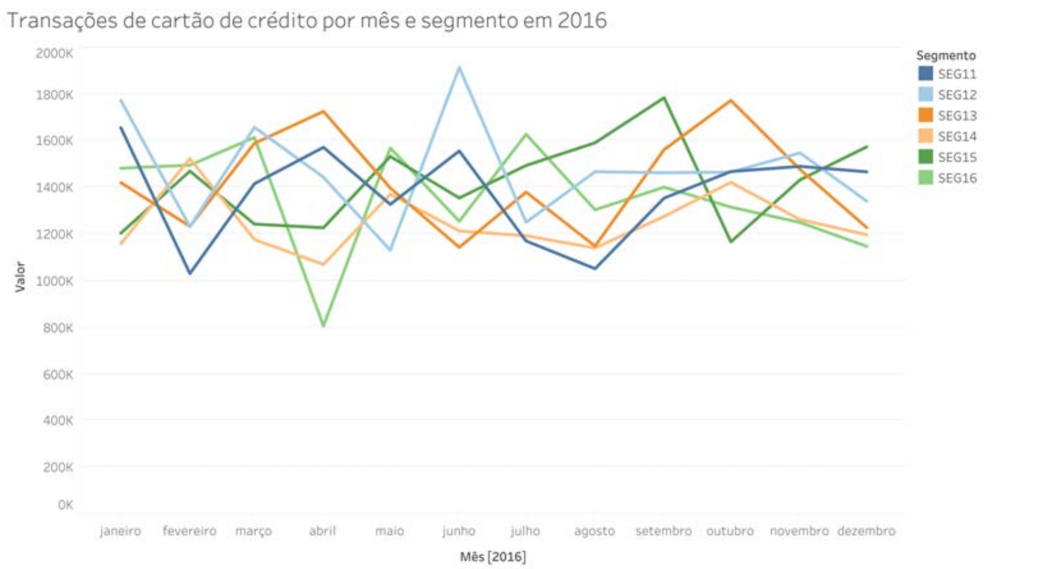


Figure 5.5: Study Chart FP.2.

5.3 Procedure and Participants

The main instrument of the study was an online questionnaire, which was available for two weeks. The invited participants were students and researchers from the Department of Informatics at PUC-Rio, and Ence and professionals who work with data analysis. In all, 44 participants filled out the whole questionnaire. If necessary, the participant could save the questionnaire to continue later. The time participants took in answering the questionnaire varied widely: the mean was 51.9 minutes,with a standard deviation of 42.4.

5.4 Analysis and Results

We calculated the Cronbach's alpha index (Cronbach, 1951) to measure the internal consistency of the survey for the task-related questions. For 44 items (questions items) and 44 samples (total of respondents), we obtained an α equal to 0.84, classifying the test as *good* (see table 5.7 as reference).

Table 5.7: Cronbach's Alpha reference

Internal consistency	α
Excellent	$\alpha > 0.9$
Good	$0.8 < \alpha <= 0.9$
Acceptable	$0.7 < \alpha <= 0.8$
Questionable	$0.6 < \alpha <= 0.7$
Poor	$0.5 < \alpha <= 0.6$
Unacceptable	$\alpha <= 0.5$

Most participants (61.3%) were between 18 and 24 years old. Only 0.04% did not have higher education, and another 50% had some complete specialization. Most participants (93.1%) fell into the STEM category, others were from the Humanities area.

About their learning background on data visualization, the majority (68.1%) had already had at least one class on the subject, but most of them (70.4%) did not subscribe to blogs about it. Most of them are used to reading details about visualizations they come across on the Web (54.5%), but not all read educational material (50%).

To assess their previous knowledge, we asked how each participant self-assessed their knowledge level on various items. We asked their knowledge of selecting chart types for exploration and communication, mapping attributes onto visual variables, evaluating visualizations with users, and reading and building all charts that we used in the study. The charts were single bars, grouped and stacked, single and multiple lines, boxplots, tables, and pie charts. We calculated the median of these items, and the majority (68.2%) received a score above 5.

5.4.1 Quantitative Data Analysis

Two data visualization researchers defined which answers we should consider as expected for each question. Later, we compared these answers with those by another data visualization researcher. In this way, we created a gold standard for each question that helped us analyze the results.

Tables 5.8 and 5.9 show the raw distribution of responses to each question. We calculated the percentage of expected responses and marked those below 60% with "**", for further analysis. We also marked in bold the expected responses by the gold standard established by the experts.

In the first part, five questions presented unexpected results. In the second part, this happened to nine questions.

Table 5.8: Part 1 Raw Results

Question	**Answer**			**Below 60%**
	1-3	**4**	**5-7**	
P1.1.1	**88.64%**	0%	11.36%	
P1.1.2	**84.09%**	4.55%	11.36%	
P1.1.3	**56.82%**	11.36%	31.82%	**
P1.1.4	**61.36%**	9.09%	29.55%	
P1.1.5	22.73%	11.36%	**65.91%**	
P1.1.6	22.73%	15.91%	**61.36%**	
P1.2	**56.82%**	11.36%	31.82%	**
P1.3.1	6.82%	2.27%	**90.91%**	
P1.3.2	13.64%	6.82%	**79.55%**	
P1.3.3	20.45%	13.64%	**65.91%**	
P1.3.4	**86.36%**	2.27%	11.36%	
P1.3.5	**90.91%**	2.27%	6.82%	
P1.3.6	**54.55%**	9.09%	36.36%	**
P1.4.1	0%	2.27%	**97.73%**	
P1.4.2	4.55%	4.55%	**90.91%**	
P1.4.3	29.55%	18.18%	**52.27%**	**
P1.4.4	**86.36%**	2.27%	11.36%	
P1.4.5	**77.27%**	2.27%	20.45%	
P1.4.6	**47.73%**	13.64%	38.64%	**

One of the questions that had unexpected results, question P1.1.3 (*What are the top-10 states with the most sales?*), is the question we had chosen for the next steps (P1.2-P1.3). The gold standard considered it hard to answer using the proposed visualization. However, many people considered it suitable for answering both in questions P1.1.3 and P1.2. Another noteworthy point was that many people considered the table suitable for answering in questions P1.3.6 and P1.4.6.

For question P1.4.3 (*Bar chart with labels, visual suggestion*), we understand that the excess of information brought by the labels may have made people consider the visualization unsuitable for answering the question, contrary to the gold standard. This same suggestion, presented in a textual form only (P1.3.3), had a different result, although expected by the gold standard.

Table 5.9: Part 2 Raw Results

Question	**Answer**				**Below 60%**
	1-3	**4**	**5-7**	**NA**	
P2.1.1	**81.8%**	4.5%	13.6%	0%	
P2.1.2	**59.1%**	6.8%	34.1%	0%	**
P2.1.3	40.9%	6.8%	**52.3%**	0%	**
P2.1.4	**47.7%**	15.9%	36.4%	0%	**
P2.1.5	**61.4%**	4.5%	34.1%	0%	
P2.1.6	**43.2%**	15.9%	40.9%	0%	**
P2.2	**59.1%**	18.2%	22.7%	0%	**
P2.3.1	25%	13.6%	**56.8%**	4.5%	**
P2.3.2	18.2%	11.4%	**70.6%**	0%	
P2.3.3	47.7%	9.1%	**43.2%**	0%	**
P2.3.4	**61.4%**	9.1%	25%	4.6%	
P2.3.5	**59.1%**	22.7%	13.6%	4.6%	**
P2.3.6	**63.6%**	15.9%	18.2%	2.3%	
P2.4.1	6.8%	11.4%	**81.8%**	0%	
P2.4.2	0%	4.6%	**95.5%**	0%	
P2.4.3	11.4%	6.8%	**81.8%**	0%	
P2.4.4	**45.5%**	13.6%	40.9%	0%	**
P2.4.5	**88.6%**	2.3%	6.8%	2.3%	
P2.4.6	**65.9%**	9.1%	25%	0%	

Of the six questions presented in P2.1.*, only P2.1.3 (*What is the unemployed number ratio between 25 to 34 years in 2008 to all unemployment for that year?*) could be answered readily with the given visualization. For the others, it would be necessary to do some calculations separately (*e.g.*, calculate the ratio, sum periods). It would be necessary to have a data value label or tooltip to ensure the value at the point. So, we expected the participants to choose the answers 1-3 for them. However, this only happened in two cases (P2.1.1 and P2.1.5).

The specific visualization given could be used to answer question P2.1.3, but, typically, this type of question cannot always be answered with this type of visualization. For example, if we changed the word least to most in the question, there is no clear answer just by comparing the lengths of the bars in each year; we would have to sum all values from each category over the years, which would make it hard to answer.

For question P2.2 (assess and make suggestions in free text), although many people agreed with the gold standard (59.1%), some (18.2%) chose the neutral option. P2.2 was difficult to answer with the suggested visualization because one would have to analyze each category separately, counting the years of continuous decrease and choosing the longest period.

Many participants considered the first suggestion inadequate for answering the question when analyzing it in its textual format (P2.3.1 - *I would change the chart type to line chart, but I would split it into small multiples, one chart for each age group.*), but they changed their minds when visualizing the result of applying that suggestion (P2.4.1). The same happened for the third suggestion, similar to the first one, for grouping the bars by age (P2.3.3 - *I would keep the chart type (bar), but I would group bars by the same age range, where each bar would be a year.* and P2.4.3) and fifth suggestion (P2.3.5 - *I would change the chart to a boxplot, distributing the number of unemployed by age group.* and P2.4.5).

Surprisingly, some participants changed their minds regarding the suggestion to stack the bars between the textual and visual form of the recommendation (P2.3.4 and P2.4.4).

The result of question P3.1 (FP.1 in Appendix D) was surprising, as many people considered the visualization adequate for answering the question, contrary to the gold standard (table 5.10). In fact, for the reduced number of states (20), it is not very costly to read the unordered list for searching a state. However, the robust suggestion for this question would be to use a bar chart ordered by category instead of frequency.

In question P1.2, if the participant did not consider the visualization fully adequate to answer the question, we asked them to write a suggestion to make it adequate (P1.2b). Each participant wrote one to three suggestions, and we analyzed and categorized each one. Table 5.11 shows the frequency of the suggestions. Some participants wrote something like "Missing average

Table 5.10: Part 3 Raw Results

Question	**Answer**			**Below 60%**
	1-3	**4**	**5-7**	
P3.1	**34.09%**	18.18%	47.73%	**
P3.2	**63.64%**	18.18%	18.18%	

sales", which we did not consider as a suggestion, and classified it as such.

Table 5.11: P1.2 (participants' suggestions)

Suggestion	**Count**	**Type**
sort	30	good (expected)
highlight	7	good (expected)
label	8	neutral
label (order)	1	neutral
other (not a suggestion)	6	bad (unexpected)
group	2	bad (unexpected)
change the chart	1	bad (unexpected)
mean line	1	bad (unexpected)
split into three charts	1	bad (unexpected)

From the suggestions provided, we considered two good suggestions to help answer the question quickly: sort by frequency and highlight the top-10. Two other suggestions – value labels and rank labels – bring some improvement, but it still requires an effort from the analyst to answer the question, so we classified it as neutral. The analyst would still need to make comparisons in both options until finding the answer. We did not consider the remainder of the suggestions as good options to answer the question. Overall, we had a significant number of expected suggestions (80%, see table 5.12).

Table 5.12: P1.2 (summarized participants' suggestions)

Expected	**Unexpected**
80%	20%

We performed the same analysis for question P2.2 (table 5.13). For this question (P2.2b), we did not consider any suggestions as neutral. Despite some good suggestions, the number of unexpected suggestions was significantly higher (55.8%, table 5.14).

Part of the unexpected group was composed of answers that we did not consider as a suggestion, for example: "Only in chart it is not possible to identify the reason for the increase in unemployment."("Apenas em Gráfico

Table 5.13: P2.2 (participants' suggestions)

Suggestion	**Count**	**Type**
line chart	17	good (expected)
clustered bar	2	good (expected)
label	12	bad (unexpected)
other (not a suggestion)	11	bad (unexpected)
stacked bar	2	bad (unexpected)
change the chart	1	bad (unexpected)
difference chart	1	bad (unexpected)
group values	1	bad (unexpected)
highlight	1	bad (unexpected)
highlight the change	1	bad (unexpected)
regression line	1	bad (unexpected)
show/hide	1	bad (unexpected)
sort	1	bad (unexpected)
tendency line	1	bad (unexpected)

não é possível identificar a razão pelo aumento de desempregado."). The other part (27.3% of respondents) suggested including the value. 9.1% of these respondents gave this same suggestion in the previous question.

Table 5.14: P2.2 (summarized participants' suggestions)

Expected	**Unexpected**
44.2%	55.8%

We performed the same analysis on the last two questions that required an improvement suggestion. For P3.1 (FP.1b in Appendix D), a significant number of people made a good suggestion (82.9%, table 5.16). We expected more respondents to suggest ordering the bars by category in this specific case (table 5.15). However, as the question had an analysis task to return a value for a specific category, it is expected that the people would suggest labeling the bars for a non-interactive chart. Again, 9.1% of the participants suggested "labeling" in the three questions about improvement.

As for P3.2 (FP.2b in Appendix D), we had a large unexpected number of suggestions (67.4%, table 5.17). We did not consider 27% of the comments as suggestions, for example: "Claramente essa junção de curvas sobrepostas e de coloridas estão uma bagunça." However, 18.7% of the suggestions changed the chart to a boxplot (table 5.18). Perhaps the misunderstanding here was that the participants did not consider "peaks" to be monthly fluctuations, but the difference between the highest and lowest value of the period.

Table 5.15: P3.1 (participants' suggestions)

Suggestion	Count	Type
label	32	good (expected)
sort	5	good (expected)
highlight	3	good (expected)
other (not a suggestion)	9	bad (unexpected)
text-only	2	bad (unexpected)
group	1	bad (unexpected)
mouse over	1	bad (unexpected)
reference line	1	bad (unexpected)
table	1	bad (unexpected)

Table 5.16: P3.1 (summarized participants' suggestions)

Expected	Unexpected
83%	17%

Table 5.17: P3.2 (summarized participants' suggestions)

Expected	Unexpected
32.6%	67.4%

Table 5.18: P3.2 (participants' suggestions)

Suggestion	Count	Type
small multiples	14	good (expected)
clustered bar	1	good (expected)
other (not a suggestion)	13	bad (unexpected)
boxplot	9	bad (unexpected)
bar chart	2	bad (unexpected)
area chart	1	bad (unexpected)
average value line	1	bad (unexpected)
change the chart	1	bad (unexpected)
color scale	1	bad (unexpected)
interactive chart	1	bad (unexpected)
split y-axis	1	bad (unexpected)
table with variance	1	bad (unexpected)
variance	1	bad (unexpected)
variation chart	1	bad (unexpected)

Interestingly, we can highlight that a small multiple is a good suggestion to improve P1.2, but no participant made this suggestion in that question. After the tutorial, where we presented this term and usage, most good suggestions (14 out of 15) pointed to this type of visualization. Another interesting fact is that there was no suggestion for labeling, even by those participants who had always made this suggestion before.

5.4.2
Qualitative Pairwise Data Analysis

Questions P1.1.3 and P1.2 are similar, with P1.1.3 providing textual suggestions and, P1.2, visual. In this way, we can compare the results of each participant to know if there was any further change in the choice of answers. Table 5.19 shows the result when the choice in P1.2 was better, worse, or equal to P1.1.3 in terms of expected values. The number of participants who provided better or worse suggestions was the same (22.7%). Counting those who maintained an expected response with those that improved, we have a larger number of participants (58.8%, see table 5.20: same+same (expected)).

Table 5.19: P1.1.3 and P1.2 comparison (textual and visual comparison)

Better	**Same (expected)**	**Neutral**	**Worse**	**Same (unexpected)**
22.7%	34.1%	2.3%	22.7%	18.2%

Table 5.20: Summarized P1.1.3 and P1.2 comparison (textual and visual comparison)

Better + Same (expected)	**Neutral**	**Worse + Same (unexpected)**
56.8%	2.3%	40.9%

Similarly, we can compare questions P2.1.4 and P2.2 from the second part. Despite the high number of participants who provided an expected response and maintained their response (43.2%), the number of participants who provided a better response (18.2%) was higher than those who provided a worse response (6.8%) (table 5.21).

Table 5.21: P2.1.4 and P2.2 comparison (textual and visual comparison)

Better	**Same (expected)**	**Neutral**	**Worse**	**Same (unexpected)**
18.2%	43.2%	11.4%	6.8%	20.5%

Again, in general, the number of participants who provided expected responses to the visual suggestion (61.4%) was higher than those who gave unexpected responses (27.3%) (table 5.22).

Table 5.22: P2.1.4 and P2.2 comparison (textual and visual comparison)

Better + Same (expected)	Neutral	Worse + Same (unexpected)
61.4%	11.4%	27.3%

We can compare the answers to question groups 3 and 4 (6 suggestions to change in the visualization to make it best suited to answer the question) in pairs for each of the first two parts of the survey since they are the same suggestions, one in textual form (group 3) and the other in visual form (group 4).

Table 5.23: P1.3 and P1.4 comparison (textual and visual comparison)

Suggestion	Better	Same (expected)	Neutral	Worse	Same (unexpected)
1	9.1%	88.6%	0%	2.3%	0%
2	18.2%	72.7%	0%	6.8%	2.3%
3	6.8%	47.7%	6.8%	22.7%	15.9%
4	9.1%	77.3%	0%	9.1%	4.5%
5	2.3%	75%	0%	15.9%	6.8%
6	20.5%	34.1%	0%	25%	20.5%

In general, in the first part, all the suggestions had more expected answers after the visual question because most had already got the textual question right and kept their answers (table 5.24 and table 5.23.).

Table 5.24: P1.3 and P1.4 comparison (textual and visual comparison)

Suggestion	Better+same (expected)	Neutral	Worse+same (unexpected)
1	97.7%	0%	2.3%
2	90.9%	0%	9.1%
3	54.5%	6.8%	38.6%
4	86.4%	0%	13.6%
5	77.3%	0%	22.7%
6	54.5%	0%	45.5%

Except for suggestion 4, all other suggestions maintained the expected responses or improved in the second part (table 5.26). The suggestions had a higher percentage of good (expected) answers than bad (unexpected) (table 5.25).

Returning to P1.2, when we compared the participants' evaluation of the adequacy of the visualization to answer the question and the textual suggestions they made for improvement, we obtained significant satisfactory results (table 5.27). 55% of the participants rated the visualization as bad and gave good suggestions for improvement. 25% evaluated it as neutral or good; even so, they gave good suggestions for improvement. However, 7.5%

Table 5.25: P2.3 and P2.4 comparison (textual and visual comparison)

Suggestion	Better	Same (expected)	Neutral	Worse	Same (unexpected)
1	31.8%	56.8%	0%	9.1%	2.3%
2	27.3%	68.2%	2.3%	2.3%	0%
3	50%	38.6%	0%	4.5%	6.8%
4	9.1%	36.4%	2.3%	29.5%	22.7%
5	29.5%	59.1%	2.3%	2.3%	6.8%
6	9.1%	56.8%	4.5%	13.6%	15.9%

Table 5.26: Summarized P2.3 and P2.4 comparison (textual and visual comparison)

Suggestion	Better+Same (expected)	Neutral+Same neutral	Worse+Same (unexpected)
1	88.6%	0%	11.4%
2	95.5%	2.3%	2.3%
3	88.6%	0%	11.4%
4	45.5%	2.3%	52.3%
5	88.6%	2.3%	9.1%
6	65.9%	4.5%	29.5%

of the participants did not rate the visualization as bad, nor did they give good suggestions for improvement. Just as 12.5% of the participants rated the visualization as neutral or bad, but they could not suggest good improvements.

Table 5.27: P1.2 and P1.2.rec comparison

Better	Same (expected)	Worse	Same (unexpected)
25%	55%	12.5%	7.5%

Table 5.28: P2.2 and P2.2.rec comparison

Better	Same (expected)	Worse	Same (unexpected)
18.6%	25.6%	44.2%	11.6%

We performed the same analysis for P2.2 (table 5.28). Initially, most participants provided unexpected suggestions for improvements (80%, table 5.12). 18.6% of the participants gave good suggestions, even though they had given unexpected evaluations before, and 25.6% of the participants provided the expected answers in the two stages. However, 44.2% of the participants had previously evaluated the visualization as bad (as expected) or neutral but could not give reasonable solutions. Moreover, 11.6% of the participants gave unexpected answers in the two stages.

Again, the same analysis for P3 (after reading the tutorial), we had a significant number of good results for the first case (P3.1), see table 5.29. 48.8% of respondents made a good suggestion even though they previously stated that

Table 5.29: P3.1 and P3.1.rec comparison

Better	Same (expected)	Worse	Same (unexpected)
48.8%	34.1%	4.9%	12.2%

the question could be easily answered with visualization. 34.1% of participants made good choices in both questions. Only 12.2% of the participants did not get good choices in the two questions, and 4.9% of the participants did not know how to give a good suggestion despite stating that the visualization did not help answering the question.

Table 5.30: P3.2 and P3.2.rec comparison

Better	Same (expected)	Worse	Same (unexpected)
9.3%	23.3%	51.2%	16.3%

For the second case, most of the participants could not make a good suggestion when they identified that it needed to be improved (51.2%), nor when it was not possible to identify this fact (16.3%, table 5.30). Only 9.3% made a good suggestion even though they did not identify the problem earlier. Moreover, 23.3% provided an expected response at both times.

Table 5.31: P1.2 and P3.1 comparison

Better	Same (expected)	Worse	Same (unexpected)
7.3%	29.3%	34.1%	29.3%

Table 5.32: P2.2 and P3.2 comparison

Better	Same (expected)	Neutral	Same (unexpected)
29.3%	46.3%	2.4%	4.9%

Questions P1.2 and P3.1 are similar because the visualization of the first would result from a suggestion for improvement of the second to answer that question. Furthermore, the second visualization would improve the first one to answer the first question. 34.1% of the participants could not draw a parallel between the problem and the solution presented in the first question and the tutorial to propose a solution for the final question. However, 7.3% made this association (table 5.31).

Likewise, 7.3% of people could not bring the information presented in the first question and tutorial to solve the final question in the second part. 29.3% of the participants succeeded (table 5.32).

5.5 Discussion

We analyzed the data quantitatively and qualitatively. We wanted to understand how participants behave when asked about the suitability of specific visualizations and suggestions for improvement to help answer data analysis questions. We defined some hypotheses related to the study that we tested against the data collected. Next, we list the hypotheses and our conclusions.

H1: For basic charts, people can identify whether the visualization is appropriate to answer an analysis question.

Of the 44 questions, 40 were of the type of assessment of whether a visualization is suitable for answering an analysis question. Sometimes the visualization was shown in an image format and sometimes in text. Among the responses, some visualizations were adequate, and others were not. We analyzed expected and unexpected answers. 93.2% of the participants gave expected answers in more than 50% of the questions (47.7% above 70% of the questions). Only 6.8% of respondents gave expected answers to less than 50% of the questions. A sign test showed that the difference between correct and incorrect answers was significant ($N = 1760, x = 1195, p < 0.001$) so we can accept H1, *i.e.*, participants can identify when a simple chart is suitable for answering an analysis question.

H2: For two selected visualizations, among the items in which the participants identified that there was a problem (and for which they were asked to recommend solutions), there was a significant difference between those who gave adequate or inadequate solutions.

Four questions asked for an improvement suggestion. Analyzing the cases in which the participant identified as expected that the visualization was inadequate, 60.6% of the items suggested an expected improvement. A sign test showed that the difference between adequate and inadequate improvements was not significant ($N = 94, x = 57, p = 0.05003$) so we cannot accept H2.

H3: Visual suggestions promote more expected responses than textual suggestions.

There were 14 dual questions: the same question presented in textual and visual format. Overall, without analyzing the performance by participants and questions, the visual format had 73.4% (452 out of 616) of expected responses against 66.5% (410 out of 616) of expected responses in the textual form. When we compared the results for each question, the number of expected

Table 5.33: Summary of expected responses in the comparison between textual and visual suggestions by question

Question	Textual	Visual	greater?
1	56.8%	56.8%	
2	47.7%	59.1%	*
3	90.9%	97.7%	*
4	79.5%	90.9%	*
5	65.9%	52.3%	
6	86.4%	86.4%	
7	90.9%	77.3%	
8	54.5%	47.7%	
9	61.4%	81.8%	*
10	70.5%	95.5%	*
11	43.2%	81.8%	*
12	61.4%	45.5%	
13	59.1%	88.6%	*
14	63.6%	65.9%	*

responses in the visual suggestions was higher than in the textual suggestions in only half of the questions (table 5.33). One may note that as the study progressed, the number of improvements increased. However, the difference was not statistically significant. In future work it would be interesting to investigate whether this is related to the types of visualization involved, or the passing of time and acquisition of knowledge about the visualizations accumulated from the tutorials.

H4: After a short tutorial, people can better identify a problem in a related visualization.

We can compare two questions before the tutorial to two questions after the tutorial, one for each part (P1.2 with P3.1 and P2.2 with P3.2). In the first part, comparing each participant's responses, only 6.8% improved their response (going from an unexpected to an expected suggestion). In the second part, this happened with 11.4% of them. However, the differences were not significant, so we cannot declare that the tutorial had a positive effect on improving the identification of a problem.

H5: After a short tutorial, people can better provide good suggestions for improving the visualizations addressed in the tutorial.

Similarly, we can compare the improvement suggestions given before and after the tutorial for each of the two parts of the questionnaire. In the first part, only 9.1% gave expected suggestions after the tutorial, when they did not give an expected suggestion before it. In the second part, this happened with 6.8% of the participants. Therefore, we cannot say that the tutorial had a positive effect in helping participants provide better suggestions.

5.6 Concluding Remarks

This chapter introduced our study to investigate whether people can identify when a visualization is suitable for answering a particular analysis question. We presented participants with an analysis question and a visualization and asked them to rate the adequacy of the data visualization to support answering the analysis question. We asked them to write suggestions to make the visualization more appropriate. We also presented some suggestions in both the textual and visual forms and asked them to assess to what extent the suggestions would make the visualizations more suitable (or not) for answering the question. Finally, we presented a short text in a tutorial format to explain the inadequacies in the visualization and how to make it more adequate. We repeated these tasks for another pair of <question, visualization>. Finally, we made a round of questions to assess the effect of the tutorial on the answers.

The results of our study revealed that analysts are generally able to identify when a visualization is not fully adequate to answer an analysis question. However, they cannot always identify reasonable solutions or suggest good improvements to make them more suitable.

Visual suggestions did not perform better than textual suggestions for a number of questions. This fact indicates that recommender systems that make textual recommendations for improvement can be ineffective in most cases.

The short tutorial did not have a positive effect in either case: identifying problems or suggesting changes. In this case, a traditional teaching system based simply on examples and counter-examples (or "do this, don't do that") may not be an adequate tool for teaching visualization. One suggestion for analysis tools is to suggest fixes on demand, highlighting errors and changes and explaining them.

6
Assessing data visualization literacy

The data visualization literature brings several works that involve the assessment of visualization data literacy with further analysis' goals: how people understand, create, teach, and make sense of visualizations (Rodrigues et al., 2021). Our search for related works was not exhaustive, but we found some relevant works referencing and contributing to this area (*e.g.*, Boy et al., 2014; Börner et al., 2016; Lee et al., 2017).

VLAT (Lee et al., 2017) is a questionnaire-based test composed of multiple-choice questions to quantify how fluently people understand charts in general. It focuses on perceptual tasks such as retrieving values or comparing means, comprising 12 chart types covering eight visualization tasks and resulting in 53 multiple-choice test items. Test items are specific questions about the data represented visually. The response options ranged from three or four alternatives, or true or false. Test items received a rating according to the difficulty and discrimination index.

For our purposes, one of the shortcomings of VLAT is that the assessment test is based only on applied questions and not about charts' conceptual aspects and purpose. In this way, we aimed to extend this test by including other visualizations, trick questions, and questions about the visualizations' structure.

This chapter presents the procedure we followed for designing a new visualization literacy assessment test. It also presents the results of the test application with a group of 68 participants.

6.1
Goal

This study aimed to create a data visualization literacy test that could cover different visualizations with applied suitable and unsuitable analysis questions and conceptual questions about chart structure and affordances.

6.2 Study Design

Our questionnaire starts with pre-study questions to characterize each participant's profile so as to relate this information to their performance in the analysis stage. We show participants 15 different visualizations concerning the same domain (IMDB data set). The selected visualizations are a clustered bar chart, single bar chart (ordered by category and frequency), stacked bar chart (single and 100%), boxplot, bubble chart (and colored), histogram, line chart (single and multiple), scatterplot (and colored), table, and pie chart.

For each visualization, we asked three questions: (i) a question about concepts related to the visualizations, *e.g.*, what a participant could represent or extract from the visualization; (ii) an applied, domain-related question, which the visualization was suitable to answer; and (iii) an applied, domain-related question, but which the current visualization unsuitable to answer. Table 6.1 exemplifies those questions in the order described, but to participants they were presented in random order. The remaining visualizations can be found in appendix E.

We took from the literature both the conceptual and the applied analytical questions. As we had previously identified the novices' common mistakes while making sense of the underlying data in a visualization (Rodrigues et al., 2021), we wanted to include some questions to explore these misconceptions. Therefore, we translated these common mistakes onto unsuitable questions (those which the did not help or allowed to answer), testing the participant's ability to detect improper task-visualization couplings.

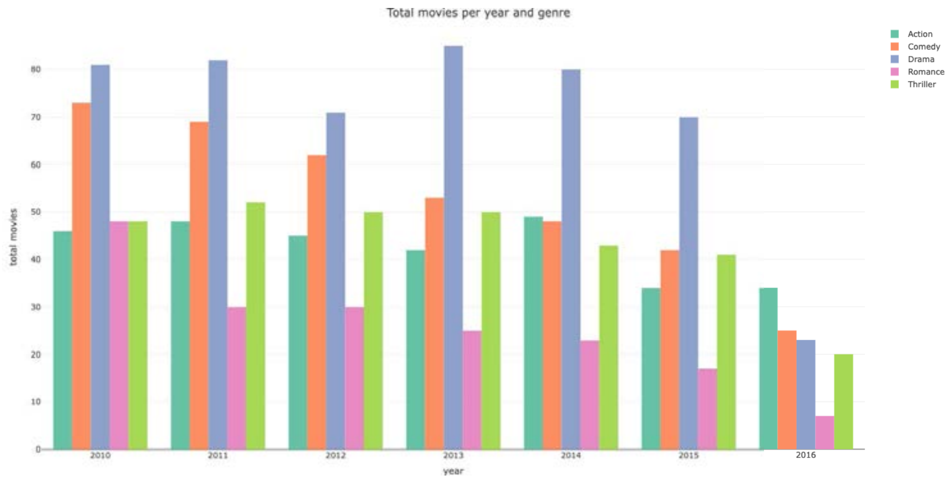
All questions have six answer options, only one of which was correct. One of them is a "I don't know" option. It allows respondents to declare that they do not know how to answer a specific question. By choosing this option, the participant misses the question but, as they are not forced to choose between specific answers, this reduces the risk of just guessing an answer when they do not feel confident they would get it right.

In addition, the applied questions had an option "This type of chart does not allow or help to answer the question". This option was the wrong answer for the suitable (applied) question and the right one for the unsuitable (applied) question. The answer choices and the order of the questions were random.

There was only one task scenario in which the participant received one type of visualization and the three questions about it at once. After answering the three questions, they would go to the next visualization. With the 15 visualization types and 3 question types, our study ended with 45 test items. The visualizations were developed in R scripting language using the ggplot2

Table 6.1: Study 4 task example

Task: For each question, choose one of the following answers:



This type of chart best allows for:	
( )	defining which types of genre are most frequent over the years
( )	I don't know
( )	comparing the number of movies across genres in each year
( )	calculating how many genres are analyzed each year
( )	analyzing the difference of years in each genre
( )	explaining why there are peaks in a specific genre

Approximately how many Drama movies were produced in 2010?	
( )	107
( )	This type of chart does not allow or help to answer the question
( )	81
( )	94
( )	I don't know
( )	48

What is the relationship between Drama and Action movies?	
( )	Correlation
( )	Inversely proportional
( )	I don't know
( )	It depends on the year of production
( )	This type of chart does not allow or help to answer the question
( )	Directly proportional

and ggplotly libraries, making them interactive.

6.3 Procedure and Participants

The online survey, including the test questions, was available for 15 days. We invited two groups of participants. The first group had 36 participants, including graduate students (Master's and Doctorate) and researchers from the Department of Informatics at PUC-Rio. The second group had 32 participants, including undergraduate students of the Data Visualization class, at PUC-Rio. The majority of these (45.6%) were between 25 and 34 years old, 29.4% were between 18 and 24 years old, and 25% were 35 years old or older. Regarding their degrees, 17.6% of the participants had Doctorate degrees, 32.35% Master degrees, 8.8% some specialization, 28% had completed Higher Education, and 13.2% had completed High school.

Profile questions included self-assessment questions about the frequency in which they: read or interpret charts (table 6.2), build charts to explore data and get insights, and build charts to communicate data insights. We asked participants about their knowledge of mean, standard deviation/variance, median, interquartile range, quartiles, linear correlation, and outliers (table 6.3). We also asked them about their experience with some visualization tools, programming languages, or libraries (table 6.4). In general, most participants stated they analyzed or created charts at least monthly, and used charts in communication at least quarterly. Participants self-assessed as having moderate to almost expert knowledge in all chart concepts, except in outliers, of which the vast majority said they had little or no knowledge. The tools that the participants demonstrated to have moderate to expert experience were general spreadsheet applications (Microsoft Excel, Numbers, Google Sheets) and Python libraries.

Table 6.2: General knowledge about charts

Activity	Never	Every year	Every semester	Every trimester	Every month	Every week	Every day
Read or interpret charts	10.29%	7.35%	10.29%	11.76%	**29.41%**	**30.88%**	0%
Build charts to explore data and get insights	11.76%	5.88%	17.65%	16.18%	**30.88%**	**17.65%**	0%
Build charts to communicate data insights	1.47%	0%	27.94%	**32.35%**	**38.24%**	0%	0%

Table 6.3: Charts concept knowledge

Concept	No knowledge	Little knowledge	Moderate knowledge	Almost expert	Expert
mean (average)	1.47%	8.82%	**36.76%**	**32.35%**	20.59%
standard deviation / variance	1.47%	2.94%	**32.35%**	29.41%	**33.82%**
median	14.71%	16.18%	**25%**	**26.47%**	17.65%
interquartile range	10.29%	14.71%	**27.94%**	**26.47%**	20.59%
quartiles	13.24%	13.24%	**27.94%**	**30.88%**	14.71%
linear correlation	13.24%	7.35%	**27.94%**	**35.29%**	16.18%
outliers	**58.82%**	**22.06%**	16.18%	1.47%	1.47%

Table 6.4: Visualization tools knowledge

Tool	No experience	Little experience	Moderate experience	Almost expert	Expert
Tableau	**58.82%**	22.06%	16.18%	1.47%	1.47%
Flourish	**91.18%**	4.41%	1.47%	2.94%	0%
Microsoft Excel / IOS Numbers / Google Sheets	5.88%	11.76%	**48.53%**	**32.35%**	1.47%
Python libraries (ex: Matplotlib)	22.06%	16.18%	**33.82%**	**22.06%**	5.88%
R libraries (ex: ggplot)	**51.47%**	17.65%	16.18%	10.29%	4.41%
Javascript libraries (ex: D3.js)	**60.29%**	17.65%	16.18%	5.88%	0%

6.4 Basic Statistics, Reability Evaluation and Item Analysis

We performed the same statistical analysis as defined for VLAT, to compare the results and to eliminate unsatisfactory items. We also applied Item Analysis based on the classical test theory (CTT) (Thorndike et al., 1991) to measure the difficulty of items and test takers' ability. In VLAT, they proposed to discard a test item if it shows little variation within the sample, it is strongly correlated with one or more other items, or it is weakly correlated with the totality of the remaining items. The latter is reflected in an increase in Cronbach's alpha if we eliminate the item from the test.

6.4.1 Basic Statistics

We included an answer option to prevent participants from guessing the answer, as we mentioned earlier. This option, if chosen honestly, prevents the participant from guessing and distorting the results. Despite this, we applied the correction-for-guessing score in the raw score of each participant using eq. (6-1):

$$S = R - \frac{W}{k-1} \tag{6-1}$$

where S is the corrected score, R is the number of items marked correctly, W is the number of items marked incorrectly, and k is the number of choices for each item.

The test takers' raw scores ranged from 0 to 40 ($M = 28.9, SD = 6.07$). The corrected scores ranged from -9 to 39 ($M = 25.68, SD = 7.28$). After the adjustments, the test takers' scores dropped an average of 3 points.

We also calculated the mean (equivalent to item difficulty) and standard deviation (a measure of the dispersion of participants' scores on that item) for each item (table 6.5).

