INGRID FERNANDES RIBEIRO FABIO
RAISSA VIEGAS SIFFERT GIRUNDI

COVID-19: Uma aplicação de Data Science, por meio de Clustering em Python, para analisar a relação entre os dados sociodemográficos e as medidas de contenção com a evolução do número de casos confirmados da doença

PROJETO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO
APRESENTADO AO DEPARTAMENTO DE ENGENHARIA INDUSTRIAL
DA PUC-RIO, COMO PARTE DOS REQUISITOS PARA OBTENÇÃO
DO TÍTULO DE ENGENHEIRO DE PRODUÇÃO

Orientadora: Fernanda Araujo Baião

Departamento de Engenharia Industrial
Rio de Janeiro, 11 de junho de 2021.

AGRADECIMENTOS

Agradecemos, primeiramente, à dedicação e suporte oferecidos pela nossa orientadora Fernanda Baião. Temos certeza que escolhemos a professora certa para nos acompanhar no desenvolvimento do trabalho. Ela representa com excelência a PUC-Rio!
Agradecemos a todos os professores do curso de Engenharia de Produção pelo conhecimento compartilhado durante todos esses anos em que a PUC-Rio foi a nossa segunda casa.
Agradecemos uma à outra pela nossa parceria durante esses meses intensos de trabalho.

- INGRID FERNANDES RIBEIRO FABIO

Agradeço aos meus pais, Rosangela e Jorge Fabio, por toda a educação, carinho e suporte durante a minha vida, pois tudo o que eu sou, eu devo a eles. Meu pai sempre foi uma inspiração pra mim e minha mãe sempre foi minha melhor amiga. Agradeço à minha avó Joaquina, e aos meus dindinhos, Maria Emília e Carlos, por me acompanharem e ajudarem. Agradeço aos demais familiares que sempre estiveram presentes e aos amigos que fui construindo ao longo da minha jornada. Agradeço, principalmente, ao Tiago Montalvão por todo o auxílio oferecido neste TCC. E, por último, agradeço ao ISMART pelo apoio e financiamento no âmbito acadêmico.

- RAISSA VIEGAS SIFFERT GIRUNDI

Primeiramente, agradeço a Deus por nunca soltar a minha mão e por me presentear com uma base tão sólida que é a minha Família! Agradeço ao meu Pai Antônio José por ser meu maior incentivador e espelho de luta e persistência. Agradeço à minha mãe Maria José por todo companheirismo e por ser a minha inspiração como mãe, mulher e melhor amiga. Agradeço aos meus irmãos Bernnard e Felippi por serem meus grandes parceiros e me apoiarem em todos os momentos. Agradeço aos meus sobrinhos Giovanna, Luanna e Arthur por serem o motivo da minha alegria e por tornar tudo mais leve e incrível. Agradeço à minha família Viegas e Girundi por me ensinarem a importância do amor e união e por todo o apoio durante a minha trajetória. Agradeço aos meus amigos por estarem sempre ao meu lado. E por fim, agradeço aos meus avós que do céu irão comemorar essa conquista junto comigo.

RESUMO

A COVID-19 foi decretada pela Organização Mundial da Saúde (OMS) como uma pandemia global em 11 de março de 2020. Por ser uma doença nova no mundo, não existiam vacinas ou medicamentos eficazes e, por isso, a luta dos países para contê-la começou com a adoção de diversas intervenções não farmacêuticas, ou seja, medidas de contenção como fechamento de escolas e locais de trabalho. Após um ano do início da pandemia causada pelo coronavírus, este trabalho busca analisar se existem correlações entre os dados sociodemográficos dos países e as medidas de restrições adotadas por eles com a evolução do número de casos da doença. Através do ciclo de vida de Data Science, foram utilizadas técnicas de Clusterização em Python (K-means, Agglomerative, DBSCAN e HDBSCAN) para encontrar as semelhanças e/ou discrepâncias entre os países agrupados. Na visualização dos resultados, obtiveram-se indícios de confirmação da hipótese de pesquisa inicial, havendo algumas exceções de países que se destacaram dentro dos seus clusters por motivos diversos.

Palavras-chave: COVID-19; Data Science; Clusterização; HDBSCAN.

ABSTRACT

COVID-19 was decreed by the World Health Organization (WHO) as a global pandemic on March 11, 2020. As it is a new disease in the world, there were no vaccines or medicines and, therefore, the struggle of countries to contain it began with the adoption of several non-pharmaceutical interventions, that is, containment measures such as closing schools and places of work. After a year of the Coronavirus, this work seeks to analyze whether there are correlations between the sociodemographic data of the countries and the restrictions measures adopted by them with the evolution of the number of cases of the disease. Through the Data Science life cycle, Python Clustering techniques (K-means, Agglomerative, DBSCAN, and HDBSCAN) were used to find similarities and/or discrepancies between the grouped countries. When viewing the results, the initial research question is confirmed, with only a few exceptions from countries that stood out within their clusters for different reasons.

Keywords: COVID-19; Data Science; Clustering; HDBSCAN.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

1. Introdução

A COVID-19 é a doença que mais ameaça a saúde pública do mundo desde que ocorreu a pandemia da Gripe Espanhola em 1918, esta última causada pelo vírus H1N1. A COVID-19 foi identificada pela primeira vez em dezembro de 2019 em Wuhan na China e em menos de 15 dias já havia se espalhado para outros países asiáticos como Tailândia e Japão, o que evidencia a sua alta velocidade de contaminação (VICK, 2020).

A sigla COVID significa *"COrona VIrus Disease"* (Doença do Coronavírus), já o número 19 se refere ao ano de 2019, ano o qual teve o primeiro caso da doença abertamente publicada. Em 11 de março de 2020 essa doença foi declarada como uma pandemia global pela Organização Mundial de Saúde (OMS) e, em 23 de março, o vírus já tinha sido detectado em 172 países dos 195 existentes (UNA-SUS, 2020).

À medida que o vírus se espalhou pelo mundo e dados sobre esta expansão passaram a ser publicamente disponibilizados, pode-se começar a estudar melhor a doença e obter dados relevantes para a tomada de decisões, tanto públicas como privadas. A partir dos dados coletados, tomadores de decisão de diversas esferas (gestores públicos e diretores de hospitais, por exemplo) tinham maior embasamento para planejar e atuar no sentido de cessar a proliferação do vírus. Dentre tais ações tomadas por governantes em diversos países, pode-se citar o estabelecimento de medidas não farmacêuticas, também denominadas medidas de contenção.

Cada país tomou suas próprias medidas de contenção e estas foram estabelecidas de acordo com diversos critérios: muitos se basearam nos dados históricos de pandemias passadas, outros se basearam em recomendações feitas por especialistas e outros, ainda, basearam-se na experiência dos demais países que já haviam realizado medidas de contenção e observaram indícios de sucesso no declínio da proliferação do vírus.

Os estudos sobre a COVID-19 são impactados de forma direta pela metodologia dos registros que os países realizam. A falta de um padrão dos registros para todos os países acaba

dificultando a comparação e a relação dos diferentes cenários de aumento (ou queda) do número de casos confirmados. Com base nessa dificuldade de análise, este trabalho apresenta um estudo dos dados disponibilizados pelos países acerca do impacto dos dados sociodemográficos e das medidas de contenção na evolução dos casos confirmados da COVID-19 no mundo. Para a realização desse estudo, utilizou-se uma base de dados pública e, para a mineração de padrões, foi aplicada a técnica de Clusterização (Clustering) por meio da linguagem Python. A metodologia do estudo seguiu as etapas do ciclo de vida de Data Science (SHCHERBAKOV *et al.*, 2014).

Com este estudo, pode-se, assim, reconhecer a tamanha importância dos registros dos dados, sejam eles sobre os casos confirmados ou as medidas de contenção tomadas pelos países de forma a viabilizar uma comparação eficaz entre os dados. Para essa análise, foi indispensável o uso de dados confiáveis e temporais a fim de que fosse possível explorar a evolução da pandemia ao longo dos dias. Neste trabalho, serão apresentadas a fundamentação da teoria base para o estudo, a descrição das metodologias utilizadas, o desenvolvimento das análises e seus efeitos, as conclusões obtidas e as considerações finais. Ao final, encontram-se as referências bibliográficas citadas.

2. Descrição do contexto da pandemia no Brasil e no Mundo

2.1. Cenário da COVID-19

O coronavírus apareceu em 1937 pela primeira vez no mundo e em 1965 foi classificado como "corona", levando esse nome pelo seu formato ser similar ao de uma coroa. Segundo o Ministério da Saúde, o coronavírus é uma família de vírus na qual provoca infecções das vias respiratórias (ROCHE, 2020).

A COVID-19 (*COrona VIrus Disease 2019*) surgiu em 2019 quando o primeiro caso da doença foi identificado. A doença apresentava como vítimas em comum pessoas que frequentavam o Mercado Atacadista de Frutos do Mar de Wuhan, na China. Ela é uma infecção respiratória provocada pelo Coronavírus da Síndrome Respiratória Aguda Grave 2 (SARS-CoV-2) (SCHUCHMANN *et al.*, 2020).

Em janeiro de 2020, de acordo com a Organização Mundial de Saúde (OMS), o surto da doença causada pelo Coronavírus se tornou uma Emergência de Saúde Pública de Importância Internacional e, em 11 de março, dois meses depois, a OMS caracterizou a COVID-19 como uma pandemia (SCHMIDT *et al.*, 2020).

Os sintomas mais comuns da COVID-19 são febre, cansaço e tosse seca, além disso, algumas pessoas ainda apresentam dores, congestão nasal, perda de olfato e paladar, dor de cabeça, entre outros (FOLHA, 2020). Uma em cada seis pessoas infectadas pela COVID-19 acabam ficando gravemente doentes e desenvolvendo alguma dificuldade de respirar. Pessoas idosas e com comorbidades como pressão alta, problemas cardíacos e do pulmão, diabetes ou câncer, possuem um risco maior de ter a doença agravada. No entanto, qualquer pessoa pode ser vítima da COVID-19 e ficar gravemente doente (MARKHAM, 2020).

2.2. Intervenções Não Farmacêuticas

As intervenções não farmacêuticas (NPI's) são medidas de contenção/restrição, diferentes de vacina e medicação, que ajudam a retardar a propagação de uma doença. As NPI's são conhecidas como uma estratégia de mitigação da comunidade. Como o vírus da pandemia é

novo, a população humana tem pouca ou nenhuma imunidade contra ele, permitindo que o vírus se espalhe rapidamente de um em um para todo o mundo. As intervenções não farmacêuticas estão entre as melhores maneiras de controlar uma pandemia quando as vacinas ainda não estão disponíveis (CDC, 2021).

Na maioria dos países, uma série de intervenções não farmacêuticas foram adotadas para alcançar o "distanciamento social" como forma de bloquear as principais vias de transmissão do vírus. O objetivo dessas intervenções era desacelerar a pandemia, restringindo a mobilidade e, assim, permanecendo dentro da capacidade dos sistemas de saúde (ASKITAS *et al.*, 2021). O objetivo principal de todos os países é que a pandemia seja controlada desacelerando a transmissão do contágio e, assim, reduzindo as mortalidades causadas pelo vírus. Apesar do objetivo ser o mesmo, as formas como estão sendo aplicadas as medidas de contenção variam de acordo com o país. Podemos notar essa diferença através da coleta realizada pela Oxford COVID-19 Government Response Tracker (UNIVERSITY OF OXFORD, 2020), em que intervenções não farmacêuticas são adotadas de diferentes maneiras, como: fechamentos totais ou parciais de escolas, medidas isoladas ou em todo território, e políticas de testagem apenas em pessoas que apresentam sintomas ou em toda população.

As medidas mais eficazes para evitar aglomeração incluem fechar e restringir a maioria dos lugares onde as pessoas se reúnem em números menores ou maiores por longos períodos de tempo (empresas, bares, escolas e assim por diante) (HAUG *et al.*, 2020).

Os principais projetos que acompanham as intervenções não farmacêuticas dos países são o de Oxford COVID-19 Government Response Tracker, citado anteriormente nesta seção, que rastreia como as medidas estão sendo implementadas diariamente em cada país, e o Complexity Science Hub COVID-19 Control Strategies List (CCCSL), que visa gerar um conjunto de dados estruturado sobre as respostas do governo a COVID-19, incluindo os respectivos cronogramas de sua implementação. Ambos foram utilizados como embasamento ao longo do nosso estudo.

Uma forma de medir a eficácia de um conjunto de medidas de contenção é através do chamado *Stringency Index*, que é uma medida criada por Oxford composta com base em nove

indicadores de resposta, incluindo fechamentos de escolas, fechamentos de locais de trabalho e proibições de viagens (UNIVERSITY OF OXFORD, 2020). Este índice será mais bem explorado no capítulo Análise de Resultados.

2.3. Índice de Mobilidade

O índice de mobilidade nos permite acompanhar o deslocamento dos indivíduos através dos seus dispositivos móveis. Dessa forma, o Google disponibilizou através do Google Maps um conjunto de dados agregados e anônimos dos usuários que ativaram a configuração "histórico de localização", sendo possível gerar *insights* para tomadas de decisões críticas no combate a COVID-19. Além disso, foi viável gerar gráficos com tendências de deslocamento ao longo do tempo por região e em diferentes categorias de locais, como varejo e lazer, mercados e farmácias, parques, estações de transporte público, locais de trabalho e áreas residenciais (GOOGLE, 2021).

Os dados oriundos do índice de mobilidade tem o objetivo de fornecer o que mudou em função das políticas de contenção implementadas. Logo, também viabilizou acompanhar a efetividade das medidas, pois a teoria é diferente da prática, já que em muitos países as pessoas acabam por não cumprir as intervenções não farmacêuticas. Este índice será mais explorado na Análise dos Resultados.

3. Ciclo de Vida de Data Science

Graças a conceitos como *Big Data* e Internet das Coisas (IoT), a quantidade de dados gerados pela sociedade aumenta diariamente no mundo todo. Essa nova realidade tem causado um novo desafio, pois, além de armazenar e recuperar esses dados, precisamos ser capazes de analisar e interpretar esse grande volume de dados e informação. Para conseguir lidar com todo esse acervo, é preciso investir em áreas como Data Science, que é capaz de limpar, tratar e analisar essas grandes massas de dados, transformando-as em *insights* úteis (CANAL COMSTOR, 2021).

No setor de Saúde, os dados desempenham um papel importante para monitorar e gerenciar operações, e viabilizar inovação. Os dados de Saúde fornecem a base para realizar a avaliação e produzir medicamentos mais eficazes, além de estabelecer uma melhor comunicação entre pacientes e médicos, melhorar a qualidade geral dos cuidados de saúde e dar uma visão mais profunda do relatório de saúde de um paciente ou falar como um medicamento específico está respondendo. Um caso de sucesso é a Assistência Virtual para Pacientes, já que, considerando a pandemia de COVID-19, houve um direcionamento de esforços para limitar as visitas dos pacientes ao hospital, transferindo tudo para plataformas virtuais. As ferramentas *Al-Integrated* permitem que os pacientes obtenham visitas virtuais e interajam com médicos online por meio de chamadas de voz e vídeo, além de ser possível interagir com *chatbots* que são programados para fornecer soluções de saúde eficientes para os pacientes (LIAM, 2021).

Para obter-se conhecimento novo e útil a partir dos dados, como *insights* de qualidade, recomenda-se seguir as etapas do ciclo de vida de Data Science, que consiste em Entendimento do Problema, Coleta de Dados, Ciclo Interno de Pesquisa, Visualização de Resultados, Criação de Ações baseadas nos Resultados e Receber Feedbacks das Ações, conforme ilustrado na Figura 1.

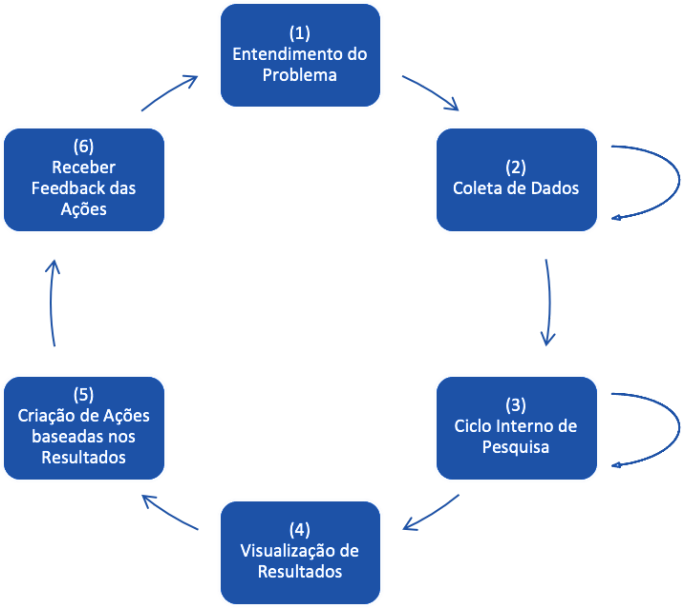

Figura 1: Ciclo de Vida de Data Science adaptado (SHCHERBAKOV *et al.*, 2014)

As etapas 2 e 3 da Figura 1 são cíclicas e feitas diversas vezes até obtermos um resultado satisfatório. Por exemplo, no Ciclo Interno de Pesquisa (etapa 3) detalhado na Figura 2, é realizado, primeiramente, o Pré-Processamento dos dados coletados onde ocorre a preparação, organização e estruturação destes dados. Sequencialmente, realiza-se a mineração dos dados em que é aplicado um algoritmo que procura efetivamente padrões/relações e regularidades em um determinado conjunto de dados. A última etapa do Ciclo Interno de Pesquisa consiste no pós-processamento, onde verifica-se a qualidade do conhecimento (padrões) descobertos, identificando se ele auxilia na resolução do problema original que motivou o estudo, definido na etapa de Entendimento dos Dados (CALIL *et al.*, 2008).

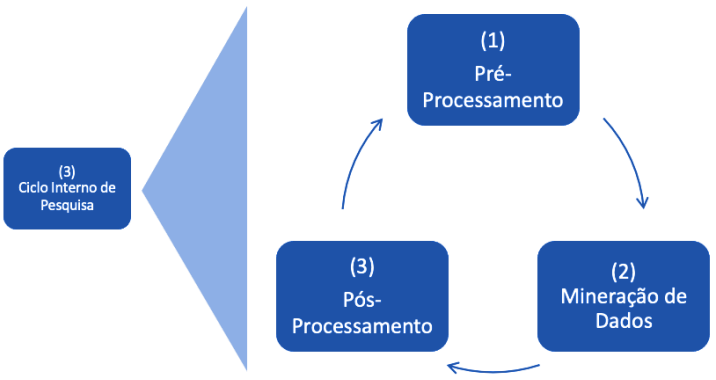

Figura 2: Ciclo Interno de Pesquisa detalhado

Nos próximos itens, explica-se melhor as etapas cobertas pelo presente trabalho, que são: Entendimento do Problema, Coleta de Dados e Ciclo Interno de Pesquisa, que serão passos fundamentais para a Visualização dos Resultados.

3.1. Entendimento do Problema

Antes de iniciar um projeto de Data Science, é crucial entender o problema ao qual se está buscando resolver. Nessa primeira etapa, deve-se determinar os objetivos que o projeto busca atingir, o problema/questão de pesquisa ou a hipótese que será investigada e qual será o tipo de pesquisa realizada.

Um problema pode ter muitas hipóteses, que são soluções possíveis para a sua resolução. As hipóteses são suposições colocadas como respostas plausíveis e provisórias para o problema de pesquisa. As hipóteses são provisórias porque poderão ser confirmadas ou refutadas com o desenvolvimento da pesquisa (VILELA JUNIOR, [20--?]).

