**PEDRO MEDEIROS HAMACHER**

# Recommending player signings to football teams: A data-driven optimization approach

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

Orientador: Silvio Hamacher

Departamento de Engenharia Industrial
Rio de Janeiro, 15 de dezembro de 2023.

## Resumo

**Recomendação de contratação de jogadores para times de futebol: uma abordagem data-driven**

Times de futebol gastam bilhões de dólares anualmente na contratação de novos jogadores para melhorar seu elenco e endereçar pontos fracos. Apesar de possuírem estatísticas completas de jogadores de vários países, muitos times não utilizam esses dados em toda sua capacidade para identificação de potenciais novas contratações. Visando preencher essa lacuna, este trabalho apresenta modelos que visam sugerir quais jogadores devem ser comprados e vendidos para atender as necessidades particulares de cada time, formar um elenco adequado e ganhar vantagem competitiva. Um modelo estocástico de dois estágios de Programação Inteira Linear Mista é apresentado para otimizar as escolhas de montagem de elenco de uma equipe considerando seu orçamento, atributos desejados e ausências ao longo da temporada devido a lesões ou suspensões. Ainda é sugerido um framework de ciência de dados para coletar, tratar e imputar dados reais para construir um modelo data-driven. O framework é aplicado em dados reais das principais ligas e alguns estudos de caso são apresentados para demonstrar os resultados do modelo e as sugestões de elenco.



**Palavras-chave:** Programação Inteira, Data Analytics para Esportes, Decisões Voltadas a Dados, Montagem de Elenco

## Abstract

**Recommending player signings to football teams: A data-driven optimization approach**

Football teams spend billions of dollars yearly signing new players to improve their squad and to fill identified areas of need. Despite having available complete statistics from players all around the globe, teams often do not use this data at its total capacity to identify potential signings. Looking to fill this void, this work presents models intended to suggest to teams which players should they buy and sell to fulfill their specific needs, adequately assemble the roster and gain a competitive edge. A stochastic two-stage Mixed Integer Linear Programming model is presented to optimize a team's roster choices considering their budget, desired attributes and absences throughout the season due to injury or suspensions. A data-science framework is also proposed for data collection and treatment to input it into a data-driven model. The framework is applied to real-world data from top leagues and some case studies are presented in order to showcase its results and roster suggestions.



**Keywords:** Mixed Integer Programming, Data Analytics for Sports, Data Decision Driving, Roster Building

# Table of Contents

## Contents

## List of Figures

## List of Tables

## List of Charts

## Introduction

The football transfer market is a massive business where slight competitive margins can sum up to a big difference, both in financial terms and in-field performance. Player transfer fees have been growing and reached a record high in 2019, where 7.35 billion euros were spent on fixed, conditional and release fees (FIFA, 2023) - those numbers took a dip due to the Covid pandemic, but are rising again and expected to break new records.

The huge sums of money invested can be explained by the impact that it has on the club's results on the pitch: 89% of the variance in a club's position at the end of a season can be explained only by the club's total payroll (KUPER and SZYMANSKI, 2009). Throughout studies performed in many leagues during different periods, the picture is always clear: clubs that invest heavier in their squad undoubtedly have more success. It is clear, therefore, why clubs are willing to spend that much on improving their roster, as it is a very important predictor of success.

Another reason that clubs make sure to invest in their roster building is the uncertainty derived both from players' injuries and suspensions. Through the course of a season, an average player will sustain 2 injuries and miss 37 days, or 12% of the total season (EKSTRAND; HÄGGLUND; WALDÉN, 2011). Players can also miss matches due to suspensions, from receiving red cards, accumulating yellow cards or off-field issues such as doping – that last case won't be treated by this paper. These events that lead to player absences can turn a team's whole season around, in particular when those occur immediately before important matches. In fact, Hägglund et al. (2013) have found that higher injury rates and lower players' match availability are associated with decreased team performance and results.

Because of that, clubs want to have more adequate players than only their starting eleven, as some quality backups could be crucial both as a safeguard against bad streaks from the starters and in maintaining the team's record during the referred absences and potentially even salvage their season. However, there exists a trade-off between how much should be invested in the starting eleven and how much in the backups which is not easy to quantify, as one has to consider the degree of uncertainty of said absences. Therefore, the questions of how many players should be on the roster and how much should be invested in players who won't be starters are simultaneously crucial and not trivial for a club to solve.

In the era of big data, football clubs have at their disposal a huge amount of statistics of players all around the globe, via both paid and open-access software. Some teams have focused on data-driven strategies that have been extremely beneficial for them, as countless teams emerged as examples both from public perception and academic papers, such as Brighton, Brentford and Liverpool (LICHTENTHALER, 2022). This data-oriented approach can help a team spend their money more wisely and potentially overachieve relative to the aforementioned financial expectations.

However, many clubs do not use this huge amount of data to their full capacity to their advantage. Kaplan (2010) agrees with this statement and argues that soccer is far behind in terms of statistics usage compared to other major sports. Those clubs would benefit from tools that suggest which players they should hire and sell. In particular, the club could identify areas of need on the pitch and the tool would automatically return roster moves to achieve that goal.

Moreover, if open access, such a tool would be extremely beneficial for fans to provide insights and help in discussions. As even clubs that tend to be more data-driven generally keep their advancements and state-of-the-art tools private, in order to secure their competitive advantage, fans around the globe don't have access to this kind of technology. Therefore, the development and easy access to such a tool would benefit multiple stakeholders in the football world environment, from clubs seeking to optimize their transfer strategy to fans interested in delving deeper into some of those topics.

In this work, a model with such characteristics is proposed, using real-world data to suggest the optimal transfers a club should make. The model consists of a two-stage stochastic Mixed-Integer Linear program where a club seeks to optimize its on-field attributes with specific constraints such as its budget. Furthermore, the model considers multiple scenarios to determine which and how many players should constitute the roster and how big should it be, considering the possibilities of players' injuries and suspensions.

Alongside the mathematical formulation, this work also proposes and implements a data-driven framework that connects the theoretical mathematical model with real-world applicable data. This framework starts with the data collection process, which involves web scraping data from player’s performance and market value respectively from FBREF (2023) and TRANSFERMARKT (2023) using the R package worldfootballR, developed by Zivkovic (2022). The data is then treated, normalizing each attribute by how many minutes a player has played in the season and filtering only players with a minimum number of matches completed

to avoid distortions. Each team's performance is also calculated to be used as reference values in the optimization model. The formulation is finally applied using the treated data as input and guaranteeing the model's plausibility with real-world data.