6.4.2 Reliability Evaluation

The test's internal consistency was high, with a Cronbach's alpha value of 0.81. This value indicates good reliability for the achievement test (table 5.7). We also calculated the reliability of the test when withdrawing an item. In all cases, the reliability remained above 0.80 (see table 6.5). Four items increased alpha when removed from the test: 36, 38, 41, and 44. We discuss this in detail in section 6.5.

6.4.3 Item Analysis: Item Difficulty and Discrimination

Item difficulty is the percentage of participants who answer an item correctly, ranging from 0 to 1; the lower the value, the more difficult the question. Item discrimination indicates the extent to which success on an item corresponds to success on the whole test, ranging from -1 to 1. If we split the participants into two groups, high-scored test takers and low-scored test-takers, it measures the ratio of the difference between the total correct answers in each group.

We performed the analysis according to CTT using the ShinyItemAnalysis R package. The CTT determined each item's difficulty and discrimination index. It showed that 51% of the questions were of moderate difficulty. The remainder were hard (29%) and easy (20%) questions. The profile questions showed that the participants had moderate to expert knowledge in most chart concepts, so the result also suggests they could answer difficult questions.

To calculate the item difficulty, with a sample larger than 30 participants, we used $k = 3$ (groups), $l = 1$ (first group) and $u = 3$ (last group), separating the participants into 27% superior and 27% inferior groups. We calculated

Table 6.5: Traditional item analysis

#	Chart	Type	Mean	SD	P		D		Item-total correlation	α without item
1		Conceptual	0.65	0.48	0.65	○	0.32	●	0.10	0.81
2	Bar (clustered)	Suitable	0.96	0.21	0.96	●	0.18	○	0.10	0.81
3		Unsuitable	0.44	0.50	0.44		0.18	○	0.12	0.81
4		Conceptual	0.87	0.34	0.87	●	0.36	●	0.22	0.80
5	Bar (ordered by category)	Suitable	0.91	0.29	0.91	●	0.18	○	0.09	0.81
6		Unsuitable	0.90	0.31	0.90	●	0.32	●	0.15	0.81
7		Conceptual	0.90	0.31	0.90	●	0.18	○	0.09	0.81
8	Bar (ordered by frequency)	Suitable	0.99	0.12	0.99	●	0.09		0.07	0.81
9		Unsuitable	0.46	0.50	0.46		0.45	●	0.18	0.81
10		Conceptual	0.79	0.41	0.79	○	0.27	○	0.14	0.81
11	Bar (stacked)	Suitable	0.82	0.38	0.82	○	0.27	○	0.13	0.81
12		Unsuitable	0.18	0.38	0.18		0.23	○	0.07	0.81
13		Conceptual	0.72	0.45	0.72	○	0.45	●	0.21	0.80
14	Bar (100% stacked)	Suitable	0.81	0.40	0.81	○	0.50	●	0.21	0.80
15		Unsuitable	0.25	0.44	0.25		0.18	○	0.11	0.81
16		Conceptual	0.66	0.48	0.66	○	0.68	●	0.28	0.80
17	Boxplot	Suitable	0.76	0.43	0.76	○	0.45	●	0.20	0.80
18		Unsuitable	0.19	0.40	0.19		0.32	●	0.12	0.81
19		Conceptual	0.78	0.42	0.78	○	0.41	●	0.22	0.80
20	Bubble chart	Suitable	0.60	0.49	0.60	○	0.23	○	0.10	0.81
21		Unsuitable	0.10	0.31	0.10		0.00		0.02	0.81
22		Conceptual	0.79	0.41	0.79	○	0.50	●	0.19	0.80
23	Bubble chart (color)	Suitable	0.81	0.40	0.81	○	0.14	○	0.12	0.81
24		Unsuitable	0.37	0.49	0.37		0.36	●	0.17	0.81
25		Conceptual	0.75	0.44	0.75	○	0.32	●	0.13	0.81
26	Histogram	Suitable	0.84	0.37	0.84	○	0.32	●	0.12	0.81
27		Unsuitable	0.51	0.50	0.51	○	0.45	●	0.19	0.81
28		Conceptual	0.88	0.32	0.88	●	0.27	○	0.14	0.81
29	Line (single)	Suitable	0.94	0.24	0.94	●	0.09		0.09	0.81
30		Unsuitable	0.59	0.50	0.59	○	0.41	●	0.17	0.81
31		Conceptual	0.87	0.34	0.87	●	0.23	○	0.13	0.81
32	Line (multiple)	Suitable	0.82	0.38	0.82	○	-0.09		0.07	0.81
33		Unsuitable	0.68	0.47	0.68	○	0.59	●	0.24	0.80
34		Conceptual	0.75	0.44	0.75	○	0.50	●	0.23	0.80
35	Scatterplot	Suitable	0.74	0.44	0.74	○	0.45	●	0.18	0.81
36		Unsuitable	0.12	0.32	0.12		0.00		0.01	0.82
37		Conceptual	0.68	0.47	0.68	○	0.45	●	0.18	0.81
38	Scatterplot (color)	Suitable	0.40	0.49	0.40		0.14	○	0.07	0.82
39		Unsuitable	0.16	0.37	0.16		0.05		0.04	0.81
40		Conceptual	0.82	0.38	0.82	○	0.50	●	0.22	0.80
41	Table	Suitable	0.79	0.41	0.79	○	0.05		0.01	0.82
42		Unsuitable	0.47	0.50	0.47		0.18	○	0.11	0.81
43		Conceptual	0.75	0.44	0.75	○	0.14	○	0.09	0.81
44	Pie chart	Suitable	0.19	0.40	0.19		0.09		0.04	0.82
45		Unsuitable	0.44	0.50	0.44		0.41	●	0.15	0.81

the index rankings according to the Office of Educational Assessment at the University of Washington.[1] We rated an item as easy (light green) if the difficulty index is greater than 0.85, moderate (medium green) if it is between 0.5 and 0.85, and hard (dark green) if it is less than 0.5. We also rate the item as good (dark brown) if the discrimination index is greater than 0.3, fair (medium brown) if it is between 0.1 and 0.3, and poor (light brown) if it is less than 0.3.

Of the 45 test items, we consider only 8 (18%) poor items. Thirty-seven (82%) of the items were either fair or good items, meaning that the test should truly represent the test takers' learning ability, *i.e.*, the items can discriminate well between the high and low-performing groups. Table 6.5 shows the results of the indexes, sorted by chart and question types.

Two questions were easy and did not discriminate, as expected, because they were suitable questions for more standard charts: Bar (ordered by frequency) and Line (single). The questions also concerned a trivial task for these charts: retrieving a bar value and finding trends in a line. For these charts and these question types, it might be more interesting to define analysis questions that involve less trivial tasks, such as making comparisons or determining ranges, respectively. However, this might raise the issue as to whether these charts are suitable for answering these questions.

The questions considered difficult with poor discrimination were unsuitable questions for more complex charts such as the Bubble chart, Scatterplot, Scatterplot (color), and a suitable question for simple charts, such as Line (multiple), Table, and Pie chart. In the latter case, very similar slices can induce errors, and perhaps this was why only 19% of the participants got it right, making the question difficult for those who scored higher and those who scored lower.

6.4.4
VLAT Comparison

Our test had eight chart types in common with the VLAT. The suitable questions in our study were similar to the VLAT questions for these chart types. Despite the question domain being different, we chose the same analysis task and built similar charts. We then compared these similar questions. The comparison between the difficulty and discrimination indices is in table 6.6.

In our study, two items discriminated better (Bubble chart and Scatterplot), and both had moderate difficulty. One item did not discriminate better

[1]http://www.washington.edu/assessment/scanning-scoring/scoring/reports/item-analysis/, last visited in February

Table 6.6: Comparison between ours and VLAT results

		P		D	
Chart	**Type**	**Ours**	**VLAT**	**Ours**	**VLAT**
Bar (ordered by category)	Conceptual	0.87	-	0.36	-
	Suitable	0.91	0.88	0.18	0.21
	Unsuitable	0.90	-	0.32	-
Bar (stacked)	Conceptual	0.79	-	0.27	-
	Suitable	0.82	0.38	0.27	0.66
	Unsuitable	0.18	-	0.23	-
Bar (100% stacked)	Conceptual	0.72	-	0.45	-
	Suitable	0.81	0.49	0.50	0.57
	Unsuitable	0.25	-	0.18	-
Bubble chart	Conceptual	0.78	-	0.41	-
	Suitable	0.60	0.26	0.23	0.09
	Unsuitable	0.10	-	0.00	-
Histogram	Conceptual	0.75	-	0.32	-
	Suitable	0.84	0.84	0.32	0.26
	Unsuitable	0.51	-	0.45	-
Line (single)	Conceptual	0.88	-	0.27	-
	Suitable	0.94	0.98	0.09	0.03
	Unsuitable	0.59	-	0.41	-
Scatterplot	Conceptual	0.75	-	0.50	-
	Suitable	0.74	0.85	0.45	0.27
	Unsuitable	0.12	-	0.00	-
Pie chart	Conceptual	0.75	-	0.14	-
	Suitable	0.19	0.98	0.09	0.03
	Unsuitable	0.44	-	0.41	-

than the VLAT (Bar (stacked)), but it also had fair discrimination and moderate difficulty in our study. The remaining items had similar discrimination in both studies. From these, in one case, the VLAT item was more difficult (Bar (100% stacked)), and in another, ours was more difficult (Pie chart). Neither of the studies achieved good discrimination for the Pie Chart and Line (single) for the questions used. However, our study achieved good discrimination for these charts types with the conceptual and unsuitable questions, being the unsuitable items with a higher difficulty index.

We may notice that some indices differed between one test and another, even using similar questions and charts, changing only the domain concerning the questions (and, of course, the group of participants). However, our study had two more answer options than the VLAT (one always wrong: "I do not know" and another right only for unsuitable questions: "This type of chart does not allow or help to answer the question'), which may have caused these differences.

Overall, our results were similar or better than those of the VLAT, considering the conceptual and unsuitable question types for certain chart types. We demonstrated that the analysis questions with a correct answer among the alternatives are not enough to create a robust and complete test.

It is worth mentioning that item analysis is a process to assess the quality of the items test and the test as a whole. So we do this to choose items as final candidates, ones for improvement or elimination. In VLAT, despite the item analysis, all items were kept in the final test. The following section reports our procedure for choosing test items for our final test version.

6.5 Visualization Literacy Final Test

We listed test items according to their degrees of difficulty (easy, moderate, hard) and discrimination (good, fair, poor), see table 6.7. These distributions provide a quick overview of the test and identify items that are not performing well and possibly be improved or discarded.

Table 6.7: Distribution of Questions by Difficulty and Discrimination

	P		
D	**Easy**	**Moderate ○**	**Hard •**
poor	8, 29	32, 41	21, 36, 39, 44
fair ○	2, 5, 7, 28, 31	10, 11, 20, 23, 43	3, 12, 15, 38, 42
good •	4, 6	1, 13, 14, 16, 17, 19, 22, 25, 26, 27, 30, 33, 34, 35, 37, 40	9, 18, 24, 45

We can discard items that do not discriminate (poor discrimination index). Therefore, the items chosen for the final test are the ones that had fair or good discrimination, regardless of difficulty. In this case, the test also positions the participant: succeeds in the easy questions, or moderate to difficult questions, or in a distributed way. We then discarded eight items: 8, 21, 29, 32, 36, 39, 41, and 44.

After removing these items, we recalculated the test's internal consistency and obtained a Cronbach's alpha value of 0.86 (good). Of the candidate items to be removed because they would increase the test's internal consistency, we had already removed three of them because they did not discriminate (36, 41, and 44). We decided not to withdraw item 38 as it presented fair discrimination and difficulty and the consistency of the test was already acceptable with this item.

The final test has 37 items, distributed in 15 different charts types. Some types received all three question variations (conceptual, suitable, and unsuitable), while others received only two. However, all types received at least one valid conceptual question.

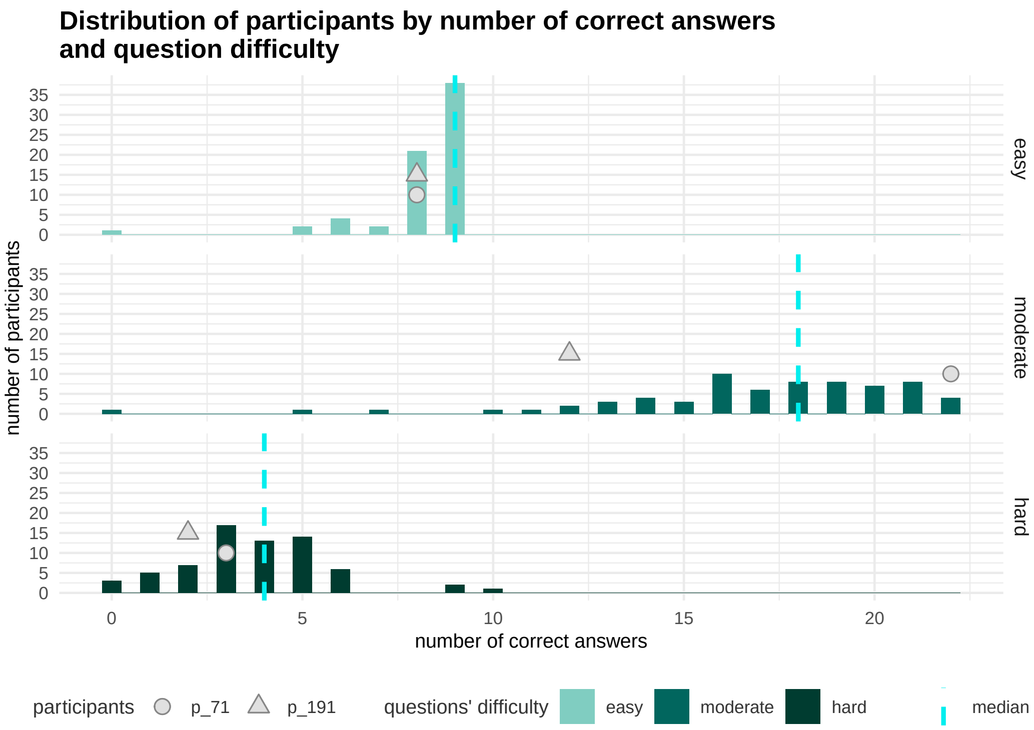


Figure 6.1: Participants' distribution according to the number of correct answers.

Figure 6.1 shows the distribution of participants according to the number of correct answers for each of the three types of question difficulty. We placed two participants (p_71 = 33 correct answers and p_191 = 22 correct answers)

Table 6.8: Final test itens (P all / D fair to Suitable)

Chart	**Type**	**P**	**D**
Bar (clustered)	Conceptual	0.65	0.32
	Suitable	0.96	0.18
	Unsuitable	0.44	0.18
Bar (ordered by category)	Conceptual	0.87	0.36
	Suitable	0.91	0.18
	Unsuitable	0.90	0.32
Bar (ordered by frequency)	Conceptual	0.90	0.18
	Unsuitable	0.46	0.45
Bar (stacked)	Conceptual	0.79	0.27
	Suitable	0.82	0.27
	Unsuitable	0.18	0.23
Bar (100% stacked)	Conceptual	0.72	0.45
	Suitable	0.81	0.50
	Unsuitable	0.25	0.18
Boxplot	Conceptual	0.66	0.68
	Suitable	0.76	0.45
	Unsuitable	0.19	0.32
Bubble chart	Conceptual	0.78	0.41
	Suitable	0.60	0.23
Bubble chart (color)	Conceptual	0.79	0.50
	Suitable	0.81	0.14
	Unsuitable	0.37	0.36
Histogram	Conceptual	0.75	0.32
	Suitable	0.84	0.32
	Unsuitable	0.51	0.45
Line (single)	Conceptual	0.88	0.27
	Unsuitable	0.59	0.41
Line (multiple)	Conceptual	0.87	0.23
	Unsuitable	0.68	0.59
Scatterplot	Conceptual	0.75	0.50
	Suitable	0.74	0.45
Scatterplot (color)	Conceptual	0.68	0.45
	Suitable	0.40	0.14
Table	Conceptual	0.82	0.50
	Unsuitable	0.47	0.18
Pie chart	Conceptual	0.75	0.14
	Unsuitable	0.44	0.41

on the chart to show that it is possible to identify which question type (regarding its difficulty level) each participant had the most difficulty. It shows that the difference between them is primarily in the moderate questions.

6.6 Concluding Remarks

This chapter introduced our study to create a data visualization literacy test following a different criterion than the tests in the literature. Tests typically include applied, suitable questions, *i.e.*, which one of the suggested answer alternatives can answer. In addition to these questions, which we considered suitable, we suggest using unsuitable ones. Also, we suggest using conceptual questions, which deal with the structure of the charts and are not applied analysis questions.

We performed an item analysis for each question, assessing the difficulty and discrimination of each one. In all cases, the conceptual-type items discriminated well. However, in four of them, the difficulty indices were not satisfactory, as they were easy questions: Bar (ordered by category), Bar (ordered by frequency), and Line (single and multiple).

In three cases, the unsuitable questions did not discriminate between the participants: Bubble chart, Scatterplot, and Scatterplot (color). For all other cases, we obtained moderate or high discrimination, but only one had easy difficulty: Bar (ordered by category)).

We obtained good discrimination and moderate to hard difficulty for the suitable type questions in eight items. In the other five items, we did not obtain good discrimination: Bar (ordered by frequency), Line (single), Line (multiple), Table, and Pie chart. The discrimination was good for the remaining items (2), but they were considered very easy (Bar (clustered) and Bar (ordered by category)).

We can conclude that the three types of questions evaluated are complementary and not exclusive. We should not just use suitable questions, or just unsuitable questions, or just conceptual questions. For each chart type, we can choose an item with good or fair discrimination and moderate or hard difficulty, except for one type: Bar (ordered by category), see table 6.8. It had its question items considered very easy, despite having high discrimination.

We also compared the result of our study with a similar study, VLAT (Lee et al., 2017), which used only suitable questions and a reduced number of visualization types. For the types in common in both studies, we compared the indexes values for the good items. In some cases, we got better combinations than they did; in others, we did not. However, for cases in which we did not obtain better results, the conceptual or unsuitable test items could be replaced as they obtained better combinations of the indices than the VLAT. Briefly, our study provided question items with better discrimination and difficulty for the same charts and questions used in the VLAT, considering a different data domain and question type.

After selecting only the items that discriminate, the final test had 37 items for 15 chart types regardless of difficulty, maintaining its good internal consistency. We also demonstrate that it is possible to position a participant according to the type of question they get/miss more concerning the median of correct answers.

7
Revisiting Visualization Task Taxonomies: Specifying Functions for the Data Transformations Stage

There are several different visualization taxonomies in the literature to guide the process of creating data visualizations. Each has its purpose, particularities, advantages, and disadvantages. These taxonomies use slightly different definitions of visualization tasks, so we need to look deeper into the definitions and not rely on the labels. They are full of overlaps, inconsistencies, and ambiguities that motivated us to create a precise specification of visualization tasks.

7.1
Goal

Data transformation involves selecting and manipulating the data for visual mapping. In this work, we discuss an approach to build visualizations and narrow the focus to define the data preprocessing operations as the first step essential to support visualization tasks.

By structuring these operations, we tried to avoid some of the ambiguity present in existing taxonomies and make it easier to implement them in executable code.

7.2
Procedure

Transforming data into a visual representation is only one of the stages of the visualization process. Ware (2019) defined this process as comprising four stages: data collection and storage, data transformations, mapping the selected data onto a visual representation, and visual and cognitive processing (fig. 7.1). We cannot map task taxonomies directly onto these stages. For example, *Find extremum* can be a visual and cognitive process if we preprocess the data neatly. It indicates that we need to decompose high-level visualization tasks into more specific actions at each stage of this visualization process.

Inspired by Ware's visualization process, we aim to define functions for three stages involved in building visualizations (fig. 7.2). In the data transformations stage (1), data-related functions prepare the data for visual

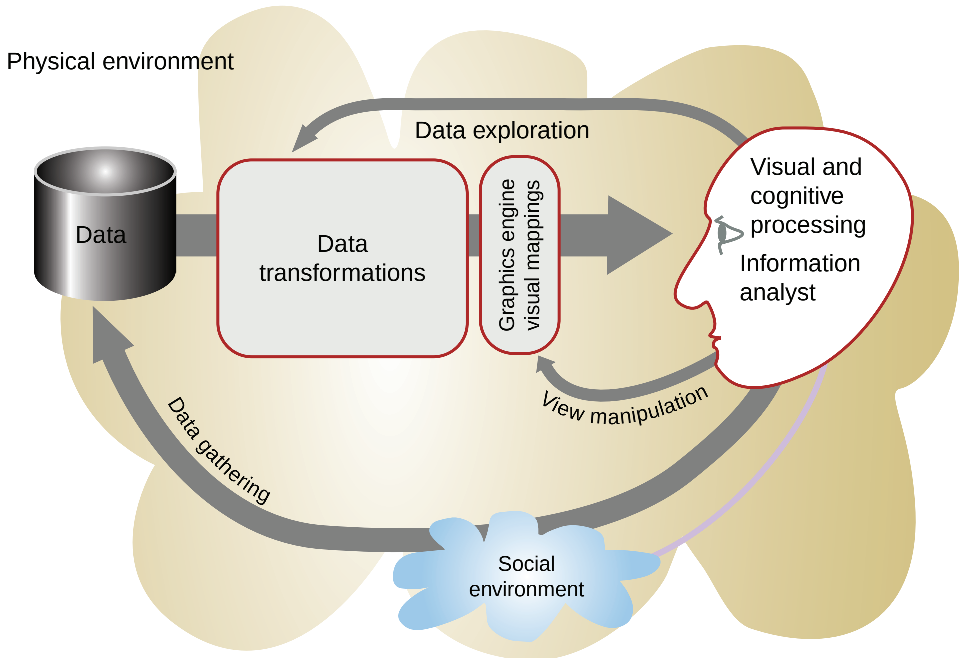


Figure 7.1: The visualization process (Ware, 2019, p.4)

encoding. It can require multiple rounds of transformations for adjusting it for the final visualization. After the data are ready to visualize, we move to the interactive exploration stage with the visual encoding (2) and the cognitive process (3). The visual encoding includes the structural mapping, which is what to present, and the selective or highlight, which is what to call attention to. In the third stage, the cognitive process occurs, where the user can see and explore the visual representation and perform mental functions to gain insights.

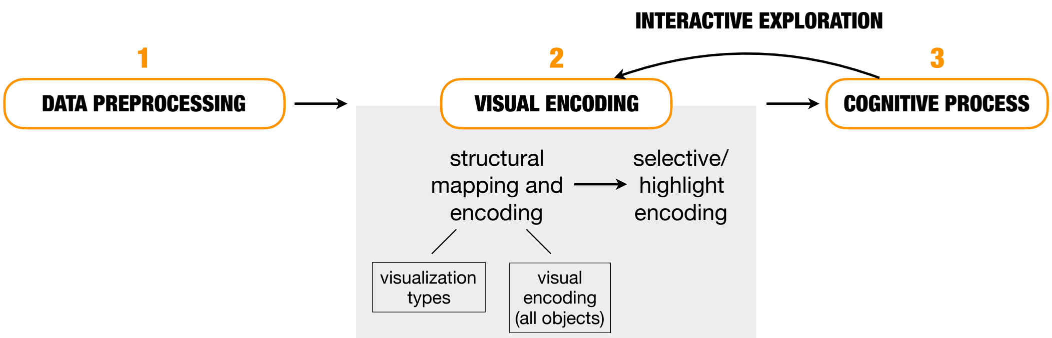


Figure 7.2: Our three-stage approach for defining visualizations.

We focus on defining functions for the first stage of the visualization process. From the investigated taxonomies, we defined ten functions related to data transformation. When structuring the functions, we noted that each one could support more than one task, even from the same taxonomy.

The final functions are: filter, identify, retrieve values, summarize, partition, map, super map, sort, find extremum, and categorize. We map characteristics such as an object, attribute, value, case, condition, and the pronoun used in the question examples for each function.

We provided a textual definition for each function, a functional notation as specification, the related tasks extracted from literature, a question template, and a sample question. We have also defined a color scheme to make it easier to identify and map features in functions and questions.

The functional notation allows us to directly describe specific high-level tasks as a composition of low-level tasks. *Determine range* is defined as a tuple of two *find extremum* applications, with minimum and maximum parameters.

Some tasks do not have a one-to-one correspondence with data transformation functions *e.g.*, *categorize* is a function of *map* and *partition*. For others, we could not associate with any functions in our model.

7.3 Data transformation functions

The taxonomies investigated comprise different stages of Ware's visualization process model. We focus on the 'data transformations' stage, *i.e.*, we focus on the data-related functions, and leave visual encoding functions and cognitive functions for future work.

The next subsections define the functions. One may note that each data function can support more than one task, even from the same taxonomy, as listed in the 'Related tasks' portion of each subsection. This means that the tasks may differ in terms of their visual encoding or visual and cognitive processing functions, but not in terms of the data transformation functions.

We have adopted the terminology objects, attributes, and values. Thinking of data in a table format, an object would be represented by a row, an attribute by a column, and a value by a cell. Note, however, that the terminology in different taxonomies differ. For instance, Amar et al. (2005) calls objects *cases*, whereas Valiati et al. (2006) calls them *items*. In our notation, uppercase letters denote sets ($e.g.$, O = objects; A = attributes; V = values) and lowercase letters denote elements of a set. We also use colors to facilitate the identification of the type of element, especially in the examples. For instance, $\boxed{O}$ denotes a set of objects, and $\boxed{o}$ denotes a single object.

7.3.1 Filter

The ***filter*** function returns a subset of objects of interest, given an input set of objects and a conditional expression on one or more attributes.

Specification: $filter(\boxed{O}, \boxed{\text{conditions on } \boxed{A}}) \rightarrow \boxed{O'}, O' \subseteq O$

Related tasks:

- filter Amar et al. (2005)
- find anomalies Amar et al. (2005)
- outliers Chen et al. (2009)
- (configure) filtering Valiati et al. (2006)

Question template: Which O satisfy boolean function(A, ...)?
Sample question: Which cities had over 500 homicides?

7.3.2 Identify

The ***identify*** function returns a single object of interest, given an input set of objects and a conditional expression on one or more attributes which uniquely identify the object. It can be considered as a special case of *filter*.

Specification: $identify(\boxed{O}, \boxed{\text{conditions on } \boxed{A}}) \rightarrow \boxed{o}, o \in O$

Related tasks:

- identify Wehrend and Lewis (1990); Valiati et al. (2006); Zhou and Feiner (1998)
- accurate value lookup Roth and Mattis (1990)

Question template: Which o has a=v?
Sample question: Which city has name São Paulo?

7.3.3 Retrieve Values

The ***retrieve_values*** function returns the values of the given attributes of a specified set of objects.

Specification: $retrieve\_values(\boxed{O}, \boxed{A}) \rightarrow \{(\boxed{o_i}, \boxed{V_i})\}, \forall i\ o_i \in O, V_i = o_i.A$

Related tasks:

- retrieve value Amar et al. (2005)
- accurate value lookup Roth and Mattis (1990)
- value Chen et al. (2009)
- (locate) values Valiati et al. (2006)

Question template: What is the attribute (value) of object?
Sample question: What is the number of homicides of the city of São Paulo?

7.3.4 Summarize

The ***summarize*** function returns a single value derived from applying a function to the set of values of a certain attribute of a set of objects.[1] Any function that receives an array of values and returns a single value may apply. Sample functions are: mean, median, min, among others. For functions that can be directly applied to the set of objects, such as *count*, the input attribute a is optional.

Specification: $summarize(\boxed{O}, \boxed{a}, \boxed{fn}, \ldots) \rightarrow \boxed{v}, v = fn(O.a, \ldots)$

Related tasks:

- compute derived value Amar et al. (2005)
- derived value Chen et al. (2009)
- (configure) derived attributes Valiati et al. (2006)
- (determine) mean, median, variance etc Valiati et al. (2006)
- characterize distribution Amar et al. (2005)
- distribution Chen et al. (2009); Wehrend and Lewis (1990)

[1] In this and all other cases that may receive a function as input, the function may also receive additional input, depicted by the ellipsis.

Question template: What is/was the summarization fn(a) of O?
Sample question: What was the average city homicide rate?

7.3.5 Partition

The ***partition*** function splits a set of object into a partition of the set according to the values of one or more attributes, which will be used to describe each set.

Specification: $partition(O, A) \rightarrow \{(O_i, V_i')\}, \forall i O_i \subseteq O, V_i' = O_{i1}.A$

Related tasks:

– cluster Amar et al. (2005); Chen et al. (2009); Wehrend and Lewis (1990); Zhou and Feiner (1998)

– (identify) clusters, (compare) clusters, (locate) clusters Valiati et al. (2006)

Question template: Which O have each value of A?
Sample question: Which cities are in each state?

7.3.6 Map

The ***map*** function returns a new attribute for O, whose values are the results of a given function applied to the value(s) of one or more original attributes of each object $o \in O$. This new attribute can then be used to partition O. Any function that receives one or more values and returns a single value may apply.

Specification: $map(O, A, fn, \ldots) \rightarrow a', \forall i\ o_i \in O, a_i' = fn(o_i.A, \ldots)$

Related tasks:

– compute derived value Amar et al. (2005)

– derived value Chen et al. (2009)

– (configure) derived attributes Valiati et al. (2006)

Question template: What are the values V = function(A) of O?
Sample question: What is the number of violent crimes of each city, considering it as the sum of homicide and manslaughter?

7.3.7
S-Map

The ***s-map*** function applies a given function fn to each set of objects O_i in a set of sets of objects (typically returned by *partition*) and associates each set O_i to a single value returned by fn. Any function that receives one or more values and returns a single value may apply. For functions that can be directly applied to the set of objects, such as *count*, the input attributes A are optional.

Specification: $smap(\{(O_i, \_)\}, A, fn, \dots) \rightarrow \{(O_i, v_i')\}, \forall i, v_i' = fn(O_i, A, \dots)$

Related tasks:

- characterize distribution Amar et al. (2005)
- distribution Chen et al. (2009); Wehrend and Lewis (1990)
- distribution of values Roth and Mattis (1990)

Question template: What is the summarization fn(Partition)?
Sample question: What is the total number of homicides per city?

7.3.8
Sort

The sort function returns a sequence of objects ordered according to a set of attributes in the specified directions.
Specification: $sort(O, A, Order) \rightarrow \langle o_i \rangle$, where $|A| = |Order| \wedge \langle o_i \rangle$ is a sequence of all the objects $o_i \in O$ ordered in terms of each attribute $a_j \in A$, following the corresponding order $order_j \in \{ascending, descending\}$.

Related tasks:

- sort Amar et al. (2005)
- rank Wehrend and Lewis (1990); Chen et al. (2009); Zhou and Feiner (1998)
- indexing Roth and Mattis (1990)
- (configure) dimensions order Valiati et al. (2006)

Question template: How are O ordered by A?
Sample question: How are cities ordered by violent crime rate?

7.3.9 Find Extremum

The ***find_extremum*** function returns a sequence of k top or bottom objects, sorted on a given set of attributes A.

Specification: $find\_extremum(O, A, k, Order) = select\_cases(sort(O, A, Order), \langle 1, \ldots, k \rangle) \rightarrow \langle o_1, \ldots, o_k \rangle$.

Note that *find_extremum* uses a supporting function defined as:

$select\_cases(O, Indices) \rightarrow \langle o_i \rangle, \forall i \in Indices \wedge Indices \subseteq \{1, \ldots, |O|\}$, *i.e.*, *Indices* is a sequence of indices to O.

Related tasks:

- find extremum Amar et al. (2005)
- extreme Chen et al. (2009)
- (identify) threshold Valiati et al. (2006)

Question template: Which are the k O with the $direction$ A?
Sample question: Which are the 5 cities with the highest homicide rate?

7.3.10 Categorize

The ***categorize*** function returns a partition of objects based on an attribute created through the application of some mapping function fn, which also returns a nominal or ordinal variable describing each set in the partition.

Specification: $categorize(O, A, fn, \ldots) = partition(O, map(O, A, fn, \ldots)) \rightarrow \{(O_i, v'_i)\}$

Related tasks:

- (identify) categories Valiati et al. (2006)
- categories Chen et al. (2009)
- categorize Zhou and Feiner (1998); Wehrend and Lewis (1990)

Question template: Which O are classified as A'= function(A) ?
Sample question: Which cities are safe or unsafe, as a function of the number of homicides ?

7.3.11
Composing functions

Some tasks are related to a composition of functions. For instance, we have seen that *categorize* is a function of *map* and *partition*.

The data operations related to other tasks in the investigated taxonomies can be similarly mapped onto an application multiple functions, or multiple applications of a single function. For instance, *determine_range* Amar et al. (2005) can be defined as a tuple of two *find_extremum* applications, with *min* and *max* parameters, respectively.

7.4
Visual Encoding and Visual and Cognitive Processing

As mentioned before, some tasks do not have a 1:1 correspondence with data transformation functions. Some tasks were not considered related to any functions in our model. In this section, we illustrate how a few tasks can be related to visual encoding functions and/or visual and cognitive processing functions.

7.4.1
Visual encoding functions

In terms of the visual encoding stage of the visualization process, we can outline two steps: structural encoding, and highlight encoding. Structural encoding is the process of selecting the type of visualization and mapping the different values into the structural slots of the selected visualization type. Highlight encoding is the process of selectively encoding a specific object's channel with one of its attribute values.

Examples of structural encoding include the mapping of objects to a column chart (vertical bar) format. This function could be described as follows:

$b1 = bar(O, a_{no}, a_q) \rightarrow vis(mark = bar, data = O, x = a_{no}, y = a_q)$, where a_{no} is a nominal or ordinal variable, a_q is a numeric (quantitative) variable

Examples of selective encoding include the mapping of specific object attributes to one of its channels, seeking to distinguish those objects from the others. This function could be described as follows:

$$distinguish(vis, O', a, channel)$$

For instance, $distinguish(b1, filter(O, a_q \geq 100), a_q, fillcolor = orange)$ changes to *orange* the $fillcolor$ channel of all objects whose values of a_q are greater than or equal to 100.

7.4.2 Visual and cognitive processing functions

Some taxonomies define tasks that rely heavily on cognitive processing, such as tasks for relating or comparing objects, including finding correlations and trends (*e.g.*, make comparisons Lee et al. (2017); association Chen et al. (2009); compare Roth and Mattis (1990); Wehrend and Lewis (1990); distinguish Wehrend and Lewis (1990); difference Chen et al. (2009); correlate Wehrend and Lewis (1990); Roth and Mattis (1990); Amar et al. (2005); finding correlations/trends Lee et al. (2017); trend Chen et al. (2009); infer Valiati et al. (2006)). Assuming that the relevant attributes have been visually encoded, these tasks can be supported by the following function:

$$compare(O, A) \rightarrow \{(r_i(o_{ip_A}, o_{iq_A}), S_i)\} : \forall i, p, q, p \neq q \wedge \\ o_{ip} \in S_i \;\wedge\; o_{iq} \in S_i \;\wedge\; S_i \subseteq O$$

These tasks can be performed on the whole dataset, without necessarily applying specific data transformation functions

7.5 Evaluation

We evaluated our set of functions' expressiveness using seventy-six empirically derived questions generated by multiple anonymous contributors using a movies domain, akin to that defined by IMDb.

It was possible to map all questions onto our data transformation functions. For each question in the study, we identified one or more (alternative) functions associated with it. Figure 7.3 shows the frequency in which each function occurred in the study questions.

We found that the most frequent function was *filter*, followed by *retrieve value(s)*. These two functions also often appeared together for a single question. The least frequent functions were *partition*, *map*, *categorize*, and *sort*. Maybe it is because they are more specific to low-level functions that may not occur as frequently in questions. Although the sort function was not frequent, one can use it indirectly inside the much more frequently *find extremum* function.