Já os possíveis tipos de pesquisa, de acordo com Nisbet *et al.* (2009), são:

- Descritivo: busca encontrar relações entre as variáveis e padrões entre os dados;
- Preditivo: busca encontrar previsões com base em algum dado;
- Prescritivo: busca, com base no descritivo e preditivo, prescrever melhores decisões a serem tomadas a partir de um determinado cenário.

Após a determinação desses pontos, pode-se ir com clareza em busca dos dados que possam ser necessários para responder o problema de pesquisa ou as hipóteses previamente definidas.

3.2. Coleta de Dados

Os projetos de Data Science seguem etapas bem definidas, logo, a etapa seguinte do estudo, após o entendimento do problema, é a coleta e integração de dados a partir de uma ou várias fontes de informação relevantes e confiáveis. O intuito em coletar esses dados é fornecer material para que se possa mapeá-los e processá-los a fim de obter respostas aos problemas elencados (SANTANA, 2019).

Como a finalidade de um projeto de Data Science é obter *insights* de negócios a partir de conjuntos de dados de grande volume, as suas fontes acabam sendo muito diversas. Há conjuntos de dados internos, ou seja, provenientes de fontes internas e conjuntos de dados externos, oriundos de lugares terceiros. Além disso, existem diferentes tipos de dados como dados estruturados proveniente de arquivos, por exemplo CSV, que seguem uma estrutura de organização e dados não estruturados como textos, vídeos e fotos (SANTANA, 2019). Por fim, a integração de todas estas fontes de dados apresenta também desafios quanto à heterogeneidade semântica, uma vez que diferentes conjuntos de dados podem se referir a um mesmo conceito utilizando termos diferentes, ou a conceitos diferentes através de um mesmo termo. Por isso, algumas abordagens de Data Science mais recentes vêm ressaltando a importância de tratar aspectos de Modelagem Conceitual para integração e tratamento de dados, inclusive com o uso de Ontologias e Ontologias de Fundamentação (AMARAL; BAIÃO; GUIZZARDI, 2021).

3.3. Pré-Processamento

O Pré-Processamento é um conjunto de atividades que envolvem preparação, organização e estruturação dos dados. Trata-se de uma etapa fundamental que precede a realização de análises e predições. Quando os cientistas de dados buscam aumentar o desempenho de seus modelos, a engenharia e a seleção dos dados costumam ser as primeiras técnicas que eles buscam utilizar para seu aprimoramento. Seguem exemplos de alterações que pode ser feito para enriquecer um dataset de acordo com Schreck (2018):

- **Seleção de Dados (*Data Selection*):** quando ocorre a filtragem das linhas segundo algum critério;
- **Seleção de Atributos (*Feature Selection*):** quando ocorre a filtragem das colunas segundo algum critério;
- **Engenharia de Atributos (*Feature Engineering*):** quando ocorre a criação de novas colunas segundo alguma regra.

Pode-se destacar, através dessas etapas, a extrema importância de olhar com atenção os registros que possam estar duplicados, inválidos ou faltantes, além de observar se os dados estão formatados de forma convencional. É preciso realizar, também, o tratamento dos

outliers, ou seja, dados que estão muito fora do padrão e que irão provavelmente afetar negativamente os resultados (GONÇALVES, 2018).

As transformações corretas dependem de muitos fatores, como: o tipo e a estrutura dos dados, o volume dos dados e, claro, os objetivos do estudo. O passo a passo do Pré-Processamento será detalhado mais adiante na Seção 4.3.

3.4. Mineração de Dados - Clustering

A Mineração de Dados (Data Mining) é uma etapa que visa explorar grandes quantidades de dados à procura de padrões consistentes. Ela é formada por um conjunto de ferramentas e técnicas, que através do uso de algoritmos de aprendizagem ou classificação baseados em redes neurais e estatística são capazes de explorar um conjunto de dados, extraindo ou ajudando a evidenciar padrões (CETAX, 2020).

Uma das técnicas da Mineração de Dados é o Clustering (Clusterização). Ele representa um método de identificação de grupos de dados semelhantes que estão em um conjunto de dados, sendo uma das técnicas mais populares de Data Science. O seu objetivo é segregar grupos com características semelhantes e agrupá-los em clusters. Ou seja, ele divide os dados em vários grupos de forma que os dados de um mesmo grupo sejam mais semelhantes aos dados do mesmo grupo do que dos demais grupos (KAUSHIK, 2016).

A Clusterização foi a técnica escolhida para o trabalho, sendo ela um passo essencial no desenvolvimento de modelos e das análises do Capítulo 5. A seguir, explora-se melhor sobre os tipos de Clustering existentes no nosso estudo.

3.4.1. Tipos de Clustering

Para a realização de uma análise, a quantidade de informação deve ser organizada de acordo com o objetivo do cientista de dados. Existem hoje, mais de 100 algoritmos de Clusterização, porém, a popularidade da maioria não é tão relevante, assim como o campo de aplicação. (SEMANTIX, 2019).

Para esse estudo utiliza-se três tipos de clusters comumente utilizados:

- **Baseado em distância:** é a Clusterização baseada no cálculo da distância entre os objetos do conjunto de dados, ou seja, na noção de que os pontos de dados mais próximos no espaço de dados apresentam mais similaridade entre si do que os pontos de dados mais distantes. Dependendo do algoritmo, o mesmo pode juntar ou dividir o conjunto de dados. Sua vantagem vem da simplicidade da sua aplicação e sua desvantagem vem da ausência de escalabilidade para lidar com um grande conjunto de dados. O tipo mais popular deste algoritmo é o aglomerativo, onde você inicia inserindo um certo número de *datapoints* que são adicionados na sequência em clusters cada vez maiores, até que seja atingido um certo limite (SEMANTIX, 2019).

- **Baseado em hierarquia:** O modelo tem como objetivo classificar cada objeto do conjunto de dados em um cluster particular. O número de clusters é determinado aleatoriamente, o que provavelmente pode ser considerado a maior fraqueza do método (SEMANTIX, 2019).

- **Baseado em densidade:** divide o conjunto de dados em clusters que tem como parâmetro a distância ε (comprimento de um ponto aos pontos vizinhos), então, se um dado é localizado dentro do círculo de raio ε, este pertence ao cluster. Ou seja, ele identifica áreas no espaço amostral onde existam concentrações de elementos e cada uma dessas áreas forma um novo cluster. Como ele funciona identificando clusters "densos" de pontos, ele permite criar clusters de forma arbitrária, tais como elíptica, cilíndrica e espiralada, além de identificar mais facilmente os *outliers* nos dados (SEMANTIX, 2019).

Alguns algoritmos utilizam mais do que um tipo de Clustering, como o HDBSCAN, que é um algoritmo que estende o DBSCAN (baseado em densidade) convertendo-o em um algoritmo de Clustering hierárquico para, posteriormente, utilizar uma técnica que, com base nas estabilidades dos clusters, extrai um Clustering simples (COMPARING, 2016). Tais algoritmos serão detalhados na Seção 3.4.2, a seguir.

3.4.2. Algoritmos de Clustering

Abaixo, segue a visão teórica dos algoritmos utilizados neste trabalho:

3.4.2.1. K-means

O K-means é um algoritmo baseado em distância, que agrupa os dados tentando separar as amostras em n grupos de variância igual, minimizando um critério conhecido como inércia ou soma dos quadrados dentro do agrupamento. Este algoritmo requer que o número de clusters seja especificado. Como ele se adapta bem a um grande número de amostras, tem sido bastante utilizado em uma grande variedade de áreas de aplicação de diferentes campos. Sua vantagem é que ele é um algoritmo simples de se implementar que garante a convergência e pode ser utilizado para uma grande variedade de dados. Já algumas das suas desvantagens vem da necessidade de escolher o número de clusters para o agrupamento de dados sem levar em consideração tamanho e densidade dos dados (SCIKIT-LEARN, 2020).

Existem várias formas de calcular a distância entre pontos, sendo a distância euclidiana (distância de Minkowski) uma das medidas de distância mais comumente usadas. A fórmula abaixo mostra como calcular a distância euclidiana entre dois pontos em um espaço bidimensional, em que o cálculo é feito usando o quadrado da diferença entre as coordenadas X e Y dos pontos (YILDIRIM, 2020).

$$distância\ euclidiana\ (a, b) = \sqrt{\left(a_x - b_x\right)^2 + \left(a_y - b_y\right)^2}$$

O K-means é um processo iterativo baseado no algoritmo de maximização da expectativa. Depois que o número de clusters é determinado, ele funciona executando as seguintes etapas descritas por Yildirim (2020):

1. Seleciona centroides aleatoriamente (centro do cluster) para cada cluster;
2. Calcula a distância de todos os pontos de dados aos centroides;
3. Atribui pontos de dados ao cluster mais próximo;
4. Encontra os novos centroides de cada cluster tomando a média de todos os pontos de dados no cluster;
5. Repete as etapas 2, 3 e 4 até que todos os pontos convirjam e os centros do cluster parem de se mover.

Para definir a melhor quantidade de clusters que podem ser encontrados, mesmo sem saber a resposta antecipadamente, podemos aplicar o Método Cotovelo (Elbow). Ele consiste em rodar o K-means para várias quantidades diferentes de clusters e dizer qual dessas quantidades é o número ótimo de clusters, testando a variância dos dados em relação ao número de clusters. O valor de *k* ideal é aquele que tem um menor Somatório do Erro Quadrático e ao mesmo tempo o menor número de clusters. Ele possui esse nome, porque a partir do ponto que seria o "cotovelo" não existe uma discrepância tão significativa em termos de variância, dizendo que a melhor quantidade de clusters *k* seria exatamente onde o cotovelo estaria, conforme a Figura 3 (GUEDES, 2019).

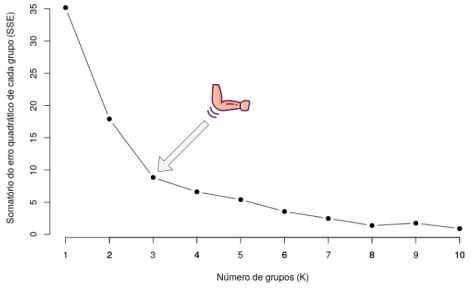

Figura 3: Exemplo de Método do Cotovelo (DATAAT, 2021)

3.4.2.2. Agglomerative Clustering

O Agrupamento Aglomerativo (Agglomerative Clustering) é um agrupamento hierárquico que usa uma abordagem de baixo para cima onde cada ponto do dataset começa em seu próprio agrupamento e os agrupamentos são sucessivamente mesclados, ou seja, cada objeto é inicialmente considerado um cluster de um único elemento (folha) e em cada etapa do algoritmo, os dois clusters mais semelhantes são combinados em um novo cluster maior (nós). Este procedimento é iterado até que todos os pontos sejam membros de apenas um único grande cluster (raiz), o resultado é uma representação dos objetos baseada em árvore, denominada dendrograma (NOVIA, [20--]). Segue abaixo esse passo a passo do agrupamento aglomerativo conforme Towards (2020):

1. Inicialmente, todos os pontos de dados são um cluster próprio;
2. Pega dois clusters mais próximos e junta-os para formar um único cluster;
3. Continua recursivamente na etapa 2 até obter o número desejado de clusters.

Suas vantagens são que sua estrutura é mais informativa para que se decida o número de clusters se comparado ao K-means e que possui uma fácil implementação. Já as desvantagens são que não é um algoritmo adequado para um grande número de dados, não permite que se desfaça a etapa anterior, é muito sensível a *outliers* e a ordenação dos registros tem impacto no resultado final (SCIKIT-LEARN, 2020).

3.4.2.3. DBSCAN

Este algoritmo vê os clusters como áreas de alta densidade separadas por áreas de baixa densidade. Devido a essa visão bastante genérica, os clusters encontrados pelo DBSCAN podem ter qualquer formato, ao contrário do K-means, que assume que os clusters têm formato convexo (clusters alongados ou de formato irregular). O componente central do DBSCAN é o conceito de amostras de núcleo, que são amostras que estão em áreas de alta densidade. Assim, dado um conjunto de pontos em algum espaço, ele agrupa pontos compactados, ou seja, pontos com muitos vizinhos próximos, marcando como *outliers* pontos que ficam isolados em regiões de baixa densidade onde os vizinhos mais próximos estão muito distantes (WIKIPÉDIA, 2021).

A seguir, seguem as etapas do algoritmo DBSCAN (KDNUGGETS, 2017):

1. Inicialmente, o algoritmo escolhe arbitrariamente um ponto no conjunto de dados (até que todos os pontos tenham sido visitados);
2. Se houver pelo menos um '*minPoint*' (número mínimo de pontos agrupados para uma região a ser considerada densa) dentro de um raio de 'ε' (medida de distância que será usada para localizar os pontos na vizinhança de qualquer ponto) para o ponto, então consideramos todos esses pontos como parte do mesmo cluster;
3. Os clusters são, então, expandidos repetindo recursivamente o cálculo de vizinhança para cada ponto vizinho.

As vantagens do DBSCAN são a formação de clusters de formato arbitrário e sua facilidade de detectar os *outliers*. Já suas desvantagens são sensibilidade em relação aos parâmetros de entrada do agrupamento e o não agrupamento de dados com grande diferença de densidade (SCIKIT-LEARN, 2020).

3.4.2.4. HDBSCAN

É um algoritmo que estende o DBSCAN convertendo-o em um algoritmo de Clustering hierárquico e, posteriormente, usando uma técnica para extrair um clustering simples com base na estabilidade dos clusters.

A seguir, seguem as etapas do algoritmo HDBSCAN (MCINNES; HEALY; ASTELS, 2016):

1. Transforma o espaço de acordo com a densidade/dispersão escolhida;
2. Constrói a árvore de abrangência mínima do gráfico de distância ponderada;
3. Constrói uma hierarquia de clusters de componentes conectados;
4. Condensa a hierarquia do cluster com base no tamanho mínimo do cluster previamente definido;
5. Extrai os cluster estáveis da árvore condensada.

É um algoritmo relativamente rápido para grandes conjuntos de dados que detecta células periféricas e, para cada uma dessas células, reporta uma probabilidade de atribuição a um cluster. É um algoritmo de Clustering baseado em densidade que é indiferente à forma de cada clusters, não requer especificação do número de clusters e é robusto em relação a clusters com diferentes densidades. Já sua desvantagem é que ele não é tão sensível como o DBSCAN criando clusters com densidades muito baixas (COMPARING, 2016).

Percebe-se assim, que não existe o melhor ou pior algoritmo de Clusterização, porém, alguns são mais apropriados dependendo da estrutura dos dados. Para selecionar o melhor em cada caso, deve-se ter um entendimento completo das vantagens, desvantagens e peculiaridades ou pode-se optar pela tentativa e erro, onde o aprendizado é oriundo dos próprios erros (SEMANTIX, 2019).

3.5. Pós-Processamento

No Pós-Processamento, avaliam-se os padrões obtidos (no caso do Clustering, os agrupamentos descobertos) para verificar se possuem qualidade suficiente para serem utilizados. Como não existe uma análise que seja absoluta para avaliação dos algoritmos de Clusterização, é possível avaliar o resultado desses algoritmos através de diferentes métricas. Essas métricas são executadas a partir de diferentes parâmetros como densidade e distância

mínima entre os clusters quando fala-se de técnicas de Clusterização e permite-se, assim, avaliar o melhor agrupamento dentro dos algoritmos que foram analisados.

3.5.1. Índice Silhouette

O primeiro método utilizado foi o **Índice Silhouette**, que é calculado, para cada ponto, a partir da distância média para os demais pontos do mesmo cluster (a) e a menor distância média para pontos de algum outro cluster (b).

O Índice Silhouette para uma amostra é calculado como:

$$CS = \frac{(b-a)}{max\,(a,b)}$$

a = distância média entre um ponto e todos os outros pontos do mesmo cluster;

b = menor distância média entre um ponto e todos os outros pontos de um outro cluster.

A função retorna o Índice Silhouette médio de todas as amostras. O melhor valor é 1 e o pior valor é -1. Valores próximos a 0 indicam clusters sobrepostos. Valores negativos geralmente indicam que uma amostra foi atribuída ao cluster errado, pois um cluster diferente é mais semelhante. O Índice Silhouette é geralmente maior para clusters convexos do que outros conceitos de clusters, como clusters baseados em densidade que é o caso daqueles obtidos através do DBSCAN (SCIKIT-LEARN, 2020).

3.5.2. Índice Calinski-Harabasz

O segundo método utilizado foi o **Índice Calinski-Harabasz**, também conhecido como Critério de Razão de Variância. O índice é a razão da soma da dispersão entre clusters e da dispersão dentro do cluster para todos os clusters, em que a dispersão é definida como a soma das distâncias ao quadrado.

Para um conjunto de dados E de tamanho n_E que foi agrupado em k clusters, a pontuação Calinski-Harabasz (s) é definida como a proporção da média de dispersão entre clusters e a dispersão dentro do cluster:

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k-1}$$

Onde $tr(B_k)$ é o traço da matriz de dispersão entre o grupo e $tr(W_k)$ é o traço da matriz de dispersão dentro do cluster definida por:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T$$

Sendo C_q o conjunto de pontos no cluster q, c_q o centro do cluster q, c_E o centro de E, e n_q o número de pontos no cluster q.

O índice Calinski-Harabasz também é geralmente mais alto para clusters convexos do que para outros conceitos de clusters, como clusters baseados em densidade. A pontuação é maior quando os clusters são densos e bem separados, o que se relaciona com um conceito padrão de um cluster. Uma pontuação mais alta de Calinski-Harabasz se relaciona com um modelo com clusters mais bem definidos (SCIKIT-LEARN, 2020).

3.5.3. Índice Davies-Bouldin

O terceiro método utilizado foi o **Índice Davies-Bouldin**. Este índice significa a média de "semelhança" entre os clusters, onde a semelhança é uma medida que compara a distância entre os clusters com o tamanho dos próprios clusters, sendo zero a menor pontuação possível. Valores mais próximos de zero indicam uma partição melhor (SCIKIT-LEARN, 2020).

O índice é definido como a semelhança média entre cada cluster C_i para $i = 1,\ldots, k$ e o mais parecido C_j. No contexto deste índice, a similaridade é definida como uma medida R_{ij}, onde:

s_i = distância média entre cada ponto do cluster i e o centroide desse cluster;

d_{ij} = distância entre os centroides do cluster i e j.

Uma escolha simples de construir R_{ij} de modo que seja não negativo e simétrico é:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

Então, o índice Davies-Bouldin é definido como:

$$DB = \frac{1}{k} \sum_{i=1}^{k} max_{i \neq j} R_{ij}$$

4. Aplicando o Ciclo de Vida de Data Science nos dados da COVID-19

De acordo com Provost (2016), “Data Science é um conjunto de princípios fundamentais que norteiam a extração de conhecimento a partir de dados”. Com a tamanha geração de dados nos últimos anos e o desenvolvimento na capacidade de processamento computacional, a área de Data Science se tornou um paradigma cada vez mais viável na prática e de extrema importância no suporte à tomada de decisão. No cenário da pandemia, os segmentos de Data Science e da Inteligência Artificial proporcionam propostas de ambientes e soluções analíticas que buscam combater o vírus.

Tendo em vista que os resultados gerados por este estudo buscam entender o efeito direto das medidas de contenção em relação aos casos confirmados de COVID-19, deve-se, assim, entender a importância e algumas dificuldades relacionadas à composição dos dados e a forma como estes foram registrados, discutidas a seguir.