The model was applied to real-world data from the top 5 leagues in Europe. After a data science approach to gathering and treating data, the model was applied using different parameters, such as budget constraint, formation, age limit and attributes of interest. Some of those case studies are shown in the paper as examples of the model's capabilities and a sensitivity analysis is applied to observe how the teams and rosters suggested by the model react to changes in the initial parameters.

This paper's main contributions are an innovative model for roster building that considers the uncertainties related to players' injuries and suspensions and decides which players to buy and sell accordingly; to make this formulation public to facilitate access from clubs and fans all around the globe; and applications of the model to real-world data, demonstrating its results and potential use cases.

The following section presents a literature review regarding models devised to optimize a team roster or lineup. In sequence, the mathematical formulation of the new proposed model is shown. The model is then applied to real-world data in a case study. Finally, final thoughts and future research are discussed in the conclusion.

## Literature Review

Integer Linear Programming models applied to the problem of player selection in sports are not uncommon in the literature. Papers such as Mahrudinda et al. (2020) apply Binary Programming Models to decide on the optimal lineup given an already assembled roster. A hybrid method that combines Integer Programming with multi-criteria analysis is presented by Nasiri et al. (2019), using a biobjective integer programming model to solve the problem. Another combined approach was published by Ozceylan (2016), where the attributes of each player are prioritized via Analytic Hierarchic Process. An important paper referenced by the literature is Boon and Sierksma (2003), which also apply Integer Optimization models to the player selection problems both in soccer and volleyball.

Whereas those papers are mainly concentrated on selecting the optimal lineup between a set of players already existing, that is, an assembled roster, some works in literature also consider the construction of the entire roster. A paper similar to the present one is Payyappalli and Zhuang (2019), which implements a data-driven approach to build and solve the Integer Programming Model of roster construction. Smith and Bickel (2023) also solve the problem of assembling the entire roster in an application focused on the U.S. Major League Soccer, with its specific rules. The question of which transfers a club should do is also tackled by Pantuso and Hvattum (2020), which implement a chance-constrained model to solve the problem.

Many papers in this literature focus on Multi-Objective Optimization, as did Zhao et al. (2021). Qader et al. (2017) also discuss the applications of multi-criteria decision-making on this problem solution. Similar problems were also discussed for other sports, such as the work Ahmed, Deb and Jindal (2013) did on cricket team selection, which is also constituted of a multi-criteria decision problem that seeks to maximize the lineup.

A sport frequently cited in the literature is American football, which has a stronger analytical branch compared to soccer, particularly in particular to fantasy football drafts. Becker and Sun (2016) propose an integer optimization formulation to solve it, whereas Fry, Lundberg and Ohlmann (2007) apply a heuristic to dynamic programming modeling. Moreover, Muniz and Flamand (2023) also propose an integer optimization, this time for basketball teams.

Stochastic programming approaches have also been studied, most notably by Pantuso (2017). This paper presents a model resembling the one proposed in this paper, although it maximizes the value of the players, not the actual performance. The stochastic approach refers to the uncertainty of each player's future performance, but it does not consider the absence of players due to injuries or suspensions.

The problem of player selection has also been tackled by methods other than Integer Linear Programming. The use of Artificial Intelligence and Neural Networks to build the optimal lineup in sports is discussed both by Al-Shboul et al. (2017) and by Vijay Fidelis and Karthikeyan (2022), where the latter applies it specifically to soccer. Jarvandi, Sarkani and Mazzuchi (2013) use Markov modeling to encapsulate the compatibility between players, simulate their performance together and ultimately decide on the selection problem. Deb and Das (2023) present a two-stage approach, where initial probabilities of player performances are derived via a multinomial logistic regression and model, which is used as an input for a GRASP-type meta-heuristic, that actually makes the team selection. Tavana et al. (2013) tackle this problem using a two-stage fuzzy inference approach.

At last, there are also papers more focused on actually evaluating a player's performance quality, which could be useful as an input for many aforementioned models, as well as the one proposed in the current paper. Pappalardo (2018) proposes a data-driven metric that encapsulates how well a given individual played into a single number. Cao et al. (2022) go even further and consider the spatial regions of player actions and how they interact with opposing players. Yu et al. (2023) expand on this idea of player interaction and propose a player selection model based on that.

This paper contributes to this literature by combining some previously proposed ideas into a brand-novel framework. The main idea of this work is similar to Payyappalli and Zhuang (2019): formulate a data-driven optimization model to suggest optimal transfers for teams in the roster-building problem using real-world data. This paper adds new concepts to this idea, such as constraints that consider the team formation, the relative quality of the suggested team compared to the rest of the league and the possibility of minimizing roster turnover from one season to another.

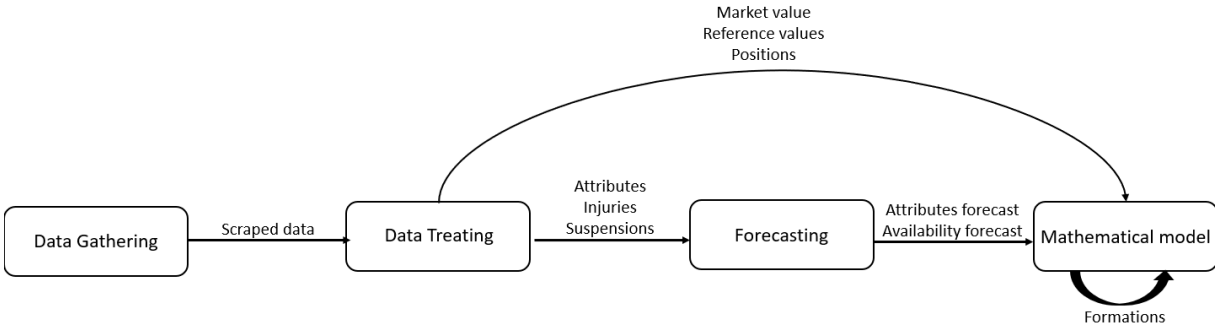
The main contribution of this work, however, is the stochastic programming approach, not present in that paper. Pantuso (2017) proposes a stochastic model for roster-building by considering the uncertainty of a player's performance in a given match, which opens up the possibility of being substituted by another player on the roster. In this paper, new sources of

uncertainty are also included in the possibility of injuries and suspensions, which force a team to use substitute players, potentially increasing the need for a bigger roster. Becker and Sun (2016) also resemble the idea proposed by this paper by considering the problem of fantasy football draft at the beginning of the season as the results of the sequences of matches throughout the season, although it doesn't propose a stochastic approach in the same sense. Therefore, the present paper provides an innovation to the sports roster-building literature with a data-driven stochastic optimization model that considers the uncertainty of player performance, injuries and suspensions.