7.6 Concluding Remarks

We have specified data functions corresponding to the data transformations stage in Ware's visualization process. We related each function to one or more tasks found in widespread task taxonomies and provided a corresponding

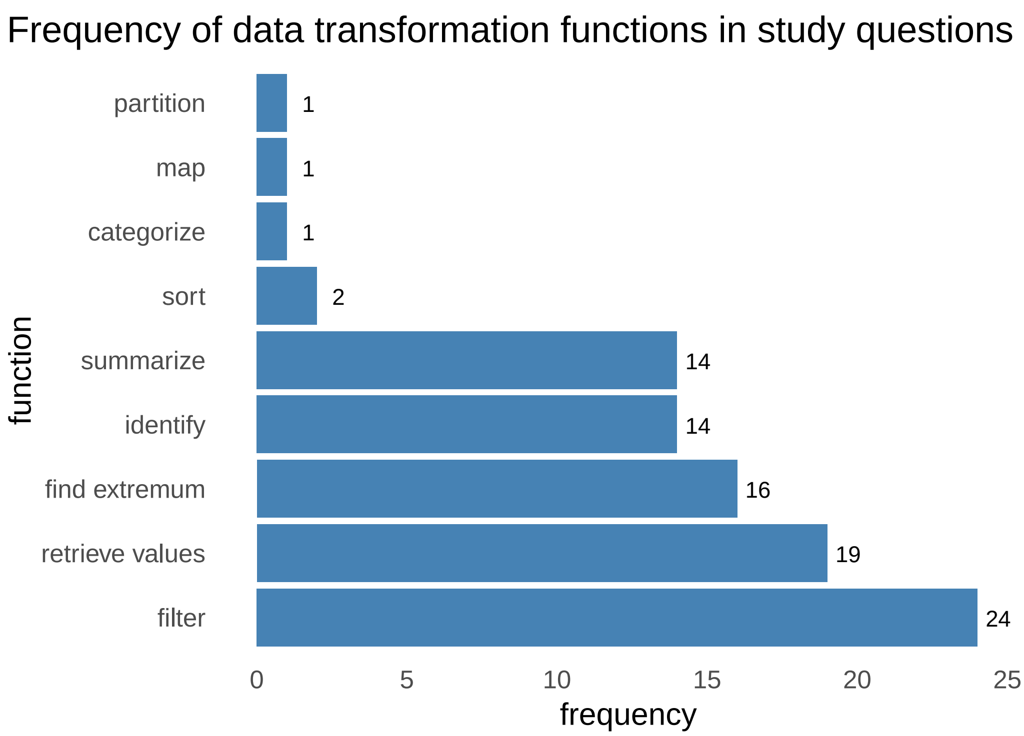


Figure 7.3: Frequency of functions found in the study questions

question template and examples to facilitate the usage of the functions. We defined these tasks using a functional notation so as to facilitate the composition of different low-level tasks into higher-level tasks.

A limitation of this work is to assume that a question template can be mapped onto a single data transformation function. Certain questions may combine different functions, requiring more advanced templates, or a way to compose our templates, to decompose them into our tasks. As future work, we plan to evaluate this assumption through an empirical study, and possibly extend our mappings between question templates and data transformation functions. We plan to extend our mappings between question templates and data transformation functions and to specify functions for both visual encoding and cognitive processing, setting higher-level visualization tasks and supporting the creation of the visualizations.

The paper "Revisiting Visualization Task Taxonomies: Specifying Functions for the Data Transformations Stage", published at HCII 2020 (Rodrigues et al., 2020), has the full description of this study.

8
Conclusion

The data-driven world requires data visualization to transform a large amount of raw and unorganized information into something functional and understandable. Despite being a subject broadly explored in the literature, we have not found more extensive studies which relate comprehension activities such as reading a chart correctly and extracting information from it. These activities are essential for teaching and learning data visualizations and data analysis tasks. This work advances data visualization education research by introducing a set of studies that identify some understanding gaps that can influence how non-experts interpret data visualizations and, so, serve as a resource in efforts to increase data visualization literacy. The following section 8.1 brings some reflections about learning and teaching Data Visualization. Section 8.2 summarizes our main contributions and section 8.3 highlights the next steps regarding this research.

8.1
Reflections about Learning and Teaching Data Visualization

Learning data visualization is not just learning how to interpret, understand, or create a data visualization. Learning data visualization is a process that consists of several steps. It is necessary to learn about the fundamentals: the characteristics of the data, the characteristics of the visualizations, color theory, how to create the visualizations, and how to interpret them. It is also necessary to learn about reliability: how to appropriately trust the data, trust that the visualization portrays the reality of the data, develop critical thinking, and, consequently, critically evaluate the visualization. Finally, one needs to learn how to deliver analytics and visualizations to an audience, especially how to know the audience's characteristics that need to be taken into account. It is necessary to learn how each visualization and information (the final insight) pair is unique to a given audience and how to make the same insight accessible to people with other profiles.

This thesis does not aim to define a data visualization teaching or learning model. The studies we have conducted call our attention to the importance of taking into account how people think of analysis questions when analyzing a visualization, how disturbances in the data can disrupt analysis, and how analysts cannot always make visualization better suited to answering a question. The main material of our studies was the analysis question. Throughout this thesis, we demonstrate several studies present in the literature that seek to identify more appropriate and even more efficient visualizations for specific problems, but little attention is given to analysis questions.

This section reflects the value of analysis questions for the entire data visualization process. Analysts can start and end with them. The starting point would be to define which questions can or should guide the analysis according to the database. Other questions may also emerge during this process. Knowing which types of visualization are most suitable to help answer each question is necessary. After this step, one can still define additional open questions, drawing attention to the need for data sources not considered at the beginning of the process.

If it were possible to choose a starting point to simplify teaching about data visualization, perhaps analysis questions would be a good choice. Among all the contributions already reported in this thesis from each of the studies carried out, an indirect contribution is the combination of analysis questions, data type, visualization task, and visualization type we use. These combinations can be used as examples of what works and does not work, what is allowed and what is not, and what is acceptable or not acceptable. Moreover, they also act as inspiration for new research in learning and teaching data visualization.

As a final product of this thesis, in addition to the factors we discovered that influence how data visualizations are interpreted by non-experts, we emphasize how essential analysis questions are for the entire data visualization process. However, there is still much to be researched and defined in this space.

8.2 Contributions

We investigated how non-experts interpret data visualizations as a first step towards devising approaches to increase data visualization literacy. To address this, we defined the research question: **What factors play a role in how novices interpret data visualizations?**. We unfolded this question into four subquestions, and conducted one empirical study to address each one.

Considering the first subquestion (**SQ1**: *What are the common novices' misinterpretations when trying to make sense of data visualizations?*), we set out to learn the data-related questions produced by 22 participants with minimal knowledge of data visualization when exposed to a set of twenty data visualizations we created (chapter 3). We collected and standardized 1,058 questions, resulting in 800 clearly answerable and 250 problematic questions. We rewrote the clear questions and generated a consolidated list of 249 unique question templates. We also used an open coded approach in the problematic questions resulting in a unified set of 20 codes, categorized into five major classes of problems: conceptual (not applied to the chart) questions (88 cases); ambiguous or unclear about the information need (41 cases); difficult to answer (with that specific chart – 43 cases), impossible to answer (that specific chart does not answer the question – 28 cases); and failure to follow the instructions, *i.e.*, did not provide a question about the chart (29 cases).

After analyzing participants' levels of effort and question order, we found a significant difference in the perceived effort to create a clear question and a problematic question, *i.e.*, they expended more effort in creating the questions that we later assessed as having lower quality. However, we expected the perceived effort level would increase with the number of questions created for each visualization, but it did not happen. Unfortunately, the study did not reveal an appropriate threshold for the number of questions we might ask participants to create effortlessly.

The study results reported can be used in teaching data visualization. They uncover and classify frequent errors people make when thinking about visually represented data. The question patterns may be used as a resource to provide more refined recommendations for creating visualizations to answer certain analysis questions. Recurring errors indicate limitations of data visualization education. These results can inform the design of visualization recommender systems, going beyond the association of the variable types with visualizations through supporting question-answering interactions more fully. They can also help users formulate better questions, providing more in-depth data analysis experience and more effective information seeking. The list of ambiguities found can aid in query-based data analysis systems, which can be designed to detect these instances and interact with users to clarify their intent.

On the one hand, researchers can use the question template as a guide to formulate questions in data visualization studies, visualization exploration tools, and literacy tests. On the other hand, each set of errors can guide new studies on data visualization literacy. **Conceptual** errors point to investigations about understanding the variables present in the charts. **Ambiguity** errors may bring misunderstandings that can affect the exploration activity, as each interpretation will provide a different answer. The **difficult-to-answer** questions suggest studies investigating how visualizations can be adapted to help answer them. Like conceptual questions, **does-not-answer** questions can also assist the development of literacy tests, just as we made in our study about visualization suitability for answering a particular analysis question (Appendix D).

To address **SQ2** (*Does the data distribution have a role in the interpretation of data visualizations?*), we investigated the visualization efficiency and effectiveness between nine visualization types and seven visualization tasks, according to two different data distributions: one clear of disturbances and a confusing one (chapter 4). We also measured the readers' perception of chart-task fit and confidence in their answers.

We found that some task-chart combinations performed better, regardless of the data distribution. Data distribution did affect participants' answers concerning efficiency, effectiveness, confidence, and perceived adequacy of the charts to certain tasks. The lowest confidence level occurred when the participants believed the chart did not answer the question. Moreover, as expected, they rated it as inadequate. We compared the correct answers across pairs of charts and ranked by effectiveness for the same task. We concluded that some charts that work for clear distributions might not work well for any distribution. This rank is a recommendation guide by chart types for each visualization task regarding the distribution. We highlighted the charts to avoid as they fared significantly worse than their counterparts.

Our study also demonstrates that the literature recommendations are inherently limited, as they do not consider the data distribution. We compared our results with three published guidelines applicable to the charts we investigated. Our results revealed the inadequacy of the existing recommendations, suggesting additional studies to explore these charts further.

To address **SQ3** (*How suitable do non-experts find certain data visualizations for a given analysis question?*), we studied how participants assess the suitability of certain data visualizations for answering specific analysis questions, before and after being exposed to related guidelines (chapter 5). We also investigated whether they could assess modifications suggestions (both textual and visual) to the visualizations to better answer the analysis questions and suggest good modifications for those they identified as unsuitable.

We discovered that, for basic charts, people could identify whether the visualization is appropriate to answer an analysis question. However, they cannot always suggest good improvements to make them more suitable. In general, the number of expected responses for visual suggestions was higher than for textual suggestions, but the difference was not significant. Furthermore, when we compare pairs of suggestions, this only occurs in half of the questions. Thus, we cannot say that participants performed better when exposed to visual suggestions than to textual ones. Our results also showed that guidelines, either in a textual or visual format, may not help novices to effectively relate analysis questions to specific chart properties. They also fall short in helping novices to improve charts to better answer specific questions. In fact, they can even confuse some novices and lead them to make mistakes they otherwise might not have made.

To address **SQ4** (*How can we assess a particular individual's data visualization literacy in detail, so as understand how to improve it?*), we devised a test for assessing people's understanding of both applied (suitable and unsuitable) questions and conceptual (which deal with the structure of the charts) aspects of data visualization. The test covered 15 different visualizations. We performed an item analysis for each question, assessing their level of difficulty and discrimination. After removing eight items that did not discriminate well, our final test ended with thirty-seven items, all with fair or good discrimination. This means that the test should genuinely represent the test takers' learning ability, *i.e.*, the items can discriminate well between the high and low-performing groups.

We obtained results similar or better than those of VLAT, considering the conceptual and unsuitable question types for certain chart types. Besides using a wider variety of charts, combining conceptual and unsuitable questions with suitable ones makes the test more comprehensive. The purpose of the test was twofold: to assess the participant's knowledge by giving them a score and by identifying where the gaps in knowledge lie. It can demonstrate the highest incidence of errors: in which question type (conceptual or applicable) and chart type combinations.

Besides the empirical studies, this work has made an additional contribution. Visualization tasks grounded all our studies. We have identified several task taxonomies in the literature. This motivated us to create a **unified list of visualization tasks in a structured format**: we specified ten data functions related to tasks found in comprehensive task taxonomies and provided a corresponding question template with examples to facilitate the usage of the functions. We defined these tasks using a functional notation to simplify the composition of different low-level tasks into higher-level ones. We focused on functions for the data transformation stage of Ware's visualization process.

Using the IMDb domain, we evaluated our set of functions' expressiveness using seventy-six empirically derived questions generated by multiple anonymous contributors. It was possible to map all questions onto our data transformation functions. Although we have not specified the functions for the other visualization process stages, we illustrated how a few tasks could be related to visual encoding functions and/or visual and cognitive processing functions.

8.3 Future Work

Upon analyzing the results of our study on how people make sense of data visualizations, we identified some limitations that can be covered in a new study. We still do not know how many questions we may ask for participants to generate before getting lower-quality results. Making a number of questions mandatory can help in this investigation. Another limitation concerns language: we left it optional for the participant to create the questions in their native language (Portuguese) or English. We do not know to what extent the language may have influenced the correctness of the questions. We did not consider grammatical errors, but we noticed that some people may have mistaken quantity concepts, such as how many/how much, which interferes with the treatment of objects and categorical and continuous variables. Considering only the formulation of questions in the participants' native language can help to sort out this issue.

We plan to apply this same questionnaire with knowledgeable participants to compare the results. We may want to identify whether the same types of error occur, whether new types of error emerge, and whether there are types of charts that cause misunderstandings regardless of the participant's level of knowledge.

As the questionnaire was anonymous, we could not conduct a more in-depth analysis. For instance, we identified the ambiguous questions, but we do not know the cause of the ambiguity, as we did not know who generated those questions. We want to delve into this issue. One alternative would be to ask participants to pose and answer questions and identify ambiguity sources. Another alternative would be to conduct interviews, in which we would be able to gather richer data.

Our investigation of the effectiveness and efficiency of data visualizations calls for further comprehensive studies. We plan to investigate more about Scatterplot and Bubble charts. We hypothesize that the visual clutter from the different sizes of the bubbles may have caused the difference in performances between these charts, but this requires further studies.

To derive more fine-grained recommendations, we wish to explore different disturbances combinations in each pair <task, chart>. Furthermore, several visualization types were not covered, such as hierarchical and georeferenced data and other distribution characteristics, such as with clusters.

We also seek to evaluate possible solutions for handling confusing distributions. For example, we can try to solve peak problems by showing two charts separately: one with an overview of the entire distribution and the other without the peak. The latter would have its scale adjusted, providing a more refined analysis. Future work must investigate and evaluate these solutions, incorporate them in a new study, and compare the results with those we already have.

As we found that short tutorials in the form of general guidelines are not enough to help people to make better decisions regarding the suitability of data visualizations and ways to improve them, we intend to investigate additional content and formats for educational material on data visualization, either static or interactive, aiming to support free exploration or answer specific analysis questions.

Our visualization literacy test study revealed that it may be more effective to define analysis questions that involve less trivial tasks. The usage of conceptual and unsuitable questions generated a satisfactory result. For future work, it would be interesting to extend the visualization literacy test using the approach described in Appendix E to other visualization types that were not covered here.

Although we developed a literacy test, we have not yet applied it in its final version. We plan to apply it in the Data Visualization courses to identify the students' literacy level and potential topics to cover more in depth during those courses.

Data storytelling is another field of interest, since questions can be essential tools to guide the data story and narrative. Assessing how and to what extent the knowledge of visualizations affects the understanding of the narrative is compelling research, which we are also interested in doing.

Bibliography

Abela, A. (2008). *Advanced presentations by design: creating communication that drives action.* John Wiley & Sons.

Ahlberg, C. (1996). Spotfire: An information exploration environment. *SIGMOD Rec.*, 25(4):25–29.

Albers, D., Correll, M., and Gleicher, M. (2014). Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 551–560. Association for Computing Machinery.

Alper, B., Riche, N. H., Chevalier, F., Boy, J., and Sezgin, M. (2017). Visualization literacy at elementary school. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5485–5497. Association for Computing Machinery.

Amar, R., Eagan, J., and Stasko, J. (2005). Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization*, pages 111–117. IEEE.

Benjamini, Y. (1988). Opening the box of a boxplot. *The American Statistician*, 42(4):257–262.

Bertin, J. (1983). *Semiology of graphics: diagrams, networks, maps.* California, Esri Press.

Bishop, F., Zagermann, J., Pfeil, U., Sanderson, G., Reiterer, H., and Hinrichs, U. (2020). Construct-a-vis: exploring the free-form visualization processes of children. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):451–460.

Börner, K., Maltese, A., Balliet, R. N., and Heimlich, J. (2016). Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization*, 15(3):198–213.

Bouali, F., Guettala, A., and Venturini, G. (2016). VizAssist: an interactive user assistant for visual data mining. *The Visual Computer*, 32(11):1447–1463.

Boy, J., Rensink, R. A., Bertini, E., and Fekete, J.-D. (2014). A principled way of assessing visualization literacy. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1963–1972.

Brehmer, M. and Munzner, T. (2013). A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385.

Bresciani, S. (2009). The risks of visualization: A classification of disadvantages associated with graphic representations of information. In *Identität und Vielfalt der Kommunikations-wissenschaft (2009)*, pages 165–178. UVK Verlagsgesellschaft GmbH.

Börner, K., Bueckle, A., and Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences*, 116(6):1857–1864.

Börner, K., Maltese, A., Balliet, R. N., and Heimlich, J. (2016). Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization*, 15(3):198–213.

Cairo, A. (2016). *The truthful art: data, charts, and maps for communication*. New Riders.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical methods for data analysis*. Chapman and Hall/Cole Publishing Company.

Chen, Y., Yang, J., and Ribarsky, W. (2009). Toward effective insight management in visual analytics systems. In *2009 IEEE Pacific Visualization Symposium*, pages 49–56. IEEE.

Chevalier, F., Henry Riche, N., Alper, B., Plaisant, C., Boy, J., and Elmqvist, N. (2018). Observations and reflections on visualization literacy in elementary school. *IEEE Computer Graphics and Applications*, 38(3):21–29.

Cleveland, W. S. and McGill, R. (1984). Graphical perception: theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554.

Cohen, R. J., Swerdlik, M. E., and Phillips, S. M. (1996). *Psychological testing and assessment: An introduction to tests and measurement, 3rd ed.* Mayfield Publishing Co. Pages: xxviii, 798.

Correll, M. and Gleicher, M. (2014). Error bars considered harmful: exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151.

Cox, K., Grinter, R. E., Hibino, S. L., Jagadeesan, L. J., and Mantilla, D. (2001). A multi-modal natural language interface to an information visualization environment. *International Journal of Speech Technology*, 4(3):297–314.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Dasgupta, A. and Kosara, R. (2010). The need for information loss metrics in visualization. In *Workshop on the role of theory in information visualization*.

de Araújo Lima, R. and Barbosa, S. D. J. (2020). A question-oriented visualization recommendation approach for data exploration. In *Proceedings of the International Conference on Advanced Visual Interfaces*, number 43, pages 1–5. Association for Computing Machinery.

de Santana, V. F., de Moraes, F. M., and de Souza, B. S. R. (2015). Investigating treemap visualization in inverted scale. In *Proceedings of the 14th Brazilian Symposium on Human Factors in Computing Systems*. Association for Computing Machinery.

de Santana, V. F., de Paula, R. A., and Pinhanez, C. S. (2017). Modeling task deviations as eccentricity distribution peaks. In *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems*. Association for Computing Machinery.

de Sousa, T. A. F. and Barbosa, S. D. J. (2014). Recommender system to support chart constructions with statistical data. In *Human-Computer Interaction. Theories, Methods, and Tools*, pages 631–642. Springer International Publishing.

Few, S. (2009). *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, 1st edition.

Fisher, R. A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94.

Fleiss, J. L., Levin, B., and Paik, M. C. (2003). *Statistical methods for rates and proportions*. John Wiley & Sons, 3rd edition.

Fujishiro, I., Ichikawa, Y., Furuhata, R., and Takeshima, Y. (2000). Gadget/iv: a taxonomic approach to semi-automatic design of information visualization applications using modular visualization environment. In *IEEE Symposium on Information Visualization 2000*, pages 77–83. IEEE.

Gnanamgari, S. (1981). *Information presentation through default displays*. PhD dissertation, The Wharton School, University of Pennsylvania.

Gotz, D. and Wen, Z. (2009). Behavior-driven visualization recommendation. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, pages 315–324. Association for Computing Machinery.

Gotz, D., When, Z., Lu, J., Kissa, P., Cao, N., Qian, W. H., Liu, S. X., and Zhou, M. X. (2010). Harvest: an intelligent visual analytic tool for the masses. In *Proceedings of the first international workshop on Intelligent visual interfaces for text analysis*, pages 1–4. Association for Computing Machinery.

Grammel, L., Tory, M., and Storey, M.-A. (2010). How information visualization novices construct visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):943–952.

Hanrahan, P. (2006). Vizql: A language for query, analysis and visualization. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, page 721. Association for Computing Machinery.

Heer, J. and Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 203–212. Association for Computing Machinery.

Heer, J., Kong, N., and Agrawala, M. (2009). Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1303–1312. Association for Computing Machinery.

Huang, W., Eades, P., and Hong, S.-H. (2009). Measuring effectiveness of graph visualizations: a cognitive load perspective. *Information Visualization*, 8(3).

Hullman, J., Qiao, X., Correll, M., Kale, A., and Kay, M. (2018). In pursuit of error: a survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):903–913.

Huron, S., Jansen, Y., and Carpendale, S. (2014). Constructing visual representations: investigating the use of tangible tokens. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2102–2111.

Keim, D. A., Mansmann, F., Oelke, D., and Ziegler, H. (2008). Visual analytics: combining automated discovery with interactive visualizations. In *Proceedings of the 11th International Conference on Discovery Science*, pages 2–14. Springer-Verlag.

Keller, P. R., Keller, M. M., Markel, S., Mallinckrodt, A. J., and McKay, S. (1994). Visual cues: practical data visualization. *Computers in Physics*, 8(3):297–298.

Keller, T., Gerjets, P., Scheiter, K., and Garsoffky, B. (2006). Information visualizations for knowledge acquisition: the impact of dimensionality and color coding. *Computers in Human Behavior*, 22(1):43–65.

Key, A., Howe, B., Perry, D., and Aragon, C. (2012). Vizdeck: self-organizing dashboards for visual analytics. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 681–684. Association for Computing Machinery.

Kim, D. H., Hoque, E., and Agrawala, M. (2020). *Answering questions about charts and generating visual explanations*, pages 1–13. Association for Computing Machinery.

Kim, Y. and Heer, J. (2018). Assessing effects of task and data distribution on the Eeffectiveness of visual encodings. *Computer Graphics Forum*, 37(3):157–167.

Klein, G., Phillips, J. K., Rall, E. L., and Peluso, D. A. (2007). A data-frame theory of sensemaking. In *Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making*, volume 113. Lawrence Erlbaum Associates New York.

Kodagoda, N., Wong, B. W., Rooney, C., and Khan, N. (2012). Interactive visualization for low literacy users: From lessons learnt to design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1159–1168. Association for Computing Machinery.

Koedinger, K. R., Baker, R., and Corbett, A. (2001). Toward a model of learning data representations. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, pages 45–50.

Krekhov, A., Michalski, M., and Krüger, J. (2019). Integrating visualization literacy into computer graphics education using the example of dear data. *Eurographics 2019 - Education Papers*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lee, B., Plaisant, C., Parr, C. S., Fekete, J.-D., and Henry, N. (2006). Task taxonomy for graph visualization. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, pages 1—-5. Association for Computing Machinery.

Lee, S., Kim, S., and Kwon, B. C. (2017). VLAT: Development of a visualization literacy assessment test. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):551–560.

Lee, S., Kim, S.-H., Hung, Y.-H., Lam, H., Kang, Y.-A., and Yi, J. S. (2016). How do people make sense of unfamiliar visualizations?: a grounded model of novice's information visualization sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):499–508.

Lee, S., Kwon, B. C., Yang, J., Lee, B. C., and Kim, S.-H. (2019). The correlation between users' cognitive characteristics and visualization literacy. *Applied Sciences*, 9(3):488.

Liu, S., Cui, W., Wu, Y., and Liu, M. (2014). A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393.

Livingston, M. A., Brock, D., Decker, J. W., Perzanowski, D. J., Van Dolson, C., Mathews, J., and Lulushi, A. S. (2019). A query generation technique for measuring comprehension of statistical graphics. In *International Conference on Applied Human Factors and Ergonomics*, pages 3–14. Springer.

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions On Graphics*, 5(2):110–141.

Mackinlay, J., Hanrahan, P., and Stolte, C. (2007). Show me: automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144.

Maltese, A. V., Harsh, J. A., and Svetina, D. (2015). Data visualization literacy: Investigating data interpretation along the novice—expert continuum. *Journal of College Science Teaching*, 45(1):84–90.

Munzner, T. (2009). A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928.

Ondov, B., Jardine, N., Elmqvist, N., and Franconeri, S. (2019). Face to face: evaluating visual comparison. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):861–871.

Ribeiro, F. C., Costa, L., de Paula, M., and de Souza, J. M. (2014). Uma proposta para classificação baseada em contexto para técnicas de visualização. In *Proceedings of the 13th Brazilian Symposium on Human Factors in Computing Systems*, pages 401–404. Sociedade Brasileira de Computação.

Rodrigues, A. M. B., Barbosa, G. D. J., Araújo Lima, R. d., Jack Freire Braga, D., Côrtes Vieira Lopes, H., and Barbosa, S. D. J. (2020). Revisiting visualization task taxonomies: specifying functions for the data transformations stage. In *International Conference on Human-Computer Interaction*, pages 655–671. Springer.

Rodrigues, A. M. B., Barbosa, G. D. J., Lopes, H., and Barbosa, S. D. J. (2019). Comparing the effectiveness of visualizations of different data distributions. In *32nd SIBGRAPI Conference on Graphics, Patterns and Images*, pages 84–91. IEEE.

Rodrigues, A. M. B., Barbosa, G. D. J., Lopes, H. C. V., and Barbosa, S. D. J. (2021). What questions reveal about novices' attempts to make sense of data visualizations: Patterns and misconceptions. *Computers & Graphics*, 94:32–42.

Roth, S. F. and Mattis, J. (1990). Data characterization for intelligent graphics presentation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 193–200. Association for Computing Machinery.

Ruchikachorn, P. and Mueller, K. (2015). Learning visualizations by analogy: Promoting visual literacy through visualization morphing. *IEEE Transactions on Visualization and Computer Graphics*, 21(9):1028–1044.

Saket, B., Endert, A., and Demiralp, Ç. (2019). Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2505–2512.

Sarikaya, A. and Gleicher, M. (2017). Scatterplots: tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):402–412.

Satyanarayan, A., Moritz, D., Wongsuphasawat, K., and Heer, J. (2017). Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350.

Schulz, H.-J., Nocke, T., Heitzler, M., and Schumann, H. (2013). A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2366–2375.

Setlur, V., Battersby, S. E., Tory, M., Gossweiler, R., and Chang, A. X. (2016). Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 365–377. Association for Computing Machinery.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343. IEEE.

Siegrist, M. (1996). The use or misuse of three-dimensional graphs to represent lower-dimensional data. *Behaviour & Information Technology*, 15(2):96–100.

Skau, D., Harrison, L., and Kosara, R. (2015). An evaluation of the impact of visual embellishments in bar charts. In *Computer Graphics Forum*, volume 34, pages 221–230. Wiley Online Library.

Spence, I. and Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology*, 5(1):61–77.

Srinivasan, A., Drucker, S. M., Endert, A., and Stasko, J. (2019). Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):672–681.

Srinivasan, A. and Stasko, J. (2017). Natural language interfaces for data analysis with visualization: Considering what has and could be asked. In *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers*, page 55–59. Eurographics Association.

Szafir, D. A. (2018). Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):392–401.

Tableau (2003). Tableau software. `http://www.tableausoftware.com/`, last accessed on July, 2020.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., and Hagen, E. P. (1991). *Measurement and evaluation in psychology and education*. Macmillan Publishing Co, Inc.

Toker, D., Conati, C., Carenini, G., and Haraty, M. (2012). Towards adaptive information visualization: on the influence of user characteristics. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 274–285. Springer.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Pr, 2nd edition.

Tufte, E. R., Goeler, N. H., and Benson, R. (1990). *Envisioning information*, volume 126. Graphics press Cheshire.

Tufte, E. R. and Graves-Morris, P. R. (1986). The visual display of quantitative information. *Information Design Journal*, 4(3):235–236.

Valiati, E. R. A., Pimenta, M. S., and Freitas, C. M. D. S. (2006). A taxonomy of tasks for guiding the evaluation of multidimensional visualizations. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, pages 1–6. Association for Computing Machinery.

Vartak, M., Huang, S., Siddiqui, T., Madden, S., and Parameswaran, A. (2017). Towards visualization recommendation systems. *SIGMOD Rec.*, 45(4):34–39.

Viegas, F. B., Wattenberg, M., Van Ham, F., Kriss, J., and McKeon, M. (2007). Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128.

Voigt, M., Pietschmann, S., Grammel, L., and Meißner, K. (2012). Context-aware recommendation of visualization components. In *The Fourth International Conference on Information, Process, and Knowledge Management*, pages 101–109. Citeseer.

Wang, Z., Sundin, L., Murray-Rust, D., and Bach, B. (2020). Cheat sheets for data visualization techniques. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery.

Ware, C. (2019). *Information visualization: perception for design*. Morgan Kaufmann.

Wehrend, S. and Lewis, C. (1990). A problem-oriented classification of visualization techniques. In *Proceedings of the 1st Conference on Visualization*, pages 139–143. IEEE Computer Society Press.

Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J., Howe, B., and Heer, J. (2016). Voyager: exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):649–658.

Zhou, M. X. and Feiner, S. K. (1998). Visual task characterization for automated visual discourse synthesis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 392–399. Association for Computing Machinery.

A
Study on Data Visualization: Terms and conditions

In this chapter, we present the informed consent form for the Visualization Literacy Study. All four studies described in this thesis were designed with a similar form, changing only the specific information for each test.

Welcome to our study on Data Visualization. Please read the Terms and Conditions of the study and inform your consent below. If you have any questions that you would like to clarify before deciding whether to participate in this study, please contact us at arodrigues@inf.puc-rio.br.

1. Purpose

This study is part of a broader research project led by Prof. Simone D. J. Barbosa at the Department of Informatics of the Pontifical Catholic University of Rio de Janeiro. This particular study is being developed by the Ph.D. candidate Ariane M. B. Rodrigues, and Master's students Gabriel D. J. Barbosa and Marisa do Carmo Silva. This study aims to evaluate how people with various degrees of knowledge in data visualization interpret different charts.

2. Procedures

There are minimal risks associated with your participation in this study, regarding possible discomfort at having to look at the display and interacting with input devices for the duration of the study.

3. Potential Benefits

This research is not designed to benefit you directly. We hope that, in the future, other people may benefit from this study through improved data visualization tools and educational material. The data collected will be anonymized and released exclusively in compiled technical reports, teaching materials, and/or scientific papers. It may be shared online to promote open research.

4. Confidentiality

All of the data collected in this survey will be anonymized in order to guarantee your anonymity. We will not store information about your location, software, or hardware. Your recorded answers will be anonymous and may be made available as a public resource for current and future research. All other collected information, such as any demographic information, will be kept confidential.

5. Ethical considerations

This research project has been approved by the Pontifical Catholic University of Rio de Janeiro's Institutional Review Board (Câmara de Ética em Pesquisa da PUC-Rio) for research involving human subjects.

Before you agree to participate, make sure you understand all of the terms here described. Upon completing this survey, you will be asked to confirm or revoke your consent. After completing the study, you may contact the researchers (via e-mail at arodrigues@inf.puc-rio.br) to learn more about the results. As all data collected will be anonymized, once you conclude the questionnaire, we may not be able to locate your answers to discard them later, unless you provide the specific date and time of the data collected. You will have two weeks to make such a request before we publicize the compiled results in a technical report.

6. Statement of Consent

By affirming your consent below, you indicate that:

- You have read this consent form or had it read to you;
- Your questions have been answered to your satisfaction;
- You voluntarily agree to participate in this research study;

Please save a copy of this page for your records.

Do you consent to the terms described above?
( ) I agree with the terms described above
( ) I do not agree with the terms described above

B
Making sense of Data Survey

In this chapter, we present the questions included in our study to assess the quality of data-related questions produced by people with minimal knowledge of data visualization, when exposed to different kinds of visualizations. We requested participants to ask up to five questions about the underlying data that could be answered by examining each visualization, presented in random order, one at a time. They were also asked to indicate the level of effort required to generate the question, on a 7-point scale, with 1 meaning "no effort", and 7 meaning "excessive effort".