Em maioria, os países registram óbito por COVID-19 quando os pacientes se encontram hospitalizados no hospital e tiveram o teste positivo para o vírus. Porém, na Bélgica, o registro também considera óbito por COVID-19 de pessoas que não estavam hospitalizadas, mas que suspeitavam de contaminação pelo vírus. Em países mais pobres que possuem baixa capacidade de realizar testes, essa disparidade no registro deve ser ainda maior indicando um número menor de casos confirmados do que a realidade (THE ECONOMIST, 2020).

A veracidade e a precisão dos dados geram um efeito direto em qualquer estudo em relação aos números de casos e medidas de contenção da COVID-19. Os números de casos vêm aumentando a cada dia, o que gera a necessidade de controle e atualização dos dados de forma em que as análises embasadas neles reflitam a realidade que o mundo está vivendo. Dados desatualizados ou decisões equivocadas podem gerar ações ineficazes em relação ao declínio da proliferação do vírus e a ineficiência destas ações normalmente só são percebidas após pelo menos 14 dias (tempo de incubação do vírus).

Diversos países usaram a tecnologia ao seu favor para combater a COVID-19. A China por exemplo realizou o rastreamento de contágio, utilizou da localização por meio de celulares, utilizou da identificação fácil como controle da temperatura além de um “health code” onde

as pessoas infectadas pelo vírus são identificadas. Com esses dados, pôde-se fazer um mapeamento on-line onde as pessoas podem evitar aglomerações em locais onde há muitas pessoas infectadas (SHAW, 2020).

Já a Coreia do Sul, de acordo com Park (2020), utiliza dados de câmeras de vigilância para obter a localização da população através dos seus celulares e também das transações bancárias. Com esses dados é possível mapear prováveis conexões entre a população que não contraiu o vírus e também mapear casos suspeitos, para que ocorra a testagem desses. Assim, a Coreia do Sul utilizou da análise de dados e do cruzamento destes para descobrir onde poderiam estar localizadas as pessoas possivelmente contaminadas e testar quem teve contato com elas, se destacando entre os demais países.

O caso de sucesso da China e Coreia do Sul atestam o quanto a tecnologia é fundamental no combate ao coronavírus. É importante perceber, assim, que diferentes métodos e uso de Data Science geraram informações precisas, coerentes e que variam de acordo com os dias e de como a pandemia se comporta em cada país. Inteligência Artificial e Data Science são essenciais para ajudar os gestores com informações em tempo real e ajudá-los a definir a melhor medida a ser adotada para diminuir o surto em seu país.

Neste sentido, este capítulo descreve a aplicação do ciclo de vida de Data Science no domínio da Gestão de Epidemias, em particular da pandemia de COVID-19, buscando *insights* e conhecimento útil que possam apoiar a tomada de decisões.

4.1. Entendimento do Problema

Com base em informações sociodemográficas/populacionais de diferentes países e no *Stringency Index* das medidas de contenção, a hipótese de pesquisa deste projeto é de que existe uma relação destes dados com a evolução do número de casos confirmados da COVID-19.

Com a hipótese definida, busca-se os dados que servirão para comprovar ou não o que propõe este trabalho.

4.2. Coleta de Dados

Para embasar a hipótese de pesquisa, buscamos diversos dados relacionados à COVID-19 e aos países e, por isso, tivemos que criar um dataset unificado com os seguintes dados:

- As medidas de contenção, o *Stringency Index* e o número de casos confirmados da doença foram obtidos no "Oxford COVID-19 Government Response Tracker", com informações do dia 21/01/2020 até 15/09/2020 de 185 países e territórios (UNIVERSITY OF OXFORD, 2020);
- O índice de Gini foi obtido através do "The World Bank", assim como o PIB per capita, a taxa de desemprego mensal, a porcentagem da população vivendo em favelas, em cidades, vivendo em aglomerações com mais de um milhão de habitantes e vivendo na maior cidade do país, região ou território (THE WORLD BANK, 2021);
- A porcentagem da população fumante acima dos 15 anos foi retirada do "Our World in Data" (RITCHIE; ROSER, 2019);
- As diferentes densidades populacionais foram obtidas através do documento "Demographia World Urban Areas" (DEMOGRAPHIA, 2021);
- O IDH foi obtido no "UNITED NATIONS DEVELOPMENT PROGRAMME-Human Development Reports" (UNDP, 2020);
- O índice de movimentação foi retirado do "Google" através do Our World in Data (2021);
- A população média existente nos países em 2020 foi retirada do Kaggle - Population by Country - 2020 (PRABHU, 2020).

Abaixo, na Tabela 1, apresenta-se o dicionário de dados do dataset utilizado neste trabalho, ou seja, a listagem das *features* existentes no dataset e sua respectiva descrição dos campos, com seu significado.

Features	**Descrição dos dados**
Nome do País	Nome por extenso de cada país
Código do País	Nome abreviado com 3 letras representando cada país
Nome da Região	(EUA tem abertura estado e UK tem abertura por país)
Código da Região	Código da região

Data	Dados entre 21/01/2020 e 15/09/2020
Fechamento de escolas	0 – Sem fechamento 1 – Fechamento recomendado 2 – Fechamento requerido de apenas algum nível ou categoria (Ex: Somente educação pública ou somente ensino fundamental) 3 – Fechamento total Em branco: Sem dados
Flag Fechamento de escolas	0 – Localizada 1 – Em todo o território Em branco: Sem dados
Fechamento de locais de trabalho	0 – Sem fechamento 1 – Fechamento recomendado 2 – Fechamento requerido de apenas alguns setores ou categorias 3 – Fechamento total das atividades não essenciais Em branco: Sem dados
Flag Fechamento de locais de trabalho	0 – Localizada 1 – Em todo o território Em branco: Sem dados
Cancelamento de eventos em locais públicos	0 – Sem cancelamento 1 – Cancelamento recomendado 2 – Cancelamento requerido Em branco: Sem dados
Flag Cancelamento de eventos em locais públicos	0 – Localizada 1 – Em todo o território Em branco: Sem dados
Restrição de aglomerações em locais privados	0 – Sem restrições 1 – Restrições de grandes aglomerações (>1000 pessoas) 2 – Restrições de médias aglomerações (Entre 101 e 1000 pessoas) 3 – Restrições de pequenas aglomerações (Entre 11 e 100 pessoas) 4 – Restrições até 10 pessoas ou menos Em branco: Sem dados
Flag Restrição de aglomerações em locais privados	0 – Localizada 1 – Em todo o território Em branco: Sem dados
Controle de viagens internacionais	0 – Sem restrições 1 – Triagem no desembarque 2 – Quarentena para indivíduos que chegaram de todas ou alguma região 3 – Banimento de vindas de algumas regiões 4 – Fechamento das fronteiras Em branco: Sem dados
Fechamento do Transporte Público	0 – Sem fechamento 1 – Fechamento recomendado ou redução significativa 2 – Fechamento requerido Em branco: Sem dados
Flag Fechamento do Transporte Público	0 – Localizada 1 – Em todo o território Em branco: Sem dados

Obrigatoriedade de ficar em casa	0 – Sem fechamento 1 – Recomendações para não sair de casa 2 – Saídas liberadas apenas para exercícios, compra de alimentos e atividades essenciais 3 – Obrigatoriedade de ficar em casa com muito poucas exceções (Ex: Apenas uma saída por semana) Em branco: Sem dados
Flag Obrigatoriedade de ficar em casa	0 – Localizada 1 – Em todo o território Em branco: Sem dados
Restrição de movimentações internas entre cidades e regiões	0 – Sem restrições 1 – Recomendação para não haver movimentação interna 2 – Restrição de movimentação interna Em branco: Sem dados
Flag Restrição de movimentações internas entre cidades e regiões	0 – Localizada 1 – Em todo o território Em branco: Sem dados
Ajuda de custos para quem perdeu o emprego ou não pode trabalhar	0 – Sem ajuda de custos 1 – Governo está garantido menos de 50% do salário perdido 2 – Governo está garantindo 50% ou mais do salário perdido Em branco: Sem dados
Flag ajuda de custos para quem perdeu o emprego ou não pode trabalhar	0 – Apenas para trabalhadores da economia formal 1 – Para trabalhadores da economia formal e informal Em branco: Sem dados
Alívio de dívidas e contratos (Ex: Proibição de corte de luz)	0 – Sem alívio 1 – Alívio limitado a um tipo específico de contrato 2 – Amplo alívio Em branco: Sem dados
Estímulos Fiscais (USD)	Valor monetário liberado para o estímulo fiscal
Ajuda Internacional (USD)	Valor monetário dedicado a ajudar outro país
H1: Campanhas públicas de conscientização	0 – Sem campanhas de conscientização 1 – Agentes públicos informando sobre os cuidados necessários 2 – Campanha coordenada de conscientização Em branco: Sem dados
Flag Campanhas públicas de conscientização	0 – Localizado 1- Em todo o território Em branco: Sem dados
H2: Política de testagem	0 – Sem política de testagem 1 – Apenas para aqueles que apresentam sintomas e atingem um critério específico (Ex: Contato com uma pessoa contaminada) 2 – Testagem em todos que apresentam sintomas 3 – Testagem aberta para a população em geral Em branco: Sem dados

H3: Rastreamento de contágio	0 – Sem rastreamento 1 – Rastreamento limitado (Não realizado para todos os casos) 2 – Rastreamento para todos os casos confirmados Em branco: Sem dados
H4: Investimento emergencial em saúde (USD)	Valor monetário investido emergencialmente
H5: Investimento em vacina (USD)	Valor monetário investido no desenvolvimento de uma vacina para COVID19
Casos confirmados	Dados acumulados do número de casos confirmados
Índice de Gini	Instrumento para medir o grau de concentração de renda em determinado grupo apontando a diferença entre os rendimentos dos mais pobres e dos mais ricos.
% da população urbana vivendo na maior cidade do país, região ou território	Percentual da população urbana vivendo na maior cidade do país, região ou território
% da população vivendo em aglomerações com mais de 1M de hab.	Percentual da população vivendo em aglomerações com mais de 1 milhão de habitantes
% da população total vivendo em cidades	Percentual da população total vivendo em cidades
% da população urbana vivendo em favelas	Percentual da população urbana vivendo em favelas
# Cidades com população > 17M	Quantidade de cidades com população maior que 17 milhões
Densidade populacional média das cidades com população > 17M	Densidade média das cidades com população maior que 17 milhões
# Cidades com população entre 13M e 17M	Quantidade de cidades com população entre 13 milhões e 17 milhões
Densidade populacional média das cidades com população entre 13M e 17M	Densidade média das cidades com população entre 13 milhões e 17 milhões
# Cidades com população entre 9M e 13M	Quantidade de cidades com população entre 9 milhões e 13 milhões
Densidade populacional média das cidades com população entre 9M e 13M	Densidade média das cidades com população entre 9 milhões e 13 milhões
# Cidades com população entre 1M e 9M	Quantidade de cidades com população entre 1 milhão e 9 milhões
Densidade populacional média das cidades com população entre 1M e 9M	Densidade média das cidades com população entre 1 milhão e 9 milhões

# Cidades com população entre 500K e 1M	Quantidade de cidades com população entre 500 mil e 1 milhão
Densidade populacional média das cidades com população entre 500K e 1M	Densidade média das cidades com população entre 500 mil e 1 milhão
IDH	LOW – baixo MEDIUM – médio HIGH – alto VERY HIGH – muito alto
IDH discretizado	0 – baixo 1 – médio 2 – alto 3 – muito alto
PIB per Capita	Valor de todos os bens e serviços finais produzidos dentro de uma nação em um determinado ano.
Taxa de Desemprego Mensal	Taxa de desemprego mensal nos países
% da população fumante acima de 15 anos	% da população fumante acima de 15 anos nos países
Índice de movimentação do Google	Dados de localização dos usuários do Google que mostram a movimentação das pessoas em diferentes tipos de locais, como parques, lojas, locais de trabalho e residências.
PIB no 1º e no 2º Quadrimestre	PIB no 1º e no 2º Quadrimestre dos países
Índices de Bolsas de Valores	Fechamento diário do principal índice das bolsas NYSE (NYSE Composite), NASDAQ (NASDAQ Composite), LSE (FTSE 350), TSE (Nikkei), SSE (SSE 180) e B3 (IBOVESPA) de 01/01/2020 até 19/09/2020
Stringency Index	Cálculo da maturidade do conjunto das medidas de restrição, a definição da metodologia se encontra no Github[1].
População	Número médio de habitantes vivendo no país no ano de 2020
Casos/população	Número de casos confirmados dividido pelo número médio de habitantes vivendo no país no ano de 2020

Tabela 1: Dicionário de dados do dataset utilizado

[1] https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/index_methodology.md

4.3. Pré-Processamento

A partir do dataset coletado e integrando todas as fontes de dados, foi realizado o pré-processamento e limpeza dos dados, para que sejam exploradas apenas as variáveis relevantes à pesquisa.

4.3.1. Imputação de Dados

Primeiramente, fizemos algumas imputações de dados conforme listagem abaixo:

- Células vazias na coluna Flag Fechamento de escolas constavam como "sem fechamento" na coluna Fechamento de escolas. Logo, foi imputado o valor "sem fechamento em todo o território" (Flag=1);
- Células vazias na coluna Flag Fechamentos de locais de trabalho constavam como "sem fechamento" na coluna Fechamento de locais de trabalho. Logo, foi imputado o valor "sem fechamento em todo o território" (Flag=1);
- Células vazias na coluna Flag Cancelamento de eventos em locais públicos constavam como "sem cancelamento" na coluna Cancelamento de eventos em locais públicos. Logo, foi imputado o valor "sem cancelamento em todo o território" (Flag=1);
- Células vazias na coluna Flag Restrição de aglomerações em locais privados constavam como "sem restrição" na coluna Restrição de aglomerações em locais privados. Logo, foi imputado o valor "sem restrição em todo o território" (Flag=1).
- Células vazias na coluna Flag Fechamento Do Transporte Público constavam como "sem fechamento" na coluna Fechamento Do Transporte Público. Logo, foi imputado o valor "sem fechamento em todo o território" (Flag=1);
- Células vazias na coluna Flag Obrigatoriedade de ficar em casa constavam como "sem obrigatoriedade" na coluna Obrigatoriedade de ficar em casa. Logo, foi imputado o valor "sem obrigatoriedade em todo o território" (Flag=1);
- Células vazias na coluna Flag Restrição de movimentações internas entre cidades e regiões constavam como "sem movimentações" na coluna Restrição de movimentações internas entre cidades e regiões. Logo, foi imputado o valor "sem movimentações em todo o território" (Flag=1);
- Células vazias na coluna Flag ajuda de custos para quem perdeu o emprego ou não pode trabalhar constavam como "sem ajuda de custos" na coluna Ajuda de custos para quem

perdeu o emprego ou não pode trabalhar. Então, foi imputado como Flag=3 que seria para "nenhuma ajuda tanto para o trabalhador formal quanto informal".

- Células vazias na coluna Flag Campanhas públicas de conscientização constavam como "sem campanhas" na coluna H1: Campanhas públicas de conscientização. Logo, foi imputado o valor "sem campanha em todo o território" (flag=1).
- Nas células das medidas de restrições que estavam vazias (total de N registros), o valor imputado foi calculado interpolando os valores das linhas anteriores e/ou posteriores. Como exemplo, temos a Argentina no dia 02/07/2020 onde na coluna ajuda_de_custos_para_quem_perdeu_emprego_ou_não_pode_trabalhar a célula estava vazia, mas suas células posteriores e anteriores constavam como 1 (Governo está garantido menos de 50% do salário perdido). Logo, neste caso, imputou-se o valor 1.

4.3.2. Limpeza de Dados

Após a imputação de dados, notou-se a necessidade de retirar do dataset alguns países cujos dados estavam demasiadamente incompletos, além de retirar algumas *features* cuja maioria dos países não possuíam dados.

Primeiramente, retiramos cinco países (Pitcaim Island, Solomon Island, Turcomenistão, Taiwan, Vanatu) pois eles não tinham o número de casos confirmados, o que é de suma importância para o estudo. Logo em seguida, foram retiradas três *features* (Índice de Gini, % da população urbana vivendo em favelas e % da população vivendo em aglomerações com mais de 1M de habitantes), visto que mais de sessenta países (33,33% da amostra) não tinham os dados destas *features* preenchidos. Além disso, foram retirados mais quatro países (Anguila, Falkland Island, Kosovo e Montserrat) porque eles não possuíam dados sociodemográficos (futuramente considerados para construir o modelo de clusterização).

4.3.3. Engenharia de Atributos (*Feature Engineering*)

Em seguida, foram acrescentados dois novos atributos ao dataset: um foi substituindo a *feature* de IDH original (a qual estava como *string*) para valores numéricos, pois seria importante usá-la na clusterização e o outro foi acrescentando o *Stringency Index* obtido através da Universidade de OXFORD (2021), que é um índice que reflete o quão restritas as medidas de contenção estão sendo no país. Logo, adicionando o *Stringency Index*, também foi

preciso remover oito países que não tinham este índice disponível na fonte de dados utilizada (British Virgin Islands, Cayman island, Czech Republic, Gibraltar, Kyrgyz Republic, Slovak Republic, Timor-Leste, Turks and Caicos Islands).

A seguir, a Figura 4 representa todo o passo a passo de Pré-Processamento no dataset:

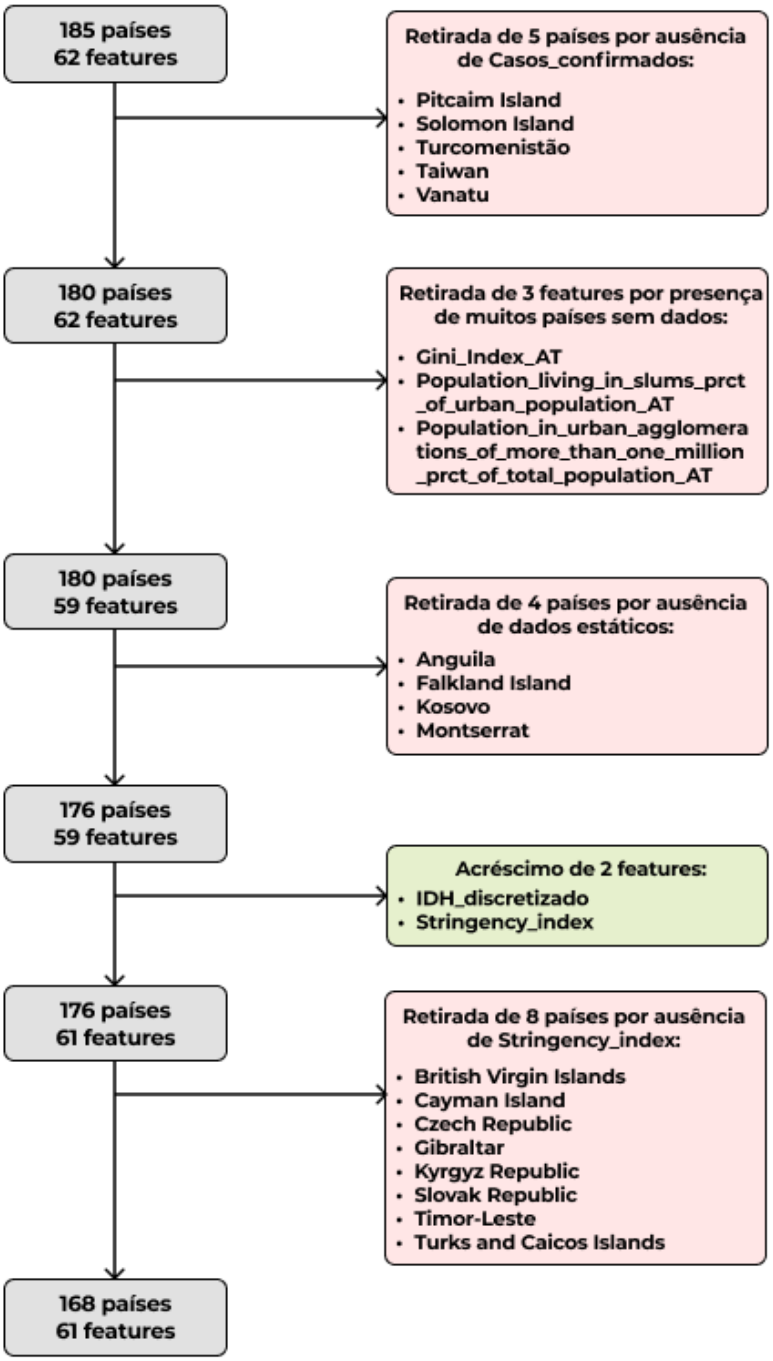

Figura 4: Etapas do Pré-Processamento dos dados

4.3.4. Análise Exploratória

Nessa etapa foi realizada a inspeção dos dados e suas propriedades buscando construir, assim, uma narrativa de acordo com as informações observadas que permitissem uma melhor compreensão dos dados. Para essa análise, utilizou-se a linguagem Python, que passou a se destacar como uma das linguagens mais importantes em Data Science nos últimos anos (SANTANA, 2019). Explorando, assim, as diversas medidas de contenção comumente tomadas por diferentes países e presentes na base de dados do estudo, foi possível compreender quais medidas foram mais impostas e quais foram menos.