## Methodology

In this section, the framework of the data-driven optimization is presented. It compromises two main steps: the data science one, used for collecting and treating data, described later in this section, and the stochastic model. The main idea of the process is to return an optimal roster for a club given its specific parameters and the real-world data from previous seasons. The stochastic model observes several simulated matches, with different player's performance and availability, and decides based on these matches what is the best roster that can be assembled at the beginning of the season as well as the ideal formation and lineup for each match. Figure 1 shows the aforementioned framework.

Figure 1: Data Science Framework




Source: Created by the author

### A) Data Science Framework

The first step in the data-driven framework is real-world data collection and treatment. Three sets of data should be collected to be used in the optimization model: players' performance in the last seasons, measured by their statistics in 184 different on-field attributes previously measured – usually, a subset of those is selected in the optimization phase; market valuation of each player at the beginning of the new season; players' preferred position; and readiness data, both from each players' injury history and the number of yellow and red cards received, which cause disciplinary suspension. All data is treated and merged to link each player's performance, market value and availability.

Using directly the raw data for each attribute, however, can lead to distortions. Players who played several matches in one season will have more opportunities to boost their statistics even if their on-field performance wasn't as good. Conversely, players could be prejudicated if they played too few matches for reasons such as injuries, which the model already treats separately. For that reason, the on-field attributes are normalized by the number of minutes they played in the season, such that all statistics are inputted into the model on a

per-game basis. A new distortion could arise from this normalization where players with too few minutes could have disproportionally high statistics, so players' seasons with less than 450 minutes or 5 entire matches completed aren't considered and players who don't have any seasons that qualify in this criterion are not included in the model.

Reference values for each attribute are also calculated in the pre-processing stage. For each team in the dataset, the average for each on-field attribute is calculated between all players on the roster from the previous season, weighted by the playing time. The output of this procedure is a matrix of average values for each team and on-field attribute. Those values will be used as reference values in a constraint in the mathematical model to ensure a team is in a top quantile of each attribute defined by the parameters.

It is also necessary to predict the probability distribution of variables for the next season, namely the on-field attributes and the probability a player will be available in a given game, considering injuries and suspensions. Those predictions can be performed via any desired method, as only the output matters for the mathematical model. The output of this step is the availability of each player in each possible match, equivalent to being both healthy and not suspended and the player's performance in each match that he is available, which can be represented either via probability distributions or a set of scenarios.

The last step in the framework is the actual optimization model, which receives the forecasted distributions of each player's performance, availability, market value and preferred positions, as well as the reference values from the previous season and uses them to suggest an optimal roster. The model has two stages: in the first stage it decides which players to sell and buy to assemble the roster for the upcoming season according to its budget; the second stage decides which players to lineup in each simulated match – those scenarios influence the first-stage decision, as a player can only be on the lineup if he was hired in the first stage.

The second-stage scenarios are generated according to the forecasted distribution for each player's on-field performance and availability considering injuries and suspensions, as mentioned previously. The scenarios are simulated via random draws from those independent distributions and the recourse decision of which players to start a simulated game depends on which ones are available, as well as their performance in that instance.

The model also allows the definition of other parameters that the club may wish to include. For instance, the age limit may be set to an arbitrary value, where the assembled roster's average age must be lower than this limit. A minimum number of players that are already on the previous season's roster can also be set when looking to minimize roster

turnover. An alternative that accomplishes this same idea is to define the value of players that already belong to the club as a fraction of their actual market value – the model formulation considers that you have to hire your own players again – although that intervention is made on the data science steps, not in the mathematical model.

The on-field attributes that will be optimized are also defined beforehand according to the specific needs of each club. This selection of a subset of attributes tends to work better than just optimizing for hundreds of statistics simultaneously, as many of them won't be of interest to the club. For each selected attribute, it is also possible to set a minimum value that the roster should achieve, based on the reference values alluded to earlier. The club could be interested in being in a top quantile on a determined attribute relative to all teams in the previous season, and this constraint will be incorporated into the model.

The formation used by the club can also be set so that the model uses it in all simulated matches. The club may select between a pre-defined set of formations, which are encoded into the model with a vector indicating how many players of each position should be starting at each match. Alternatively, the model can select the optimal formation for the defined parameters by running the mathematical formulation multiple times, one for each possible formation, and selecting the one that returns the best objective value. All those parameter options ensure flexibility and a model that adapts to the needs and possibilities of each specific club, providing a more realistic solution to the real-world problem.

**B) Mathematical Formulation**

In this subsection, the proposed two-stage integer programming model is presented. The notation of the sets, parameters and decision variables used is presented first, followed by the model formulation:

| Sets | Index |
|---|---|
| Player | i |
| Player already on the team | j |
| Foreign player | k |
| In-game attribute | a |
| Position on the field | p |
| Injury and suspension scenario | s |

| Parameters | Description | Unit |
|---|---|---|
| $attribute_{ias}$ | Player **i**'s attribute a performance in scenario s | /game |
| $value_i$ | Market value of player **i** | millions of euros |
| $budget$ | Total team budget | millions of euros |
| $quantile_a$ | Quantile of attribute **a** out of all teams in the previous season | /game |
| $available_{is}$ | If player **i** is healthy and not suspended in scenario s | Unitless |
| $age_i$ | Age of player **i** | Years |
| $age\_limit$ | Maximum roster's average age | Years |
| min_keep | Minimum number of players that must be kept from the previous season | Unitless |
| $bool\_pos_{ip}$ | Boolean indicating if player **i** can play in position **p** | Unitless |
| $qt\_pos_p$ | How many players start in position **p** in the selected formation | Unitless |

| Decision variables | Description | Unit | Domain |
|---|---|---|---|
| $x_i$ | Decision of hiring player i for the roster | Unitless | {0,1} |
| $y_{is}$ | Decision of lining up player i in scenario s | Unitless | {0,1} |