Clustered bar chart

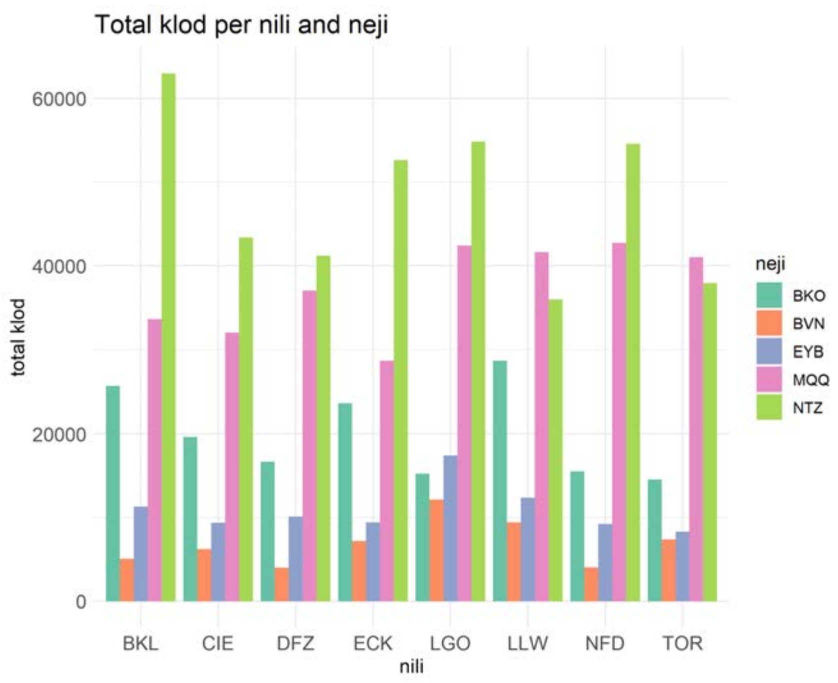


Figure B.1: Clustered bar chart

Frequency ordered bar chart

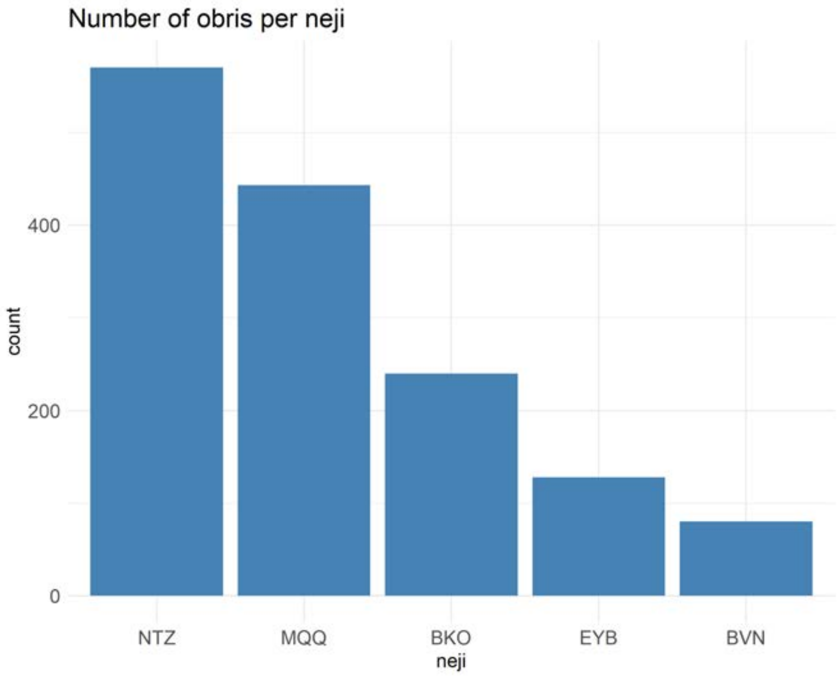


Figure B.2: Frequency ordered bar chart

Category ordered bar chart

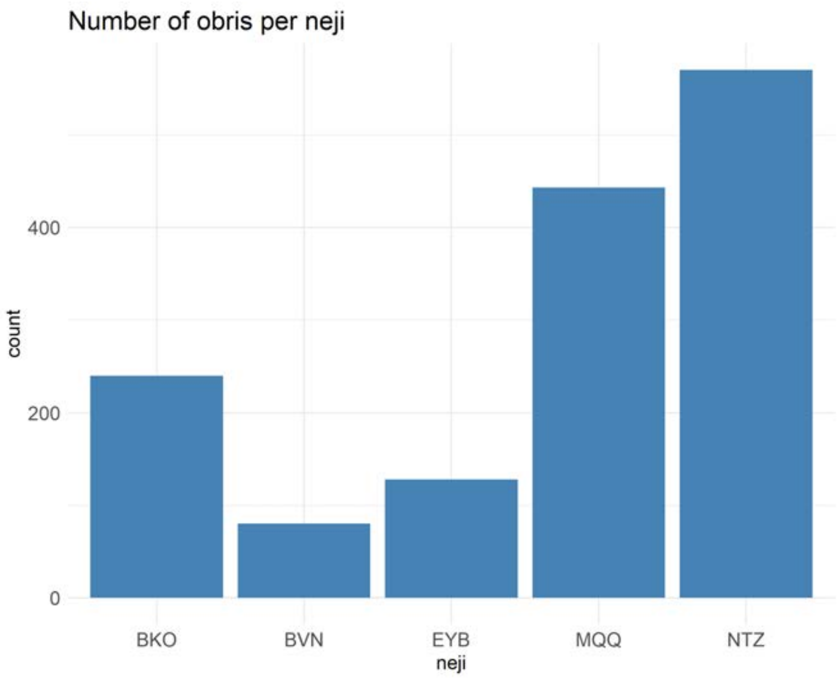


Figure B.3: Category ordered bar chart

Stacked bar chart

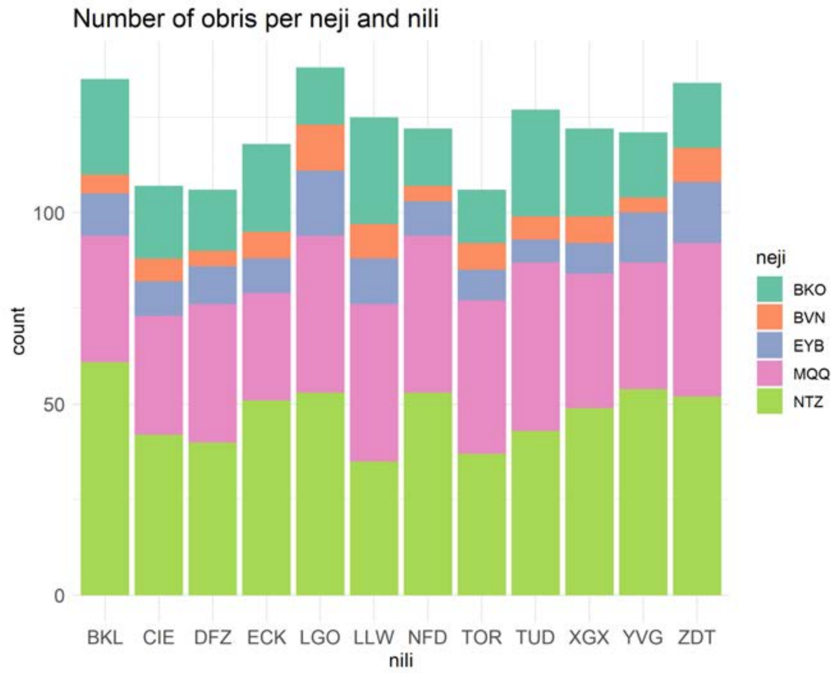


Figure B.4: Stacked bar chart

Boxplot

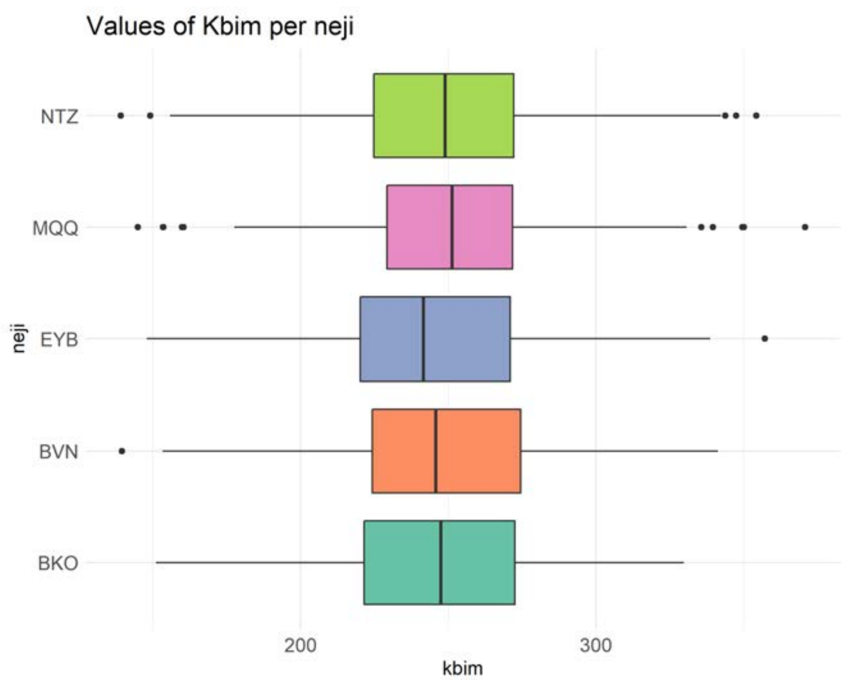


Figure B.5: Boxplot

Colored bubble chart

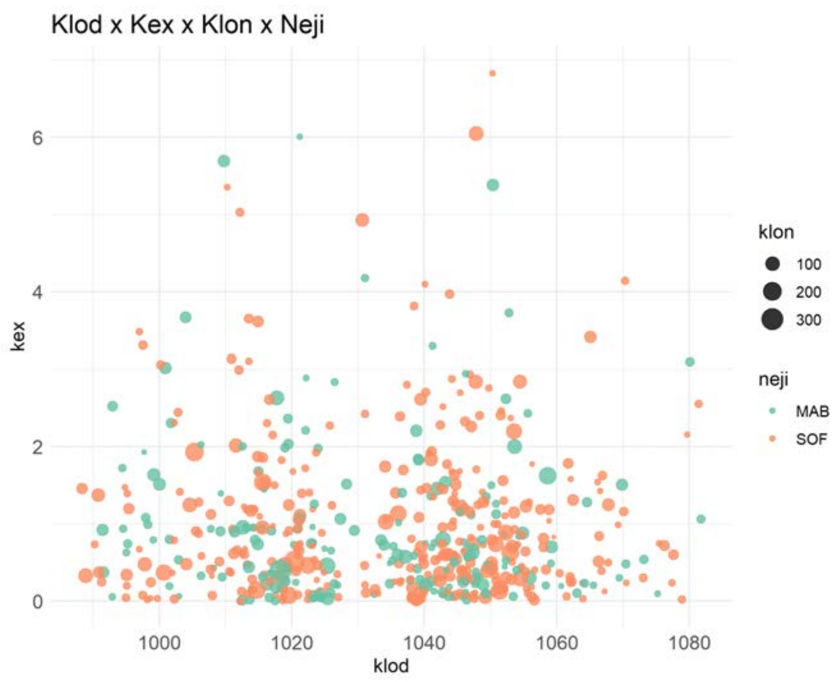


Figure B.6: Colored bubble chart

Bubble chart

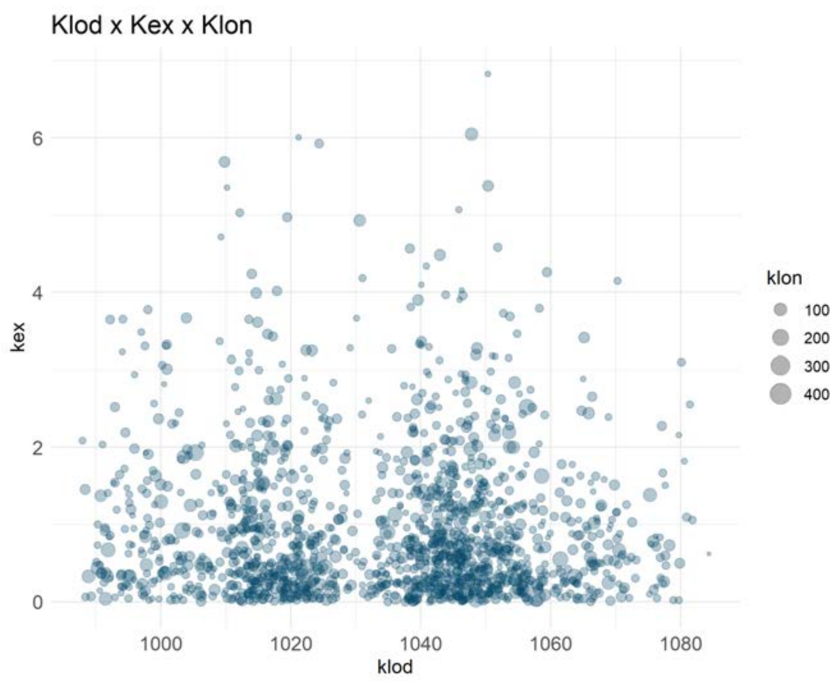


Figure B.7: Bubble chart

Chord diagram

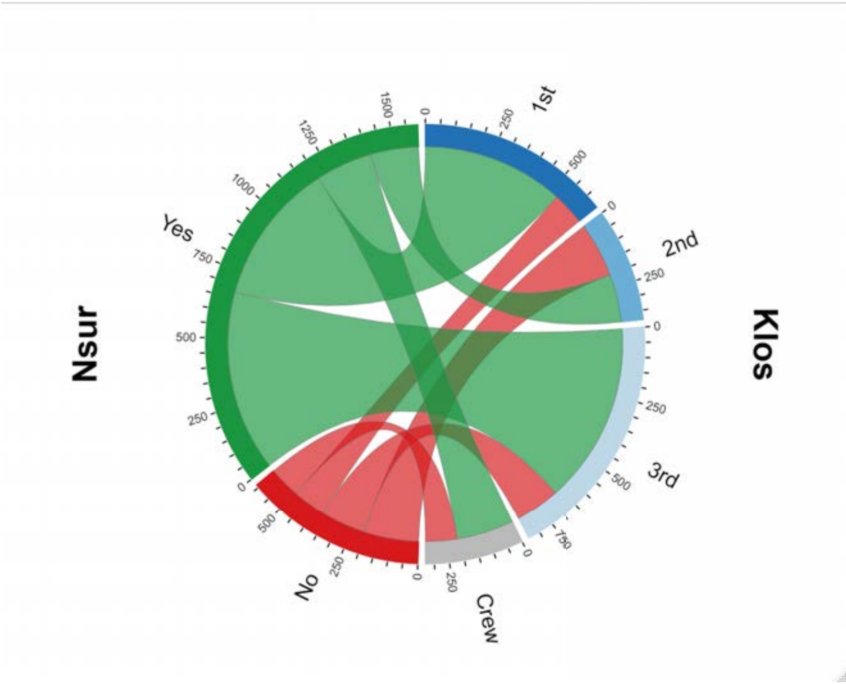


Figure B.8: Chord

Heatmap

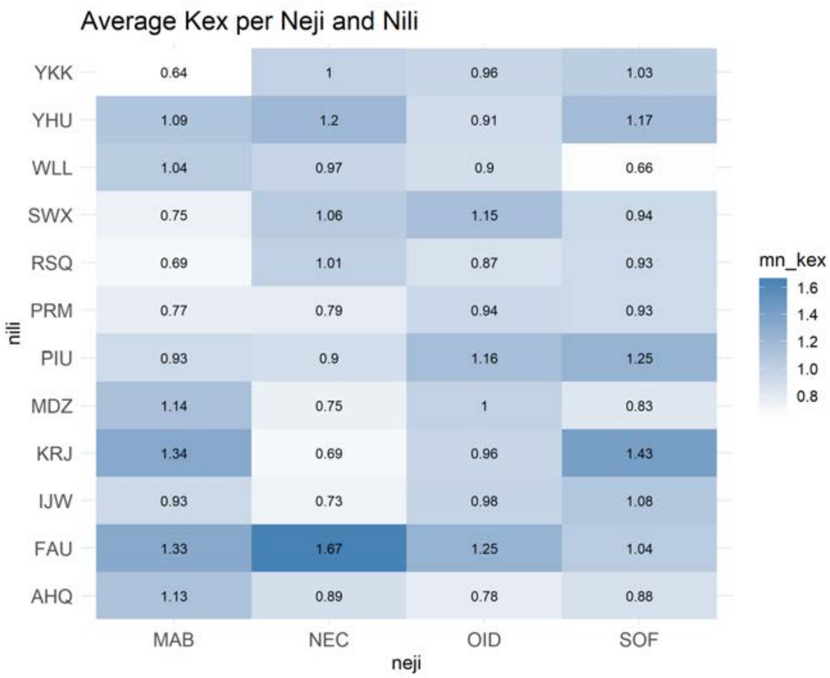


Figure B.9: Heatmap

PUC-Rio - Certificação Digital Nº 1712675/CA

Histogram

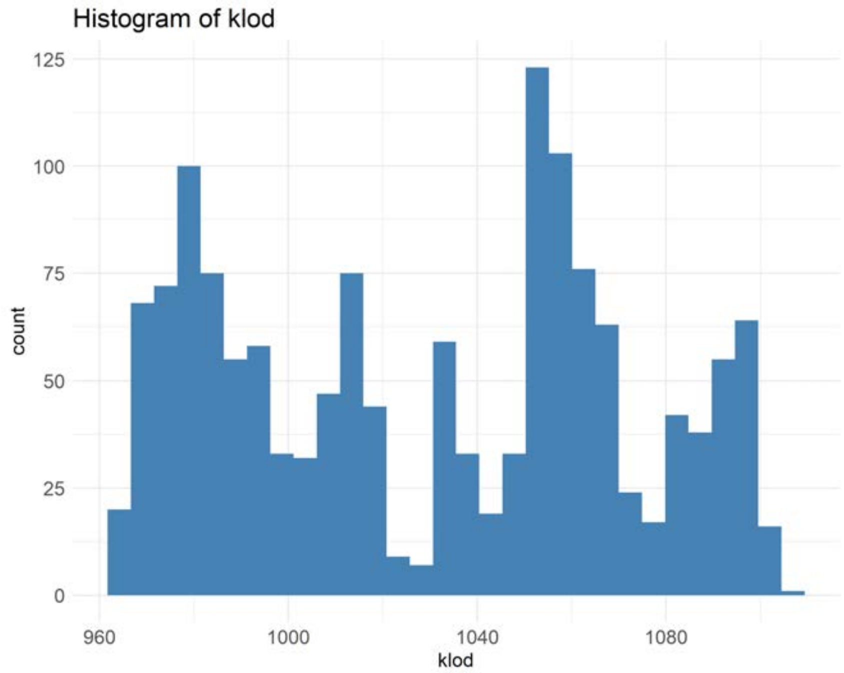


Figure B.10: Histogram

Multiple line chart

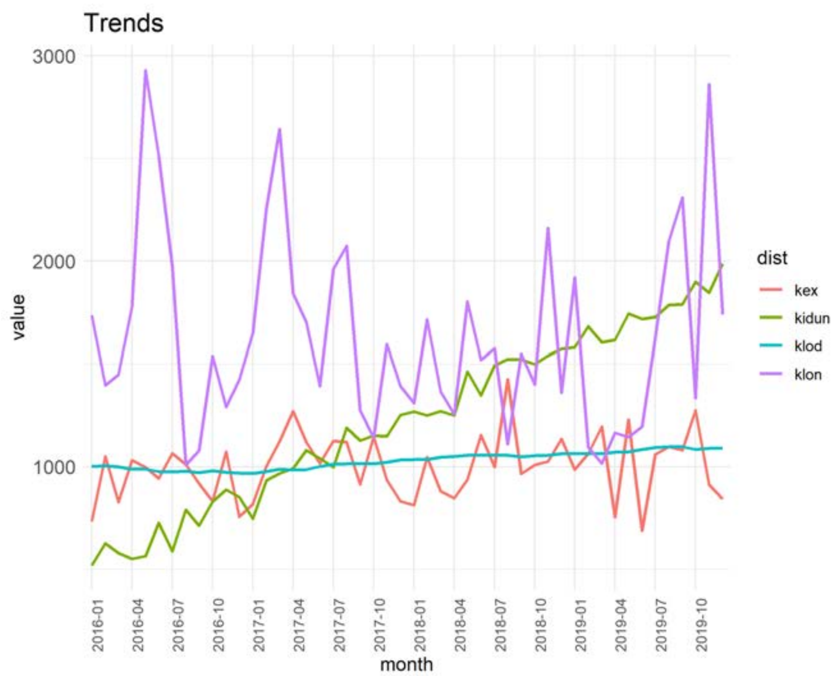


Figure B.11: Multiple line chart

Line chart

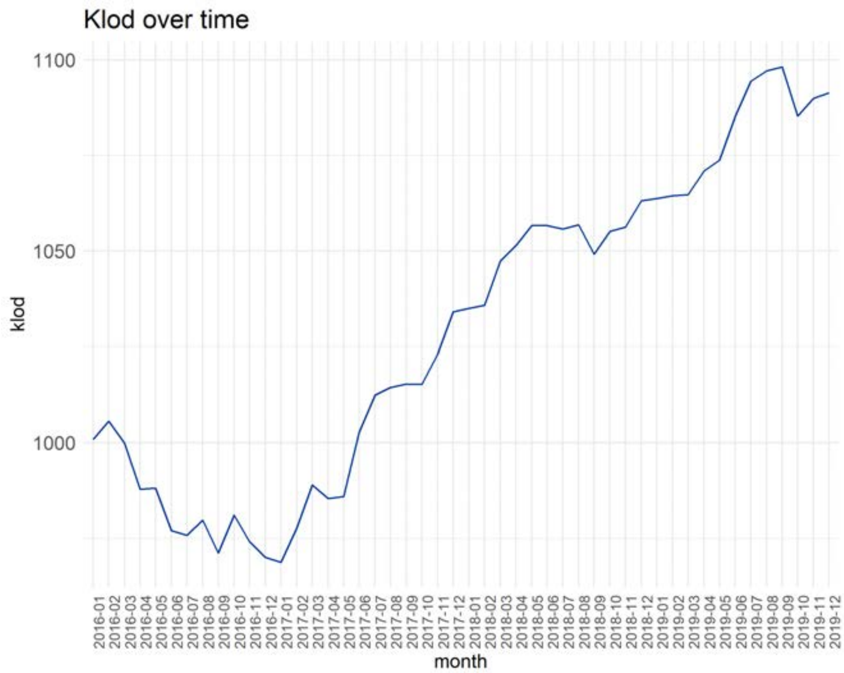


Figure B.12: Line chart

Cartogram

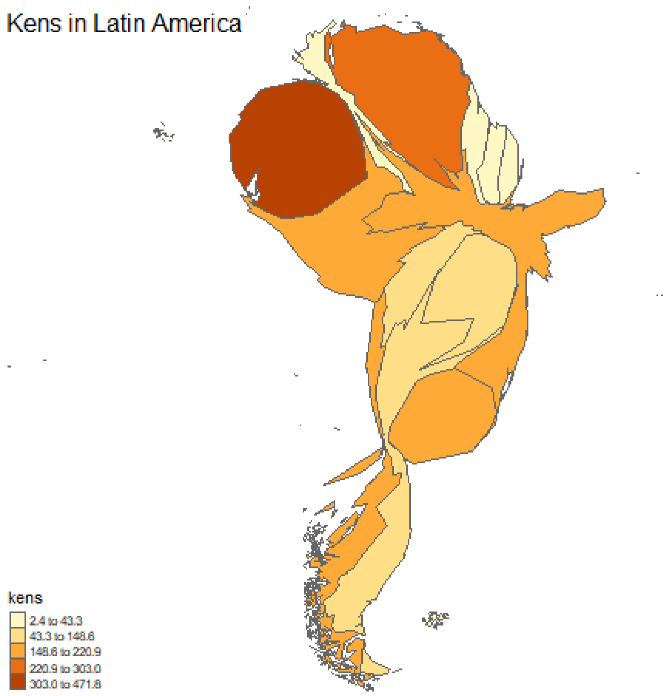


Figure B.13: Cartogram map

Choropleth map

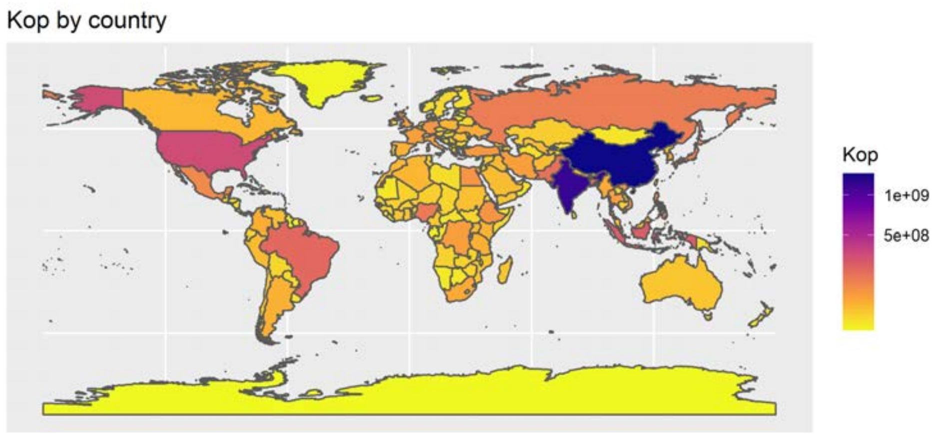


Figure B.14: Choropleth map

Network

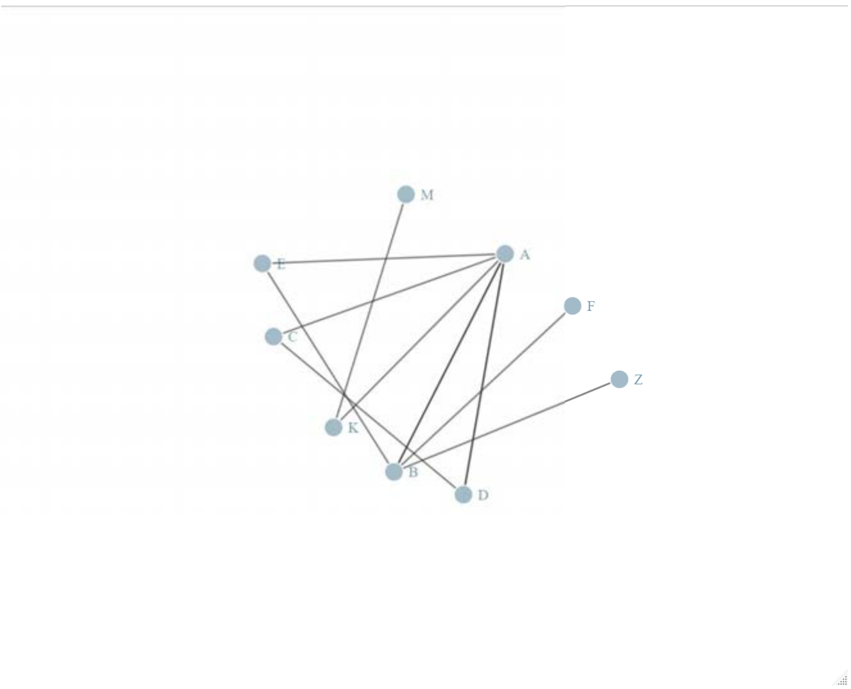


Figure B.15: Network

Ridge

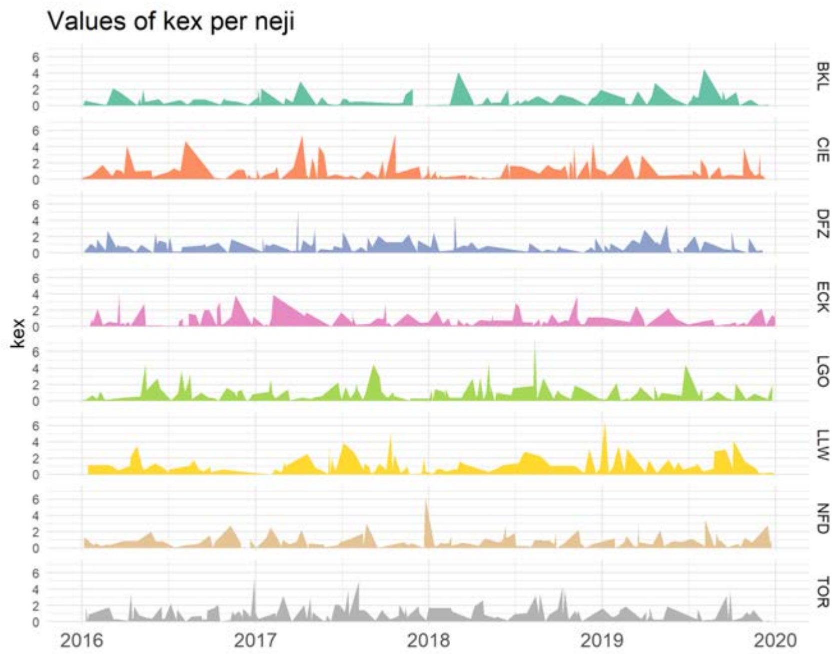


Figure B.16: Ridge

Sankey

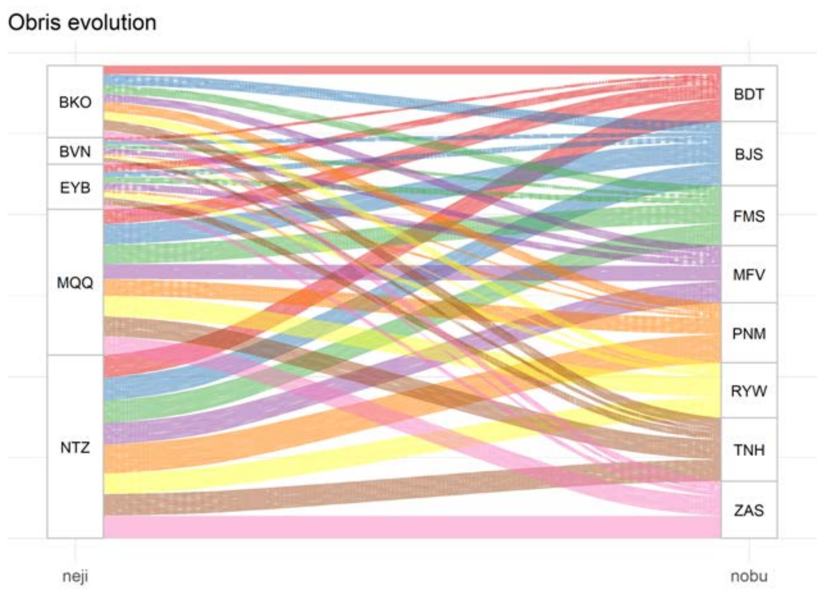


Figure B.17: Sankey

Colored scatterplot

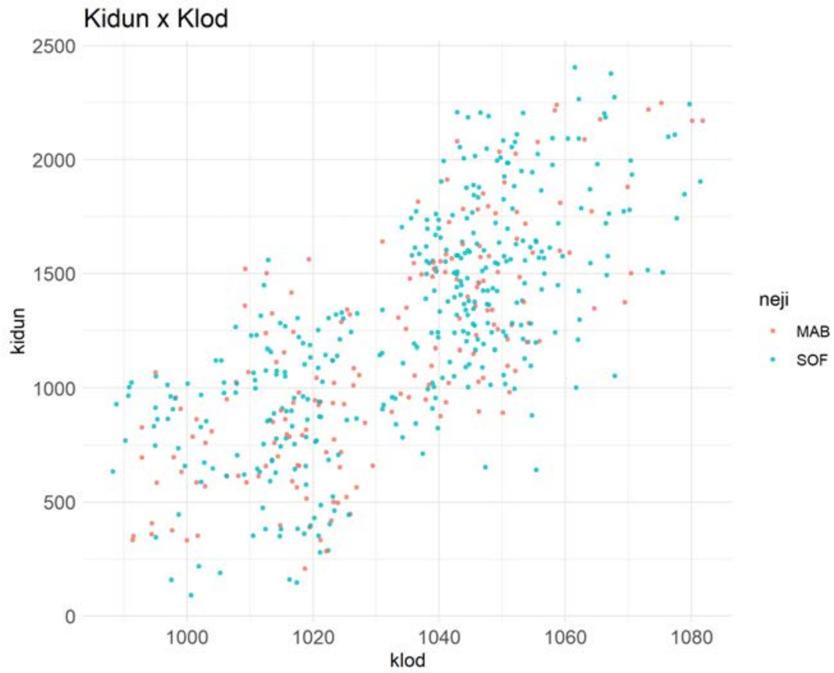


Figure B.18: Colored scatterplot

Scatterplot

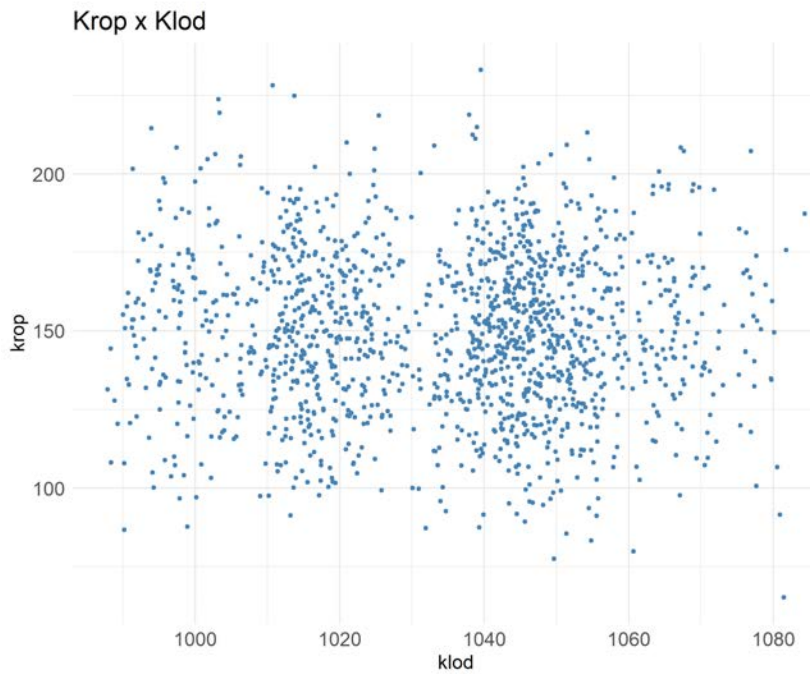


Figure B.19: Scatterplot

Table

neji	number of obris
BKO	247
BVN	80
EYB	155
MQQ	457
NTZ	522

Figure B.20: Table

C
Effectiveness of Visualizations Survey

In this chapter, we present the questions included in our study to explore how data distributions can affect the effectiveness and efficiency of data visualization. To simplify, we show a list separated by each visualization task type; the chart types specific to each task, and the data distribution type. In each of these triple, we present the resulting chart for the participant to analyze and answer the question. In the survey, we present them randomly.

At the beginning of the listing, we show an example/template question used for each one (appendix C), making the necessary modifications. The question and the answer options were the same for the same task, and we presented them at the beginning of each section. All questions included:

1. the main task-related question with the answer options to choose from and a chart to analyze and help answer the question;

2. a scale for choosing the confidence in the answer;

3. a scale for the charts' suitability in helping to answer the question; and

4. a free text field for comments.

It is important to note that for all questions, we added the option "The chart does not allow me to answer", letting us capture the participants' evaluation regarding the inadequacy of the visualization. We have also added a "None" option for multiple-choice or non-exclusive multiple-choice questions.

In which year was Comedy's raw profit minimal?

○ None ○ 2001 ○ 2002 ○ 2003 ○ 2004 ○ 2005 ○ 2006 ○ 2007 ○ 2008 ○ 2009
○ 2010 ○ 2011 ○ 2012 ○ 2013 ○ 2014 ○ 2015 ○ The chart does not allow me to answer

What is your confidence in the answer?

○ 1-None ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7-I'm sure it is right

How good is this visualization to help answer the question?

○ Terrible ○ Very bad ○ Bad ○ Not so bad ○ Good ○ Very good ○ Excellent

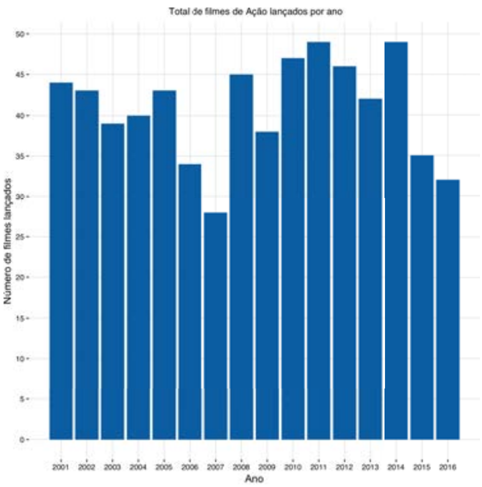
Additional comments (optional):

Retrieve Value

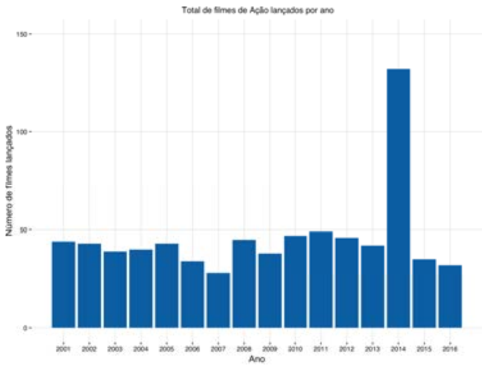
Question: Quantos filmes de Ação foram lançados em 2015? / How many Action movies were released in 2015?