A análise exploratória foi realizada em todos os dados do dataset e, abaixo, seguem apresentadas algumas medidas que permitiram inferir *insights* entre elas e compreender melhor os dados.

Plotando os Gráficos 1 e 2, consegue-se visualizar que o fechamento de escolas em relação ao fechamento dos locais de trabalho teve uma abrangência muito maior ao que se refere ao fechamento total (3). Essa grande diferença pode ser oriunda de diversos motivos como a dificuldade das empresas a se adaptar ao *home office* ou a tentativa do governo de tentar manter a economia circulando, pois medidas de restrição como o fechamento total das atividades não essenciais (3) podem gerar desemprego e a falência de diversos empresários. Apesar da dificuldade inicial das escolas de se adaptarem ao EAD, o fechamento total das escolas permite que os alunos ainda continuem estudando diretamente de casa e com maior segurança. As maiores abrangências do fechamento de locais de trabalho, tendo pouca diferença entre as duas, foi o sem fechamento (0) e o fechamento requerido de apenas alguns setores e categorias (2).

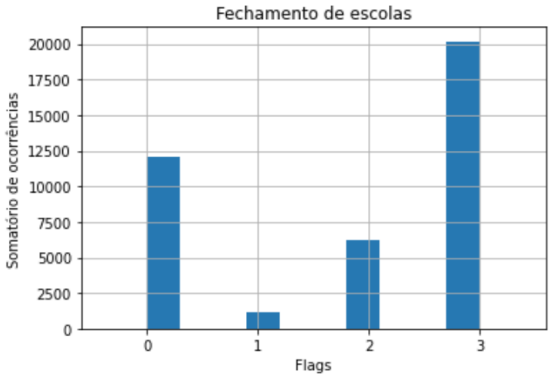

Gráfico 1: Fechamento de escolas

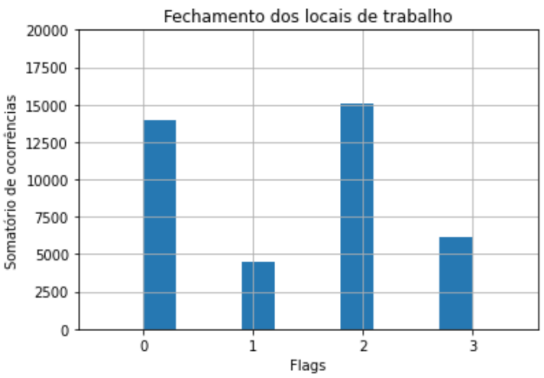

Gráfico 2: Fechamento dos locais de trabalho

A partir do histograma do cancelamento de eventos em locais públicos (Gráfico 3) observa-se uma grande abrangência para o cancelamento requerido (3) e um pouco menos da metade deste para sem cancelamento (0).

Já em relação a restrição de aglomeração em locais privados (Gráfico 4), tem-se alta adesão para sem restrições (0) e para restrições de até 10 pessoas (4) ficando com um pouco menos de adesão a restrição que permite aglomeração entre 11 e 100 pessoas (3). Através dessa análise, pode-se perceber que as medidas de restrições em relação a eventos em locais públicos foram bem mais rígidas do que as restrições de aglomeração em locais privados, pois muitos países não optaram por esse tipo de restrição.

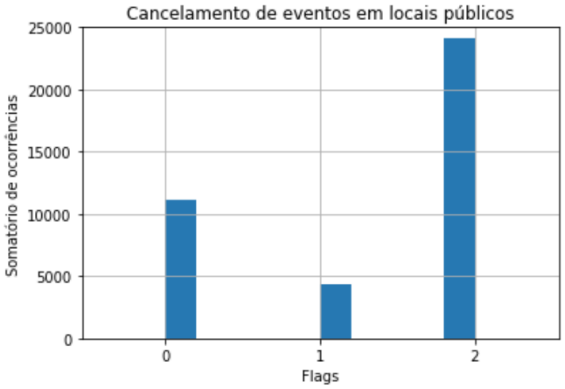

Gráfico 3: Cancelamento dos eventos em locais públicos

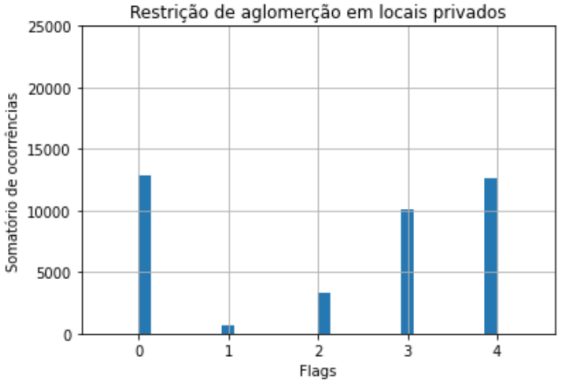

Gráfico 4: Restrição de aglomeração em locais privados

Analisando o histograma do Gráfico 5, pode-se perceber que existem diferentes tipos de abrangência. Poucos países foram adeptos à obrigatoriedade de ficar em casa com poucas exceções (1) e muitos países não foram adeptos (0). Já em relação a recomendação de não sair de casa (2) e saídas liberadas apenas para exercícios, compra de alimentos e atividades essenciais (3), houve uma quantidade relevante de adesão. Em relação ao Gráfico 6, que refere-se às movimentações internas entre cidades e regiões, poucos países recomendaram não haver movimentações internas (1) e uma grande parcela foi adepta a restrição (2). A medida de restrição mais comumente utilizada foi a não restrição de movimentações internas (0).

Assim, analisando-se tanto o Gráfico 5 como o Gráfico 6, conclui-se que em relação a restrição da movimentação de pessoas internamente ocorreram dois extremos: ou ocorria a

restrição total ou não havia nenhum tipo de restrição relacionada à movimentação. Já em relação à obrigatoriedade de ficar em casa, não foi a medida mais comumente utilizada.

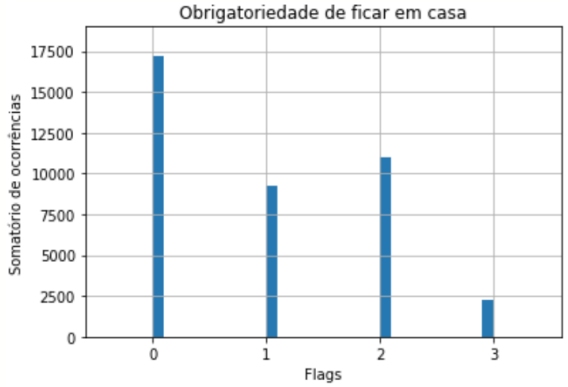

Gráfico 5: Obrigatoriedade de ficar em casa

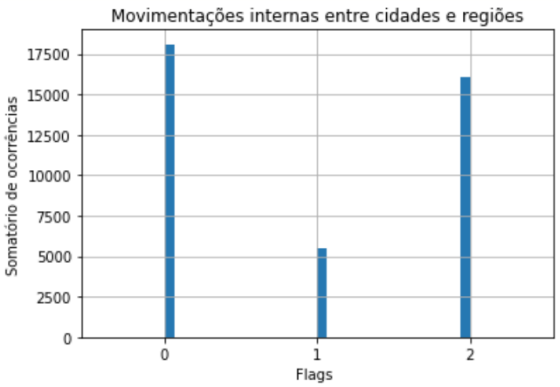

Gráfico 6: Movimentações internas entre cidades e regiões

Em relação à política de testagem do Gráfico 7, temos uma maior abrangência para a testagem de pessoas que apresentam sintomas e atingem um critério específico (ex: contato com uma pessoa contaminada) seguida posteriormente da testagem de todos que apresentam sintomas (2). A política de testagem de toda a população (3) e sem política de testagem (0) ocorreu em uma boa parte dos países por um bom tempo o que se pode especular que a distribuição de testes foi escassa no início de 2020 por ser um vírus pouco conhecido e pela pandemia ter pego todos de surpresa e que posteriormente com um aumento da fabricação dos testes foi possível realizar uma testagem maior da população. Já o Gráfico 8 plota as diferentes medidas de rastreamento de contágio. O rastreamento para todos os casos confirmados (2) foi a medida que os países mais tomaram. Realizando por muito tempo o rastreamento limitado (1) que não ocorria para todos os casos e o não rastreamento (0).

Analisando os Gráficos 7 e 8, especula-se que inicialmente pela falta de testagem não era possível realizar o rastreamento. Posteriormente com o aumento da testagem na população foi possível começar a realizar o rastreamento do contágio da doença.

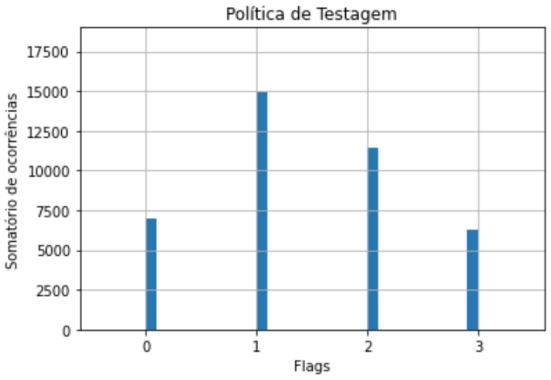

Gráfico 7: Política de testagem

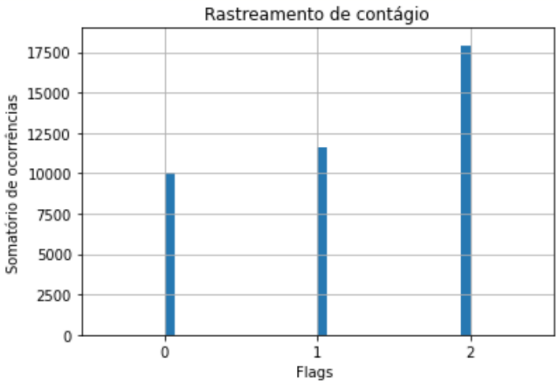

Gráfico 8: Rastreamento de contágio

4.4. Clustering

Para inicializamos a clusterização, selecionamos os dados demográficos e socioeconômicos como *features*, visto que são os dados estáticos do nosso dataset. As 13 *features* escolhidas se encontram abaixo:

- % da população total vivendo em cidades
- # Cidades com população > 17M
- Densidade populacional média das cidades com população > 17M
- # Cidades com população entre 13M e 17M
- Densidade populacional média das cidades com população entre 13M e 17M
- # Cidades com população entre 9M e 13M
- Densidade populacional média das cidades com população entre 9M e 13M
- # Cidades com população entre 1M e 9M
- Densidade populacional média das cidades com população entre 1M e 9M
- # Cidades com população entre 500k e 1M
- Densidade populacional média das cidades com população entre 500k e 1M
- IDH discretizado
- PIB per capita

O processo de cada um dos modelos se encontra destacado nos subitens a seguir.

4.4.1. K-means

A primeira escolha foi realizar a clusterização através do K-means, por ser um algoritmo mais simples de se implementar e por ser utilizado para uma grande variedade de dados, que era o caso do dataset criado. Primeiramente, realizou-se a sua aplicação sem normalizar os dados e, através do Método do Cotovelo (Elbow), o k=5 foi escolhido como ponto de partida conforme o Gráfico 9.

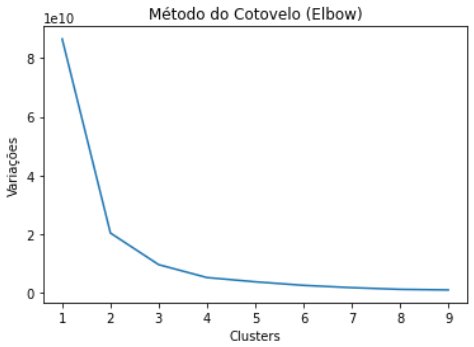

Gráfico 9: Resultado do Método do Cotovelo (Elbow)

Em seguida, variou-se a escolha do k através da comparação entre clusters de cada k, em que do 4 pro 5 foi possível separar melhor os países, mas que o mesmo não aconteceu quando variou de 5 para 6, já que apenas Bermuda e Luxemburgo se separaram para formar um único cluster, conforme o Gráfico 10 a seguir.

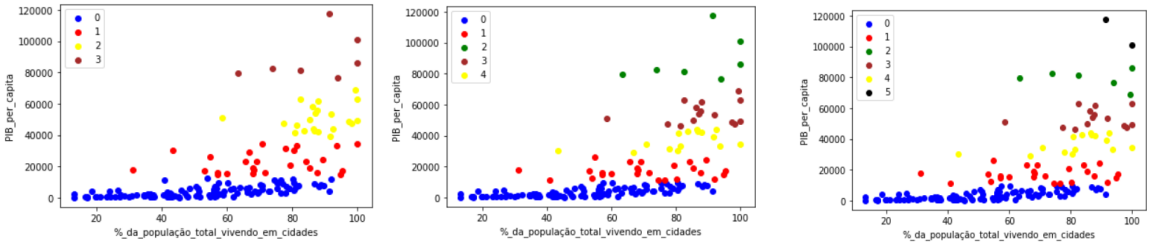

Gráfico 10: Aplicação do K-means com dados não normalizados para k=4, k=5 e k=6 respectivamente

Durante as análises, foi notório que o PIB per capita estava influenciando demais o modelo por ter números extremamente maiores que os existentes nas demais *features*, ou seja, ele estava enviesando a análise por estar dando um peso maior pro PIB per capita já que ele variava de 0 a 120000 enquanto as demais variavam de 0 a 100. Então, foi necessário optar-se por normalizar os dados e, então, obter-se, novamente pelo Método do Cotovelo (Elbow), 6 clusters. Também foi variado com k=5 e k=7, mas não obteve-se uma grande variação de países em cada diferente cluster. A partir disso, foi realizado o pós-processamento para

entender se esse algoritmo seria o mais adequado e se alguma outra configuração de k faria mais sentido. No Gráfico 11, segue a nova clusterização encontrada após a normalização dos dados.

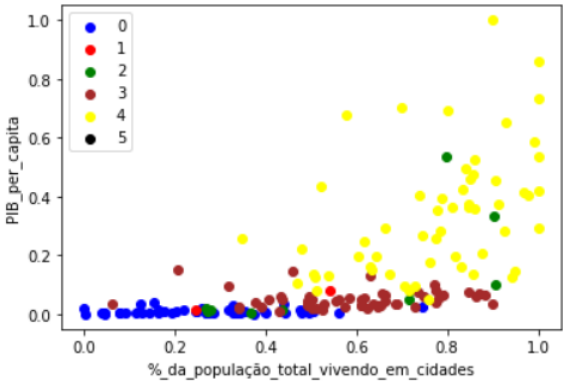

Gráfico 11: Aplicação do K-means com dados normalizados para k=6

É importante ressaltar que os gráficos estão utilizando sempre as mesmas duas *features* (PIB_per_capita e %_da_população_total_vivendo_em_cidades) para facilitar a compreensão da variação dos resultados, mas as outras análises se encontram no repositório público do Github[2]. Além disso, os outros 3 modelos a seguir já prosseguiram com a lógica dos dados normalizados.

4.4.2. Agglomerative Clustering

Como segunda escolha de algoritmo, optamos pelo Agglomerative Clustering ou Agrupamento Aglomerativo. Porém, o dendograma gerado (Gráfico 12) ficou difícil de ser compreendido devido ao grande número de países no dataset. Apesar da sua fácil implementação, desvantagens como sensibilidade a *outliers* e por não ser um algoritmo ideal para grandes números de dados fez com que esse algoritmo não fosse escolhido.

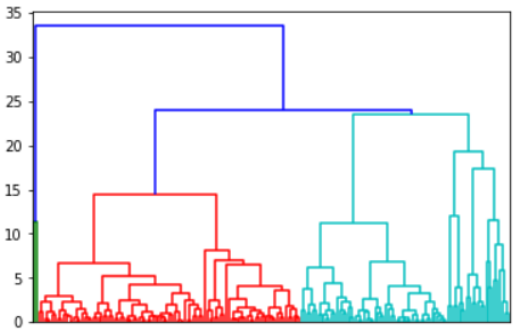

Gráfico 12: Aplicação do Agglomerative Clustering com dados normalizados

[2] https://github.com/ingridfrf/covid19-clustering-analysis

4.4.3. DBSCAN

Buscando uma nova forma de realizar a Clusterização dos dados já normalizados sem ser pela hierarquia, foi escolhido utilizar o algoritmo DBSCAN, que realiza sua clusterização através da densidade, permitindo que os clusters encontrados tenham qualquer formato e não apenas um formato convexo como o K-means. Ao realizarmos diversos testes com diferentes parâmetros de entrada neste algoritmo, não foi possível obter um número de *outliers* pequeno para serem retirados da amostra, ou seja, eram sempre encontrados muitos *outliers*, o que tornaria o estudo muito restritivo. Então, devido a sua sensibilidade em relação aos parâmetros de entrada para realizar o agrupamento e o fato dele não agrupar dados com diferença de densidade, foi optado por não seguir com esse algoritmo.

4.4.4. HDBSCAN

Por ser um algoritmo baseado em densidade, que é indiferente à forma de cada cluster, e hierarquia, não sendo necessária a especificação do número de clusters, o HDBSCAN foi visto como um ótimo algoritmo para o estudo.