$$max \sum_s \sum_a \sum_i y_{is} * attribute_{ias} \quad (1)$$

$s.t.$

$$\sum_i x_i * value_i \leq budget \quad (2)$$

$$\sum_{i,s} y_{is} * attribute_{ia} \geq quantile_a * \sum_{i,s} y_{is}, \forall a \in A \quad (3)$$

$$y_{is} \leq x_i, \forall i \in I, s \in S \quad (4)$$

$$y_{is} \leq available_{is}, \forall i \in I, s \in S \quad (5)$$

$$x_i \leq \sum_s y_{is}\,, \forall i \in I \tag{6}$$

$$\sum_{i,s} y_{is} * age_i \leq age\_limit * \sum_{i,s} y_{is} \tag{7}$$

$$\sum_{j,s} y_{j,s} \geq \sum_{i,s} y_{is} * min\_keep \tag{8}$$

$$\sum_k y_{k,s} \leq max\_foreigners, \forall s \in S \tag{9}$$

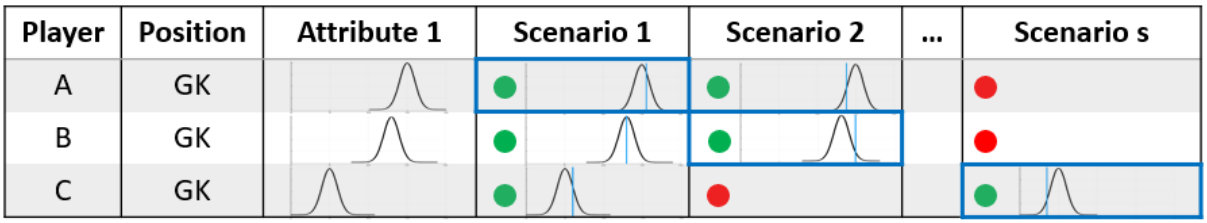
$$\sum_i y_{is} * bool\_pos_{ip} = qt\_pos_p,\ \forall p \in P, s \in S \tag{10}$$

The first stage of the optimization model consists of assembling the roster by buying and selling players, indicated by the binary decision variable $x_i$. The second stage is composed of scenarios of a generic match where each player's performance and availability (related to both injuries and suspensions) varies according to the player's distribution of probability of these events. The recourse variable is the decision of selecting a player to the starting lineup in a match in scenario **s** given his expected performance and availability in that scenario, where a player can be selected only if he was hired in the first-stage decision.

This process can be visualized in Figure 2, where three example goalkeepers are shown. Each player has a distribution probability associated with each attribute, as can be seen in the column Attribute 1 – usually there are multiple attributes being considered simultaneously – as well as for his availability. For each second-stage scenario, the player's availability is determined by a random draw and, if he is healthy and not suspended, his performance in each attribute of interest is calculated via a new random draw, according to his distribution probability. In the example below, if the model had selected those three goalkeepers in the first stage, it would start player A in the first scenario, as all of them are

available and he had the best performance (blue line) in that scenario; in the second scenario player B would be selected, as he had a better random draw than player A despite a worse probability distribution and player C is not available; in the last scenario, player C would start despite his poor performance, as he is the only available goalkeeper. This repeats for all scenarios, determining the model's second-stage decisions and consequently impacting the first-stage decisions.

Figure 2: Second-stage scenarios example




Source: Created by the author

Apart from the data-driven input derived from the data science framework, some parameters can be set. In particular, the set A of attributes can be set beforehand by selecting a subset of all possible on-field attributes as the ones of interest, as it's usually more interesting than optimizing simultaneously for all 184 possible on-field attributes. Besides that, parameters specific to the club in question are set, such as their transfer budget, the maximum average roster age accepted, the minimum number of players that should stay on the roster, the formation used on matches and the minimum value the roster should have on each on-field attribute of interest a.

The objective function (1) seeks to maximize a team's performance in the season. Therefore, it maximizes the sum of each on-field attribute of interest of each player used by the club in the totality of all simulated matches.

Constraint (2) states that the team can't spend more than their budget, where the sum of the market values of each hired player calculates their spending. Constraint (3) relates to the reference values alluded to in the data science section and states that, for each attribute of interest, its mean value in the line-up, averaged for all matches, must be bigger than a certain reference value – the 80% quantile out of all teams in the dataset, for example. Constraint (4) states that a player can only be selected as a start in a match if he is hired by the team. Constraint (5) guarantees that only players who are neither injured nor suspended in a given scenario can be selected for the match. Constraint (6) ensures that the model doesn't hire unnecessary players that won't be used. Constraint (7) states that the average roster age,

weighted by the playing time, must not surpass the limit set on the parameters. Constraint (8) minimizes roster turnover by guaranteeing that a minimum number of players from the previous season will be kept on the roster, also weighted by the playing time. Constraint (9) respects the rules stated in some leagues limiting the number of foreign players playing at the same time. At last, constraint (10) states that the starting lineup must always adhere to the team's formation.

## Results

Two case studies are presented to showcase the model capabilities, one looking for the best possible roster and one applied to a real club's specific parameters. Two data sources were used in those case studies: players' on-field performance statistics and suspensions from yellow and red cards were taken from FBREF (2023); financial data regarding the market value of each player, data regarding players' injuries and players preferred positions were taken from TRANSFERMARKT (2023). Data from both websites were scraped using the worldfootballR (ZIVKOVIC, 2022) package in the R language. All collected data – market value, injury data and different sets of on-field attributes - were then merged based on each player's ID from each site. Columns not relevant or redundant were removed such as not-numerical information.

Data was collected from the 2022/2023 season from the five main leagues in the world - England, Spain, Germany, Italy and France top-flights - for a total of 98 teams. In total, 2562 players had the minimum number of minutes played in the previous season and qualified for the model selection. Each player has data regarding 184 different on-field attributes during the season, which are used as their performance forecast to be used in the model.

For the performance forecast, multiple panel data models were compared via cross-validation. The selected one was a Dynamic Ordinary Least Squares, using as explanatory variables the player's performance on the attribute in the previous season, his age and the age squared. The injury probability forecast is performed with a simple linear regression, where the only explainable variable is how many days on average the player missed in all seasons in his career up until that point. For players with less than 4 seasons played, the forecast is just the average injury probability among this group, as the small sample size could cause distortions. The suspension frequency is the same for every player, as no significant deviations were found from one player to another. The injury and suspension probability are then summed, resulting in the final absence frequency used in the model.