Answer option: free text

Charts: figs. C.1 to C.8

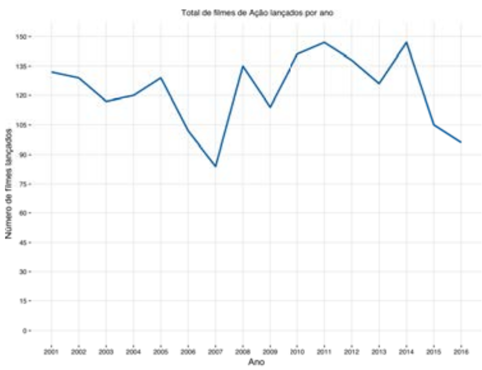


C.1(a): clear

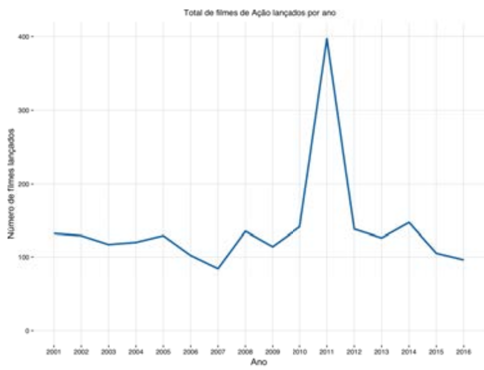


C.1(b): confusing

Figure C.1: Bar - Retrieve Value

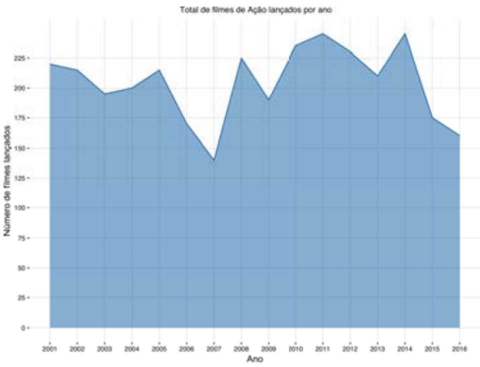


C.2(a): clear

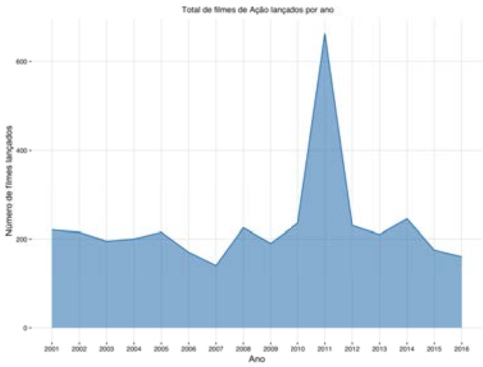


C.2(b): confusing

Figure C.2: Line - Retrieve Value

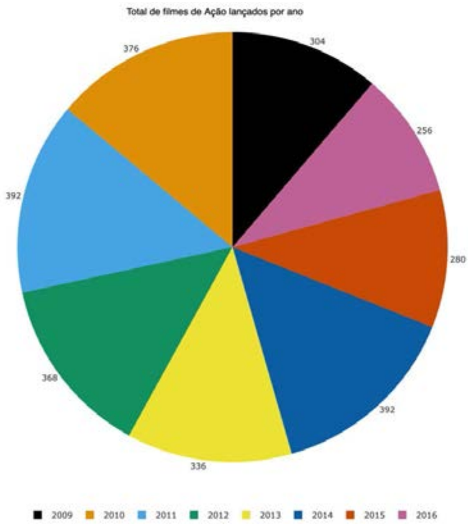


C.3(a): clear

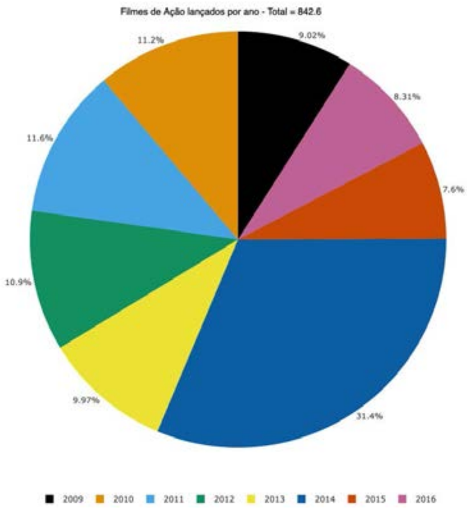


C.3(b): confusing

Figure C.3: Area - Retrieve Value

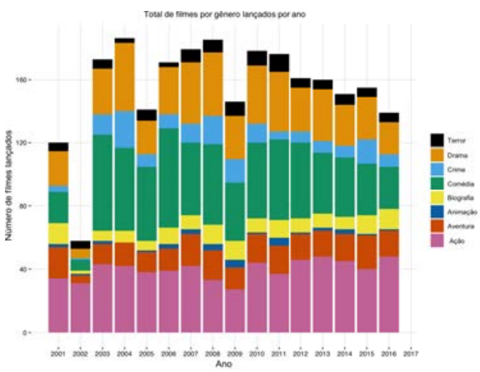


C.4(a): clear

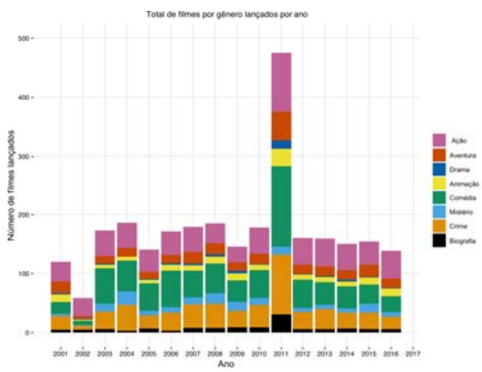


C.4(b): confusing

Figure C.4: Pie - Retrieve Value

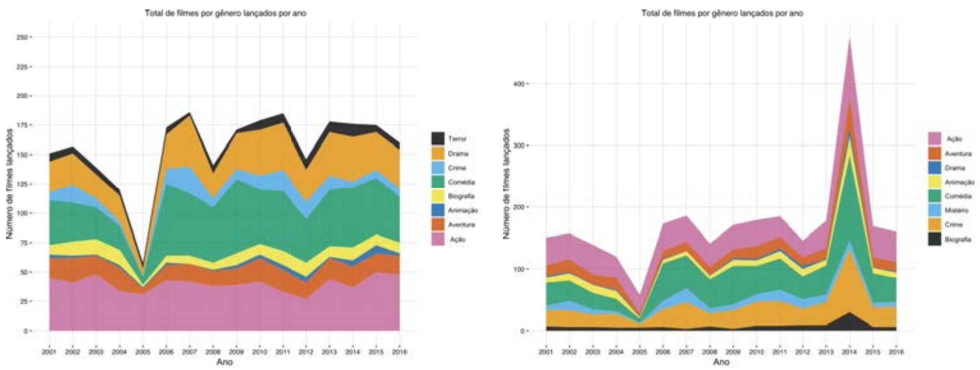


C.5(a): clear



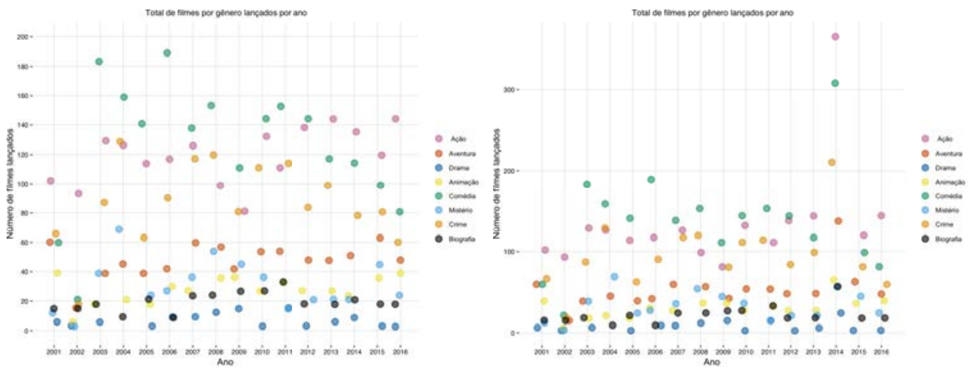
C.5(b): confusing

Figure C.5: Stacked Bar - Retrieve Value



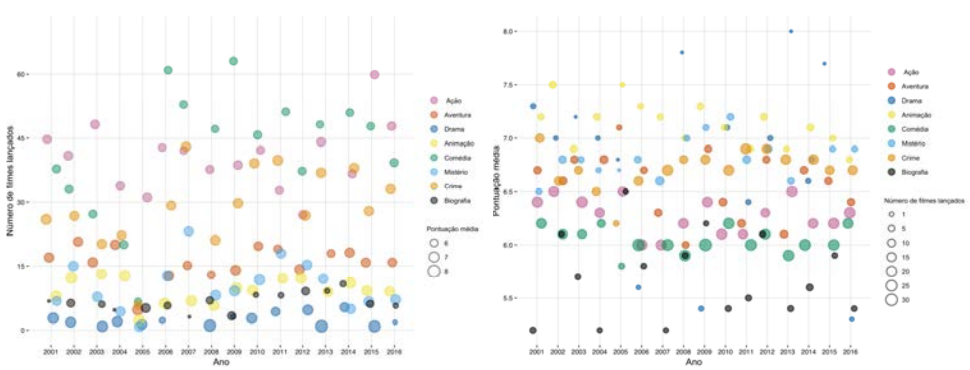
C.6(a): clear C.6(b): confusing

Figure C.6: Stacked area - Retrieve Value



C.7(a): clear C.7(b): confusing

Figure C.7: Scatterplot - Retrieve Value



C.8(a): clear C.8(b): confusing

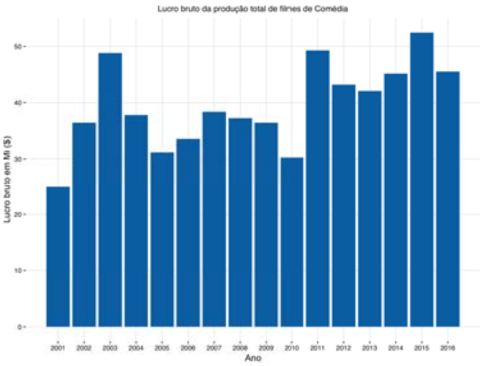
Figure C.8: Bubble chart - Retrieve Value

Find Extremum

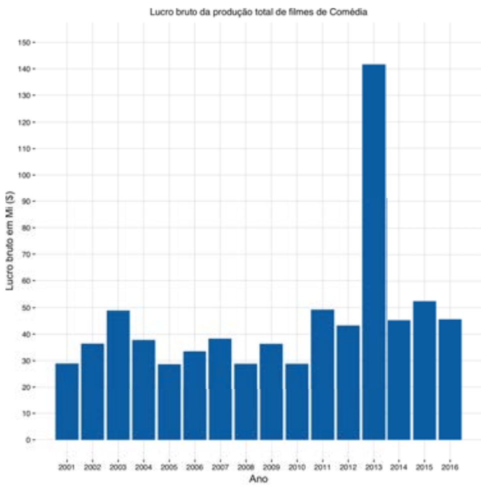
Question: Em qual ano o lucro bruto de Comédia foi mínimo? / When Comedy gross profit was minimal?

Answer options: select one (options: one year per option)

Charts: figs. C.9 to C.16

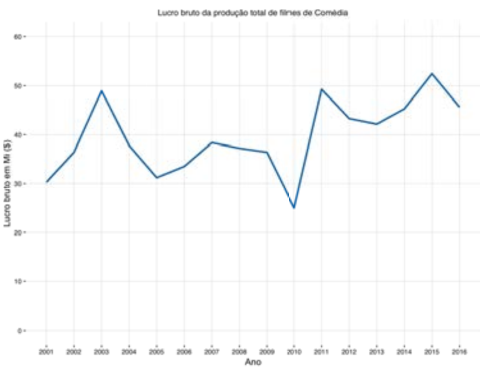


C.9(a): clear

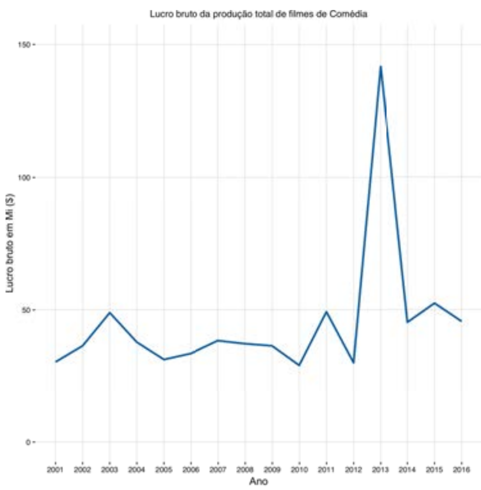


C.9(b): confusing

Figure C.9: Bar - Find Extremum

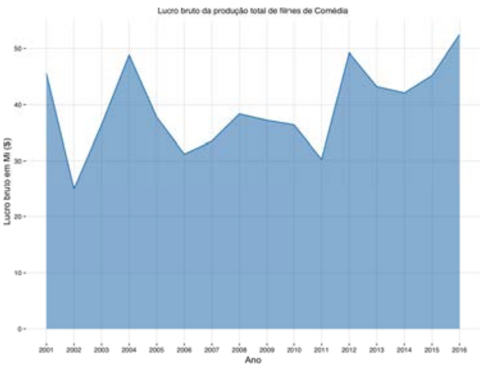


C.10(a): clear

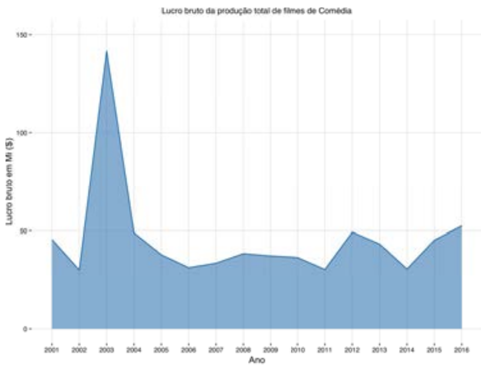


C.10(b): confusing

Figure C.10: Line - Find Extremum

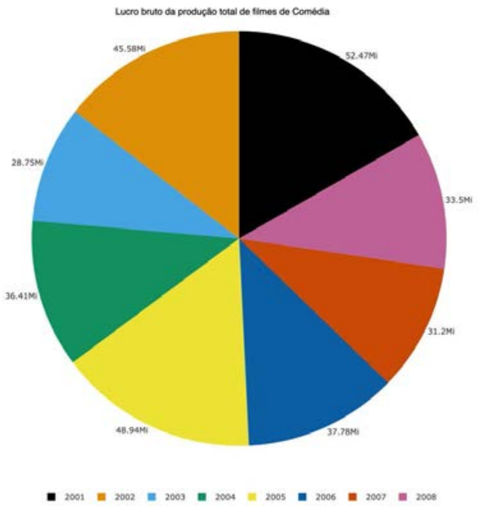


C.11(a): clear

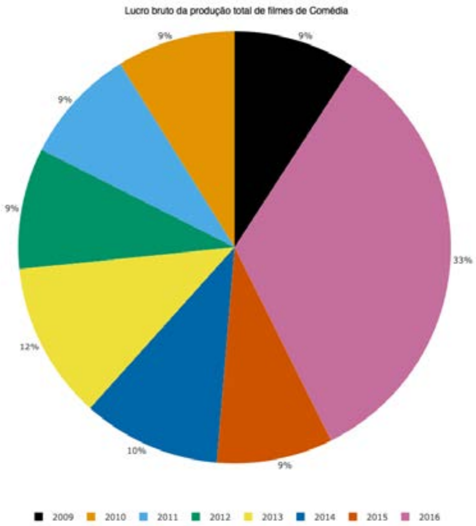


C.11(b): confusing

Figure C.11: Area - Find Extremum

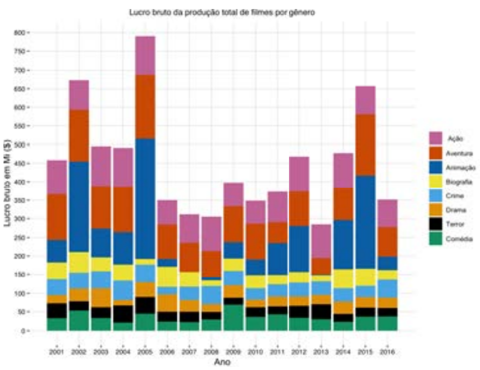


C.12(a): clear

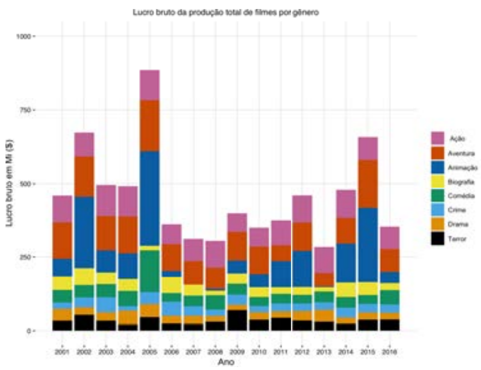


C.12(b): confusing

Figure C.12: Pie - Find Extremum

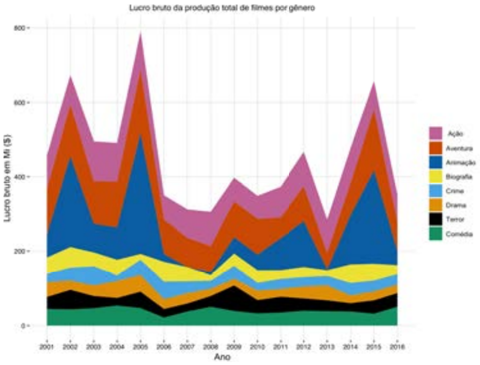


C.13(a): clear

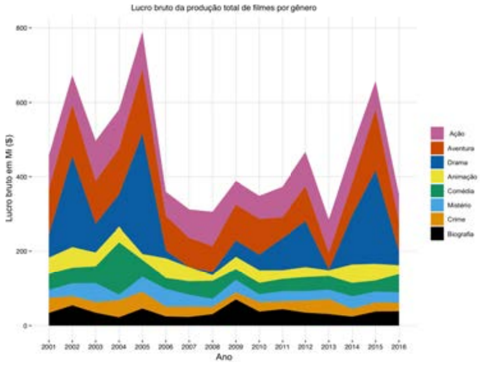


C.13(b): confusing

Figure C.13: Stacked bar - Find Extremum

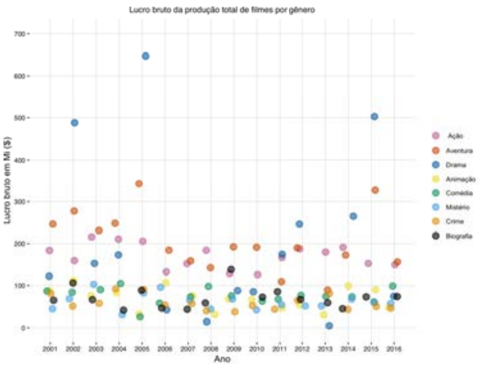


C.14(a): clear

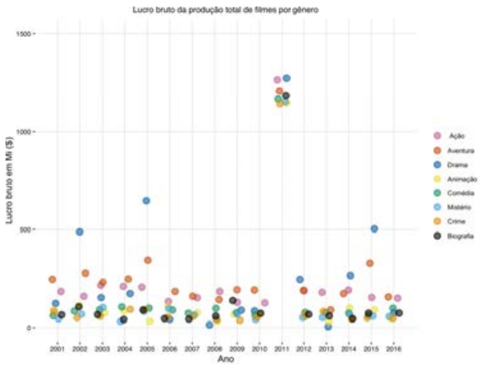


C.14(b): confusing

Figure C.14: Stacked Area - Find Extremum

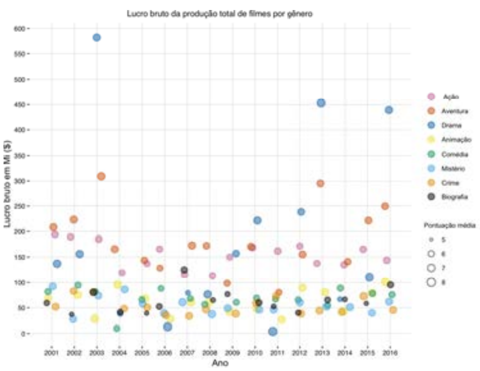


C.15(a): clear



C.15(b): confusing

Figure C.15: Scatterplot - Find Extremum



C.16(a): clear

C.16(b): confusing

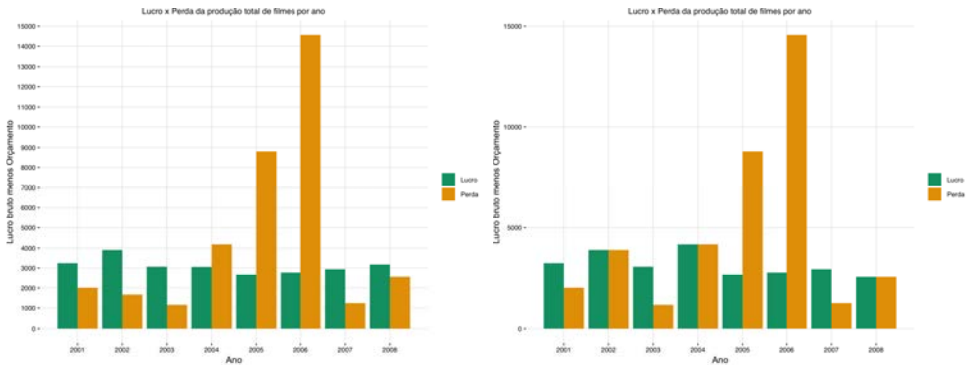
Figure C.16: Bubble chart - Find Extremum

Make Comparisons

Question: Em quais anos houve mais perda do que lucro? / In which years there were more non-profit movies than profit ones?

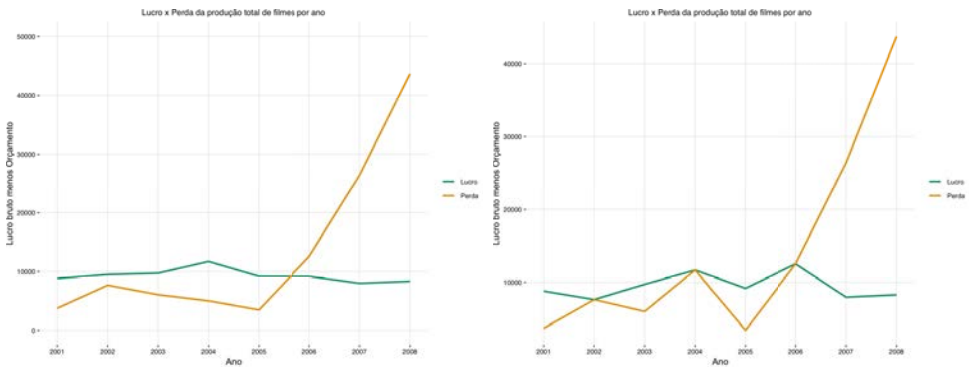
Answer options: select multiples (options: one year per option)

Charts: figs. C.17 to C.24



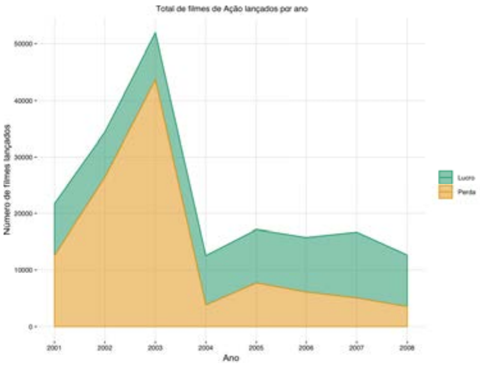
C.17(a): clear C.17(b): confusing

Figure C.17: Bar - Make Comparisons

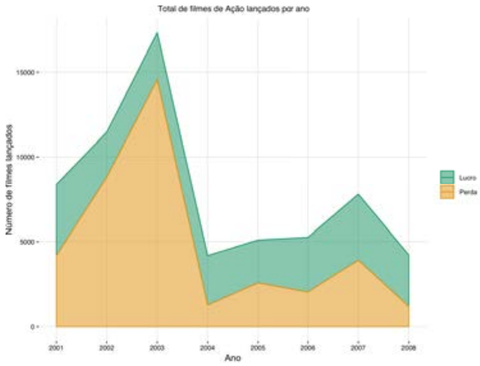


C.18(a): clear C.18(b): confusing

Figure C.18: Line - Make Comparisons

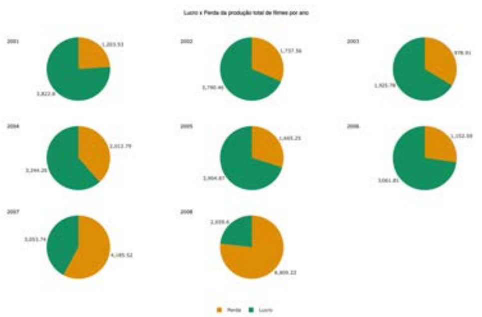


C.19(a): clear

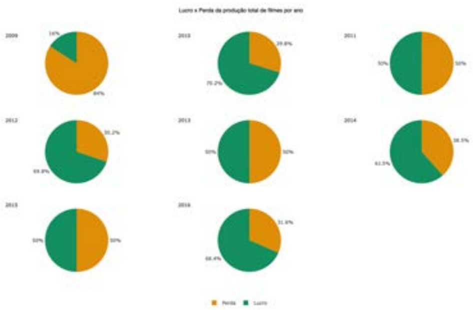


C.19(b): confusing

Figure C.19: Area - Make Comparisons

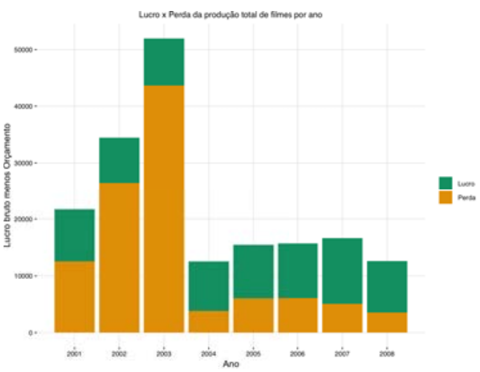


C.20(a): clear

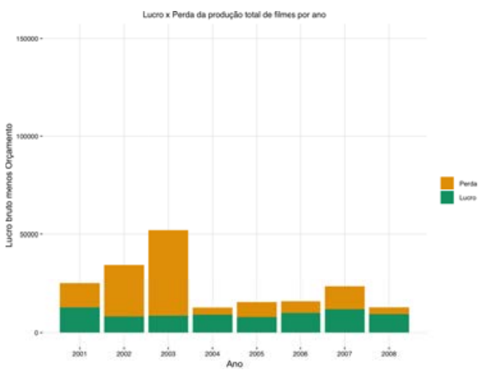


C.20(b): clear

Figure C.20: Pie - Make Comparisons

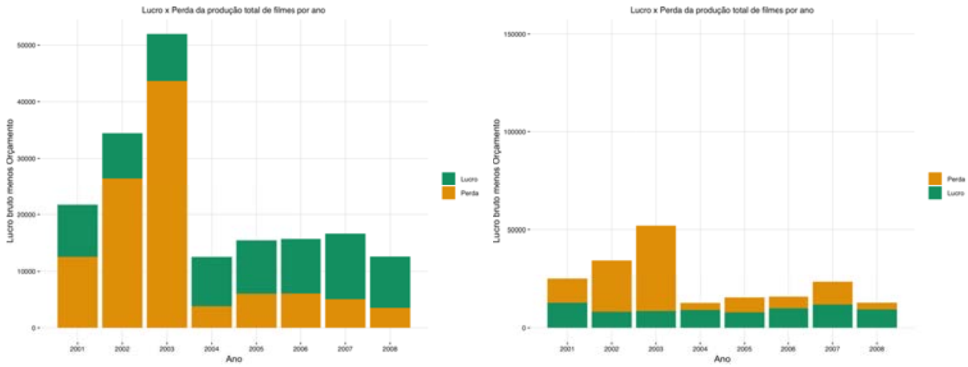


C.21(a): clear



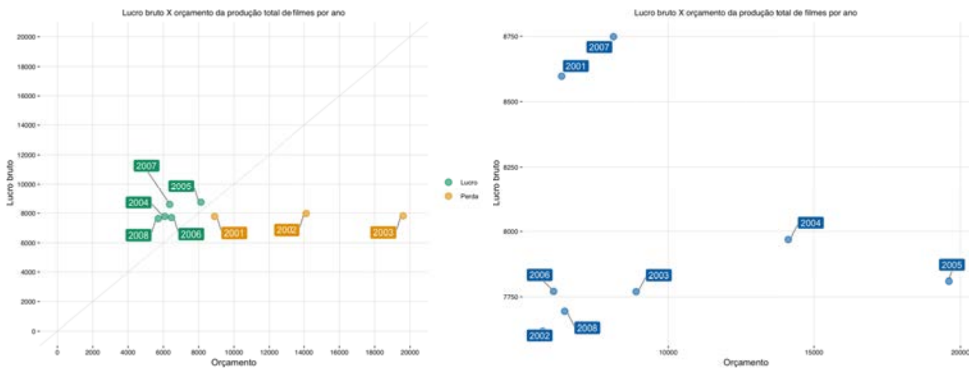
C.21(b): confusing

Figure C.21: Stacked bar - Make Comparisons



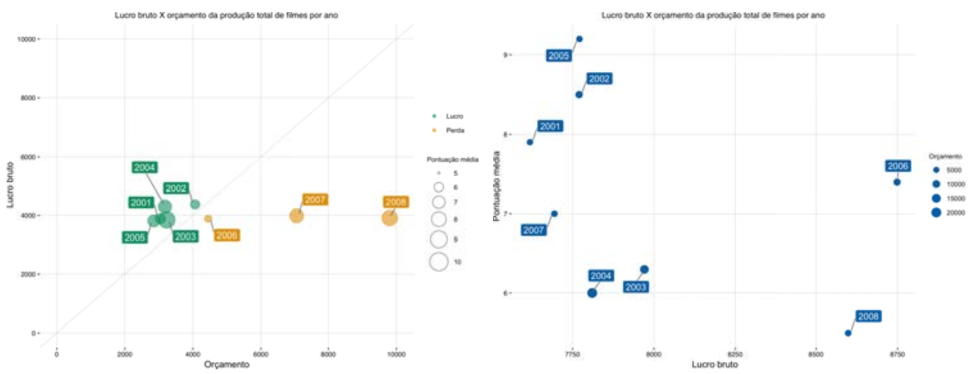
C.22(a): clear C.22(b): confusing

Figure C.22: Stacked Area - Make Comparisons



C.23(a): clear C.23(b): confusing

Figure C.23: Scatterplot - Make Comparisons



C.24(a): clear C.24(b): confusing

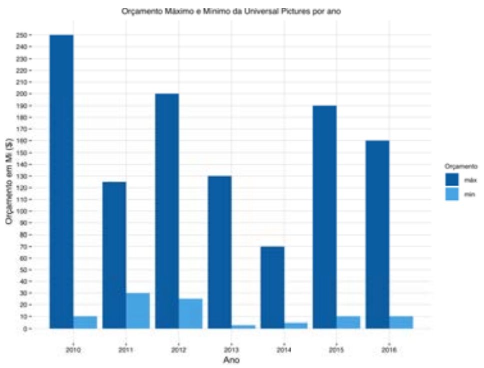
Figure C.24: Bubble Chart - Make Comparisons

Determine Range

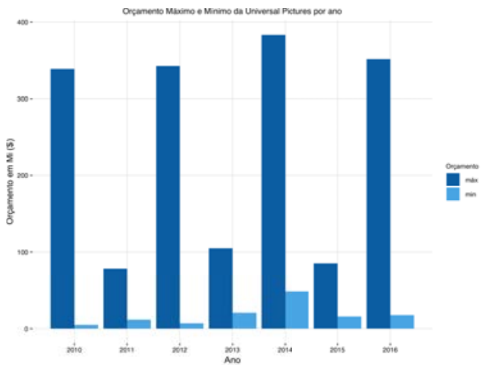
Question: Em qual ano o intervalo do orçamento (MAX menos MIN) foi o maior? / In which year was the budget range (MAX minus MIN) the highest?

Answer options: select one (options: one year per option)

Charts: figs. C.25 to C.29

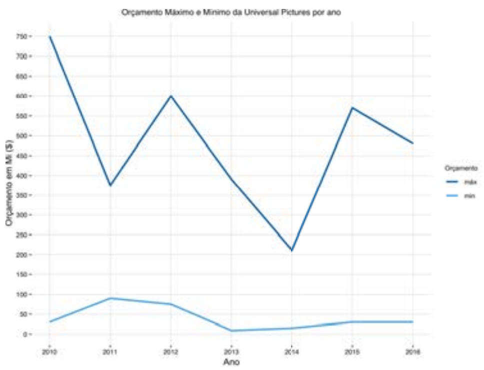


C.25(a): clear

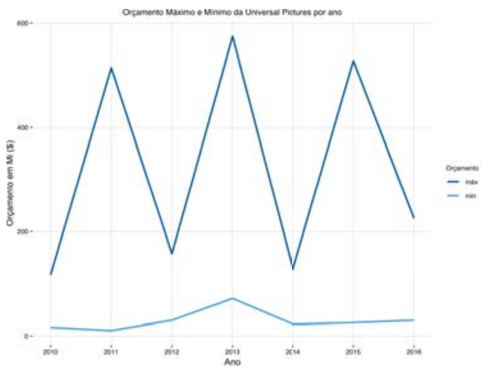


C.25(b): confusing

Figure C.25: Bar - Determine Range

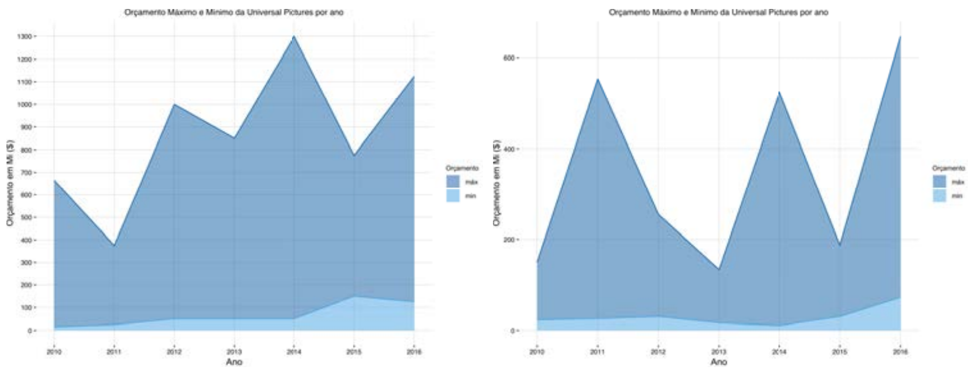


C.26(a): clear



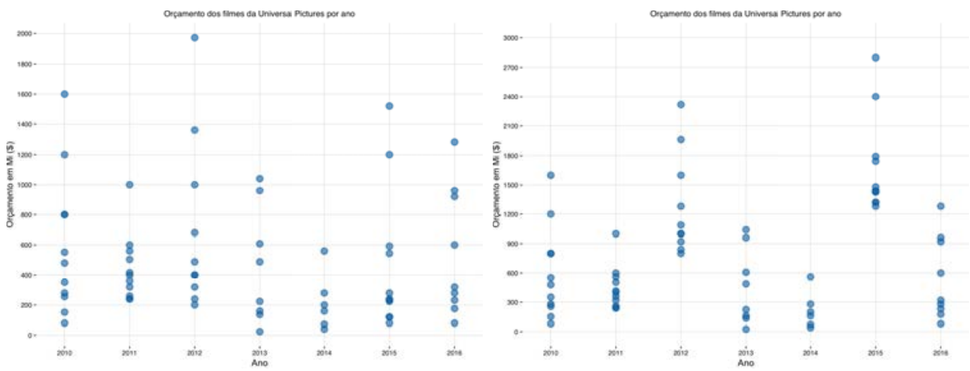
C.26(b): confusing

Figure C.26: Line - Determine Range



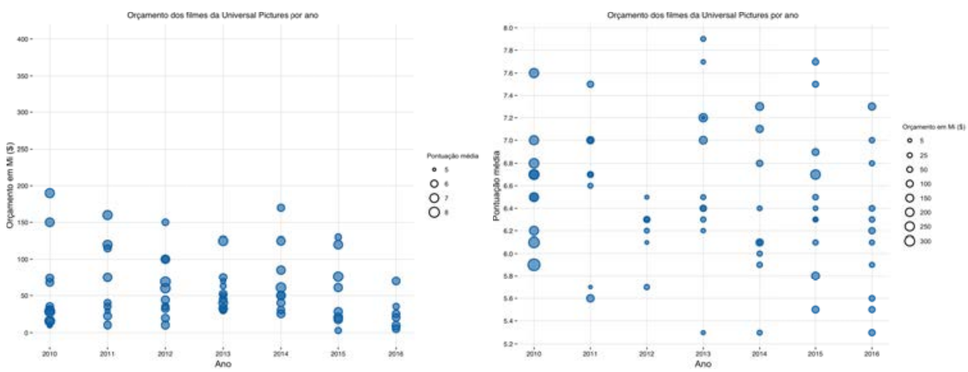
C.27(a): clear C.27(b): confusing

Figure C.27: Stacked Area - Determine Range



C.28(a): clear C.28(b): confusing

Figure C.28: Scatterplot - Determine Range



C.29(a): clear C.29(b): confusing

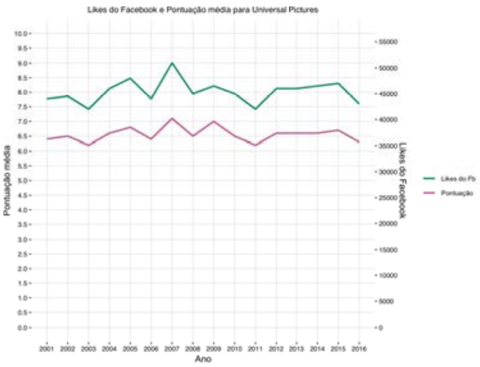
Figure C.29: Bubble chart - Determine Range

Find Correlations

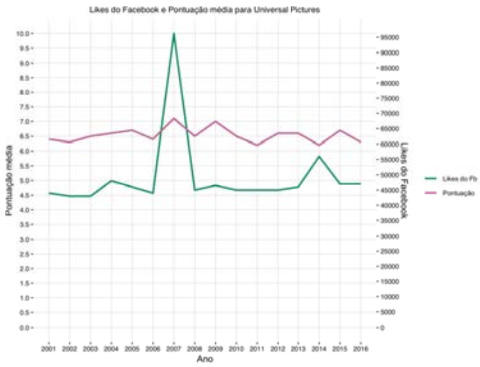
Question: Há uma relação linear entre Likes do Facebook e pontuação IMDb. / In general, the number of Facebook likes increases as the average score increases.