Pelo método HDBSCAN, após diversos testes, chegou-se em uma clusterização de 4 clusters, pois foi alcançado um número mínimo de *outliers* de 14 países. Para compreender melhor o que estava acontecendo, o Gráfico 13 foi plotado com os clusters antes e depois da retirada dos 14 países *outliers*, estes em preto (-1) na primeira metade do gráfico. Os países removidos, conforme o resultado do método, foram: *Bahrain, Bermuda, Switzerland, Estonia, Greece, Ireland, Iceland, Luxembourg, Macao, Norway, Portugal, Qatar, Saudi Arabia* e *Slovenia.*

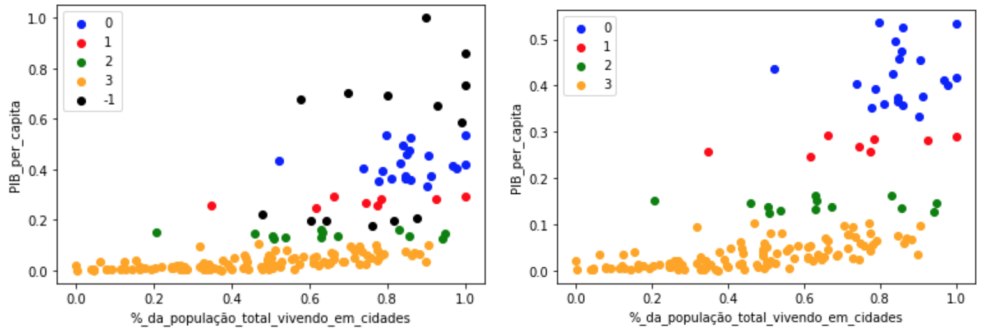

Gráfico 13: Aplicação do HDBSCAN com dados normalizados com *outlier* e sem *outlier* respectivamente

4.5. Pós Processamento

Após executar as 4 diferentes técnicas de Clusterização e suas respectivas análises empíricas com gráficos e métodos, o processo de Avaliação foi iniciado, de fato, para verificar qual foi o melhor modelo para prosseguir com as análises, de acordo com as métricas de avaliação apresentadas na Seção 3.5. Diversos cenários foram avaliados para as técnicas K-means e HDBSCAN (visto que as demais técnicas não seriam as ideais conforme explicado nas Seções 4.4.2 e 4.4.3), assim, foram registrados os valores calculados de cada métrica sobre o modelo resultante de cada cenário, conforme ilustrado na Tabela 2.

Cenário	Parâmetros de Entrada	Número de Clusters	Número de *Outliers*	Índice Silhouette	Índice Calinski Harabasz	Índice Davies Bouldin
KMEANS-1	n_clusters=2	2	-	0.379762	89.008931	1.121786
KMEANS-2	n_clusters=3	3	-	0.402671	85.161409	1.051313
KMEANS-3	n_clusters=4	4	-	0.299267	74.475242	1.214484
KMEANS-4	n_clusters=5	5	-	0.309365	71.481898	1.103795
KMEANS-5	n_clusters=6	6	-	0.287977	64.444743	1.294911
KMEANS-6	n_clusters=7	7	-	0.331700	62.491741	0.963505
KMEANS-7	n_clusters=8	8	-	0.289534	61.238270	1.135556
KMEANS-8	n_clusters=9	9	-	0.290436	60.876774	1.148809
KMEANS-9	n_clusters=10	10	-	0.281863	63.536577	1.076828
KMEANS-10	n_clusters=15	15	-	0.294993	55.997330	0.899254
KMEANS-11	n_clusters=20	20	-	0.278831	58.978385	0.773998
KMEANS-12	n_clusters=25	25	-	0.291834	62.793261	0.775739
HDBSCAN-1	min_cluster_size=9, min_samples=3, epsilon=0.3	6	49	0.187725	34.419528	1.789853
HDBSCAN-2	min_cluster_size=9, min_samples=3, epsilon=0.4	6	49	0.265090	43.047606	1.973494
HDBSCAN-3	min_cluster_size=8, min_samples=3, epsilon=0.3	4	14	0.202550	35.144002	1.829286
HDBSCAN-4	min_cluster_size=8, min_samples=3, epsilon=0.4	4	14	0.425893	23.903592	1.461430
HDBSCAN-5	min_cluster_size=7, min_samples=3, epsilon=0.3	4	14	0.202550	35.144002	1.829286
HDBSCAN-6	min_cluster_size=7, min_samples=3, epsilon=0.4	4	14	0.425893	23.903592	1.461430
HDBSCAN-7	min_cluster_size=6, min_samples=3, epsilon=0.3	12	32	0.202550	35.144002	1.829286
HDBSCAN-8	min_cluster_size=6, min_samples=3, epsilon=0.4	12	32	0.425893	23.903592	1.461430
HDBSCAN-9	min_cluster_size=5, min_samples=3, epsilon=0.3	17	23	0.202550	35.144002	1.829286

HDBSCAN-10	min_cluster_size=5, min_samples=3, epsilon=0.4	17	23	0.425893	23.903592	1.461430
HDBSCAN-11	min_cluster_size=9, min_samples=4, epsilon=0.3	3	25	0.182076	32.272445	1.824185
HDBSCAN-12	min_cluster_size=9, min_samples=4, epsilon=0.4	3	25	0.256685	40.922941	2.048071
HDBSCAN-13	min_cluster_size=8, min_samples=4, epsilon=0.3	3	25	0.182076	32.272445	1.824185
HDBSCAN-14	min_cluster_size=8, min_samples=4, epsilon=0.4	3	25	0.256685	40.922941	2.048071
HDBSCAN-15	min_cluster_size=7, min_samples=4, epsilon=0.3	4	18	0.182076	32.272445	1.824185
HDBSCAN-16	min_cluster_size=7, min_samples=4, epsilon=0.4	4	18	0.256685	40.922941	2.048071
HDBSCAN-17	min_cluster_size=6, min_samples=4, epsilon=0.3	11	45	0.182076	32.272445	1.824185
HDBSCAN-18	min_cluster_size=6, min_samples=4, epsilon=0.4	11	45	0.256685	40.922941	2.048071
HDBSCAN-19	min_cluster_size=5, min_samples=4, epsilon=0.3	13	42	0.191173	30.187271	1.835632
HDBSCAN-20	min_cluster_size=5, min_samples=4, epsilon=0.4	13	42	0.387098	22.635512	1.542123
HDBSCAN-21	min_cluster_size=9, min_samples=5, epsilon=0.4	3	27	0.229496	35.716642	2.200710
HDBSCAN-22	min_cluster_size=9, min_samples=5, epsilon=0.3	3	27	0.166993	27.884805	1.925227
HDBSCAN-23	min_cluster_size=8, min_samples=5, epsilon =0.4	3	27	0.229496	35.716642	2.200710
HDBSCAN-24	min_cluster_size=8, min_samples=5, epsilon =0.3	3	27	0.166993	27.884805	1.925227
HDBSCAN-25	min_cluster_size=7, min_samples=5, epsilon= 0.4	4	20	0.229496	35.716642	2.200710
HDBSCAN-26	min_cluster_size=7, min_samples=5, epsilon =0.3	4	20	0.166993	27.884805	1.925227
HDBSCAN-27	min_cluster_size=6, min_samples=5, epsilon=0.4	10	53	0.229496	35.716642	2.200710
HDBSCAN-28	min_cluster_size=6, min_samples=5, epsilon=0.3	10	53	0.166993	27.884805	1.925227
HDBSCAN-29	min_cluster_size=5, min_samples=5, epsilon=0.4	11	48	0.229496	35.716642	2.200710
HDBSCAN-30	min_cluster_size=5, min_samples=5, epsilon=0.3	11	48	0.166993	27.884805	1.925227

Tabela 2: Processo de experimentação (K-means e HDBSCAN)

No caso dos cenários da técnica K-means, o único parâmetro passível de variação na chamada da função é o número de clusters, que foi variado entre k=2 e k=25. Conforme destacado na Tabela 2, o maior valor do índice **Silhouette** para os cenários da técnica K-means foi **0,4026 (k=3)**. No índice **Calinski-Harabasz,** o maior valor encontrado para os cenários da técnica

K-means foi **89,0089** (k=2). Já no índice **Davies-Bouldin,** o menor valor da técnica K-means foi **0,7739** (k=20). Como não teve nenhum cenário do K-means em que dois ou mais índices tiveram o melhor valor, foi optado por não seguir com essa técnica.

Com relação à técnica HDBSCAN, os parâmetros da função que foram experimentados entre os diferentes cenários foram: tamanho mínimo do cluster, mínimo de amostras e epsilon. Dentre os cenários experimentados desta técnica, o maior valor do índice **Silhouette** foi **0,4259**. No índice **Calinski-Harabasz**, obteve-se o maior valor como **43,0476**. Já no índice **Davies-Bouldin,** o menor valor foi **1,4614**. Dessa forma, têm-se alguns cenários em que pelo menos dois índices obtiveram o melhor valor, mostrando que são os melhores cenários para a técnica HDBSCAN, considerando tanto o índice Silhouette quanto o índice Davies-Bouldin. Portanto, foi decidido seguir com os parâmetros de entrada descritos no cenário H-6 (min_cluster_size=7, min_samples=3, epsilon=0.4, número de clusters=4, número de *outliers*=14), pois dentre os cenários destacados, ele é o que apresentou menor número de *outliers* e, tendo em vista que o algoritmo HDBSCAN encontra um agrupamento removendo os *outliers* identificados, não foi desejado reduzir ainda mais a amostra de países.

5. Análise do melhor agrupamento obtido através da técnica HDBSCAN

Após a decisão de seguir com quatro clusters oriundos do HDBSCAN e com dados normalizados, há indícios que a hipótese inicial de que os dados demográficos e socioeconômicos estão correlacionados à evolução da doença. A maioria dos países apresentaram um comportamento similar na curva de casos da doença, porém, algumas exceções foram encontradas e, assim, foi necessário procurar motivos que justificassem essas curvas tão discrepantes se comparadas aos demais países de seus respectivos clusters.

Abaixo, segue a composição de países por cluster do agrupamento que obteve a melhor avaliação durante a experimentação. O Gráfico 14 ilustra os clusters em função do percentual da população vivendo em cidades (eixo X) e o PIB per capita do país (eixo Y), enquanto a Figura 5 ilustra a disposição georreferenciada dos países de cada cluster. As cores do mapa se referem ao cluster de mesma cor do Gráfico 14.

- **Cluster 0 (azul) com 21 países:**

Andorra, United Arab Emirates, Australia, Austria, Belgium, Canada, Germany, Denmark, Finland, France, United Kingdom, Greenland, Hong Kong, Israel, Japan, Netherlands, New Zealand, Singapore, San Marino, Sweden e United States;

- **Cluster 1 (vermelho) com 8 países:**

Aruba, Brunei, Cyprus, Spain, Italy, South Korea, Kuwait e Puerto Rico;

- **Cluster 2 (verde) com 13 países:**

Barbados, Chile, Guam, Croatia, Hungary, Lithuania, Latvia, Oman, Panama, Poland, Seychelles, Trinidad and Tobago e Uruguay;

- **Cluster 3 (amarelo) com 112 países:**

Afghanistan, Angola, Albania, Argentina, Azerbaijan, Burundi, Benin, Burkina Faso, Bangladesh, Bulgaria, Bosnia and Herzegovina, Belarus, Belize, Bolivia, Brazil, Bhutan, Botswana, Central African Republic, China, Cote d'Ivoire, Cameroon, Democratic Republic of Congo, Congo, Colombia, Cape Verde, Costa Rica, Cuba, Djibouti, Dominica, Dominican Republic, Algeria, Ecuador, Egypt, Eritrea, Ethiopia, Fiji, Gabon, Georgia, Ghana, Guinea,

Gambia, Guatemala, Guyana, Honduras, Haiti, Indonesia, India, Iran, Iraq, Jamaica, Jordan, Kazakhstan, Kenya, Cambodia, Laos, Lebanon, Liberia, Libya, Sri Lanka, Lesotho, Morocco, Moldova, Madagascar, Mexico, Mali, Myanmar, Mongolia, Mozambique, Mauritania, Mauritius, Malawi, Malaysia, Namibia, Niger, Nigeria, Nicaragua, Nepal, Pakistan, Peru, Philippines, Papua New Guinea, Paraguay, Palestine, Romania, Russia, Rwanda, Sudan, Senegal, Sierra Leone, El Salvador, Somalia, Serbia, South Sudan, Suriname, Eswatini, Syria, Chad, Togo, Thailand, Tajikistan, Tunisia, Turkey, Tanzania, Uganda, Ukraine, Uzbekistan, Venezuela, Vietnam, Yemen, South Africa, Zambia e Zimbabwe.

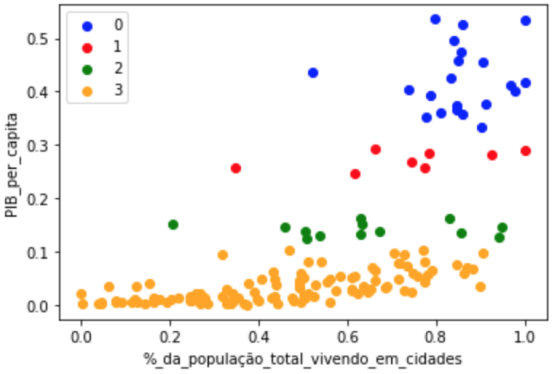

Gráfico 14: Cluster de melhor avaliação (HDBSCAN)

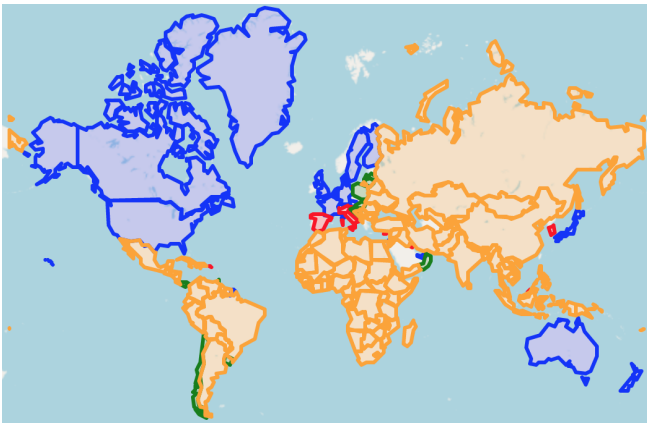

Figura 5: Distribuição dos países por cluster (HDBSCAN)

5.1. Evolução dos números de casos

Sabe-se que, por mais que as medidas de restrições tenham sido aplicadas, nem toda população aderiu. Com isso, foi preciso analisar os índices de mobilidade dos países através do Our World in Data (Our World in Data, 2021) e, com isso, verificar que a mobilidade não

foi tão reduzida nos países cuja trajetória da evolução da doença foi destoante dos demais países do cluster, conforme os gráficos apresentados ao longo desta Seção.

- **Cluster 0 - Estados Unidos**

O Gráfico 15 ilustra a curva do número de casos confirmados de COVID-19 nos países do cluster 0. A despeito da semelhança encontrada pelo algoritmo de Clustering entre os países deste cluster considerando suas características sociodemográficas, observa-se que a curva de casos dos Estados Unidos destoa negativamente dos demais países. Neste sentido, foi analisado, então, os índices de mobilidade (Gráfico 16) e a evolução do *Stringency Index* (Gráficos 17 e 18) dele neste mesmo período buscando *insights* de indícios que possam justificar a diferença na evolução da epidemia.

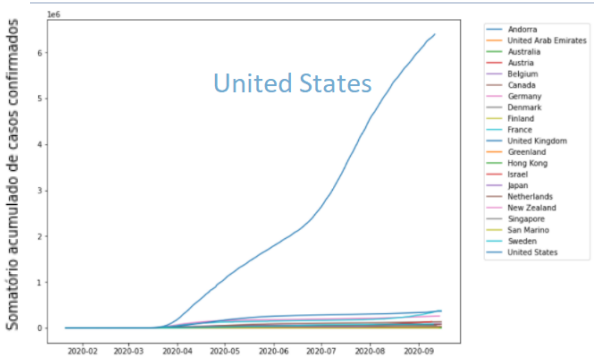

Gráfico 15: Evolução do número de casos no cluster 0

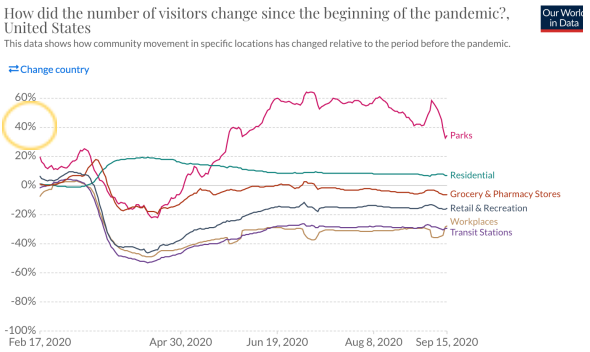

Gráfico 16: Mobilidade dos EUA desde o início da pandemia (Our World in Data, 2021)

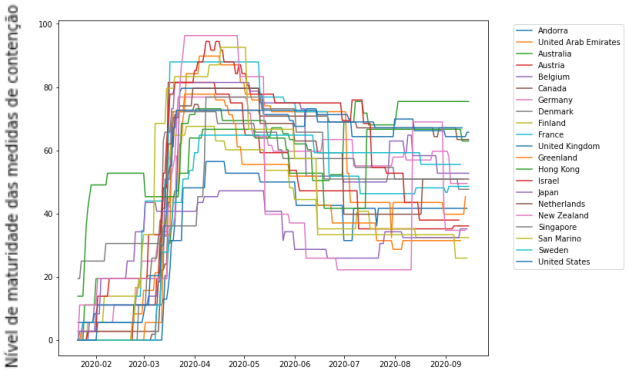

Gráfico 17: Evolução do stringency index no cluster 0

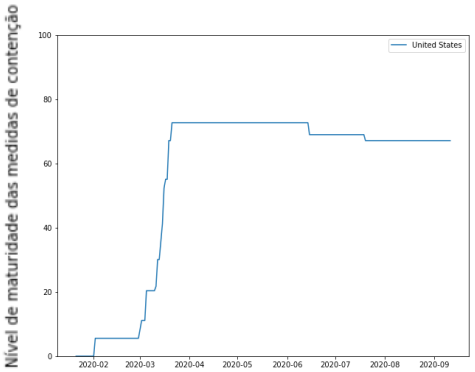

Gráfico 18: Evolução do stringency index dos EUA

De acordo com o Gráfico 16 de mobilidade, foi possível analisar que alguns índices de mobilidade diminuíram mais de 40% entre março e abril e, após esse período, voltaram a aumentar consideravelmente principalmente nos parques, o que nos mostra que poucas

pessoas aderiram ao isolamento total. Já as medidas de restrições, como podem ser vistas nos Gráficos 17 e 18, se mantiveram constantes ou com poucas mudanças a partir do mês de março. Ao analisar a discrepância da trajetória da evolução da doença nos EUA com relação aos demais países do seu cluster (Gráfico 15), pode-se citar o aumento do número de testagem a partir de março de 2020, já que antes desse período o país passou por escassez de testes(Gráfico 19). No início de março de 2020,Thomas Tsai, pesquisador de políticas de saúde na Universidade Harvard afirmou que "Grande parte da culpa pela situação atual se deve ao atraso nos testes nos EUA. Estávamos vendo a pandemia se desenrolar, sem capacidade de testar e identificar casos. Isso resultou na disseminação maciça de covid-19 nos EUA" (BBC, 2020). O problema começou quando o CDC (*Centers for Disease Control and Prevention*) decidiu que seria o único a fabricar os testes, porém, alguns testes estavam com defeito e tiveram que ser substituídos, o que resultou em muitos estados sem acesso a esses testes nos primeiros meses do ano. Com o tempo, a situação mudou e, em meados de março, o número de testes realizados diariamente no país estava crescendo exponencialmente, como também cresceram os números de casos positivos para a doença (BBC, 2020).

Gráfico 19: Política de testagem nos EUA

- **Cluster 1 - Itália e Espanha**

O Gráfico 20 ilustra a curva do número de casos confirmados de COVID-19 nos países do cluster 1. A despeito da semelhança encontrada pelo algoritmo de Clustering entre os países deste cluster considerando suas características sociodemográficas, observa-se que as curvas de casos da Espanha e da Itália destoam negativamente dos demais países. Neste sentido, foi analisado, então, a evolução do *Stringency Index* (Gráficos 21 e 22) e os índices de mobilidade (Gráficos 23 e 24) nestes dois países no mesmo período buscando *insights* de indícios que possam justificar a diferença na evolução da epidemia nesses 2 países.