### A) Best possible roster

The first case study hereby presented aims to build the optimal roster with no restrictions. Therefore, constraints related to budget, age, foreigners or a team's current players are not considered. Moreover, the model is applied to different tactical formations and the one with the best results, where the attributes are maximized, is selected. The attributes of interest were selected amongst the 184 possible choices looking to encapsulate a successful

football team in many phases of play and are described in Table 1. The selected roster must be on the top 5% of teams in each one of these attributes - otherwise the model would just focus on increasing the objective value via a subset of them, neglecting other ones, notably the defensive attributes.

Table 1: Attributes of interest and their quantiles in the best roster case study

| Abbreviation | Description | Quantile |
|---|---|---|
| Tkl+Int | Tackles plus interceptions | 95% |
| Clr | Ball clearances | 95% |
| SCA | Shot-creating actions | 95% |
| xA | Expected assists | 95% |
| Dribbles | Successful dribbles | 95% |
| PrgDist | Ball total progressive distance | 95% |
| PrgCarr | Progressive distance by carrying the ball | 95% |
| PrgRec | Progressive distance by receiving the ball | 95% |
| Goals | Goals scored | 95% |
| PSxG+- | Goals conceived by the goalkeeper relative to the expected | 95% |

Source: Made by the author

As mentioned, the purpose of those attributes is to select meaningful factors on the field that are good predictors of performance for the following season. The first two attributes on the table are included to represent a team’s defensive capabilities and its ability to recoup possession. Attributes SCA and xA measure a player's capability to create goal chances for his team. Attributes relative to dribbles and progressive distance are indicators of a player's skill in moving the ball forward and putting the team in a better position on the field. The number of goals scored by the player is also included as it is important to have reliable scorers on the team. At last, PSxG+- is a goalkeeper performance indicator, included to assure that the ideal goalkeeper is hired.

For the case study, 100 match scenarios were used, where the available players change from one to another due to stochastic injuries and suspensions. Figure 3 presents the most used players in the suggested roster. For each position in the field, the player with the most appearances throughout the scenarios is shown, as well as the number of matches he has played.

Figure 3: Most frequent players and number of matches played for the best roster case study


Source: Made by the author

The most notable characteristic of the selected players is their age: the average on the roster was 22.3 years-old and the oldest player in the starters is Mbappé, who is 24 years-old. This happened due to the performance forecast, which prioritized younger players, as well as for those who had great numbers in the previous season, which explains Mbappé being the older one. As there was no budget constraint, the model selected a numerous roster with expensive luxury players on the bench, such as Haaland and Neymar.

The roster's total value was 2.07 billion euros, including some players who started only a few matches. Out of this total, only 19.7% was spent in the starting line-up, the players shown in Figure 3, as the model opted for having a lot of options at each match to select the best one in that scenario. This can be confirmed by observing the number of matches played by the starters, as a lot of them played in less than half of the matches, as other players on the roster had better performances in those particular scenarios. One noticeable exception is Bellingham, a young player who had amazing performance in the previous season and played 84 out of 86 matches where he was available, with no injuries or suspensions.

It can also be seen in Figure 3 that the displayed formation is a 3-4-3. In fact, this was the formation that resulted in the best objective value among many formations tested – Table

2 presents this value for each one of them, as well as the total cost of each alternative. It can be noted that both 4-3-3 and 4-4-2 formations yield a similar objective value, but save hundreds of millions of Euros when compared to the selected 3-4-3.

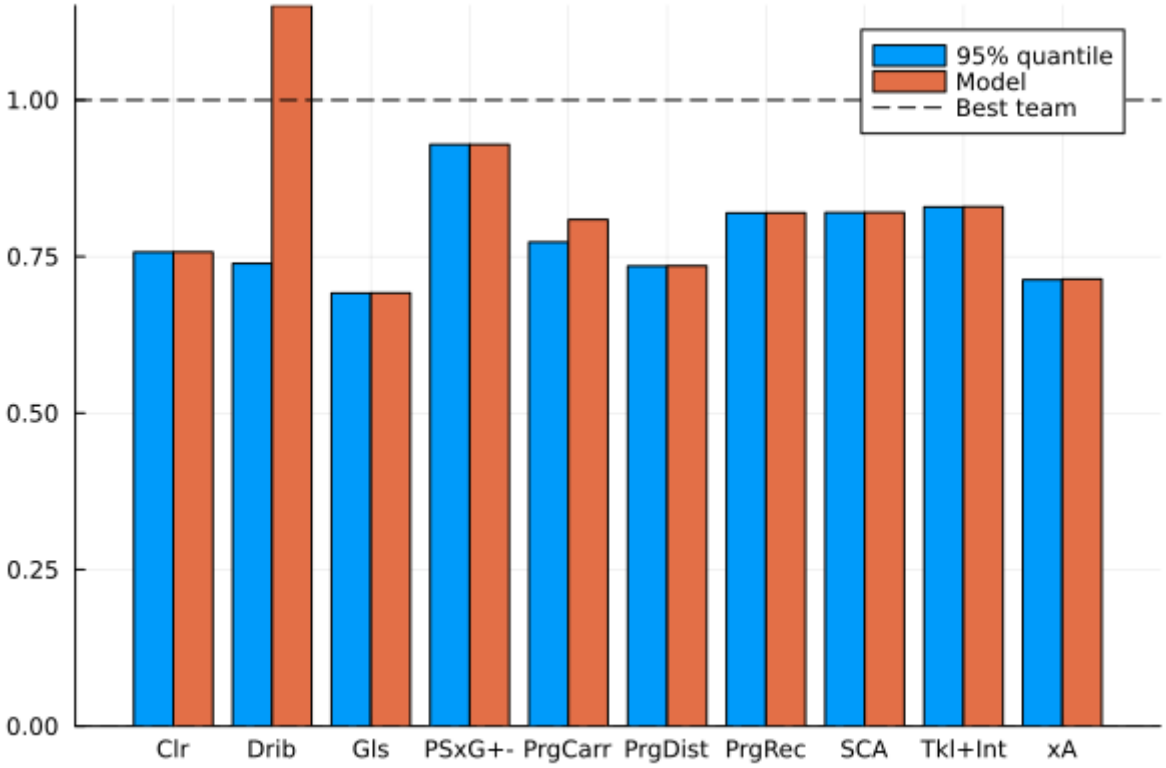
Table 2: Objective value and total cost for each formation in the best roster case study

| Formation | Objective value | Total cost (millions of Euros) |
|---|---|---|
| 4-1-4-1 | 3.044 | 1897.2 |
| 3-4-3 | 3.089 | 2068.2 |
| 4-3-3 | 3.067 | 1717.5 |
| 4-4-2 | 3.072 | 1747.5 |
| 3-5-2 | 3.023 | 1944.7 |