Answer options: select one (TRUE or FALSE)

Charts: figs. C.30 to C.33

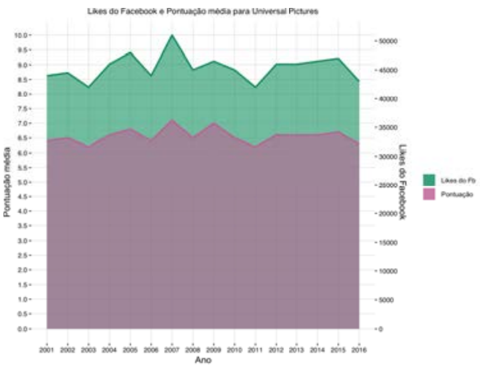


C.30(a): clear

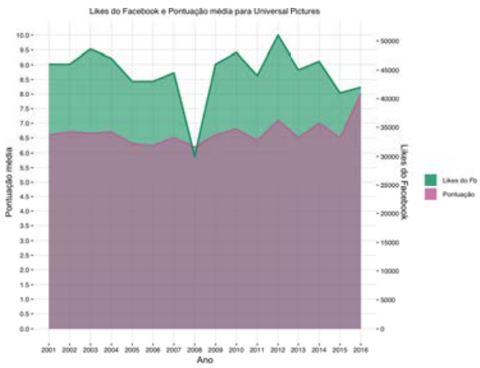


C.30(b): confusing

Figure C.30: Line - Find Correlations

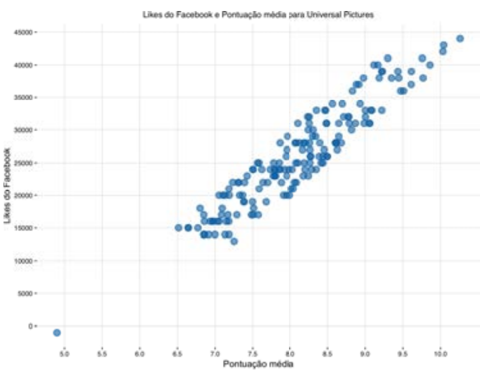


C.31(a): clear

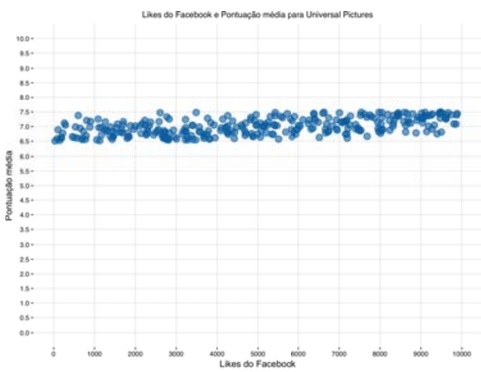


C.31(b): confusing

Figure C.31: Area - Find Correlations

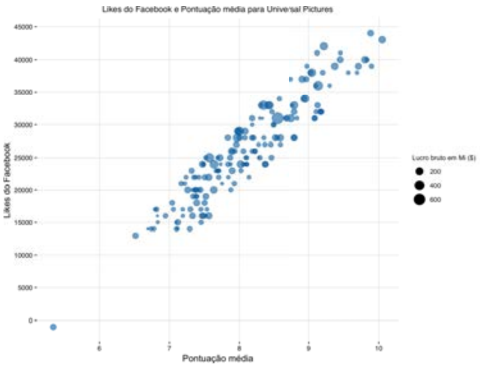


C.32(a): clear

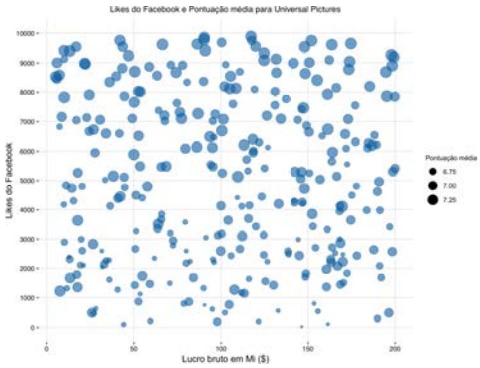


C.32(b): confusing

Figure C.32: Scatterplot - Find Correlations



C.33(a): clear



C.33(b): confusing

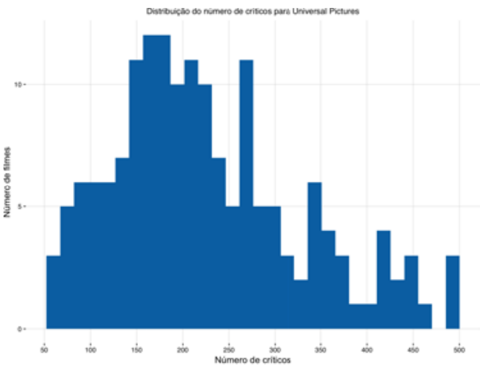
Figure C.33: Bubble Chart - Find Correlations

Characterize Distribution

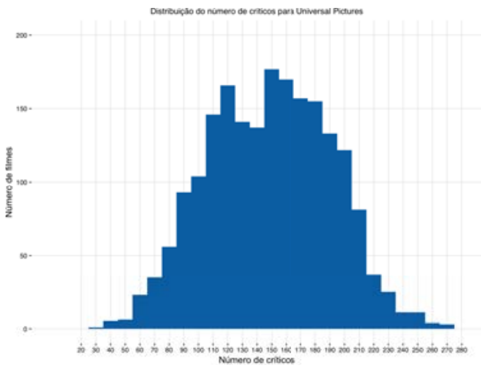
Question: Qual é o tipo de distribuição de número de críticos para a Universal Pictures? / What is the distribution of critics number for Universal Pictures?

Answer options: select one (options: normal, bimodal, uniform, skwed)

Charts: figs. C.34 and C.35

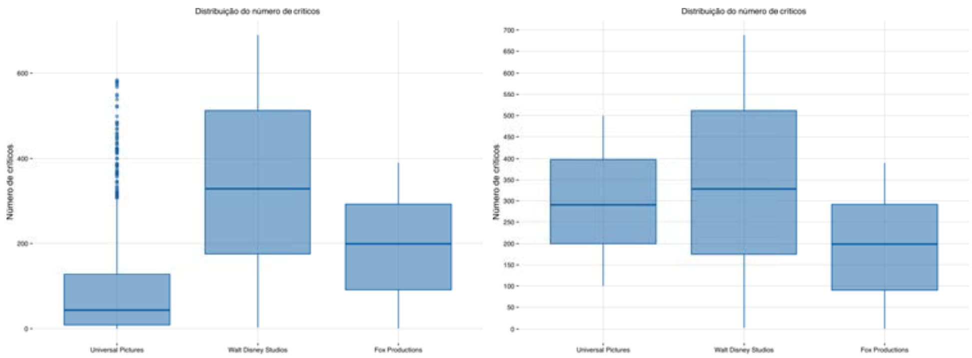


C.34(a): clear



C.34(b): confusing

Figure C.34: Histogram - Characterize Distribution



C.35(a): clear C.35(b): confusing

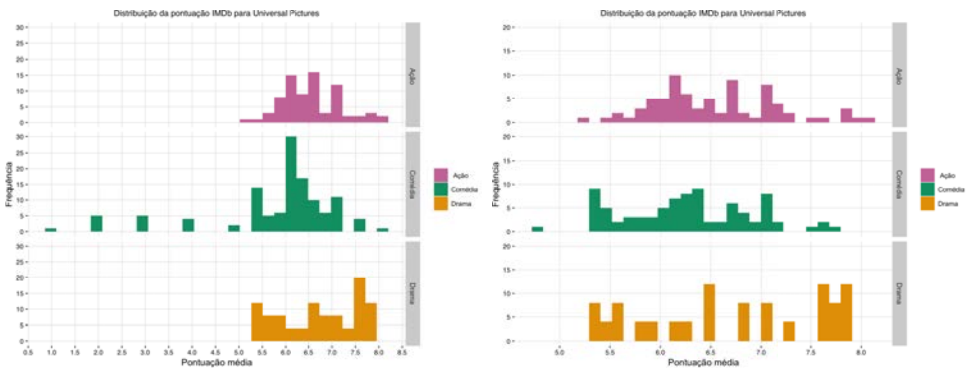
Figure C.35: Boxplot - Characterize Distribution

Find anomalies

Question: Qual gênero tem outlier(s) (pontos extremos)? / Which year (or group of years) stand out?

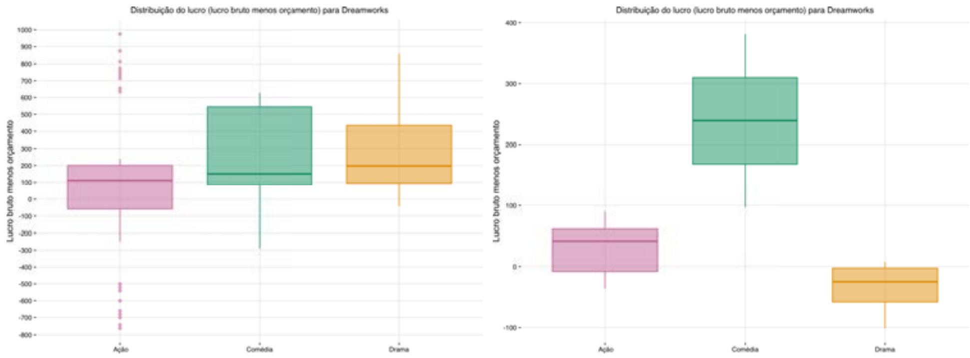
Answer options: select multiples - except for Histogram, select one (options: one year per option)

Charts: figs. C.36 to C.39



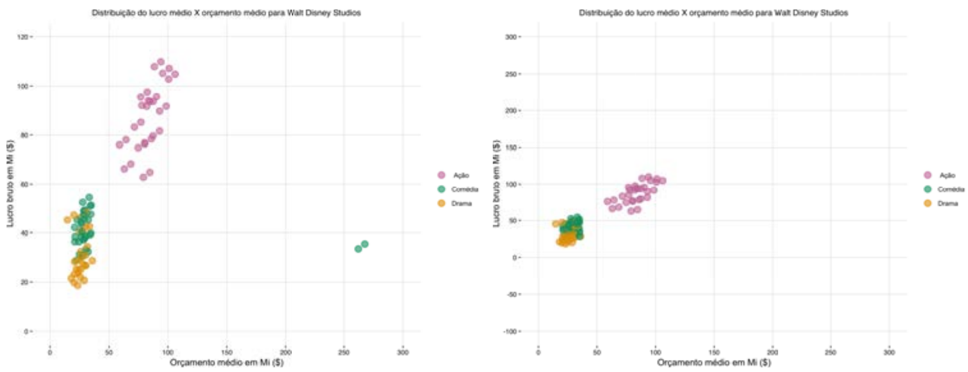
C.36(a): clear C.36(b): confusing

Figure C.36: Histogram - Find anomalies



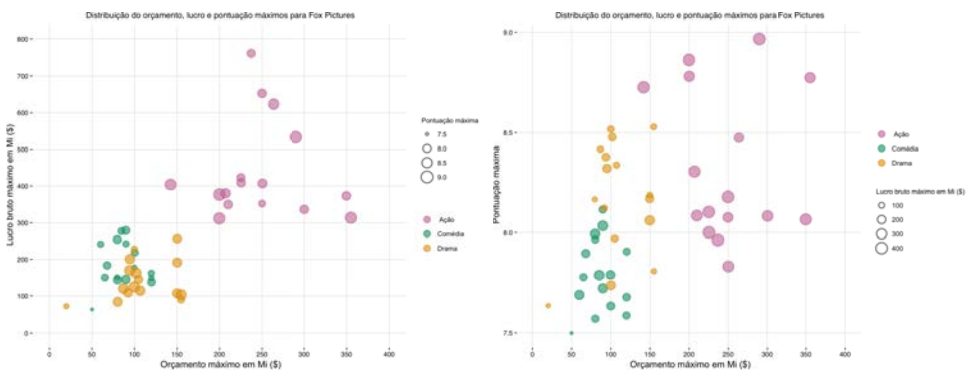
C.37(a): clear C.37(b): confusing

Figure C.37: Boxplot - Find anomalies



C.38(a): clear C.38(b): confusing

Figure C.38: Scatterplot - Find anomalies



C.39(a): clear C.39(b): confusing

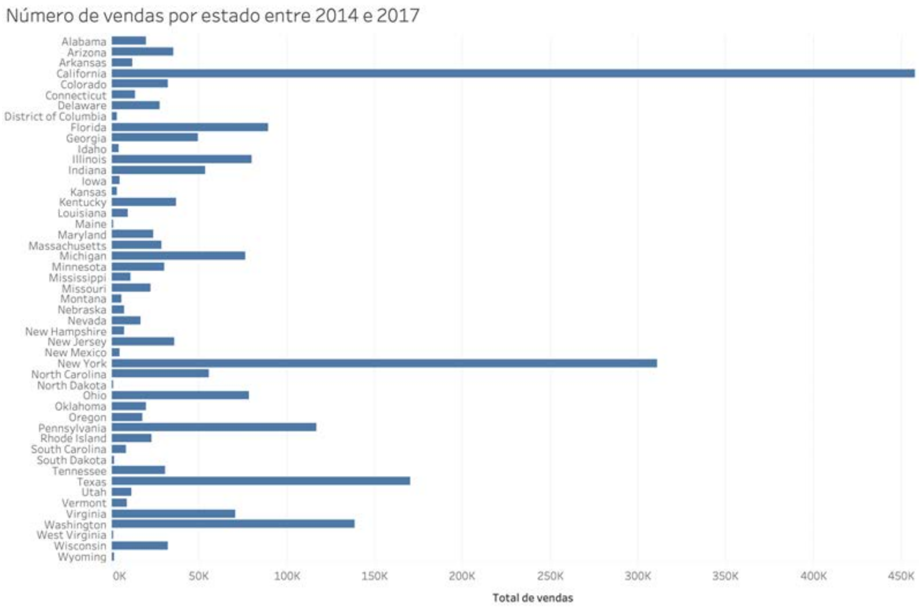
Figure C.39: Bubble Chart - Find anomalies

D
Identifying Visualizations Suitability Survey

This chapter presents the questions included in our study to identify when visualization is suitable for answering a particular analysis question. We split the questionnaire into three parts. The first two are similar, with ten questions each. We randomly presented questions 5-6, as were the sentences in questions 1 and 4. After the set of questions for each part, we presented explanations about some problems and suggestions related to ways to improve or design a better visualization for answering the proposed question. The image from question 1 (Parts 1 and 2) also appears in questions 2-4. Part 3 presents four final questions, two referring to each previous part.

Part 1

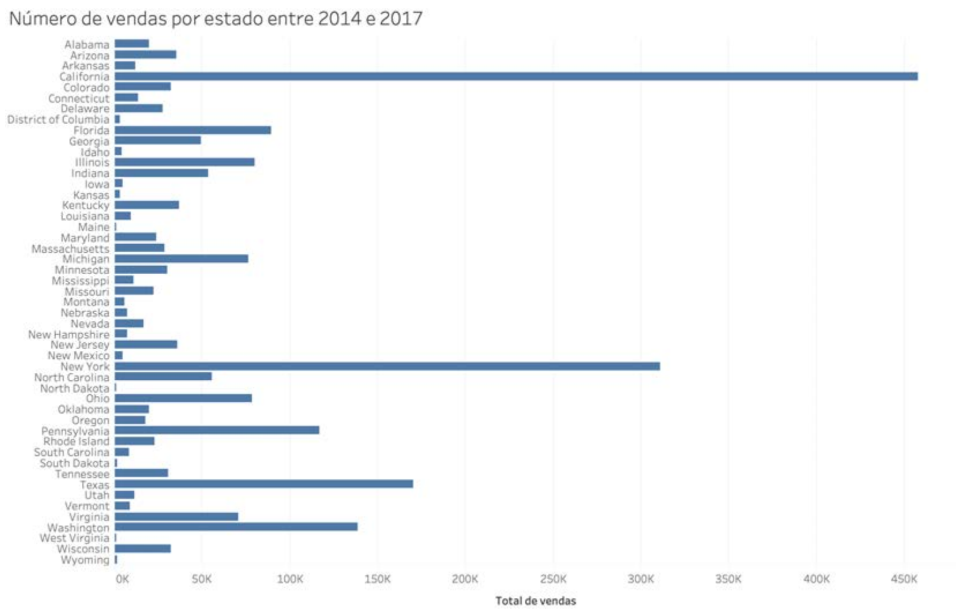
1. Um programa de criação de gráficos sugeriu o seguinte gráfico após você adicionar dois tipos de variáveis: uma nominal no eixo Y e uma quantitativa no eixo X.



Para cada pergunta a seguir, marque o quanto você a considera fácil de responder com o gráfico.

	1 - impossível	2	3	4	5	6	7 - trivial
1. North Dakoka teve mais vendas que Maine?	( )	( )	( )	( )	( )	( )	( )
2. Quais estados estão abaixo da média de vendas?	( )	( )	( )	( )	( )	( )	( )
3. Quais os 10 estados com menos vendas?	( )	( )	( )	( )	( )	( )	( )
4. Qual é o intervalo do número de vendas entre 2014 e 2017?	( )	( )	( )	( )	( )	( )	( )
5. Qual o número de vendas da Califórnia?	( )	( )	( )	( )	( )	( )	( )
6. Quantos estados superaram Illinois em vendas?	( )	( )	( )	( )	( )	( )	( )

2a. Ainda sobre o gráfico sugerido pelo programa de criação de gráficos, responda as questões:



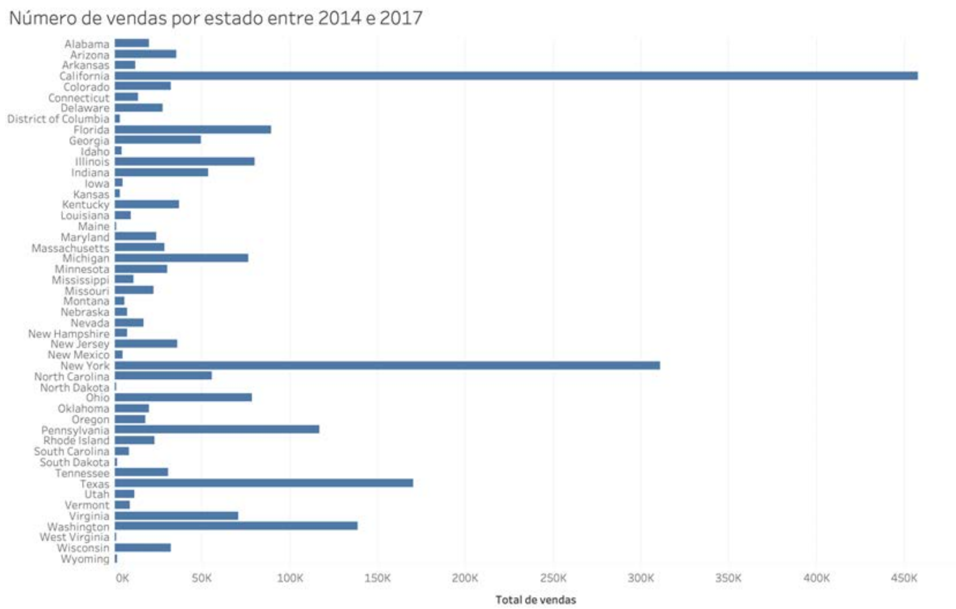
Tendo em vista que você gostaria de investigar graficamente a seguinte pergunta "Quais os top-10 estados com mais vendas?", o quanto você considera este gráfico adequado para respondê-la?

( ) 1 - totalmente inadequado ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - totalmente adequado

2b. Como você considerou que este gráfico **pode ser melhorado**, que alterações você faria para responder melhor a pergunta?

3. Analise novamente a pergunta e o gráfico sugerido.

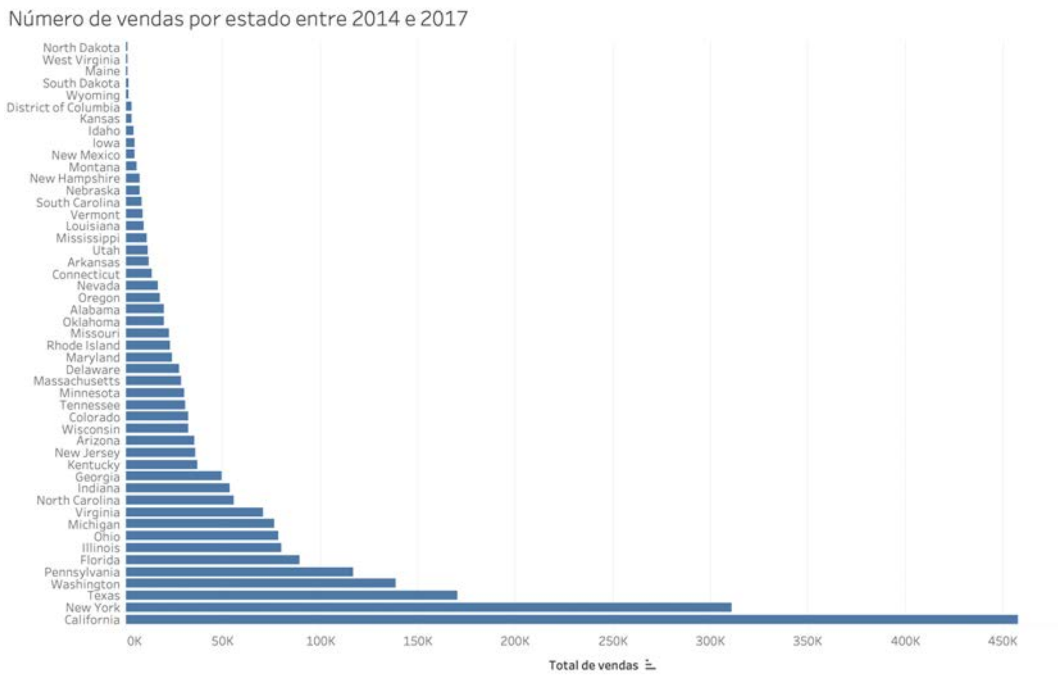
Quais os top-10 estados com mais vendas?



Marque sua concordância para cada uma das recomendações a seguir. Para tornar o gráfico <u>mais adequado</u> para responder a pergunta, eu:

	1 - discordo totalmente	2	3	4	5	6	7 - concordo totalmente
1. Manteria o tipo de gráfico (barras), mas ordenaria as barras pelo número de vendas.	( )	( )	( )	( )	( )	( )	( )
2. Manteria o tipo de gráfico (barras), mas destacaria as barras dos 10 estados com maior número de vendas.	( )	( )	( )	( )	( )	( )	( )
3. Manteria o tipo de gráfico (barras), mas incluiria o valor de vendas em cada barra.	( )	( )	( )	( )	( )	( )	( )
4. Manteria o tipo de gráfico (barras), mas usaria uma cor diferente (azul, amarelo, ...) para cada estado.	( )	( )	( )	( )	( )	( )	( )
5. Mudaria o gráfico para pizza, onde cada fatia seria um estado, ordenadas pelo nome dos estado.	( )	( )	( )	( )	( )	( )	( )
6. Mudaria para uma tabela, ordenada pelo nome dos estados, colorindo as células com um degradê de intensidade conforme o valor do total de vendas em cada um.	( )	( )	( )	( )	( )	( )	( )

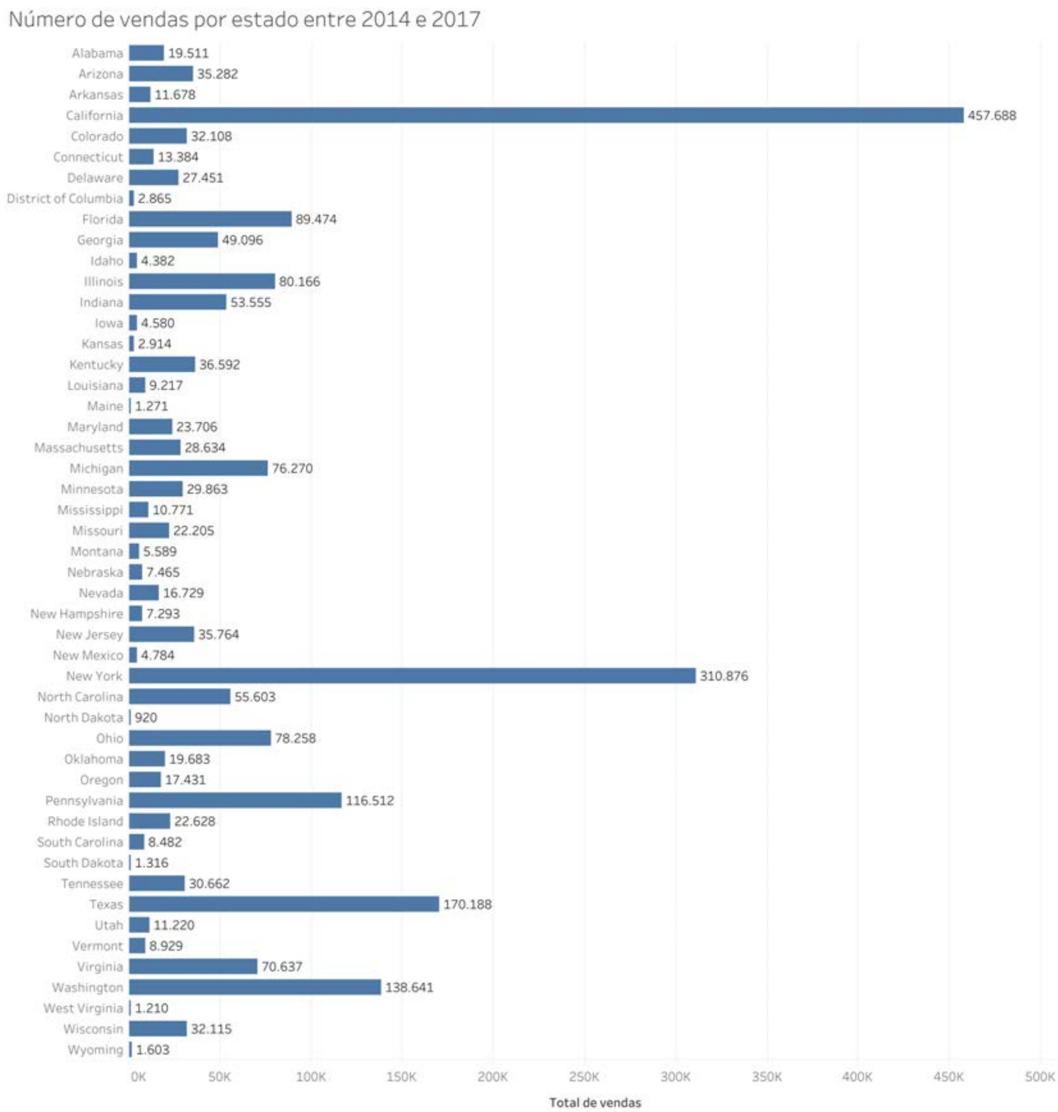
4a. Analise o gráfico a seguir para responder a seguinte pergunta: **Quais os top-10 estados com mais vendas?**



Este gráfico responde a pergunta:

( ) 1 - extremamente mal ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - muito bem ( ) Sem resposta

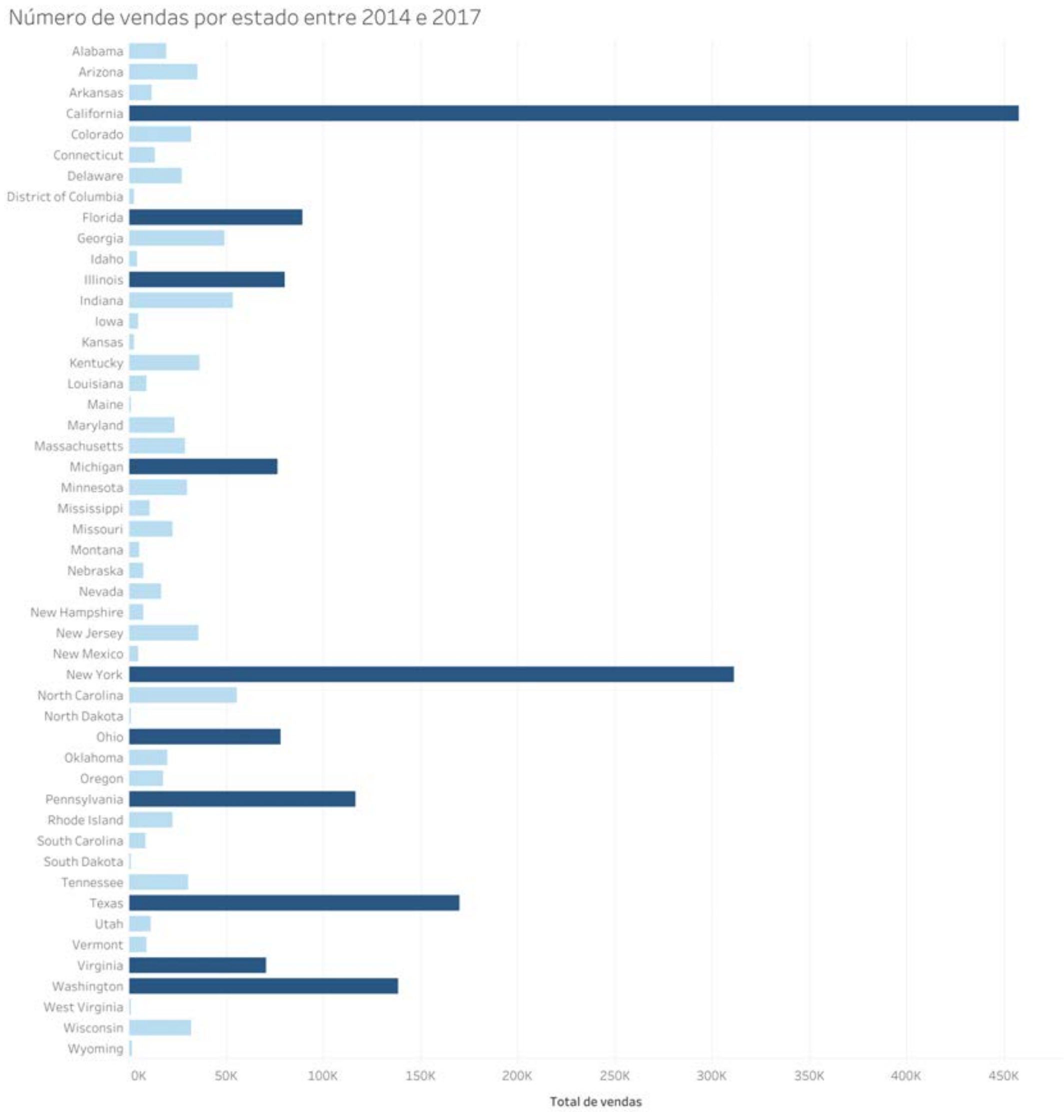
4b. Analise o gráfico a seguir para responder a seguinte pergunta: **Quais os top-10 estados com mais vendas?**



Este gráfico responde a pergunta:

( ) 1 - extremamente mal ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - muito bem ( ) Sem resposta

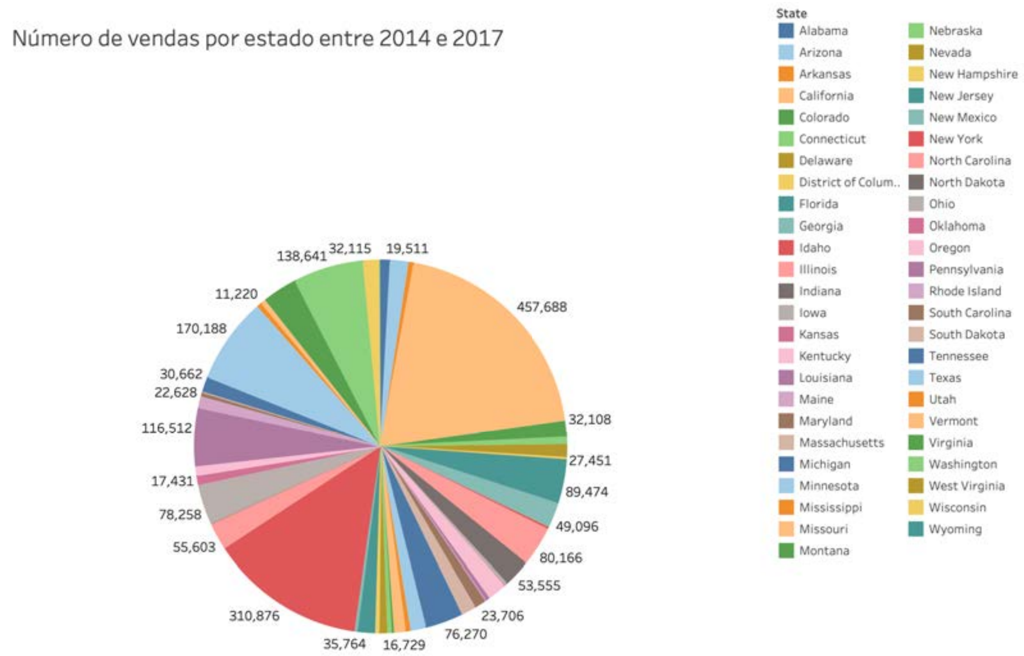
4c. Analise o gráfico a seguir para responder a seguinte pergunta: **Quais os top-10 estados com mais vendas?**



Este gráfico responde a pergunta:

( ) 1 - extremamente mal ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - muito bem ( ) Sem resposta

4d. Analise o gráfico a seguir para responder a seguinte pergunta: **Quais os top-10 estados com mais vendas?**



Este gráfico responde a pergunta:							
( ) 1 - extremamente mal	( ) 2	( ) 3	( ) 4	( ) 5	( ) 6	( ) 7 - muito bem	( ) Sem resposta

4e. Analise o gráfico a seguir para responder a seguinte pergunta: **Quais os top-10 estados com mais vendas?**


Número de vendas por estado entre 2014 e 2017

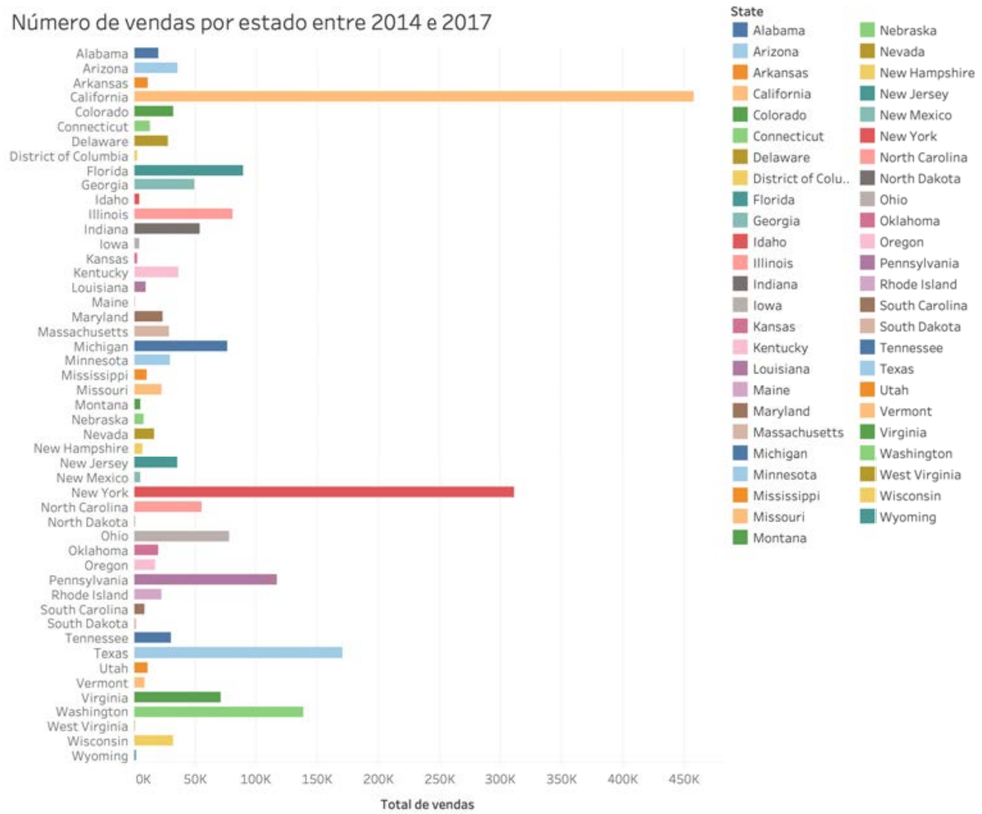
Vendas: 920 – 457.688

Alabama	19.511
Arizona	35.282
Arkansas	11.678
California	457.688
Colorado	32.108
Connecticut	13.384
Delaware	27.451
District of Columbia	2.865
Florida	89.474
Georgia	49.096
Idaho	4.382
Illinois	80.166
Indiana	53.555
Iowa	4.580
Kansas	2.914
Kentucky	36.592
Louisiana	9.217
Maine	1.271
Maryland	23.706
Massachusetts	28.634
Michigan	76.270
Minnesota	29.863
Mississippi	10.771
Missouri	22.205
Montana	5.589
Nebraska	7.465
Nevada	16.729
New Hampshire	7.293
New Jersey	35.764
New Mexico	4.784
New York	310.876
North Carolina	55.603
North Dakota	920
Ohio	78.258
Oklahoma	19.683
Oregon	17.431
Pennsylvania	116.512
Rhode Island	22.628
South Carolina	8.482
South Dakota	1.316
Tennessee	30.662
Texas	170.188
Utah	11.220
Vermont	8.929
Virginia	70.637
Washington	138.641
West Virginia	1.210
Wisconsin	32.115
Wyoming	1.603

Este gráfico responde a pergunta:

( ) 1 - extremamente mal ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - muito bem ( ) Sem resposta

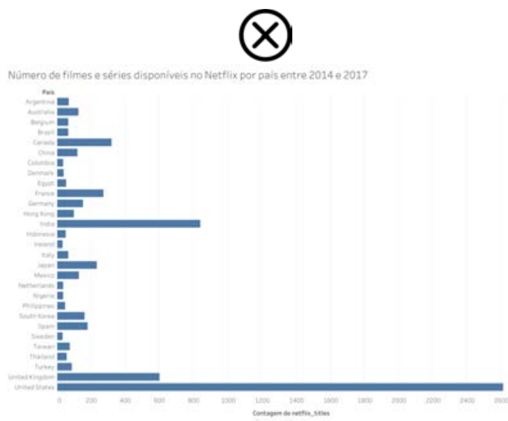
4f. Analise o gráfico a seguir para responder a seguinte pergunta: **Quais os top-10 estados com mais vendas?**

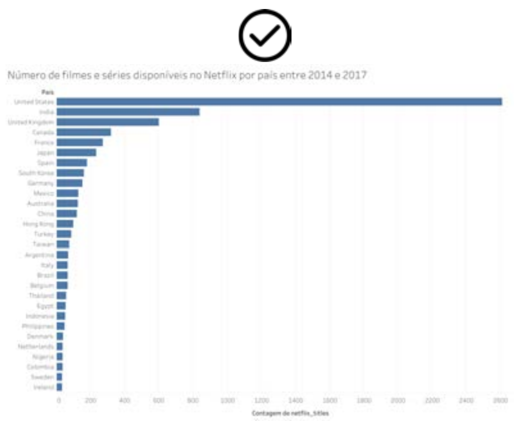


Este gráfico responde a pergunta:
( ) 1 - extremamente mal ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - muito bem ( ) Sem resposta

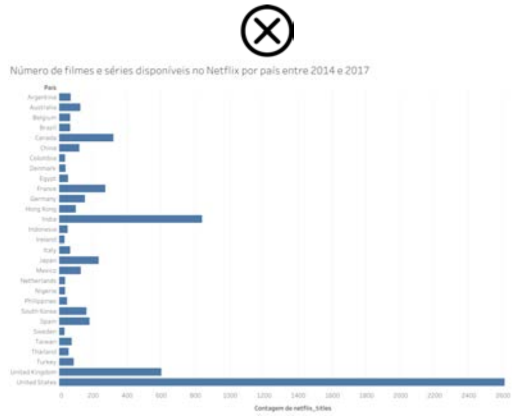
5. Tutorial 1

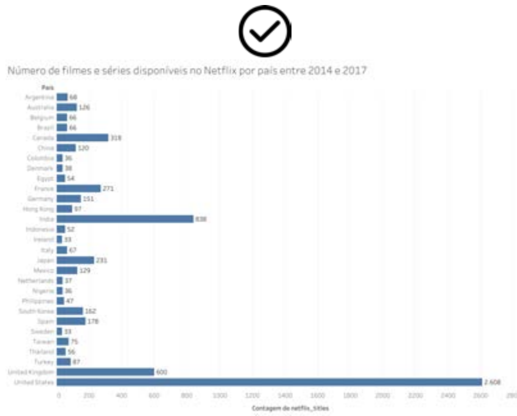
Para perguntas de análise cuja tarefa seja encontrar casos extremos, recomenda-se utilizar um gráfico de barras ordenadas pelo valor do atributo de interesse. Por exemplo, para responder a pergunta **Quais os top-10 estados com mais filmes disponíveis?**, apresentar um gráfico ordenado pelo número de filmes é melhor do que apresentá-lo ordenado pelos nomes dos filmes.