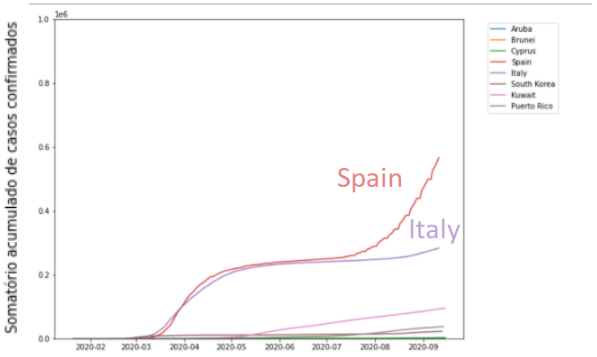

Gráfico 20: Evolução do número de casos no cluster 1

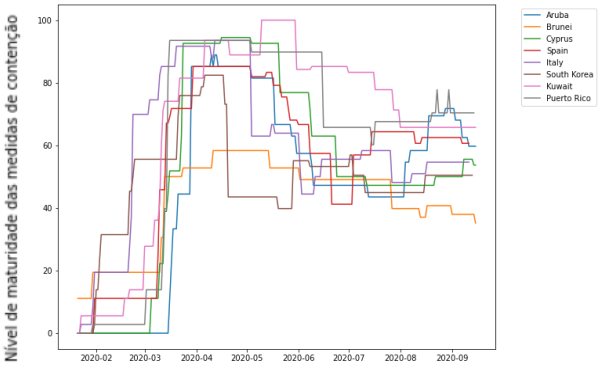

Gráfico 21: Evolução do stringency index no cluster 1

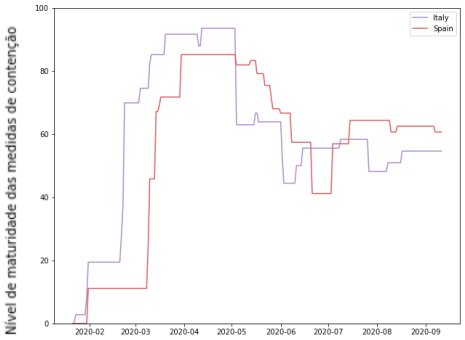

Gráfico 22: Evolução do stringency index na Itália e na Espanha

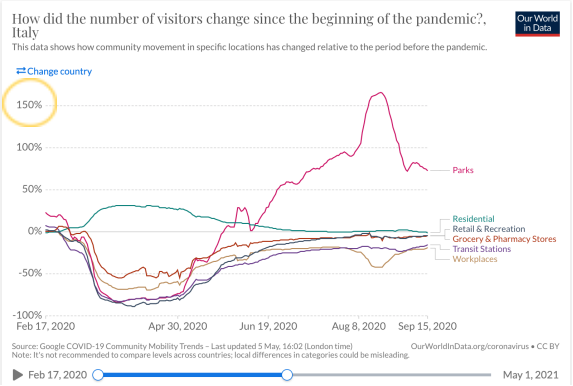

Gráfico 23: Mobilidade da Itália desde o início da pandemia (Our World in Data, 2021)

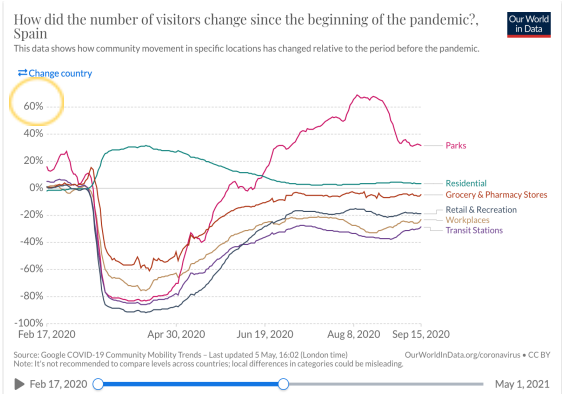

Gráfico 24: Mobilidade da Espanha desde o início da pandemia (Our World in Data, 2021)

Em março de 2020, a Itália era o segundo país com o maior número de casos confirmados de COVID-19, e especialistas buscam entender porque a doença chegou tão cedo no país. Uma das possibilidades encontradas é o intenso tráfego aéreo entre a Itália e a China, país de onde surgiu o primeiro caso, e as diferentes políticas adotadas dependendo da região, já que a administração italiana não é centralizada. Com a trajetória de evolução da doença cada vez

crescendo mais(Gráfico 20), o país tomou medidas de restrições realmente severas (Gráficos 21 e 22) onde os cidadãos tinham que comprovar a importância do seu trabalho para continuar a exercer a atividade ou o estado de saúde e outras razões que justificassem a necessidade de viajar para fora da área de residência, por exemplo. Através dessas restrições, pode-se ver no Gráfico 23 de mobilidade que o país atingiu mais de 80% de diminuição em alguns índices, o que fez com que a evolução dos casos começassem a diminuir em meados de maio (BRAUN, 2020).

Já na Espanha, no dia 13 de março de 2020 já haviam sido detectados casos de COVID-19 em todas as cinquenta províncias do país, o que pode estar relacionado diretamente aos milhões de euros destinados a compra de testes pelo governo nessa época. A partir daí, começou uma quarentena apertada em todo o território da Espanha, onde a população podia apenas deslocar-se para ir trabalhar (os que não podiam trabalhar de *home office*) ou para abastecer-se de bens essenciais, o que pode ser percebido pela queda de mobilidade de mais de 80% em alguns índices, assim como na Itália, estabilizando a evolução dos casos em meados de abril até julho (Gráfico 20). Com a diminuição das medidas de restrição (Gráficos 21 e 22) e, consequentemente, o aumento da mobilidade, como podemos ver no Gráfico 24, a evolução dos casos na Espanha voltaram a subir brutalmente em meados de julho, diferente do que aconteceu na Itália. Estudos apontam que além do maior acesso aos testes(Gráfico 26), a Espanha não estava preparada para retornar às atividades (NEVES, 2020).

Como a Espanha e a Itália foram países que tiveram picos da doença logo no início, estudos mostram que o número real de casos confirmados é ainda maior do que o exposto nos dados, visto que ainda não tinham quantidades suficientes de testes para toda a população, testando assim, apenas pessoas com sintomas mais graves, como pode ser visto nos Gráficos 25 e 26. Outro estudo interessante está relacionado aos primeiros casos terem surgido em cidades como Madrid, Catalunha e no norte da Itália, cidades mais densamente populosas, o que explica o aumento drástico logo no início de março (GUELL; MEDINA, 2020).

Gráfico 25: Política de testagem na Itália

Gráfico 26: Política de testagem na Espanha

- **Cluster 2 - Chile**

O Gráfico 27 ilustra a curva do número de casos confirmados de COVID-19 nos países do cluster 2. A despeito da semelhança encontrada pelo algoritmo de Clustering entre os países deste cluster considerando suas características sociodemográficas, observa-se que a curva do Chile destoa negativamente dos demais países. Neste sentido, analisamos então os índices de mobilidade (Gráfico 28) e a evolução do *Stringency Index* (Gráficos 29 e 30) dele neste mesmo período buscando *insights* de indícios que possam justificar a diferença na evolução da epidemia no Chile.

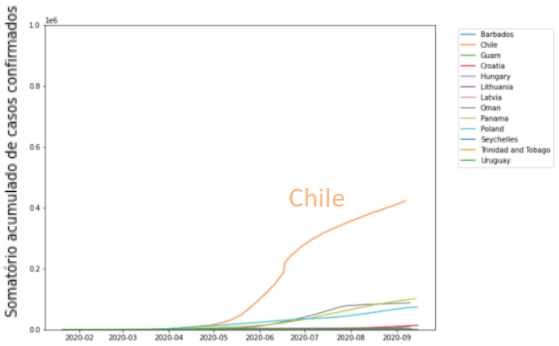

Gráfico 27: Evolução do número de casos no cluster 2

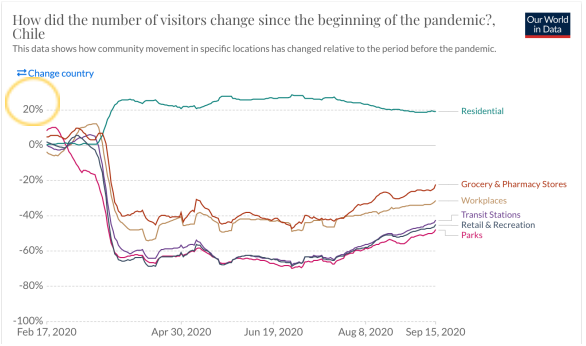

Gráfico 28: Mobilidade do Chile desde o início da pandemia (Our World in Data, 2021)

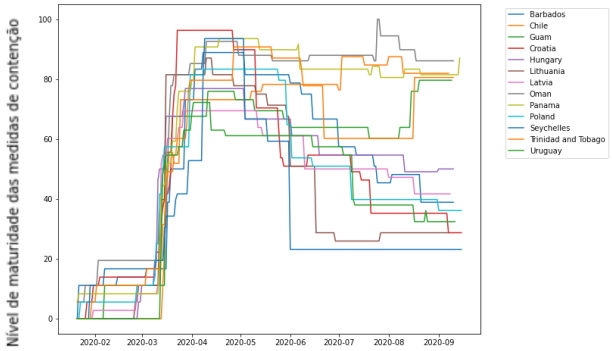

Gráfico 29: Evolução do stringency index no cluster 2

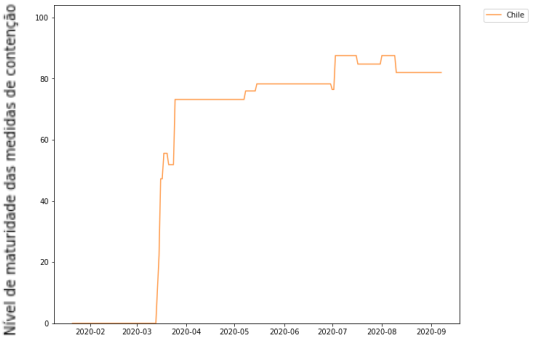

Gráfico 30: Evolução do stringency index no Chile

No Chile, a partir do primeiro caso, no começo de março de 2020, o país estabeleceu uma estratégia de “quarentenas dinâmicas”, somente nas regiões onde haviam pessoas infectadas. Com isso, registrou uma evolução lenta da quantidade de contágios (Gráfico 27). Porém, em maio, o presidente do país decretou uma “nova normalidade” reabrindo comércios e shoppings centers. Esse retorno, nada seguro, fez com que a partir das primeiras semanas de maio, a curva de contágios tivesse um salto. Como pode-se perceber no Gráfico 31, a política de testagem começou no final de março para todos que apresentavam sintomas e se manteve assim até o fim do período analisado. O Gráfico 28 de mobilidade, no entanto, nos mostra que apesar da mobilidade ter decaído bastante a partir de março, ela se manteve entre altos e baixos durante todo o restante do período analisado, o que pode ser explicado pela estratégia de “quarentenas dinâmicas" (FARINELLI, 2020).

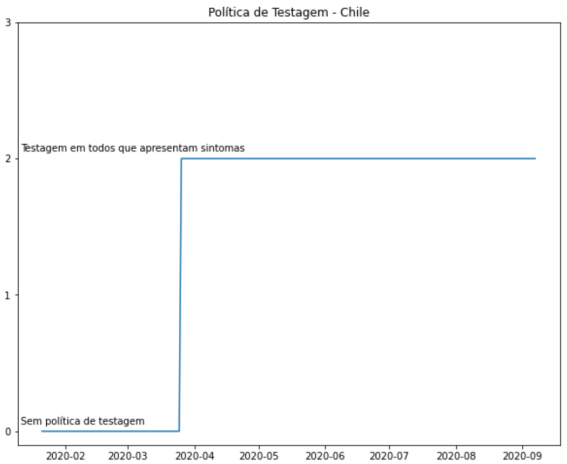

Gráfico 31: Política de testagem no Chile

- **Cluster 3 - Brasil e Índia**

O Gráfico 32 ilustra a curva do número de casos confirmados de COVID-19 nos países do cluster 3. A despeito da semelhança encontrada pelo algoritmo de Clustering entre os países deste cluster considerando suas características sociodemográficas, observa-se que a curva de casos do Brasil e da Índia destoam negativamente dos demais países. Neste sentido, foi analisado, então, a evolução do *Stringency Index* (Gráficos 33 e 34) e os índices de mobilidade nestes dois países no mesmo período (Gráficos 35 e 36), buscando *insights* de indícios que possam justificar a diferença na evolução da epidemia tanto no Brasil como na Índia.

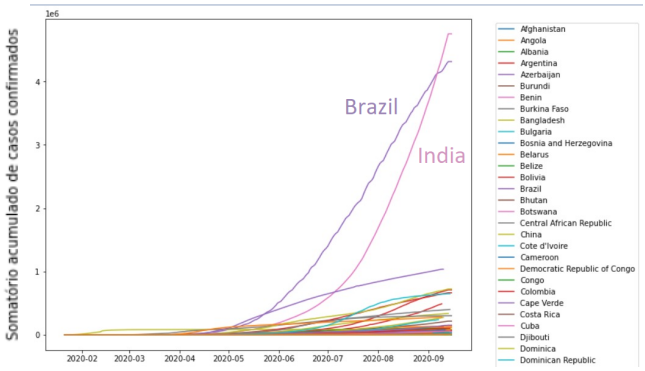

Gráfico 32: Evolução do número de casos no cluster 3

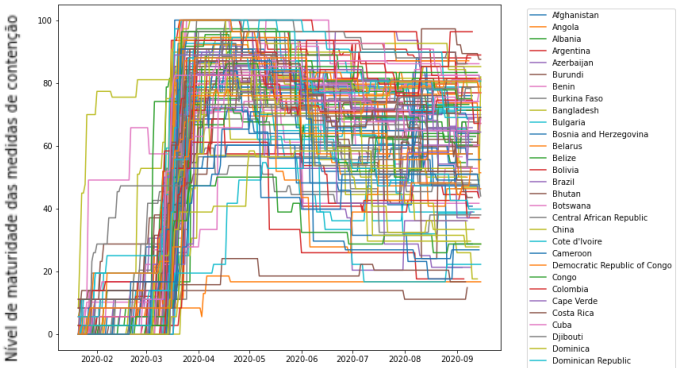

Gráfico 33: Evolução do stringency index no cluster 3

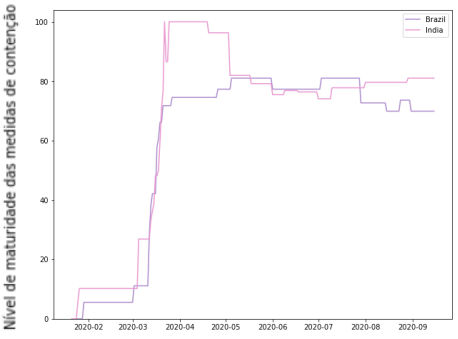

Gráfico 34: Evolução do stringency index no Brasil e na Índia

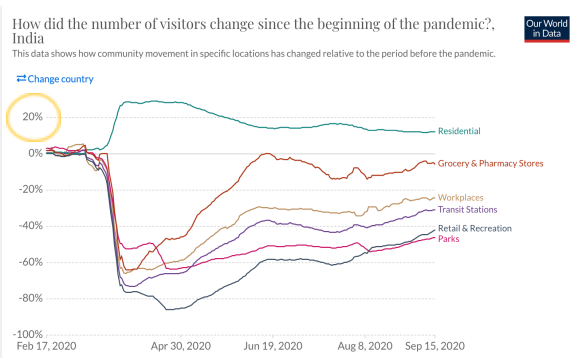

Gráfico 35: Mobilidade da Índia desde o início da pandemia (Our World in Data, 2021)

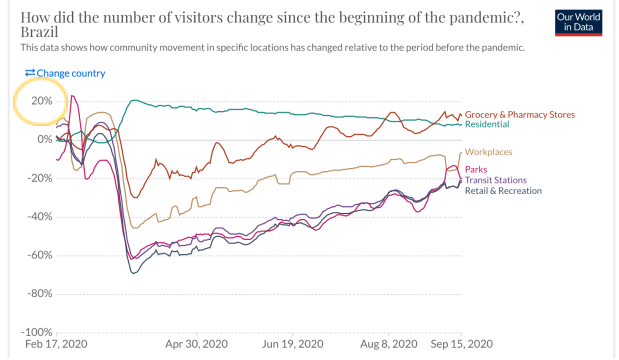

Gráfico 36: Mobilidade do Brasil desde o início da pandemia (Our World in Data, 2021)

A Índia conseguiu conter, em um primeiro momento, a expansão da COVID-19 com um rígido confinamento e altas medidas de restrições (Gráfico 33 e 34), que durou 70 dias . Mas a partir de julho de 2020, cerca de um mês depois do início da reabertura, o país registrou uma aceleração no número de casos que pode ser vista na trajetória de evolução da doença, que chegou a ultrapassar o Brasil em setembro de 2020 (Gráfico 32). Essa reabertura somada a uma maior testagem da população, como pode ser visto no Gráfico 37, do segundo país mais populoso do mundo, pode ser a causa de tamanha evolução. O Gráfico 35 de mobilidade mostra exatamente essa queda brusca inicial e nos meses seguintes já é possível observar um aumento considerável da mobilidade, chegando a menos de 50% (G1, 2020).

Já o Brasil, em março de 2020 após o surgimento dos primeiros casos, conseguiu reduzir bastante a mobilidade, porém, por pouco tempo, já que logo depois a curva subiu como exposto no Gráfico 36. Com diversos brasileiros infringindo as restrições impostas (Gráficos 33 e 34) e participando de eventos clandestinos, não teve como a trajetória de evolução da doença no Brasil não aumentar ao longo do tempo. No Gráfico 32, correspondente à trajetória de evolução dos números de casos, pode-se perceber poucos casos de fevereiro até maio que são explicados pela falta de testagem da população, como pode ser visto no Gráfico 38.

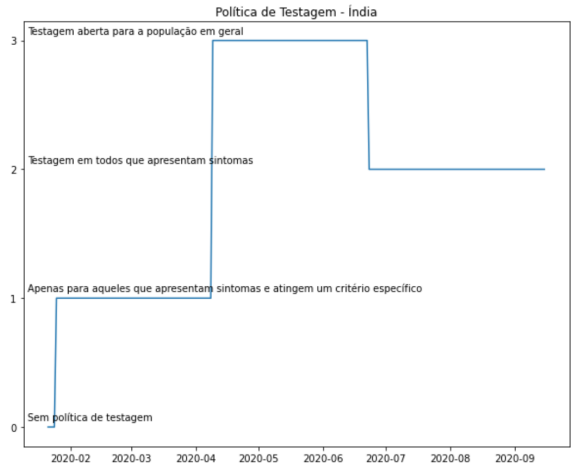

Gráfico 37: Política de testagem na Índia

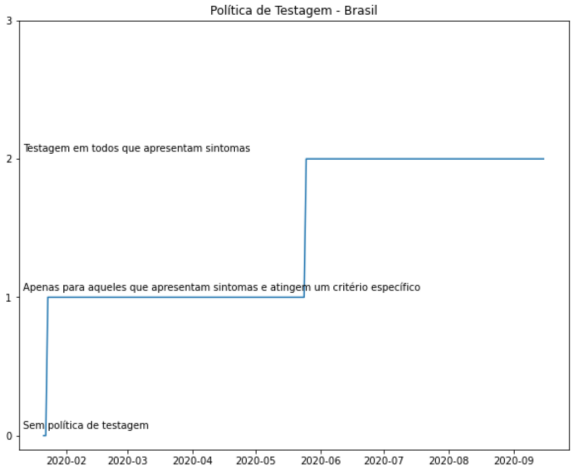

Gráfico 38: Política de testagem no Brasil

Pode-se perceber que os países que se destacaram na evolução dos casos confirmados foram países populosos que tiveram diversos tipos de influência nessa evolução, como a política de testagem adotada, índices de mobilidades em contextos variados, diferentes tipos de quarentena e consequentemente de medidas de restrição.