Source: Made by the author

To summarize this case study results, Chart 1 compares the suggested roster performance on the selected attributes to the 95% quantile team in each attribute from the previous season across the leagues. As the attributes are normalized, the horizontal line at 1.0 allows also the comparison to the best team in each attribute. The selected roster manages to beat the 95% quantile in each attribute, as specified by the model constraints. In particular, the selected roster is more than 50% better than the best team at successful dribbles, which boosted the objective value.

Chart 1: Roster attributes performance compared to the 95% and 100% quantile




Source: Made by the author

**B) Monaco case study**

The second case study presented in this section is an application of the model to a real-world club situation, which is arguably the most common use case for the model. The club selected was Monaco, as they changed their manager for the upcoming 2023/2024 season to Adi Hütter, a coach with a distinct style of play that the previous Monaco roster wasn't quite familiar with.

In particular, the new style of play is heavily focused on high pressure on the opposition's build-up. This can be confirmed by comparing Monaco's performance in the first 10 matches of the 2023/24 season to the previous season. In 2022/2023, Monaco was in the middle of the pack in Ligue 1 in high-pressure statistics, attempting 2.6 tackles per game on the offensive third of the field and creating 0.45 shots per game via defensive actions. With the new coach, Monaco became the third team in the league with the most attempted tackles in offense with 2.9 and the best in creating shots via defensive actions with 0.8 – an increase of 8.2% and 77.8%, respectively. With that in mind, the parameters described in Table 3 were

constructed, giving high priority to those aforementioned attributes and also including other ones pertinent to the new coach style.

Table 3: Attributes of interest and their quantiles in the Monaco case study

| Abbreviation | Description | Quantile |
|---|---|---|
| 3rd Tackles | Tackles attempted at the offensive third | 95% |
| Def_SCA | Shots created via defensive actions | 95% |
| Goals | Goals scored | 75% |
| PSxG+- | Goals conceived by the goalkeeper relative to the expected | 75% |
| 3rd Carries | Caries into the offensive third | 75% |
| PrgRec | Progressive distance by receiving the ball | 75% |
| PrgDist | Ball total progressive distance | 75% |
| KP | Key passes, which led to a shot | 75% |

Source: Made by the author

The other parameters for this case study were defined based on Monaco's real-life transfers. The budget is 344.2 million euros, which is the total roster value in real life. Considering every minute played by Monaco's player in the 2023/2024 season, the average age weighted by playing time is 25.8 years old and the proportion of those minutes played by players that were in the squad in the previous season was 71%. Therefore, the age limit and minimum playing time of former players were set to those values. The foreign player limit of Ligue 1 is 4 players at the same time on the field with players outside Europe or Africa being considered foreigners, so that rule is enforced by the model. At last, the new coach also changed the team formation to 3-4-3, so the case study was performed with that formation. Once again, 100 match scenarios were considered in the model and the most used players in the selected roster are presented in Figure 4:

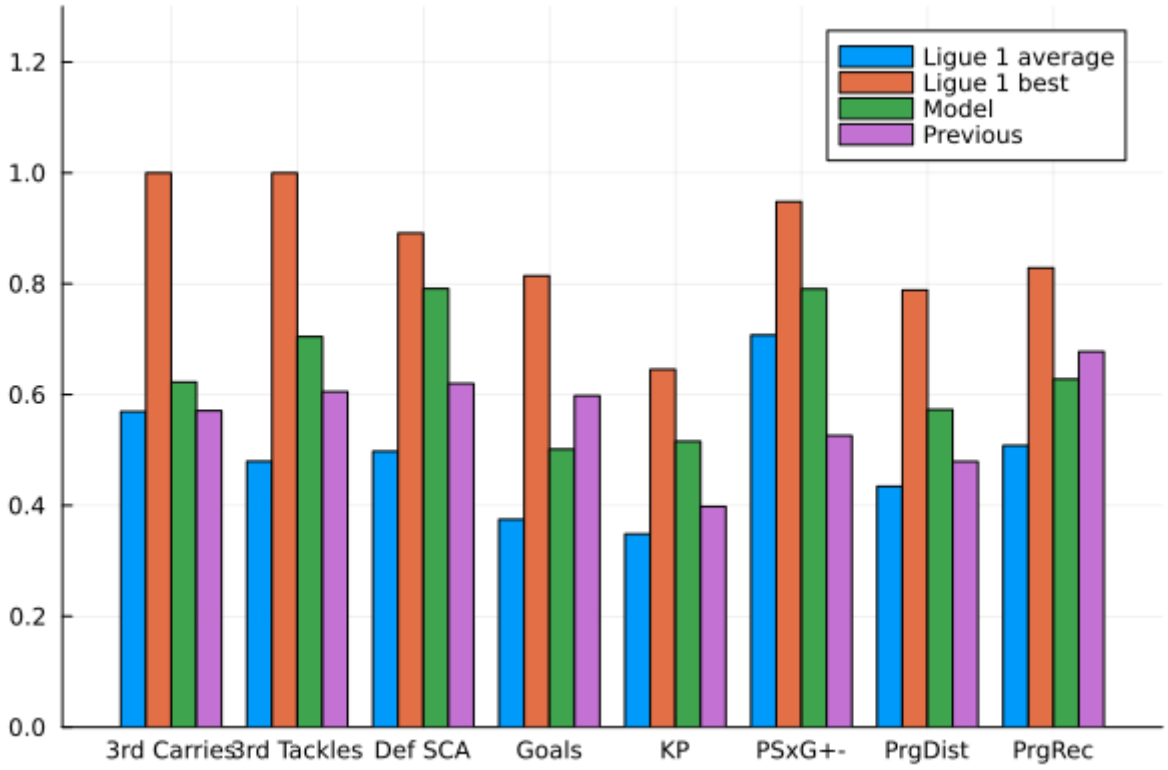
Figure 4: Most frequent players and number of matches played for the best roster case study


Source: Made by the author

Out of the most used players, the only ones who weren't on Monaco's roster in the previous season are Yehvann Diouf and Jeremie Boga. As the previous case study, the model once again opted for young players, which have growth projection in the selected attributes to next season. The average age of the players weighted by the playing time was 23.7 years-old, which is below the established limit of 25.8, meaning that this constraint wasn't active in the optimal solution, as the model prioritized by itself younger players.