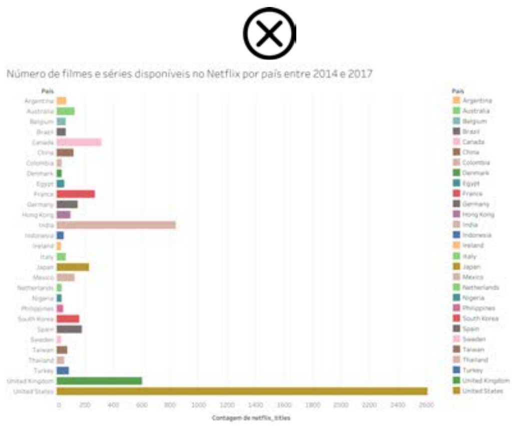


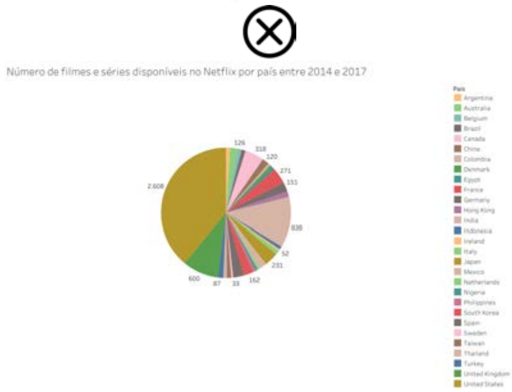
A diferença entre Turquia e México é muito pequena e não é possível identificar qual das duas disponibilizou mais filmes. Nesses casos, é recomendado acrescentar os valores nas barras.



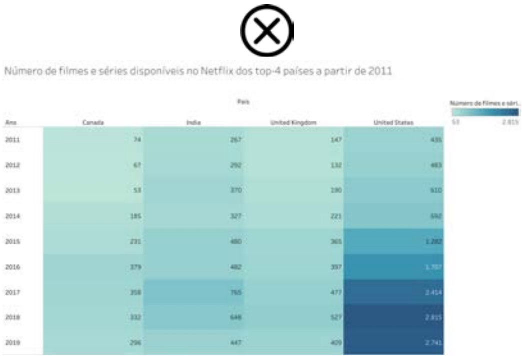


Colorir cada barra com uma cor ou mudar para um gráfico de pizza não são mudanças que ajudam o analista a responder a pergunta.



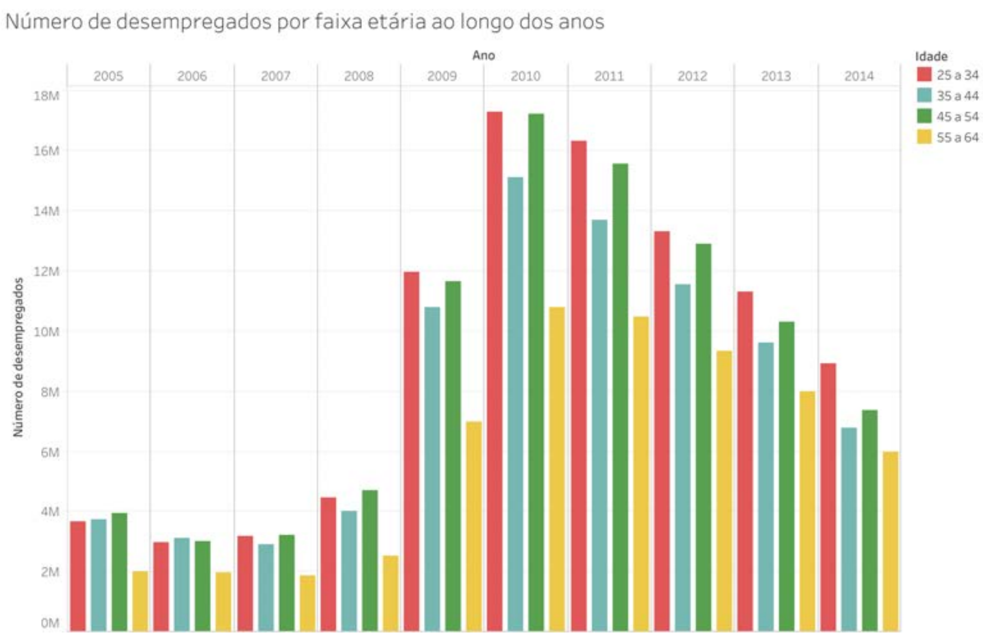


Utilizar uma tabela com degradê, mantendo a ordenação pelo nome do país, também não torna a visualização muito eficiente.



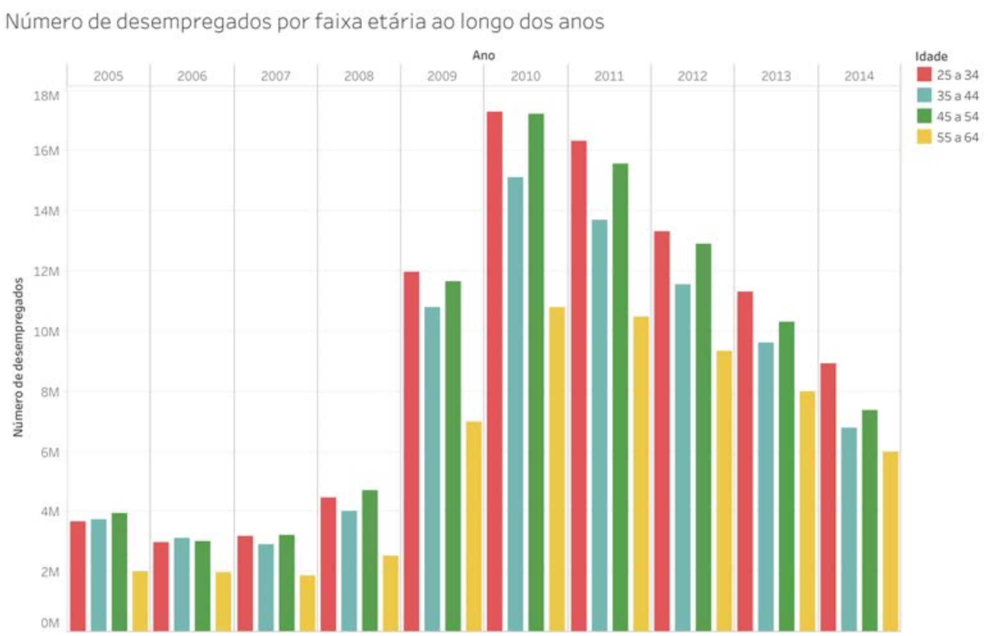
Part 2

1. Um programa de criação de gráficos sugeriu o seguinte gráfico após você adicionar três tipos de variáveis: duas nominais e uma quantitativa.



Para cada pergunta a seguir, marque o quanto você a considera fácil de responder com o gráfico.						
	1 - impossível	2	3	4	5	6	7 - trivial
1.Qual é a razão entre o número de desempregados com 25 a 34 anos em 2008 para todos os desempregos daquele ano?	( )	( )	( )	( )	( )	( )	( )
2. Em 2006, o número de desempregados com 25 a 34 anos foi menor do que o das pessoas com 45 a 54 anos?	( )	( )	( )	( )	( )	( )	( )
3. Qual é a faixa de idade com menor número de desempregados?	( )	( )	( )	( )	( )	( )	( )
4. Qual foi o período mais longo de decréscimo do número de desempregados para todas as faixas de idade?	( )	( )	( )	( )	( )	( )	( )
5. Em 2005 e 2006, o número de desempregados com 55 a 64 anos foi igual?	( )	( )	( )	( )	( )	( )	( )
6. Qual é o número de desempregados entre 55 e 64 anos em 2010?	( )	( )	( )	( )	( )	( )	( )

2a. Ainda sobre o gráfico sugerido pelo programa de criação de gráficos, responda as questões:



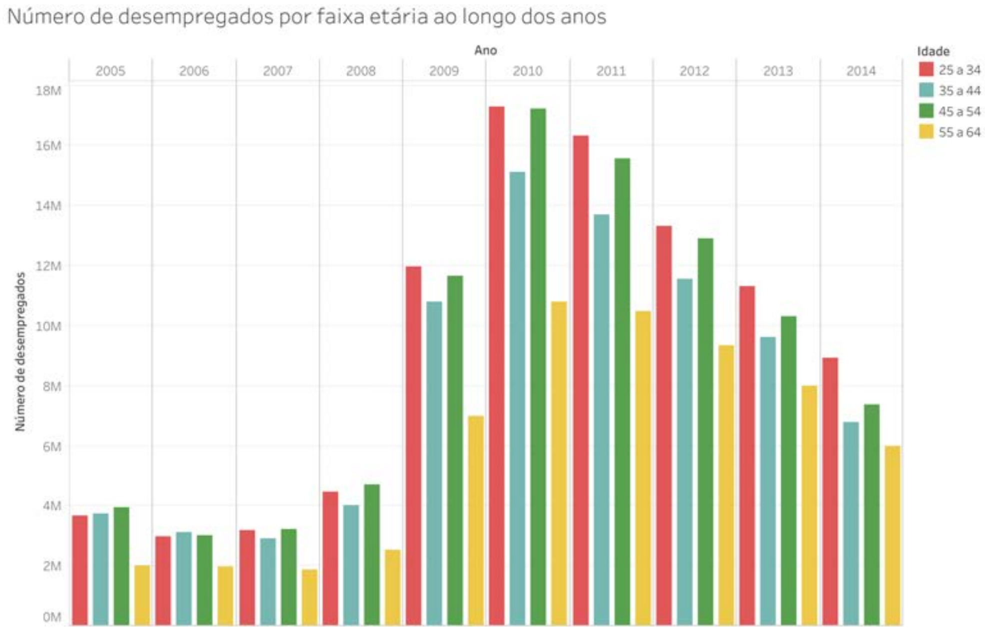
Tendo em vista que você gostaria de investigar graficamente a seguinte pergunta "Qual foi o período mais longo de decréscimo do número de desempregados para todas as faixas de idade?", o quanto você considera este gráfico adequado para respondê-la?

( ) 1 - totalmente inadequado ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - totalmente adequado

2b. Como você considerou que este gráfico **pode ser melhorado**, que alterações você faria para responder melhor a pergunta?

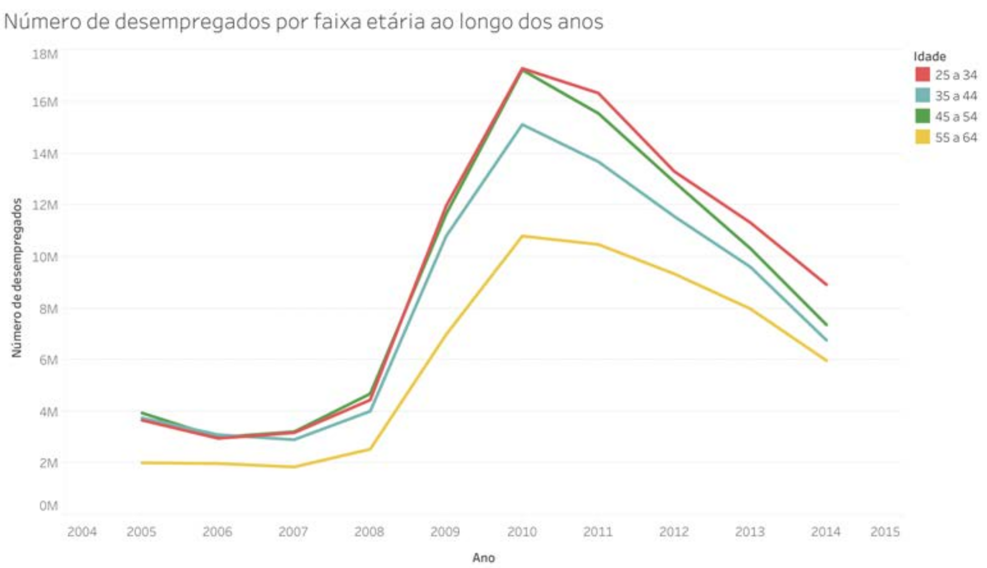
3. Analise novamente a pergunta e o gráfico sugerido.

Qual foi o período mais longo de decréscimo do número de desempregados para todas as faixas de idade?



Marque sua concordância para cada uma das recomendações a seguir. Para tornar o gráfico <u>mais adequado</u> para responder a pergunta, eu:						
	1 - discordo totalmente	2	3	4	5	6	7 - concordo totalmente
1. Mudaria o gráfico para linhas, mas separaria em pequenos múltiplos, uma faixa etária por gráfico.	( )	( )	( )	( )	( )	( )	( )
2. Mudaria o tipo de gráfico para linhas, uma cor para cada faixa de idade.	( )	( )	( )	( )	( )	( )	( )
3. Manteria o tipo de gráfico (barras), agrupando pela mesma faixa de idade, onde cada barra seria um ano.	( )	( )	( )	( )	( )	( )	( )
4. Manteria o tipo de gráfico (barras), empilhando as faixas de idade.	( )	( )	( )	( )	( )	( )	( )
5. Mudaria o gráfico para boxplot, distribuindo o número de desempregados por faixa etária.	( )	( )	( )	( )	( )	( )	( )
6. Mudaria para uma tabela, ordenada pelo ano, colorindo as células com um degradê de intensidade conforme o número de desempregos.	( )	( )	( )	( )	( )	( )	( )

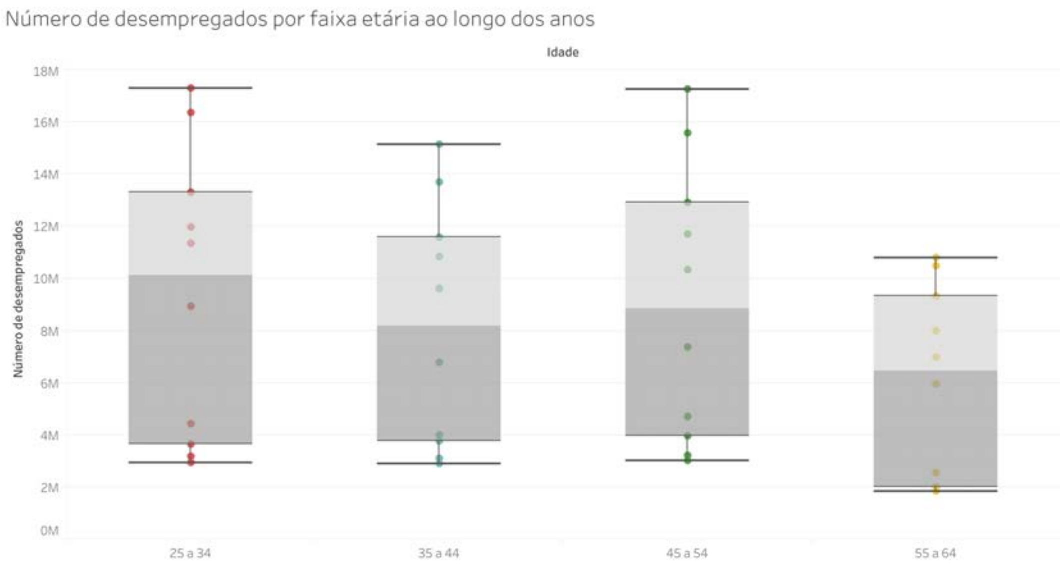
4a. Analise o gráfico a seguir para responder a seguinte pergunta: **Quais os top-10 estados com mais vendas?**



Este gráfico responde a pergunta:

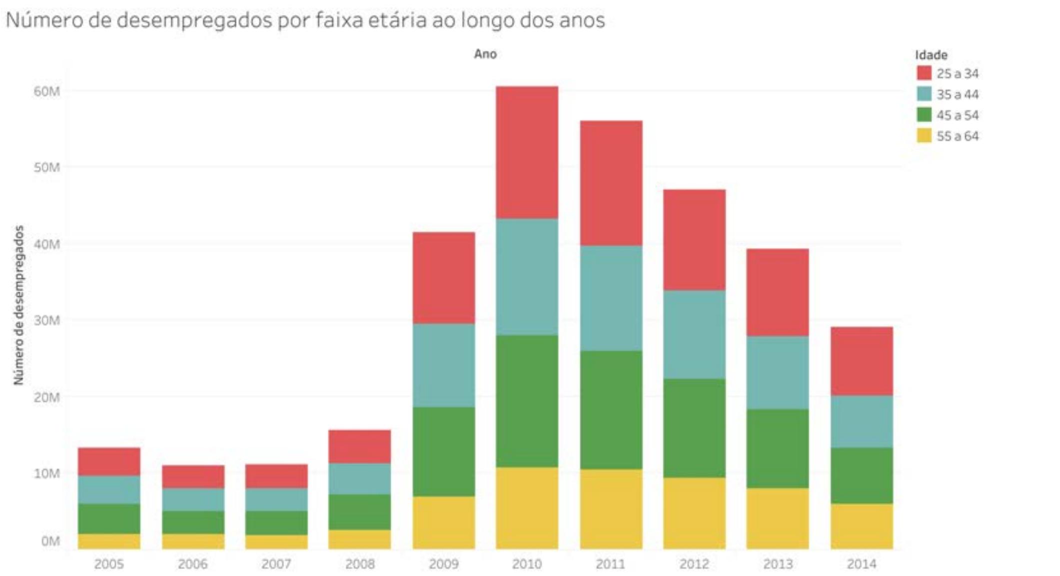
( ) 1 - extremamente mal ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - muito bem ( ) Sem resposta

4b. Analise o gráfico a seguir para responder a seguinte pergunta: **Quais os top-10 estados com mais vendas?**



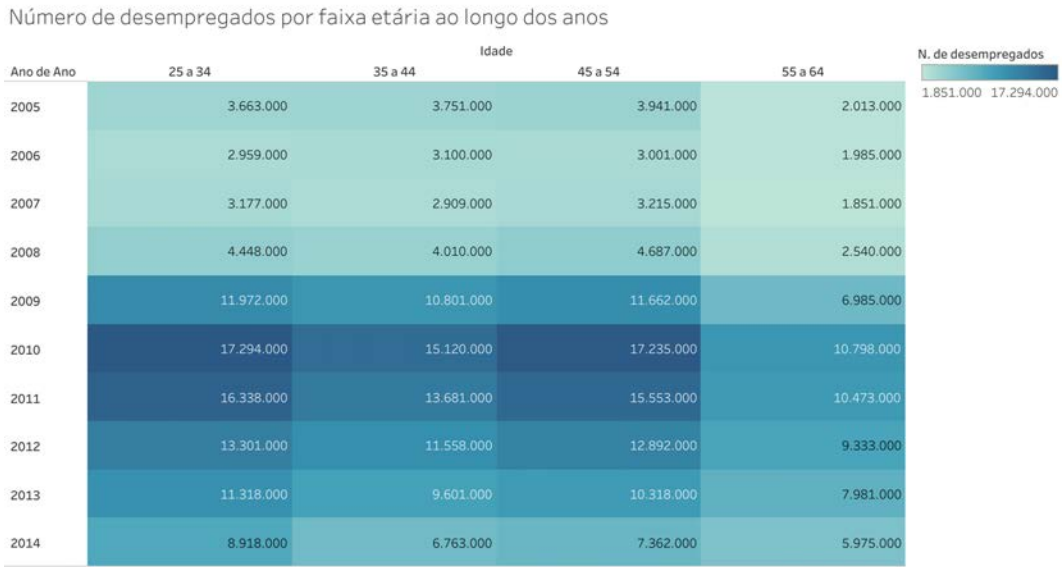
Este gráfico responde a pergunta:
( ) 1 - extremamente mal ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - muito bem ( ) Sem resposta

4c. Analise o gráfico a seguir para responder a seguinte pergunta: **Quais os top-10 estados com mais vendas?**



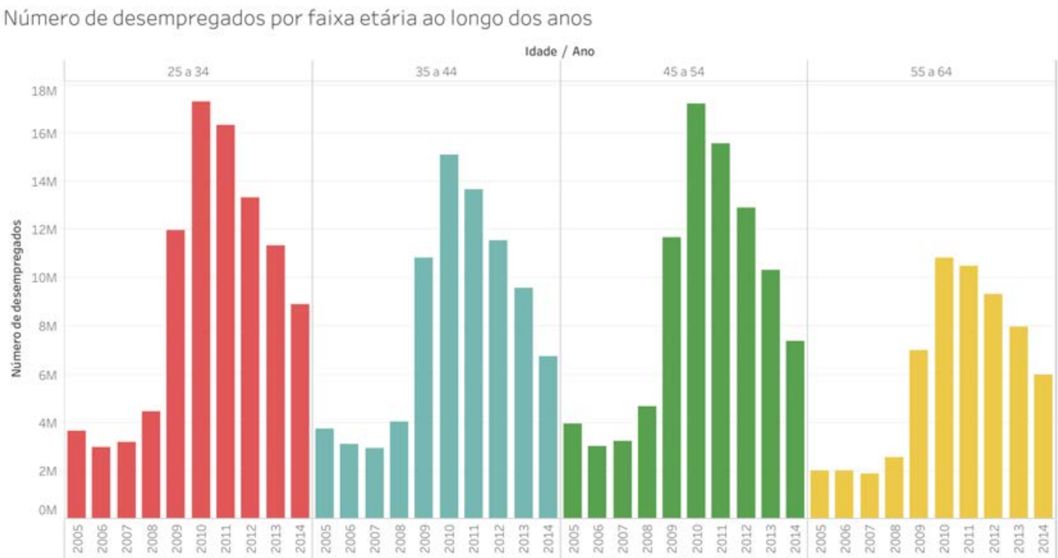
Este gráfico responde a pergunta:
( ) 1 - extremamente mal ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - muito bem ( ) Sem resposta

4d. Analise o gráfico a seguir para responder a seguinte pergunta: **Quais os top-10 estados com mais vendas?**



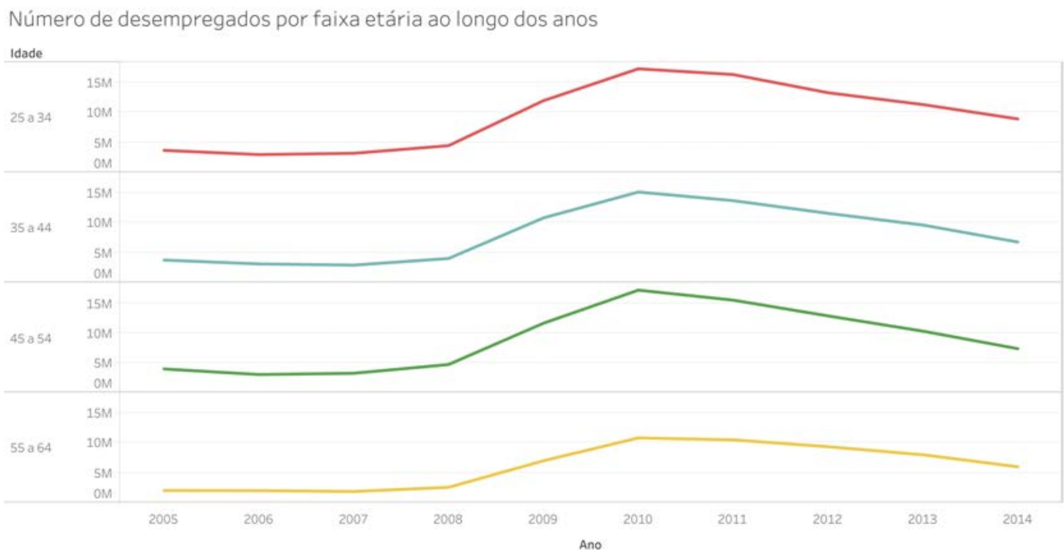
Este gráfico responde a pergunta:
( ) 1 - extremamente mal ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - muito bem ( ) Sem resposta

4e. Analise o gráfico a seguir para responder a seguinte pergunta: **Quais os top-10 estados com mais vendas?**



Este gráfico responde a pergunta:
( ) 1 - extremamente mal ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - muito bem ( ) Sem resposta

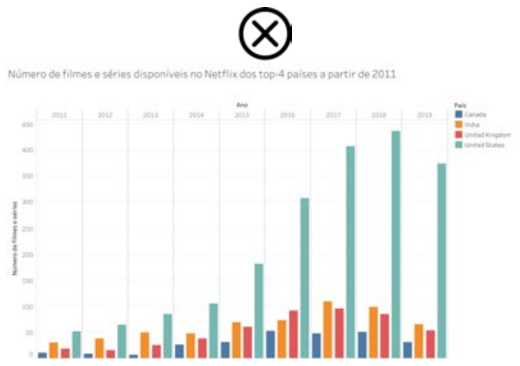
4f. Analise o gráfico a seguir para responder a seguinte pergunta: **Quais os top-10 estados com mais vendas?**

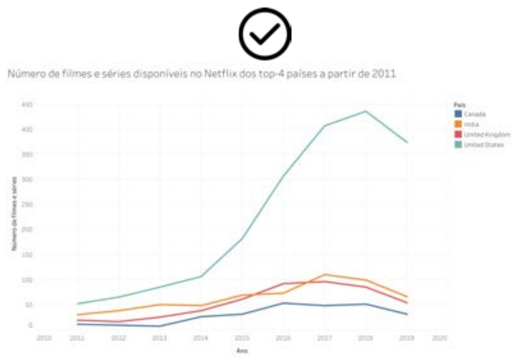


Este gráfico responde a pergunta:
( ) 1 - extremamente mal ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - muito bem ( ) Sem resposta

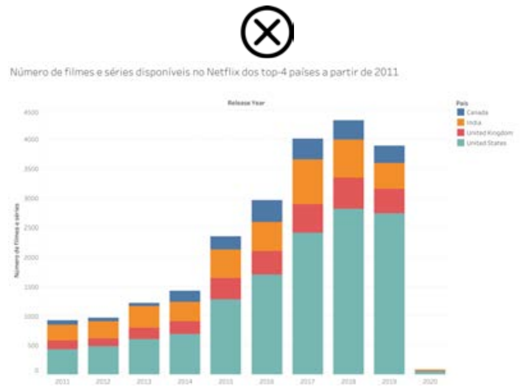
5. Tutorial 2

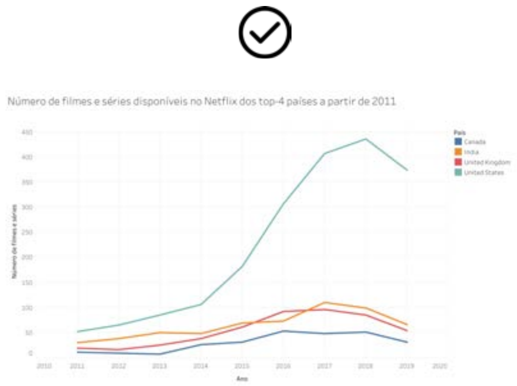
Para perguntas de análise cuja tarefa seja determinar tendências, recomenda-se utilizar um gráfico de linhas. Por exemplo, para responder a pergunta **Qual foi o período mais longo de acréscimo do número de filmes disponíveis para todos os países?**, apresentar um gráfico de barras agrupadas não é a melhor opção. Perceba como é muito mais fácil comparar a inclinação dos segmentos nas linhas.



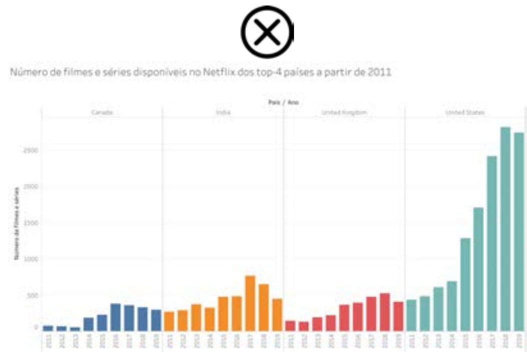


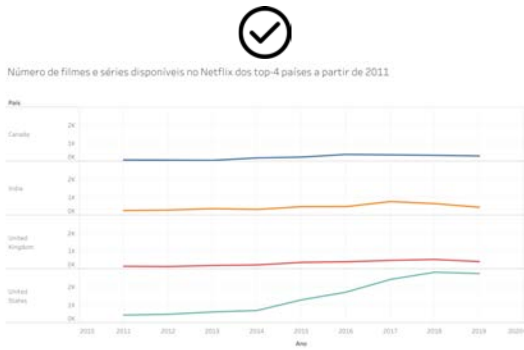
Empilhar as barras melhora um pouco este cenário, permitindo comparar os tamanhos dos segmentos de mesma categoria, ano a ano. No entanto, as faixas de valores muito próximos induzem o erro. Por exemplo, de 2011 para 2012 teve um ligeiro acréscimo na Índia (267 para 292). Sendo assim, as linhas múltiplas continua sendo a melhor opção para este tipo de pergunta.



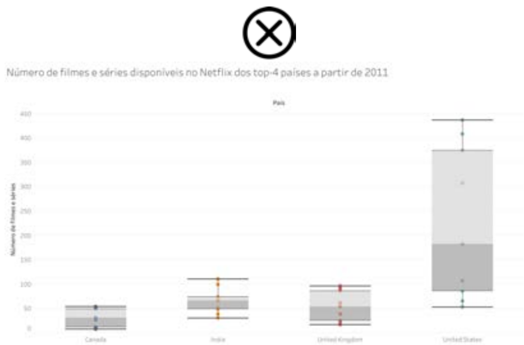


Se optar por usar pequenos múltiplos para separar os gráficos, ainda assim é recomendado usar linhas para este tipo de pergunta.

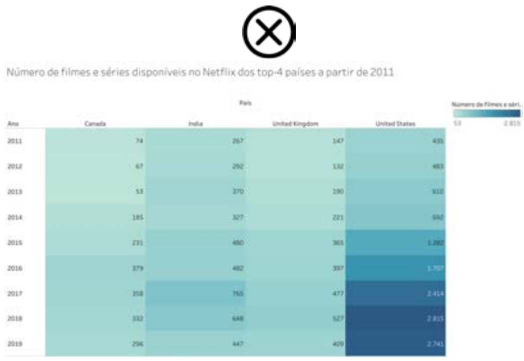




Mudar o gráfico para boxplot, distribuindo o número de filmes e séries disponíveis por País irá fazer com que se perca a dimensão dos anos e não ajudará o analista a responder a pergunta.

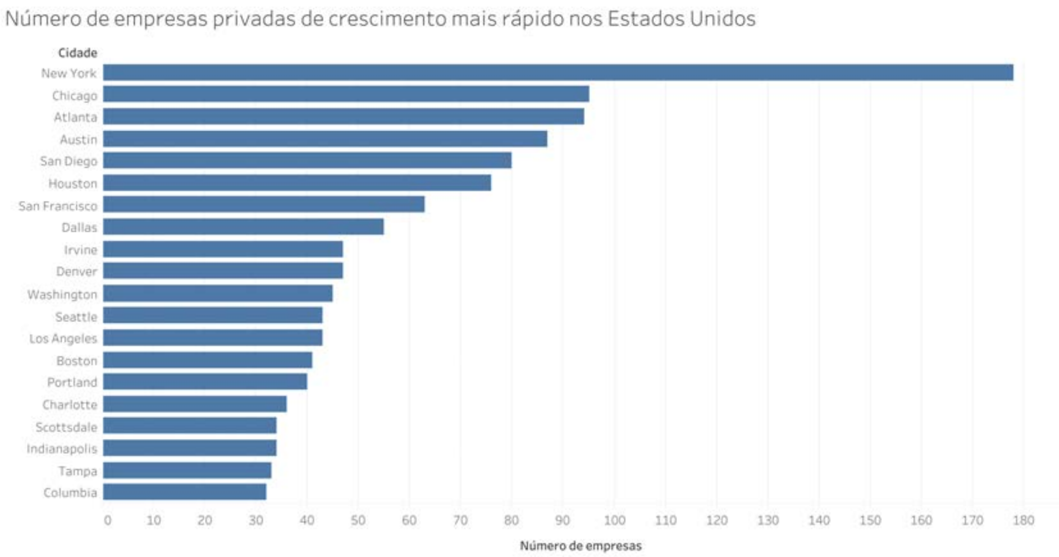


Utilizar uma tabela com degradê, mantendo a ordenação pelo anos, também não torna a visualização muito eficiente.



Part 3

FP1a. Analise o gráfico a seguir.

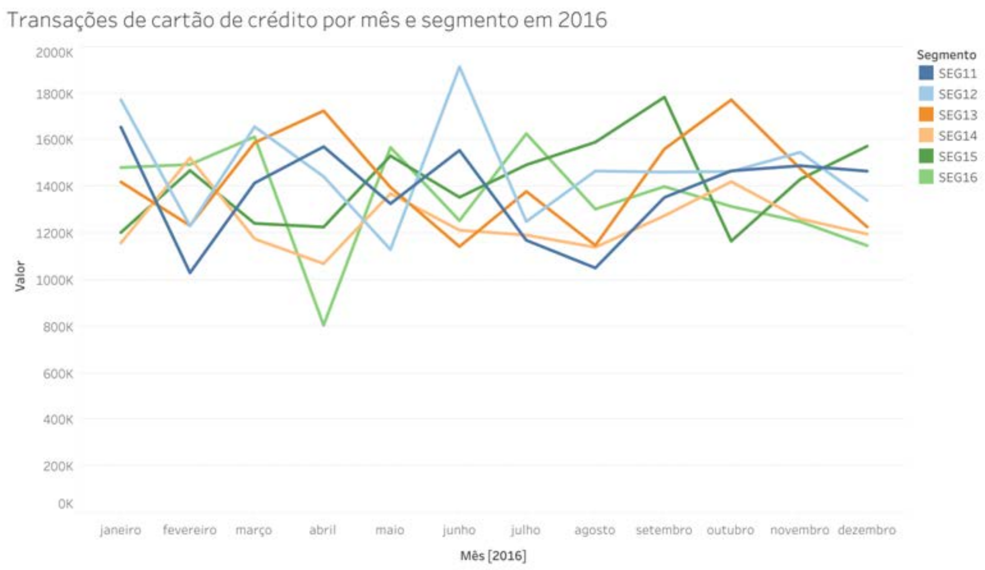


Tendo em vista que você gostaria de investigar graficamente a seguinte pergunta "Qual o número total de empresas de crescimento rápido em Charlotte?", o quanto você considera este gráfico adequado para respondê-la?

( ) 1 - totalmente inadequado ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - totalmente adequado

FP1b. Como você considerou que este gráfico **pode ser melhorado**, que alterações você faria (no tipo de gráfico ou em alguma(s) das suas características) para responder melhor a pergunta?

FP2a. Analise o gráfico a seguir.



Tendo em vista que você gostaria de investigar graficamente a seguinte pergunta "Qual segmento se manteve mais constante (com menos picos) com relação aos valores das transações no ano de 2016?", o quanto você considera este gráfico adequado para respondê-la?

( ) 1 - totalmente inadequado ( ) 2 ( ) 3 ( ) 4 ( ) 5 ( ) 6 ( ) 7 - totalmente adequado

FP2b. Como você considerou que este gráfico **pode ser melhorado**, que alterações você faria (no tipo de gráfico ou em alguma(s) das suas características) para responder melhor a pergunta?

E
Visualization Literacy Test Survey

In this chapter, we present the questions used in the literacy test. In random order, we displayed to participants the types of charts (set of 3 questions), the types of questions within the sets, and the answer options for each question. The charts were interactive and provided some information we cannot catch in the static figures presented here; for example, when hovering a bar, a tooltip reported the exact values of all corresponding variables at that point.