5.2. Evolução dos números de casos relativos à população do país

Ao analisar os gráficos anteriores, percebe-se que os países que se destacavam negativamente em relação à evolução dos números de casos da COVID-19 eram países muito populosos. Então, é necessário realizar uma comparação mais justa. Para isso, os valores da evolução dos números de casos foram divididos pela quantidade média de habitantes em 2020 no respectivo país e, agora, os países que se destacaram negativamente são diferentes dos analisados na Seção 5.1 com exceção do Chile (cluster 2) e do Brasil (cluster 3) que continuaram como países que se destacaram negativamente nessa nova análise.

- **Cluster 0 - San Marino e Andorra**

A partir dessa nova análise, considerando as características sociodemográficas dos países do cluster 0, observa-se que as curvas de casos confirmados relativos à população de San Marino e Andorra no Gráfico 39 destoam dos demais países do cluster 0. San Marino tem aproximadamente 30 mil habitantes é rodeado pela Itália, já Andorra tem quase 80 mil habitantes e está situada entre a França e a Espanha, ambos são territórios pequenos onde qualquer surto de caso faz com que as estatísticas subam muito (WIKIPÉDIA, 2020).

Destaca-se também a influência dos países próximos que são países mais populosos e que tiveram uma grande quantidade de números de casos confirmados. San Marino adotou medidas bastante restritivas a partir de março, mas como pode ser visto nos Gráficos 40 e 41 não foi possível controlar o avanço da contaminação da população. Já Andorra, optou por medidas menos restritivas e também não conseguiu controlar o aumento do número de casos confirmados relativos à população. Além disso, a política de testagem exposta nos Gráficos 42 e 43 mostra que, por muito tempo, houve uma testagem de toda a população, o que também pode acarretar em um maior número de casos confirmados relativos à população, influenciando na evolução da doença nesses territórios. Como os dados de mobilidade destes territórios não estavam disponíveis, não foi possível buscar explicações nesse âmbito.

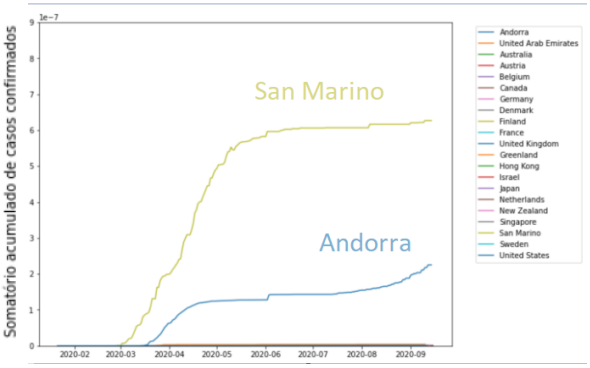

Gráfico 39: Evolução do número de casos relativos à população no cluster 0

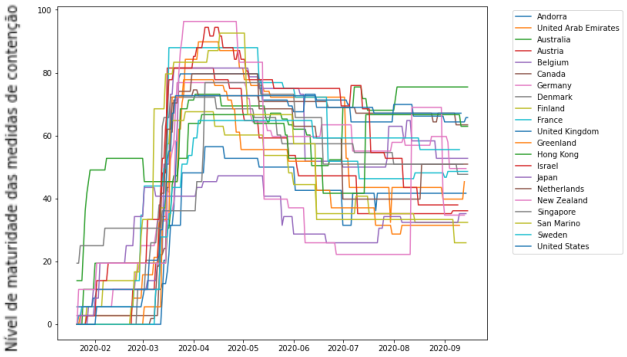

Gráfico 40: Evolução do stringency index no cluster 0

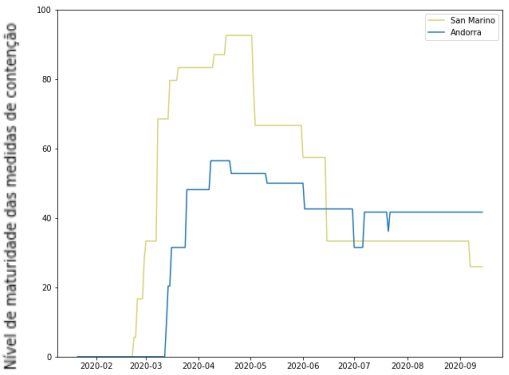

Gráfico 41: Evolução do stringency index em San Marino e em Andorra

Gráfico 42: Política de testagem em San Marino

Gráfico 43: Política de testagem em Andorra

- **Cluster 1 - Aruba e Kuwait**

Aruba também é um território pequeno com um pouco mais de 100 mil habitantes, já Kuwait tem aproximadamente 4 milhões de habitantes (WIKIPÉDIA, 2020). Aruba, como pode-se ver no Gráfico 44 de evolução do número de casos relativos à população, teve um pico em agosto próximo à mudança da política de testagem, como pode ser visto no Gráfico 49, que passou a testar mais a população. Além disso, como exposto nos Gráficos 45 e 46, em julho foi quando as medidas de restrição de Aruba mais se afrouxaram, tendo assim, um índice de mobilidade maior (Gráfico 47) nessa época. A política de testagem no Kuwait não sofreu alterações a partir do meio de março (Gráfico 50). O aumento do número de casos confirmados relativos à população pode ser explicado pelo aumento no índice de mobilidade do Kuwait no final de maio, mesmo com as altas medidas de restrições tomadas no início do mesmo mês (Gráficos 45 e 46). Essa análise mostra, assim, uma baixa aderência da população às restrições. Com o passar dos meses, houve um afrouxamento cada vez maior das medidas e um aumento gradual da circulação de pessoas (Gráfico 48), que resultou em uma evolução crescente do número de casos relativos à população (Gráfico 44).

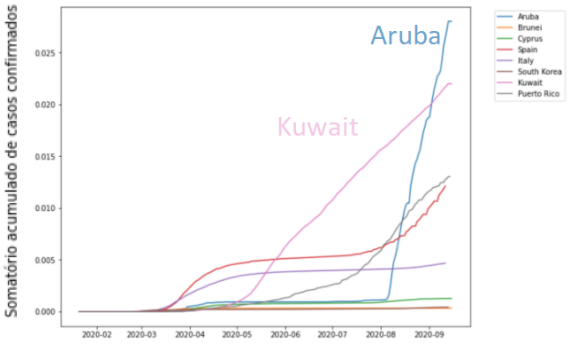

Gráfico 44: Evolução do número de casos relativos à população no cluster 1

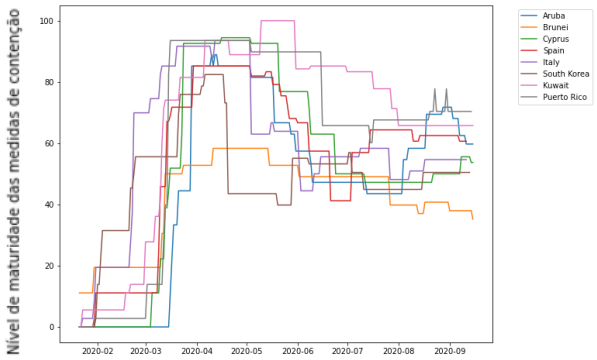

Gráfico 45: Evolução do stringency index no cluster 1

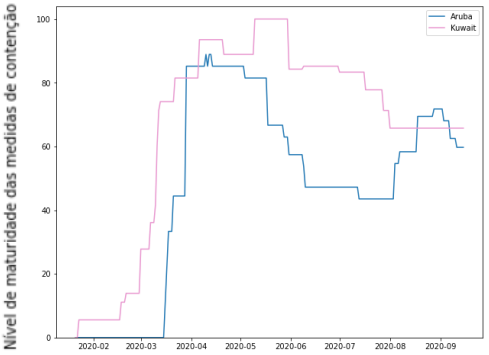

Gráfico 46: Evolução do stringency index em Aruba e no Kuwait

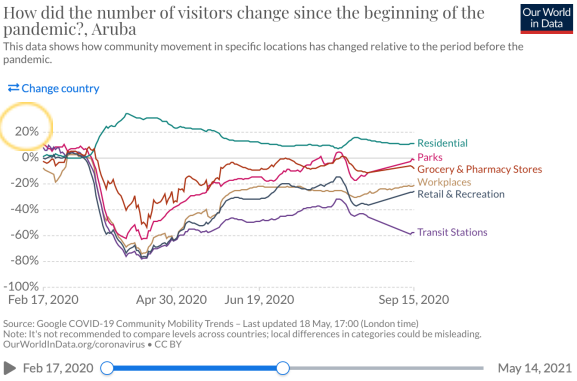

Gráfico 47: Mobilidade de Aruba desde o início da pandemia (Our World in Data, 2021)

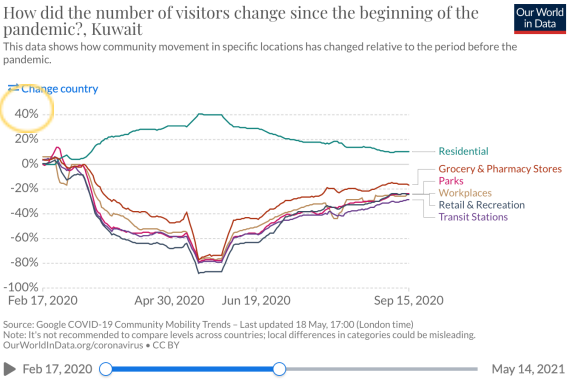

Gráfico 48: Mobilidade do Kuwait desde o início da pandemia (Our World in Data, 2021)

Gráfico 49: Política de testagem em Aruba

Gráfico 50: Política de testagem no Kuwait

- **Cluster 2 - Panamá e Chile**

No cluster 2, o Chile permaneceu sendo um país de destaque na evolução do número de casos confirmados a partir dessa nova análise, mas agora também tem-se o Panamá como pode ser visto no Gráfico 51. O Panamá possui uma população de aproximadamente 4 milhões de pessoas (WIKIPÉDIA, 2020) e teve medidas de restrições rigorosas durante o período analisado. A política de testagem, como pode ser vista no Gráfico 55, não teve mudanças a partir de março, que foi quando passaram a testar todos que apresentavam sintomas. Em junho, percebe-se um aumento drástico e incontrolável na evolução dos casos confirmados relativos à população (Gráfico 51), exatamente no momento em que as medidas de restrições afrouxaram um pouco, como pode ser visto tanto nos Gráficos 52 e 53 como também no Gráfico 54 de mobilidade, onde junho houve uma maior circulação de pessoas no país.

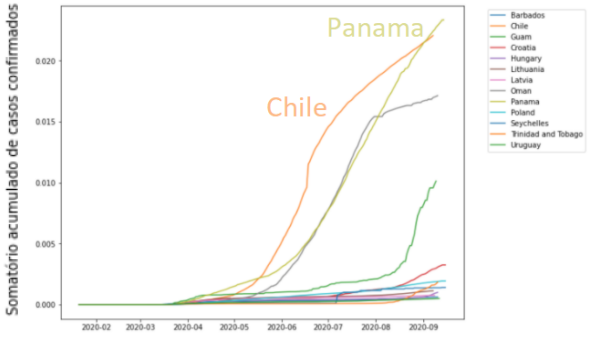

Gráfico 51: Evolução do número de casos relativos à população no cluster 2

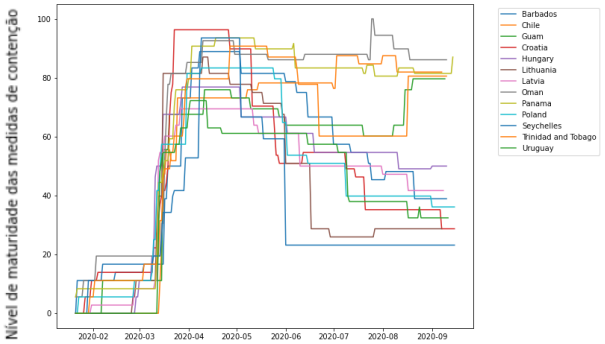

Gráfico 52: Evolução do stringency index no cluster 2

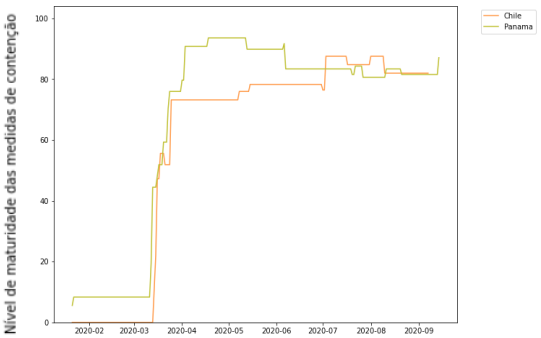

Gráfico 53: Evolução do stringency index no Panamá e no Chile

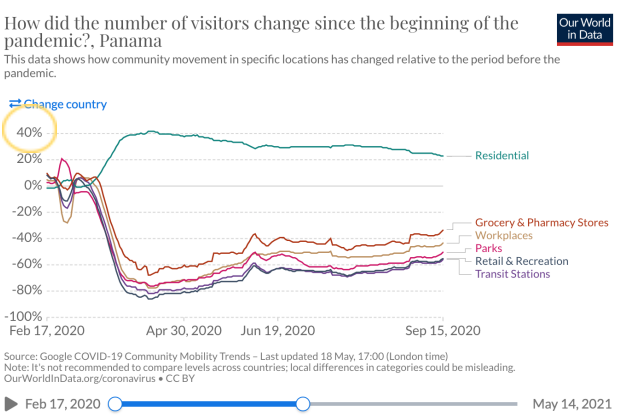

Gráfico 54: Mobilidade do Panamá desde o início da pandemia (Our World in Data, 2021)

Gráfico 55: Política de testagem no Panamá

- **Cluster 3 - Brasil e Peru**

No cluster 3, o Brasil permaneceu sendo um país de destaque na evolução do número de casos a partir dessa nova análise, mas agora ao invés da Índia tem-se o Peru (Gráfico 56). O Peru tem uma população aproximada de 35 milhões de pessoas (WIKIPÉDIA, 2020), com a evolução do número de casos relativos à população se destacando a partir de maio, que é quando ocorre uma maior testagem da população, como pode-se perceber no Gráfico 60. Porém, durante apenas o mês de junho, o país devido a quantidade de pessoas apresentando sintomas e a falta de testes disponíveis para testar todos voltou a testar menos pessoas, o que não causou grandes alterações na curva de evolução visto que a contaminação no país estava desenfreada(Gráfico 56). Esse mês em que a testagem estava mais restrita nos aponta para um número ainda maior de casos confirmados relativos à população que não foram registrados. Em relação ao índice de mobilidade do Gráfico 59, pode-se perceber que em março ocorreu uma redução considerável da população circulando no país, mas com o passar do tempo e mesmo com as medidas de restrições altas (Gráficos 57 e 58), houve uma maior circulação de pessoas, indicando assim, que a população passou gradualmente a não respeitar essas restrições, o que acarretou em um aumento desenfreado na evolução do número de casos confirmados relativos à população.

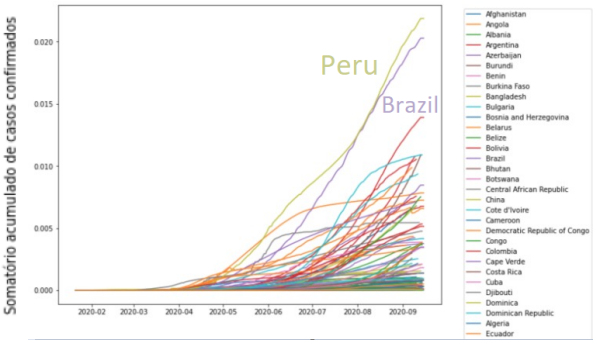

Gráfico 56: Evolução do número de casos relativos à população no cluster 3

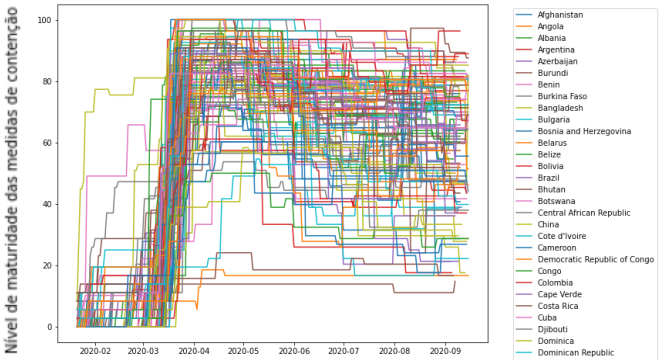

Gráfico 57: Evolução do stringency index no cluster 3

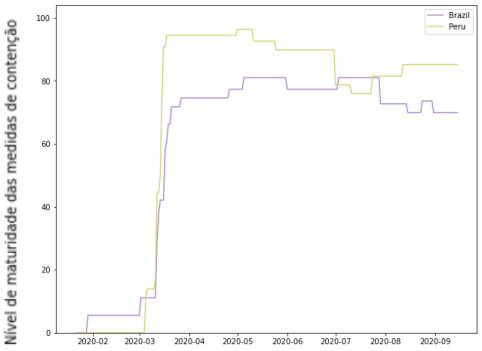

Gráfico 58: Evolução do stringency index no Brasil e no Peru

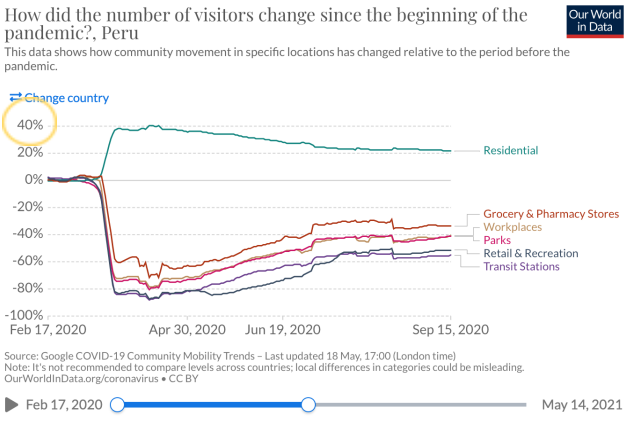

Gráfico 59: Mobilidade do Peru desde o início da pandemia (Our World in Data, 2021)

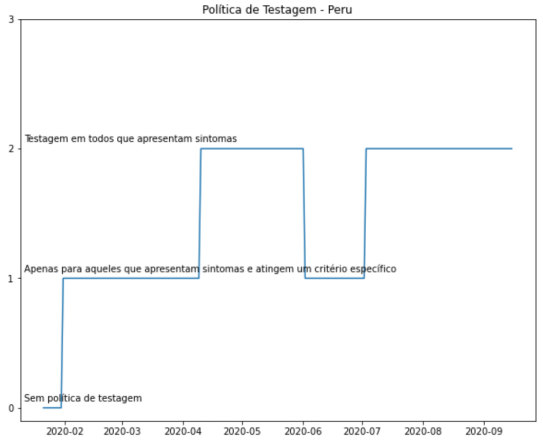

Gráfico 60: Política de testagem no Peru

De forma geral, alguns territórios como San Marino e Andorra que se destacaram nessa análise eram territórios menores que tiveram diferentes tipos de influência como os países com os quais fazem fronteira, a política de testagem e as medidas de restrição. A forma como a doença atinge locais menos populosos pode também estar ligada a uma cultura de maior aglomeração dentro desses territórios.