In total, the roster was constituted of 41 players for a total value of 343.7 million euros, almost the entire budget. The starters shown in Figure 4 composed a little over half of the total value at 198 million euros (57.6%). Former Monaco players had 71% of the total playing time, again just below the constraint, between starters from the previous season, bench players that earned many more minutes such as Matsima and Matazo. Chart 2 compares the selected roster's performance to the Ligue 1's best and average teams at each attribute, as well as to Monaco's former roster:

Chart 2: Roster attributes performance compared to the best, average and former squad

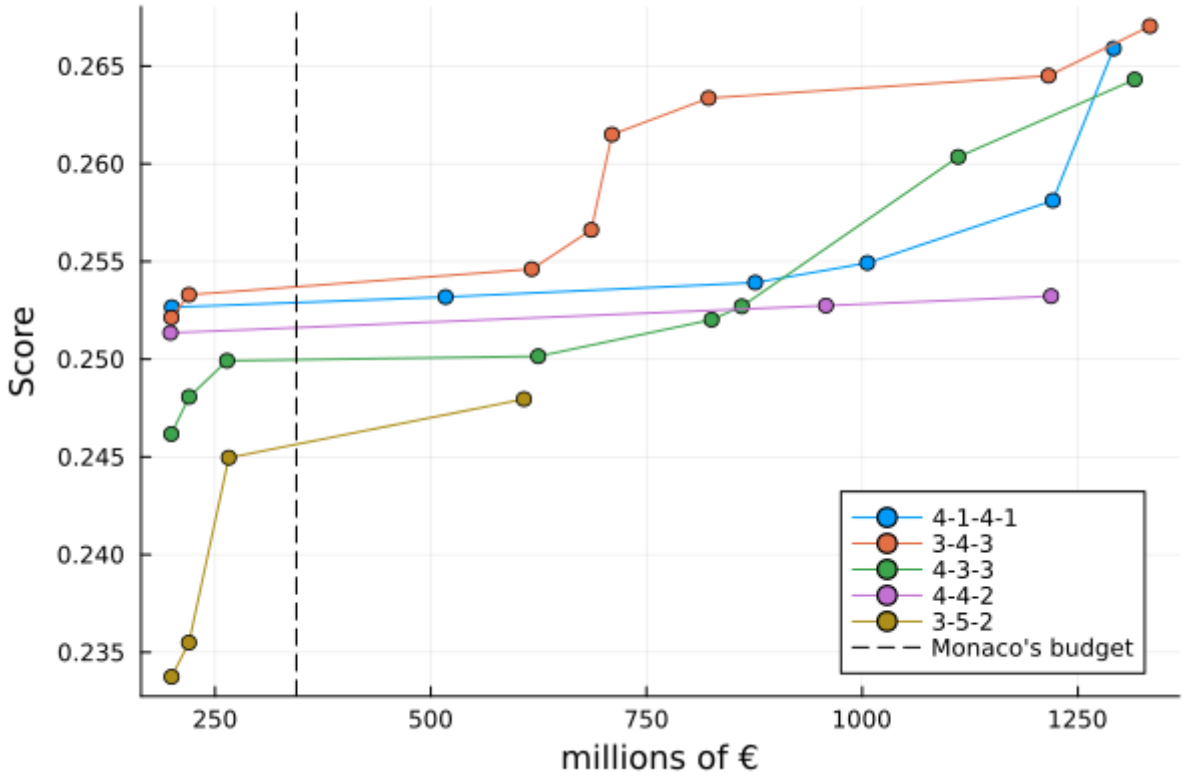



Source: Made by the author

The selected roster performed better than average and than Monaco's previous roster in most attributes, including both high-pressure attributes that were crucial to the new coach. This time, the model didn't focus in a single attribute to explore the objective value maximization, but instead composed a balanced team, always between the average and the best team in Ligue 1.

A sensibility analysis is also presented in Chart 3, where the available budget and the formation used were tested with different values. In general, 3-4-3 is the best option, with 4-1-4-1 coming close in some budget ranges. In particular, that is the best option for Monaco's case study, represented with the dashed vertical line, which shows that their formation used in real life is a good choice for their style of play and available budget.

Chart 3: Sensibility analysis for the budget and formation




Source: Made by the author

## Conclusion

The model presented in this paper is a tool that can be used by football teams and fans to debate transfer decisions, perform sensibility analysis on parameters and discuss formation options for the roster-building problem from real-world data. The mathematical model resembles Payyappalli and Zhuang (2019) but proposes a significant innovation in the inclusion of second-stage scenarios, where each player's performance and availability are uncertain, which affects the first-stage roster decisions. This stochastic approach to the roster-building problem applied to real-world data has no equivalent in the current literature and brings a realistic and accurate facet to transfer decisions.

This work presents not only a novel stochastic formulation for the roster-building problem, but also proposes a data science framework for data collection, treatment and forecasting to be inputted into the model, assuring the data-driven optimization approach. The pipeline is general and flexible, allowing it to be customized with different data sources, forecasting methods and parameters and thus providing a useful tool for football studies and discussions at any level, using any desired set of data and parameters.

Case studies were performed that showcased the model's capabilities in finding a roster adequate to the user parameters considering the players' availability and the decisions of how deep should the roster be. Case study A showed model results in a more unconstrained situation, whereas case study B was an application to a team using real-world parameters, which demonstrated how the model was able to meet all attribute requirements and optimize their value in a scenario with tighter constraints. The model is extremely general and can be equivalently applied to teams in any situation by just altering its parameters.

Some limitations in the case studies happened due to data scarcity. In particular, the model's budget constraint could also consider each player's salary impact, which isn't present in the mathematical formulation as there was no data available on that matter. Another potentially impactful change could be to consider the players' ability to play in multiple positions – with a possible downgrade in performance when playing out of his preferred position, which was also not present in the data sources used. This could be important when substituting unavailable players in some scenarios, as the model would have greater flexibility to use the players in different positions and thus be more effective.