Clustered bar chart

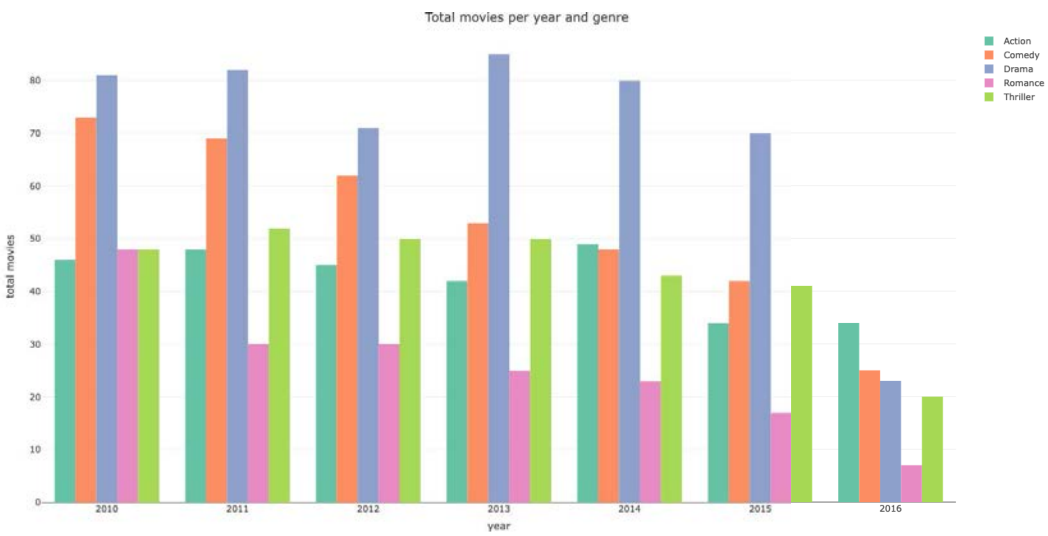


Figure E.1: Clustered bar chart

[Conceptual] This type of chart best allows for:

A ( ) comparing the number of movies across genres in each year

B ( ) analyzing the difference of years in each genre

C ( ) explaining why there are peaks in a specific genre

D ( ) defining which types of genre are most frequent over the years

E ( ) calculating how many genres are analyzed each year

F ( ) I don't know

[Suitable] Approximately how many Drama movies were produced in 2010?

A ( ) 81

B ( ) 107

C ( ) 48

D ( ) 94

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] What is the relationship between Drama and Action movies?

A ( ) This type of chart does not allow or help to answer the question

B ( ) Correlation

C ( ) It depends on the year of production

D ( ) Inversely proportional

E ( ) Directly proportional

F ( ) I don't know

Simple bar chart - ordered by name

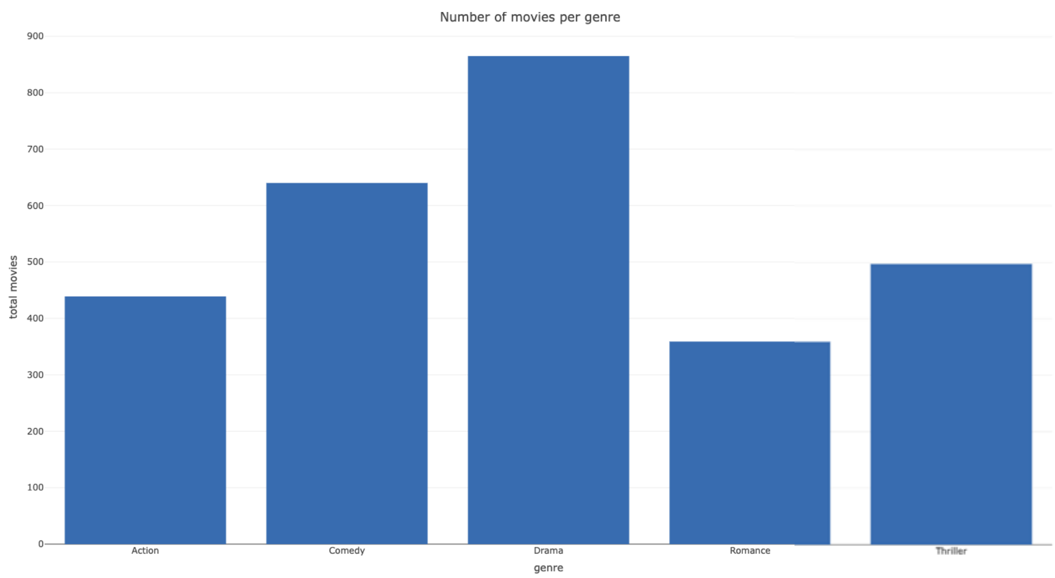


Figure E.2: Simple bar chart - ordered by name

[Conceptual] This type of chart best allows for:

A ( ) making numerical comparisons across genres

B ( ) analyzing trends between genres

C ( ) finding a relation of increase of objects with respect to genres

D ( ) identifying the genre where the number of movies starts to increase

E ( ) finding the total number of movies overall

F ( ) I don't know

[Suitable] What is the approximate number of Romance movies?

A ( ) 359

B ( ) 497

C ( ) 865

D ( ) 310

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] What is the growth ratio of the number of movies in relation to the years?

A ( ) This type of chart does not allow or help to answer the question

B ( ) Partial

C ( ) Increasing

D ( ) With peaks

E ( ) With valleys

F ( ) I don't know

Simple bar chart - ordered by frequency

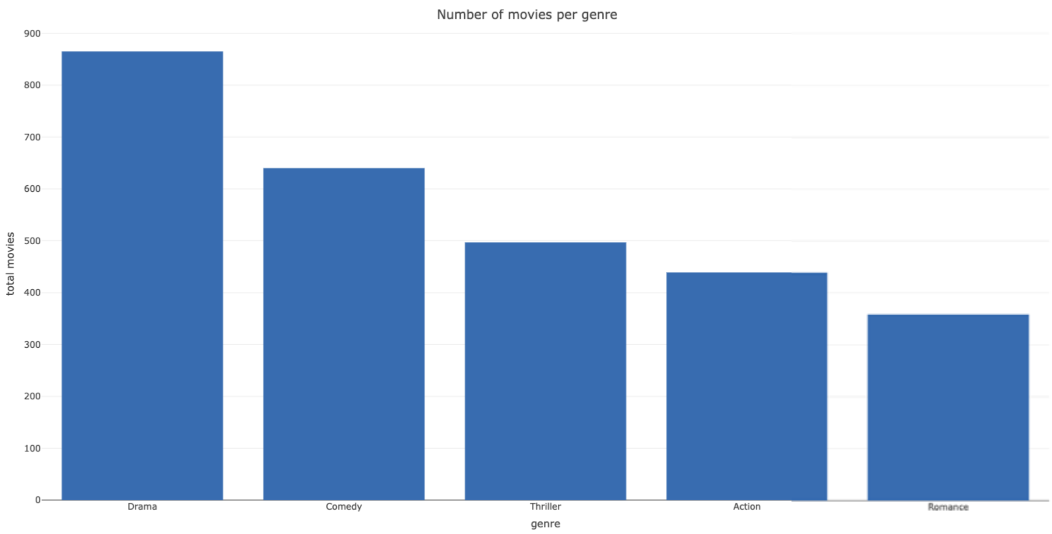


Figure E.3: Simple bar chart - ordered by frequency

[Conceptual] This type of chart best allows for:

A ( ) identifying the genres with the highest and lowest number of movies

B ( ) analyzing trends between genres

C ( ) describing how many categories there are in each genre

D ( ) revealing what is the evolution of the number of movies along the genres

E ( ) explaining why the number of movies is so disparate among most genres

F ( ) I don't know

[Suitable] Which genre has the largest number of movies?

A ( ) Drama

B ( ) Romance

C ( ) Comedy

D ( ) Thriller

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] Looking at the number of movies, what is the trend in terms of genres?

A ( ) Exponential

B ( ) Linear

C ( ) Logarithmic

D ( ) Decreasing

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

Stacked bar chart

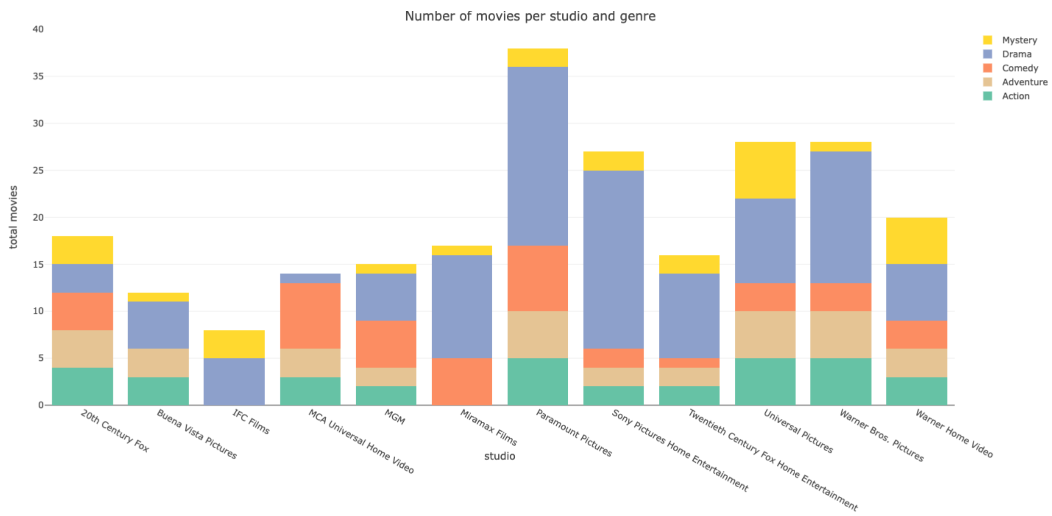


Figure E.4: Stacked bar chart

[Conceptual] This type of chart best allows for:

A ( ) comparing the relative number of movies per genre per studio, emphasizing the total number of movies per studio

B ( ) calculating the number of movies per genre

C ( ) identifying the distribution of movies in each genre

D ( ) defining which genre has the greatest influence in each studio

E ( ) revealing which genre has lower values of movies for more studios

F ( ) I don't know

[Suitable] What is the approximate number of Adventure movies published by Universal Pictures?

A ( ) 5

B ( ) 20

C ( ) 10

D ( ) 30

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] Which studio produces movies of most genres?

A ( ) This type of chart does not allow or help to answer the question

B ( ) Paramount Pictures

C ( ) All except IFC Films

D ( ) Paramount, Sony, Universal and Warner

E ( ) None

F ( ) I don't know

Stacked bar chart - 100%

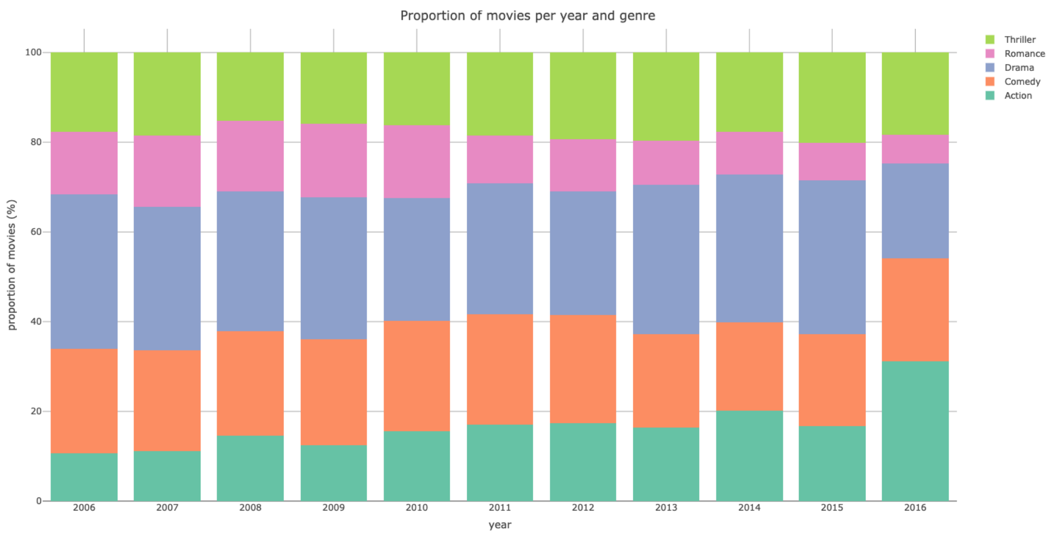


Figure E.5: Stacked bar chart - 100%

[Conceptual] This type of chart best allows for:

A ( ) seeing the relative differences between each group and over the years

B ( ) defining which genre has fewer movies in more years

C ( ) identifying which year has the most genres

D ( ) identifying trends between years

E ( ) extracting the exact value of each segment

F ( ) I don't know

[Suitable] What is the proportion of Thriller movies in 2013?

A ( ) 20

B ( ) 80

C ( ) 100

D ( ) 15

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] The overall proportion of Thriller movies produced was higher than Action movies.

A ( ) This type of chart does not allow or help to answer the question

B ( ) False, it is lower

C ( ) True, if we remove 2016

D ( ) Not sure, it cannot be said because the parts sum up to 100%

E ( ) True, if we remove 2007

F ( ) I don't know

Boxplot

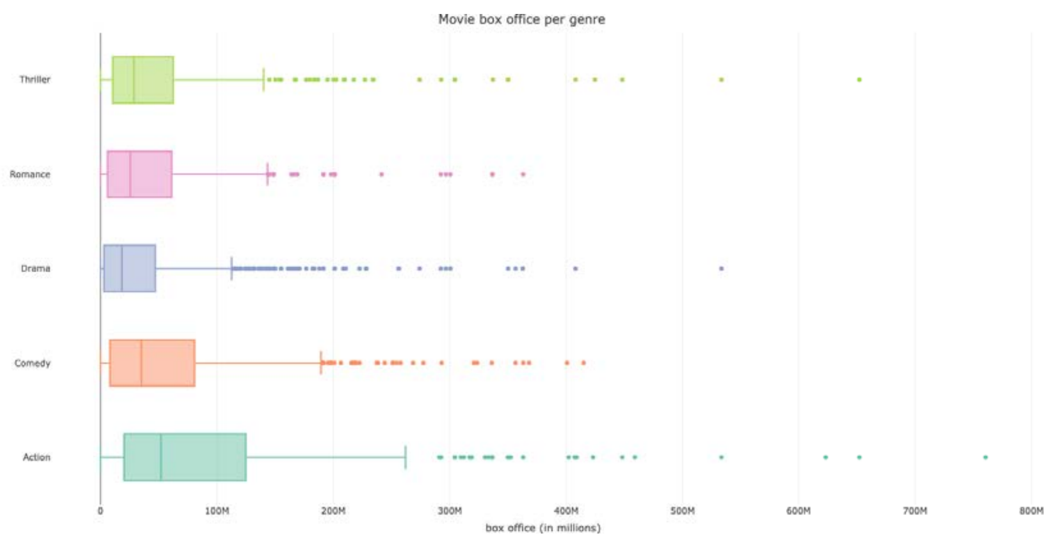


Figure E.6: Boxplot

[Conceptual] This type of chart best allows for:

A ( ) identifying the quartiles and outliers in the box office distribution of each genre

B ( ) identifying which genre has the most outliers

C ( ) analyzing which genre has box office values with the highest trend

D ( ) calculating the total box office of each genre

E ( ) identifying the distribution of box office values of all movies

F ( ) I don't know

[Suitable] Which genre has the largest box office variability?

A ( ) Action

B ( ) Romance

C ( ) Drama

D ( ) Comedy

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] Which genre has the highest number of outliers?

A ( ) This type of chart does not allow or help to answer the question

B ( ) Action

C ( ) Thriller

D ( ) Drama

E ( ) Action and Thriller

F ( ) I don't know

Bubble chart

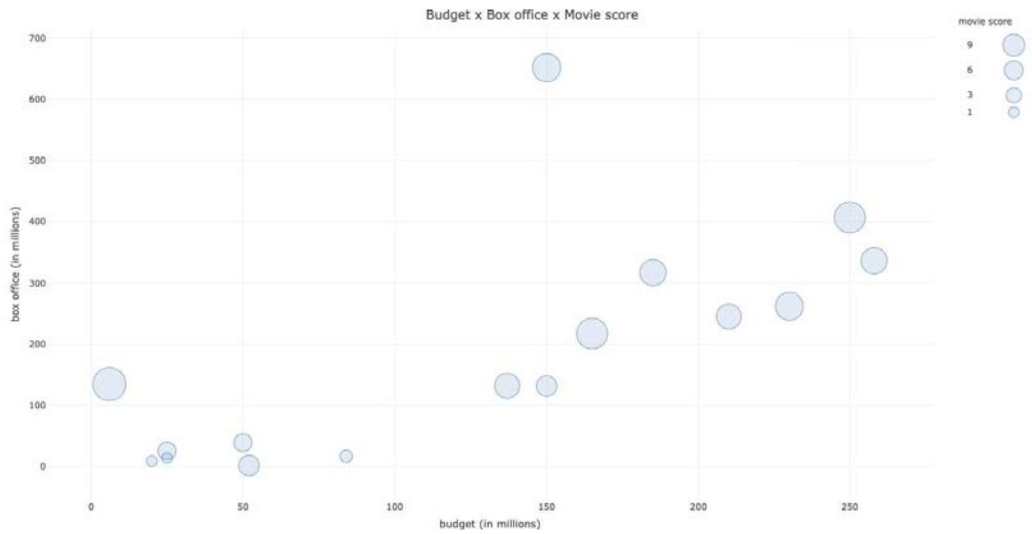


Figure E.7: Bubble chart

[Conceptual] This type of chart best allows for:

A ( ) comparing and showing the relationships between variables, through the use of positioning and size proportions

B ( ) revealing where is the highest concentration of points with respect to the movie scores

C ( ) identifying which range of values of one variable is related to the largest amount of another variable

D ( ) identifying the most frequent values of each variable

E ( ) defining if it is possible to identify clusters in relation to the increase of movie scores

F ( ) I don't know

[Suitable] What happens as the box office increases?

A ( ) In general, the score increases

B ( ) In general, the score decreases

C ( ) The score remains the same

D ( ) The budget decreases

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] What is the relationship between the three variables?

A ( ) This type of chart does not allow or help to answer the question

B ( ) They seem to have strong pairwise correlations, but it is necessary to confirm this using Cohen's d.

C ( ) They seem to have weak pairwise correlations, but it is necessary to confirm this using Pearson's r.

D ( ) Moderate correlation, with The Godfather as an outlier.

E ( ) If you remove Jurassic World, the pairwise correlations will decrease.

F ( ) I don't know

Bubble chart with color

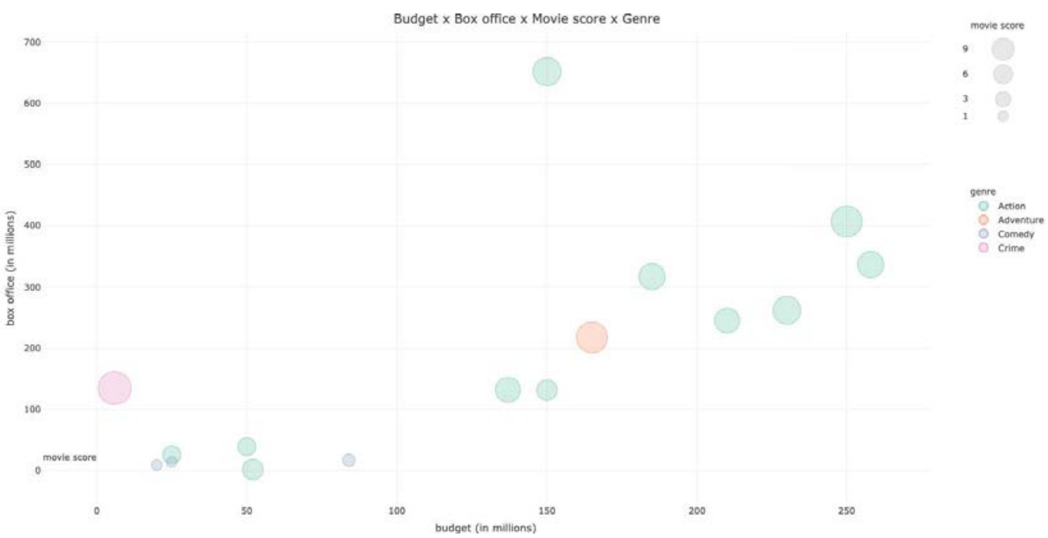


Figure E.8: Bubble chart with color

[Conceptual] This type of chart best allows for:

A ( ) comparing and showing the relationships between the box office and budget values for each genre

B ( ) identifying which genre has the highest volume of movies

C ( ) analyzing in what ranges of budget there is a smaller incidence of a certain genre

D ( ) calculating which budgets are less frequent

E ( ) identifying that movies with largest scores always have large box office values

F ( ) I don't know

[Suitable] In general, Crime movies received higher scores than Comedy movies.

A ( ) True

B ( ) False, they received lower scores

C ( ) False, they received the same scores

D ( ) False, the circles have the same sizes

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] Which budget value has the fewest number of scores?

A ( ) This type of chart does not allow or help to answer the question

B ( ) 1 M

C ( ) 4 M

D ( ) < 10 M

E ( ) 200 M

F ( ) I don't know

Histogram

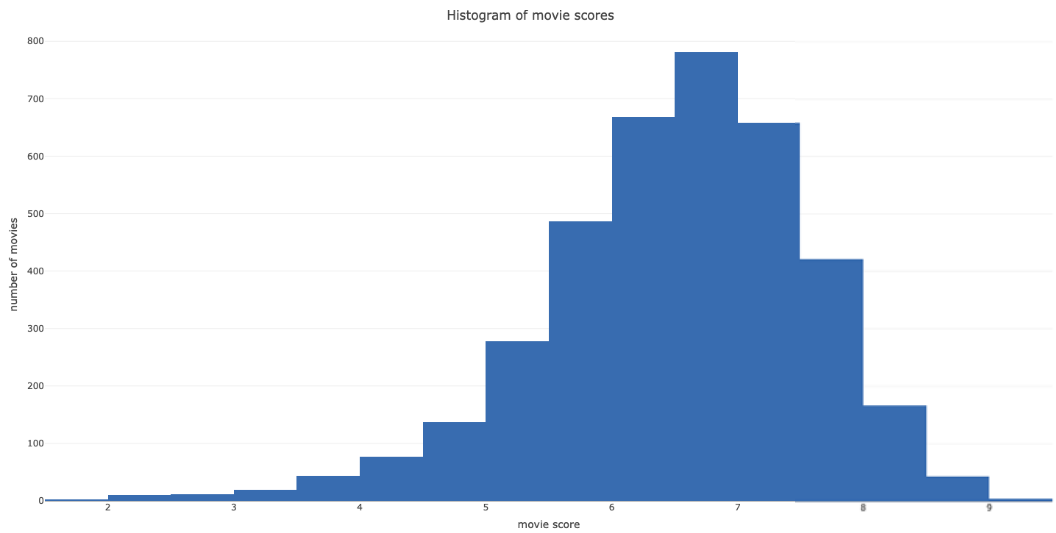


Figure E.9: Histogram

[Conceptual] This type of chart best allows for:

A ( ) characterizing the distribution of scores

B ( ) identifying the dispersion of the frequency of scores

C ( ) identifying the mean of scores

D ( ) calculating the total scores in each interval

E ( ) extracting the exact value of each score interval

F ( ) I don't know

[Suitable] Approximately how many movies have scores between 5.0 and 5.5?

A ( ) 278

B ( ) 486

C ( ) 137

D ( ) 20% of total movies

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] What is the percentage of scores with a frequency of less than 50?

A ( ) This type of chart does not allow or help to answer the question

B ( ) 50%

C ( ) 25%

D ( ) 5%

E ( ) 0

F ( ) I don't know

Line chart (single)

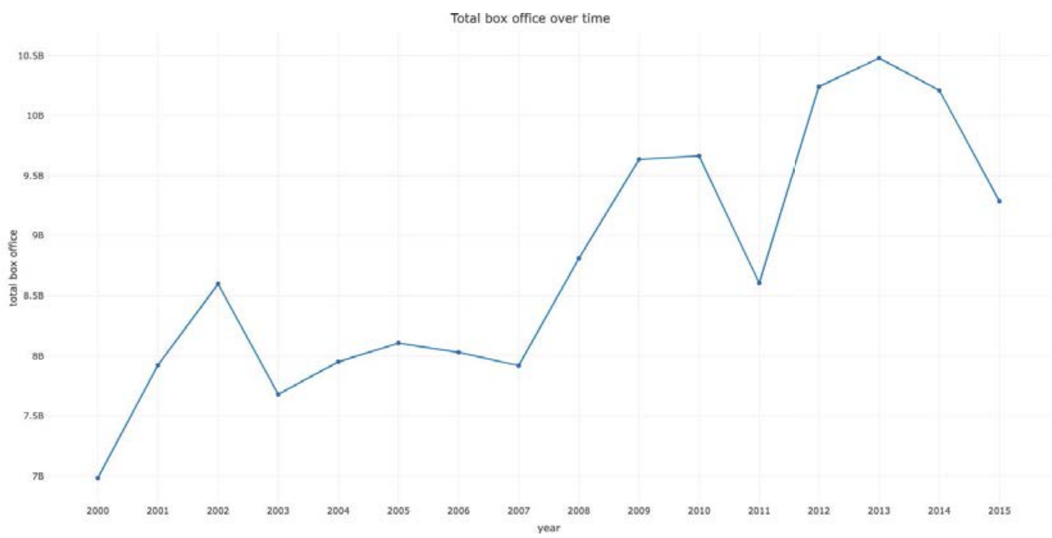


Figure E.10: Line chart (single)

[Conceptual] This type of chart best allows for:

A ( ) identifying trends and analyzing how the data has changed over time

B ( ) extracting the exact numerical value in each section of the curve

C ( ) identifying from which point on there was a gradual increase of total box office

D ( ) analyzing when there was a recovery period

E ( ) calculating the total box office in the period

F ( ) I don't know

[Suitable] Over the last two years of the period, the total box office...

A ( ) decreased

B ( ) was stable

C ( ) increased

D ( ) was injective

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] What is the mean box office in the period?

A ( ) This type of chart does not allow or help to answer the question

B ( ) 8.5B

C ( ) The same as the median

D ( ) 9B

E ( ) The maximum minus the minimum divided by 2

F ( ) I don't know

Line chart (multiple)

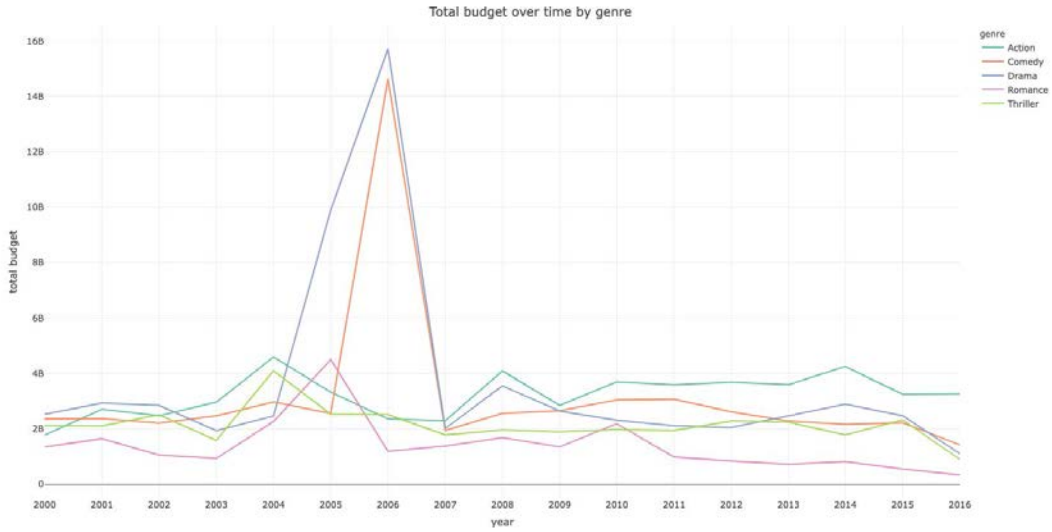


Figure E.11: Line chart (multiple)

[Conceptual] This type of chart best allows for:

A ( ) analyzing how the budget of each genre has changed over time

B ( ) identifying which categorical variable is more frequent in a specific period

C ( ) understanding why one genre is more stable than another

D ( ) specifying time points with declining data values

E ( ) displaying the frequency of continuous data of different data series

F ( ) I don't know

[Suitable] Which genre had the largest variation of total budget between 2010 and 2011?

A ( ) Romance

B ( ) Drama

C ( ) Action

D ( ) Comedy

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] Why did Romance remain stable while Drama had so many spikes?

A ( ) This type of chart does not allow or help to answer the question

B ( ) Because Drama had a larger budget

C ( ) Because Romance had a lower budget

D ( ) Because y-axis does not start at zero

E ( ) Because they are directly correlated

F ( ) I don't know

Scatterplot

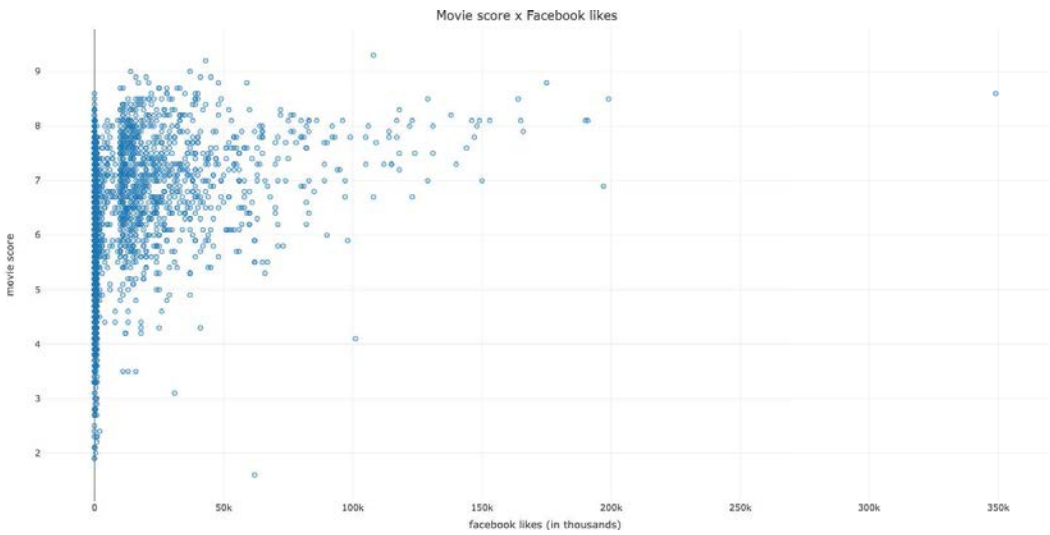


Figure E.12: Scatterplot

[Conceptual] This type of chart best allows for:

A ( ) detecting if there is a relationship or correlation between the two variables

B ( ) detecting relationships between the number of scores

C ( ) analyzing the frequency of each score value

D ( ) identifying where score values are more concentrated

E ( ) analyzing which mode of score has the highest number of likes

F ( ) I don't know

[Suitable] How many Facebook likes are there for the score of 1.6?

A ( ) 62K

B ( ) 0

C ( ) 340K

D ( ) 50K

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] Is there a relationship between the number of points and the score value?

A ( ) This type of chart does not allow or help to answer the question

B ( ) Yes

C ( ) No

D ( ) It depends on the best-fit line

E ( ) It depends on the trend line

F ( ) I don't know

Scatterplot with color

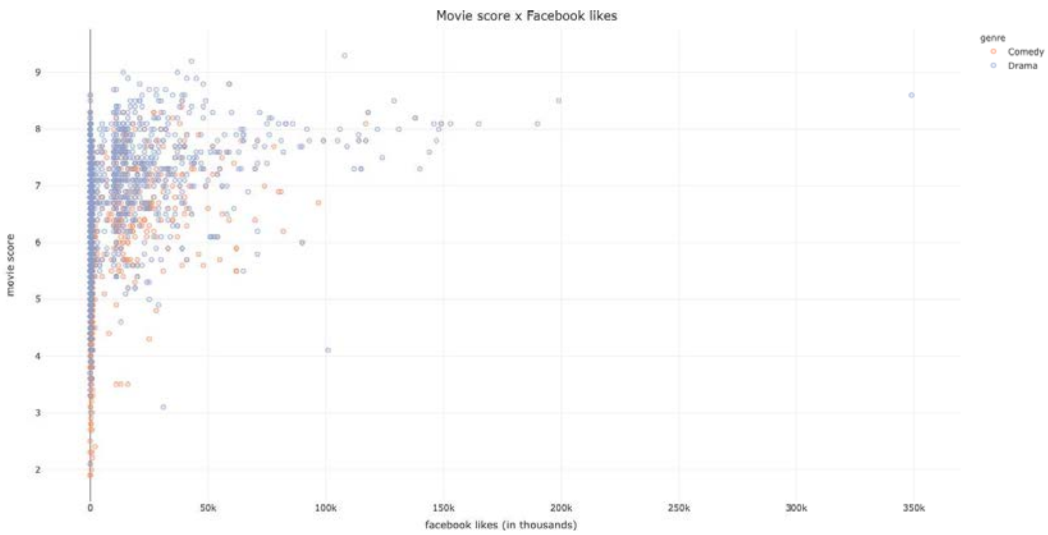


Figure E.13: Scatterplot with color

[Conceptual] This type of chart best allows for:

A ( ) detecting if there is a relationship or correlation between score and facebook likes in each genre

B ( ) analyzing the distribution of genre in relation to the growth of score

C ( ) identifying which genre has a concentration over score

D ( ) defining which score values can be considered outliers

E ( ) revealing which is the common genre in relation to one score value

F ( ) I don't know

[Suitable] Which genre has the fewest movies with a score lower than 4?

A ( ) Drama

B ( ) Comedy

C ( ) None

D ( ) They are the same

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] Which genre is most concentrated in terms of Facebook likes?

A ( ) This type of chart does not allow or help to answer the question

B ( ) Comedy

C ( ) Drama

D ( ) Both

E ( ) Neither

F ( ) I don't know

Table

IMDb Dataset

keywords	number of movies
love	151
friend	136
murder	130
death	102
police	92

Figure E.14: Table

[Conceptual] This type of chart best allows for:

A ( ) obtain the exact value for a certain keyword

B ( ) explaining why a keyword is associated to more movies than another

C ( ) defining what each keyword means

D ( ) immediately identifying the mean number of movies

E ( ) immediately identifying the median number of movies

F ( ) I don't know

[Suitable] Which keyword has the fewest movies?

A ( ) Police

B ( ) Death

C ( ) Friend

D ( ) Murder

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] Why is love so frequent when compared to police?

A ( ) This type of chart does not allow or help to answer the question

B ( ) Because people are more romantic

C ( ) Because people like romantic movies more

D ( ) Other information needs to be taken into account to claim this

E ( ) They are proportionally equal, taking into account the number of movies produced with each keyword

F ( ) I don't know

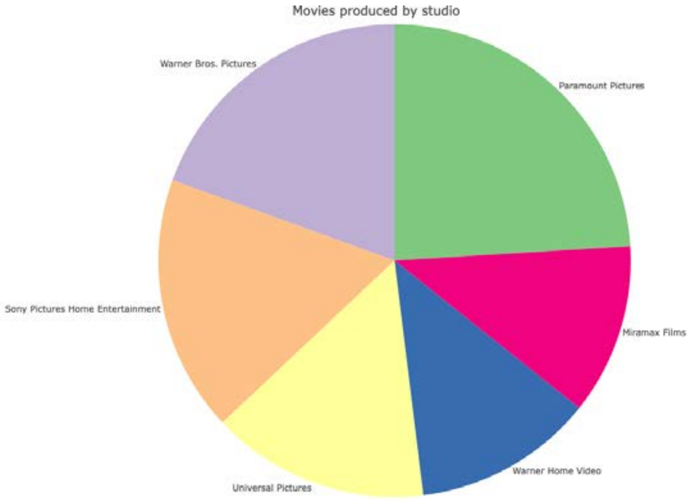


Figure E.15: Pie chart

Pie chart

[Conceptual] This type of chart best allows for:

A ( ) showing the proportions of each studio

B ( ) analyzing the distribution of movies between studios

C ( ) extracting the exact value of each studio

D ( ) identifying which studio has the lowest value

E ( ) revealing trends

F ( ) I don't know

[Suitable] Do you have any comments on the study or any problems you faced while answering the questions?

A ( ) Miramax Films

B ( ) Warner Home Video

C ( ) Universal Pictures

D ( ) Sony Pictures

E ( ) This type of chart does not allow or help to answer the question

F ( ) I don't know

[Unsuitable] Paramount produced twice as many movies as Miramax.

A ( ) This type of chart does not allow or help to answer the question

B ( ) False, it's 1,5 times

C ( ) True

D ( ) False, it’s 2,5 times

E ( ) False, it’s the same

F ( ) I don’t know