6. Conclusão

A pandemia da COVID-19 atingiu o mundo todo de forma inesperada, um vírus novo, sem estudos prévios, sem um diagnóstico certeiro e, principalmente, sem uma medida farmacêutica ou restritiva extremamente eficaz para combatê-la. O início de março de 2020 foi tomado por medo, além de excessivas e conturbadas informações sobre a doença, com o grupo de risco se alterando a cada novo estudo e as novas medidas surgindo para combater o vírus. Foram necessários alguns meses para o mundo se adaptar ao "novo normal" com distanciamento social e uso constante de máscaras.

O objetivo deste trabalho foi, assim, buscar entender melhor o comportamento da evolução da doença com relação ao número de casos em cada país, buscando extrair alguns *insights* a partir de características sociodemográficas dos países (como a população, o percentual da população vivendo em cidades, PIB per capita, IDH...), do índice de restrição (*Stringency Index*) determinado pelas medidas de contenção estabelecidas oficialmente por cada governo e catalogadas pela iniciativa "COVID-19 Government Response Tracker" da Universidade de Oxford (UNIVERSITY OF OXFORD, 2021), e dos relatórios de mobilidade da população em diferentes países com as curvas de tendências de deslocamento ao longo do tempo, disponibilizado publicamente pelo Google (GOOGLE, 2021). A integração de tais fontes de dados e a modelagem dos dados utilizando técnicas de Clusterização baseada em dados sociodemográficos/populacionais permitiram entender em até qual ponto essas informações influenciam na evolução do número de casos dos países.

Conclui-se, assim, que apesar de a curva de casos da COVID-19 de alguns países destoarem dentro dos seus clusters, a hipótese de pesquisa de que os dados sociodemográficos/populacionais influenciam na evolução do número de casos pôde ser confirmada. Em relação às medidas de contenção, pode-se perceber diferentes tipos de abordagem nos diversos países como, por exemplo, restrições de mobilidade, fechamento de escolas, locais de trabalho e das fronteiras. Porém, em todos eles nota-se que o que mais afetava a curva de evolução do número de casos da doença era a política de testagem, visto que quanto mais pessoas são testadas, maior acaba sendo os casos confirmados, já que muitas pessoas são assintomáticas ou apresentam sintomas leves.

O presente estudo foi realizado no contexto do NOIS (Núcleo de Operações e Inteligência em Saúde)[3], um grupo formado por alunos, professores e pesquisadores do Departamento de Engenharia Industrial e do Instituto Tecgraf da PUC-Rio, Fiocruz, USP e IDOR, que vêm pesquisando e desenvolvendo soluções inovadoras para monitoramento da COVID-19, provendo evidências para decisões baseadas em dados. O código fonte e o dataset estão disponíveis publicamente no Github: https://github.com/ingridfrf/covid19-clustering-analysis.

6.1. Limitações

Durante o desenvolvimento do trabalho, houve a falta de dados de países pequenos, o que resultou na redução do dataset e em algumas conclusões não tão claras. Além disso, os dados de mobilidade oferecidos pelo Google são um pouco restritos, visto que não abrangem toda população; por exemplo, a China que não autoriza produtos do Google (MACEDO, 2015) e pessoas que não habilitam o histórico de localização do Google Maps em seus dispositivos.

Além disso, a falta de testes principalmente no início da pandemia e a discrepância entre os países que conseguem testar toda a população e os que apenas testam pessoas com sintomas traz uma análise limitada em relação aos casos confirmados e sua curva de evolução.

6.2. Trabalhos Futuros

Para trabalhos futuros seria interessante realizar a Clusterização das séries temporais e não somente dos dados sociodemográficos/populacionais (dados estáticos). Através desse novo tipo de abordagem, seria possível analisar se houve grandes mudanças na evolução do número de casos de acordo com a composição das medidas de contenção ao longo das datas.

Um outro ponto que temos que levar em consideração é que nem sempre uma medida de contenção adotada em um país é abordada de uma mesma forma em outro. Por exemplo, no Brasil o uso de máscara é obrigatório, mas nem todos os cidadãos utilizam (G1 RR, 2021). Já na França, o uso de máscara é obrigatório e ela obrigatoriamente precisa ser do tipo cirúrgica FFP2 (semelhante à PFF2 brasileira e à N95) diferentemente do que ocorre no Brasil (ALEGRETTI, 2021). Vê-se, então, que entre os dois países existem diferenças na especificidade da máscara utilizada e no cumprimento às leis estabelecidas. Logo, para tentar

[3] https://sites.google.com/view/nois-pucrio/sobre-nois

contornar essas diferenças, pode-se trabalhar futuramente na modelagem conceitual das medidas de restrições, aplicando Ontologias de Fundamentação (UFO), mais especificamente a UFO-L, que é a ontologia núcleo de aspectos jurídicos construída sob a perspectiva das relações jurídicas (GRIFFO, 2018).

Estender a análise para verificar o impacto das características sociodemográficas dos países na evolução do número de pessoas vacinadas seria muito interessante também. Assim, permitiria visualizar como os países estão agindo para conter a COVID-19.

REFERÊNCIAS

ALEGRETTI, Laís. **Máscara N95 e PFF2: por que países da Europa reprovam material caseiro e agora exigem máscara profissional**. Disponível em: https://noticias.uol.com.br/saude/ultimas-noticias/bbc/2021/01/28/mascara-n95-e-pff2-por-que-paises-da-europa-reprovam-material-caseiro-e-agora-exigem-mascara-profissional.htm. Acesso em: 30 maio 2021.

AMARAL, Glenda; BAIÃO, Fernanda; GUIZZARDI, Giancarlo. Foundational ontologies, ontology‐driven conceptual modeling, and their multiple benefits to data mining. **Wires Data Mining And Knowledge Discovery**, [S.L.], e1408, 24 mar. 2021. Wiley. http://dx.doi.org/10.1002/widm.1408.

ASKITAS, Nikolaos *et al*. Estimating worldwide effects of non-pharmaceutical interventions on COVID-19 incidence and population mobility patterns using a multiple-event study. **Scientific Reports**, [S.L.], v. 11, n. 1, p. 1-13, 21 jan. 2021. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41598-021-81442-x. Disponível em: https://www.nature.com/articles/s41598-021-81442-x. Acesso em: 22 maio 2021.

BBC. **Coronavírus: 4 fatores que explicam o impacto da covid-19 nos EUA, país com maior número de infectados e mortos**. 2020. Disponível em: https://www.bbc.com/portuguese/internacional-52280762. Acesso em: 20 maio 2021.

BRAUN, Julia. **Como a Itália se tornou o segundo país mais afetado pelo coronavírus**. 2020. Disponível em: https://veja.abril.com.br/mundo/como-a-italia-se-tornou-o-segundo-pais-mais-afetado-pelo-coronavirus/ Acesso em: 20 maio 2021.

CALIL, Leonardo Aparecido de Almeida *et al*. **Mineração de Dados e Pós-processamento em Padrões Descobertos**. 2008. Disponível em: http://ri.uepg.br/riuepg/bitstream/handle/123456789/142/ARTIGO_Minera%C3%A7%C3%A3oDadosP%C3%B3s.pdf?sequence=1. Acesso em: 24 maio 2021.

CANAL COMSTOR. **5 vantagens de utilizar Data Science nos negócios**. Disponível em: https://blogbrasil.comstor.com/5-vantagens-de-utilizar-data-science-nos-negocios. Acesso em: 24 maio 2021.

CDC. **Nonpharmaceutical Interventions (NPIs)**. Disponível em: https://www.cdc.gov/nonpharmaceutical-interventions/index.html. Acesso em: 21 de maio de 2021.

CETAX. **Data Mining: O que é, conceito e definição**. 2020. Disponível em: https://www.cetax.com.br/blog/data-mining/. Acesso em: 07 jul. 2021.

COMPARING. **Python Clustering Algorithms**. 2016. Disponível em: https://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html. Acesso em: 18 maio 2021.

DATAAT, Grupo. **Agrupamento**. Disponível em: https://dataat.github.io/introducao-ao-machine-learning/agrupamento.html. Acesso em: 07 jul. 2021.

DEMOGRAPHIA. **World Urban Areas: Built Up Urban Areas or World Agglomerations - 17th ANNUAL EDITION**. 2021. Disponível em: http://www.demographia.com/db-worldua.pdf. Acesso em: 10 mar. 2021.

FABIO, Ingrid; GIRUNDI, Raissa. **Covid19-clustering-analysis**. 2021. Disponível em: https://github.com/ingridfrf/covid19-clustering-analysis. Acesso em: 30 maio 2021.

FARINELLI, Victor. **Governo chileno perde o controle da pandemia e enfrenta novos protestos sociais**. 2020. Disponível em: https://www.brasildefato.com.br/2020/05/26/governo-chileno-perde-o-controle-da-pandemia-e-enfrenta-novos-protestos-sociais. Acesso em: 20 maio 2021.

G1 RR. **Imagens mostram aglomeração e pessoas sem máscara na 'Feira do Garimpeiro' em Boa Vista**. Disponível em: https://g1.globo.com/rr/roraima/noticia/2021/02/28/imagens-mostram-aglomeracao-e-pessoas-sem-mascara-na-feira-do-garimpeiro-em-boa-vista.ghtml. Acesso em: 30 maio 2021.

G1. **Índia ultrapassa o Brasil e se torna o segundo país com mais casos de coronavírus**. 2020. Disponível em: https://g1.globo.com/mundo/noticia/2020/09/07/india-ultrapassa-o-brasil-e-se-torna-o-segundo-pais-do-mundo-com-mais-casos-de-coronavirus.ghtml. Acesso em: 20 maio 2021.

GONÇALVES, Pollyanna. **Afinal, como se desenvolve um projeto de Data Science?** 2018. Disponível em: https://medium.com/techbloghotmart/afinal-como-se-desenvolve-um-projeto-de-data-science-233472996c34. Acesso em: 10 maio 2021.

GOOGLE. **COVID-19:Relatórios de mobilidade da comunidade**. 2021. Disponível em: https://www.google.com/covid19/mobility/. Acesso em: 01 maio 2021.

GRIFFO, Cristine. **UFO-L: UMA ONTOLOGIA NÚCLEO DE ASPECTOS JURÍDICOS CONSTRUÍDA SOB A PERSPECTIVA DAS RELAÇÕES JURÍDICAS**. 2018. Disponível em: https://philarchive.org/archive/BECUUO. Acesso em: 30 maio 2021.

GUEDES, Erivelton Pires. **Clusterização**. 2019. Disponível em: https://www.kaggle.com/eriveltonguedes/7-clusteriza-o-k-means-erivelton. Acesso em: 07 jul. 2021.

GUELL, Oriol; MEDINA, Miguel Ángel. **Por que a Espanha tem a maior taxa de mortalidade pelo coronavírus do planeta?** 2020. Disponível em: https://brasil.elpais.com/internacional/2020-04-08/por-que-a-espanha-tem-a-maior-taxa-de-mortalidade-pelo-coronavirus-do-planeta.html. Acesso em: 20 maio 2021.

HAUG, Nils et al. Ranking the effectiveness of worldwide COVID-19 government interventions. **Nature Human Behaviour**, [S.L.], v. 4, n. 12, p. 1303-1312, 16 nov. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41562-020-01009-0. Disponível em: https://www.nature.com/articles/s41562-020-01009-0. Acesso em: 22 maio 2021.

KAUSHIK, Saurav. **An Introduction to Clustering and different methods of clustering**. 2016. Disponível em: https://www.datasciencecentral.com/profiles/blogs/an-introduction-to-clustering-and-different-methods-of-clustering Acesso em: 22 maio 2021.

KDNUGGETS. **DBSCAN Clustering Algorithm in Machine Learning**. 2017. Disponível em: https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html. Acesso em: 10 jun. 2021.

LIAM, Harry. **The Role Of Data Science in Healthcare Advancements: Applications and Benefits**. Disponível em: https://datafloq.com/read/role-of-data-science-healthcare-advancements-applications-benefits/8514. Acesso em: 24 maio 2021.

MACEDO, Daniela. **Confira oito sites populares que são proibidos na China**. 2015. Disponível em: https://veja.abril.com.br/tecnologia/confira-oito-sites-populares-que-sao-proibidos-na-china/. Acesso em: 30 maio 2021.

MARKHAM, Merry Jennifer. **Coronavírus e COVID-19: o que as pessoas com câncer precisam saber**. 2020. Disponível em: https://www.cancer.net/blog/2020-06/coronav%C3%ADrus-e-covid-19-o-que-pessoas-com-c%C3%A2ncer-precisam-saber. Acesso em: 14 abr. 2021.

MCINNES, Leland; HEALY, John; ASTELS, Steve. **How HDBSCAN Works**. 2016. Disponível em: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html. Acesso em: 10 jun. 2021.

NEVES, Sofia. **Covid-19: porque é que Espanha tem uma das taxas de letalidade mais altas do mundo?** 2020. Disponível em: https://www.publico.pt/2020/04/07/mundo/noticia/covid19-espanha-taxas-letalidade-altas-mundo-1911397. Acesso em: 20 maio 2021.

NISBET, Robert *et al*. **Handbook of Statistical Analysis and Data Mining Applications**. [S. L.]: Academic Press, 2009. Disponível em: https://www.elsevier.com/books/handbook-of-statistical-analysis-and-data-mining-applications/nisbet/978-0-12-374765-5. Acesso em: 24 maio 2021.

NOVIA, Data. **HIERARCHICAL CLUSTERING IN R: THE ESSENTIALS**. 20--. Disponível em: https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/ . Acesso em: 10 jun. 2021.

Our World in Data. **COVID-19: Google Mobility Trends**. 2021. Disponível em: https://ourworldindata.org/covid-google-mobility-trends. Acesso em: 22 maio 2021.

PRABHU, Tanu N. **Population by Country - 2020**. 2020. Disponível em: https://www.kaggle.com/tanuprabhu/population-by-country-2020. Acesso em: 01 maio 2021.

RITCHIE, Hannah; ROSER, Max. **Smoking**. 2019. Disponível em: https://ourworldindata.org/smoking. Acesso em: 10 mar. 2021.

ROCHE. **Cenário da COVID-19 no Brasil**. 2020. Disponível em: https://www.roche.com.br/pt/por-dentro-da-roche/cenario-da-COVID-19-no-Brasil.html. Acesso em: 14 abr. 2021.

SANTANA, Felipe. **Guia passo a passo de como um projeto de Data Science é desenvolvido**. 2019. Disponível em: https://minerandodados.com.br/guia-passo-a-passo-de-como-um-projeto-de-data-science-e-desenvolvido/. Acesso em: 17 maio 2021.

SANTANA, Felipe. **Por que o Python é a Linguagem mais adotada na área de Data Science?** 2019. Disponível em: https://minerandodados.com.br/por-que-o-python-e-a-linguagem-mais-adotada-na-area-de-data-science/. Acesso em: 07 jul. 2021.

SCHMIDT, Beatriz *et al*. **Impactos na Saúde Mental e Intervenções Psicológicas Diante da Pandemia do Novo Coronavírus (COVID-19).** SciELO Preprints, 1(1), 1–26. doi: https://doi.org/10.1590/SCIELOPREPRINTS.58.

SCHRECK, Ben. **Feature Engineering vs Feature Selection**. 2018. Disponível em: https://innovation.alteryx.com/feature-engineering-vs-feature-selection/. Acesso em: 25 maio 2021.

SCIKIT-LEARN. **Clustering**. 2020. Disponível em: https://scikit-learn.org/stable/modules/clustering.html. Acesso em: 10 mar. 2021.

SEMANTIX. **Escolhendo o algoritmo de clusterização apropriado para seus dados**. 2019. Disponível em: https://www.semantix.com.br/escolhendo-o-algoritmo-de-clusterizacao-apropriado-para-seus-dados/. Acesso em: 18 maio 2021.

SHAW, Rajib; KIM, Yong-kyun; HUA, Jinling. **Governance, technology and citizen behavior in pandemic: Lessons from COVID-19 in East Asia.** Progress in disaster science, p. 100090, 2020. Disponível em:Governance, technology and citizen behavior in pandemic: Lessons from COVID-19 in East Asia (nih.gov). Acesso em: 22 março de 2021.

SHCHERBAKOV, M. *et al*. 2014. **"Lean Data Science Research Life Cycle: A Concept for Data Analysis Software Development"**, Knowledge-Based Software Engineering, v. 466, pp. 708-716, Springer, 2014.

The Economist (2021). **TRACKING covid-19 excess deaths across countries**, 15 abr. 2020. Disponível em: https://www.economist.com/graphic-detail/coronavirus-excess-deaths-tracker. Acesso em: 13 abril 2021.

THE WORLD BANK. **GDP per capita (current US$)**. 2018. Disponível em: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD. Acesso em: 10 mar. 2021.

THE WORLD BANK. **Gini index (World Bank estimate)**. 2018. Disponível em: https://data.worldbank.org/indicator/SI.POV.GINI. Acesso em: 10 mar. 2021.

THE WORLD BANK. **Population in the largest city (% of urban population)**. 2018. Disponível em: https://data.worldbank.org/indicator/EN.URB.LCTY.UR.ZS. Acesso em: 10 mar. 2021.

THE WORLD BANK. **Population in urban agglomerations of more than 1 million (% of total population)**. 2018. Disponível em: https://data.worldbank.org/indicator/en.urb.mcty.tl.zs. Acesso em: 10 mar. 2021.

THE WORLD BANK. **Population living in slums (% of urban population)**. 2018. Disponível em: https://data.worldbank.org/indicator/en.pop.slum.ur.zs. Acesso em: 10 mar. 2021.

THE WORLD BANK. **Urban population (% of total population)**. 2018. Disponível em: https://data.worldbank.org/indicator/SP.URB.TOTL.in.zs. Acesso em: 10 mar. 2021.

TOWARDS. **Agglomerative Clustering and Dendrograms — Explained**. 2020. Disponível em: https://towardsdatascience.com/agglomerative-clustering-and-dendrograms-explained-29fc12b85f23. Acesso em: 10 jun. 2021.

UNA-SUS. **Organização Mundial de Saúde declara pandemia do novo Coronavírus.** 2020. Disponível em: https://www.unasus.gov.br/noticia/organizacao-mundial-de-saude-declara-pandemia-de-coronavirus. Acesso em: 14 abr. 2021.

UNDP. **Human Development Data Center**. 2020. Disponível em: http://hdr.undp.org/en/data. Acesso em: 10 mar. 2021.

UNIVERSITY OF OXFORD. **COVID-19 GOVERNMENT RESPONSE TRACKER**. 2020. Disponível em: https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker. Acesso em: 10 mar. 2021.

VICK, Mariana. **Pandemia: origens e impactos, da peste bubônica à covid-19**. 2020. Disponível em: https://www.nexojornal.com.br/explicado/2020/06/20/Pandemia-origens-e-impactos-da-peste-bub%C3%B4nica-%C3%A0-covid-19. Acesso em: 14 abr. 2021.

VILELA JUNIOR, Guanis de Barros. **Um problema pode ter muitas hipóteses, que são soluções possíveis para a sua resolução.** Campinas: Cpaqv, 20--?. Color. Disponível em: http://www.cpaqv.org/metodologia/o_problema_e_a_hipotese.pdf. Acesso em: 24 maio 2021.

YILDIRIM, Soner. **K-Means Clustering — Explained**. 2020. Disponível em: https://towardsdatascience.com/k-means-clustering-explained-4528df86a120. Acesso em: 09 jun. 2021.

WIKIPÉDIA. **Lista de países por população**. 2020. Disponível em: https://pt.wikipedia.org/wiki/Lista_de_pa%C3%ADses_por_popula%C3%A7%C3%A3o. Acesso em: 25 maio 2011.

WIKIPÉDIA. **DBSCAN**. 2021. Disponível em: https://en.wikipedia.org/wiki/DBSCAN. Acesso em: 09 jun. 2021.