Besides such improvements in the data science framework and mathematical formulation, a great extension to the model would be to develop an app that automatizes every step. This model would be connected to the mathematical formulation and allow users online

to customize the parameters to their specifications and observe the model results, with its suggested roster and statistics and comparisons of the new squad. Therefore, the model could be easily accessed by interested parts across the world, popularizing it as a tool to aid decision-making in football transfers.

## References


AHMED, F.; DEB, K.; JINDAL, A. Multi-objective optimization and decision making approaches to cricket team selection. **Applied Soft Computing**, v. 13, n. 1, p. 402–414, jan. 2013.

AL-SHBOUL, R. et al. Automated Player Selection for a Sports Team using Competitive Neural Networks. **International Journal of Advanced Computer Science and Applications (ijacsa)**, v. 8, n. 8, 23 fev. 2017.

BECKER, A.; SUN, X. A. An analytical approach for fantasy football draft and lineup management. **Journal of Quantitative Analysis in Sports**, v. 12, n. 1, 1 jan. 2016.

BOON, B. H.; SIERKSMA, G. Team formation: Matching quality supply and quality demand. **European Journal of Operational Research**, v. 148, n. 2, p. 277–292, jul. 2003.

CAO, A. et al. Team-Builder: Toward More Effective Lineup Selection in Soccer. **IEEE Transactions on Visualization and Computer Graphics**, p. 1–16, 2022.

DEB, S.; DAS, S. Optimal selection of the starting lineup for a football team. 2023.

EKSTRAND, J.; HÄGGLUND, M.; WALDÉN, M. Injury incidence and injury patterns in professional football: the UEFA injury study. **British Journal of Sports Medicine**, v. 45, n. 7, p. 553–558, 1 jun. 2011.

FBREF. **Estatísticas e histórico do futebol**. Disponível em: fbref.com. Acesso em: 8 out. 2023.

FIFA. **Global Transfer Report**, 2023.

FRY, M. J.; LUNDBERG, A. W.; OHLMANN, J. W. A Player Selection Heuristic for a Sports League Draft. **Journal of Quantitative Analysis in Sports**, v. 3, n. 2, 19 jan. 2007.

HÄGGLUND, M. et al. Injuries affect team performance negatively in professional football: an 11-year follow-up of the UEFA Champions League injury study. **British Journal of Sports Medicine**, v. 47, n. 12, p. 738–742, ago. 2013.

JARVANDI, A.; SARKANI, S.; MAZZUCHI, T. Modeling team compatibility factors using a semi-Markov decision process: a data-driven approach to player selection in soccer. **Journal of Quantitative Analysis in Sports**, v. 0, n. 0, p. 1–20, 14 jan. 2013.

KAPLAN T. When It Comes to Stats, Soccer Seldom Counts. **The New York Times**, 2010

KUPER, S.; SZYMANSKI, S. **Soccernomics**. Nation Books, 2009.

LICHTENTHALER, U. Mixing data analytics with intuition: Liverpool Football Club scores with integrated intelligence. **Journal of Business Strategy**, v. 43, n. 1, p. 10–16, 3 jan. 2022.

MAHRUDINDA, M. et al. Optimization of The Best Line-up in Football using Binary Integer Programming Model. **International Journal of Global Operations Research**, v. 1, n. 3, p. 114–122, 5 ago. 2020.

MUNIZ, M.; FLAMAND, T. Sports analytics for balanced team-building decisions. **Journal of the Operational Research Society**, v. 74, n. 8, p. 1892–1909, 3 ago. 2023.

NASIRI, M. M. et al. A novel hybrid method for selecting soccer players during the transfer season. **Expert Systems**, v. 36, n. 1, p. e12342, fev. 2019.

OZCEYLAN, E. O. A MATHEMATICAL MODEL USING AHP PRIORITIES FOR SOCCER PLAYER SELECTION: A CASE STUDY. **South African Journal of Industrial Engineering**, v. 27, n. 2, ago. 2016.

PANTUSO, G. The Football Team Composition Problem: a Stochastic Programming approach. **Journal of Quantitative Analysis in Sports**, v. 13, n. 3, p. 113–129, 26 set. 2017.

PANTUSO, G.; HVATTUM, L. M. Maximizing performance with an eye on the finances: a chance-constrained model for football transfer market decisions. **TOP**, v. 29, n. 2, p. 583–611, jul. 2021.

PAPPALARDO, Luca. A comprehensive data-driven evaluation of soccer players performance, 2018.

PAYYAPPALLI, V. M.; ZHUANG, J. A data-driven integer programming model for soccer clubs' decision-making on player transfers. **Environment Systems and Decisions**, v. 39, n. 4, p. 466–481, dez. 2019.

QADER, M. A. et al. A methodology for football players selection problem based on multi-measurements criteria analysis. **Measurement**, v. 111, p. 38–50, dez. 2017.

SMITH, Z. J.; BICKEL, J. E. A roster construction decision tool for MLS expansion teams. **Journal of Quantitative Analysis in Sports**, v. 19, n. 1, p. 1–14, 28 mar. 2023.

TAVANA, M. et al. A fuzzy inference system with application to player selection and team formation in multi-player sports. **Sport Management Review**, v. 16, n. 1, p. 97–110, 1 jan. 2013.

TRANSFERMARKT. **Mercado de transferências, rumores, valores**. Disponível em: fbref.com. Acesso em: 8 out. 2023.

VIJAY FIDELIS, J.; KARTHIKEYAN, E. Optimization of Artificial Neural Network Parameters in Selection of Players for Soccer Match. Em: AURELIA, S. et al. (Eds.). **Sustainable Advanced Computing**. Singapore: Springer Singapore, 2022. v. 840p. 275–288.

YU, S. et al. Discovering a cohesive football team through players' attributed collaboration networks. **Applied Intelligence**, v. 53, n. 11, p. 13506–13526, jun. 2023.

ZHAO, H. et al. Multi-Objective Optimization for Football Team Member Selection. **IEEE Access**, v. 9, p. 90475–90487, 2021.

ZIVKOVIC, J. et al. **worldfootballR: Extract and Clean World Football (Soccer) Data**. , 26 nov. 2022. Disponível em: <https://cran.r-project.org/web/packages/worldfootballR/index.html>. Acesso em: 8 out. 2023