DAT-3113 PROJECT

FIFA North America

AUTHOR
Jose Almanza

PUBLISHED
January 1, 2023

INTRODUCTION

My position as the FIFA organization's data analyst is particularly relevant to the North American region, which includes leagues such as Major League Soccer (MLS) and Liga MX. Both leagues have rich soccer tradition and a great fan base, this region creates a deep base of players, clubs, and competitions.The Major League Soccer, has evolved into a creation of foreign talent, promoting a diversified and competitive playing field.Also, the Liga MX, Mexico's top-tier league, is a powerhouse in the North American football ecosystem, known for its fervent fan culture and technical brilliance on the field. Therefore, it is understandable the interest on getting a basic profile of the players from the region, and understand how the variables associated with a player's wage and their value. We will try to understand not only the complexities of player profiles, but also indicate aspects impacting players earnings and values.As we go into the area of data analytics, our goal is to uncover patterns and insights that not only suit our client's immediate requirements but also contribute to the ongoing conversation about player dynamics in the MLS and Liga MX.

PART A: DATASET

This dataset includes players from the Major League Soccer (MLS) and Liga MX, two of North America's most famous professional soccer leagues. It is a detailed portrayal of players active in these leagues, taken from the FIFA 22 video game. It includes a large roster that closely resembles the actual player profiles found in MLS and Liga MX. This dataset is a good resource for people interested in investigating the digital representation of North American soccer players in the context of FIFA 22.This dataset includes the following variables:

short_name - Name of the player player_positions - Positions the player can play overall - Player's overall rating (1-100) potential - Highest expected overall rating (1-100) value_eur - Player's value in euros wage_eur - Player's wage in euros age - Player's age height_cm - Player's height (cm) weight_kg - Player's weight (kg) club_name- Current club league_name- Name of the league league_level - Level of the league (1= first division, 2= second division, etc.) club_position - Player's position within the club club_joined - Date when the player joined the current club nationality_name - Nationality of the player preferred_foot - Favorable foot skill_moves - Agility rating (1-5) international_reputation - Player's international reputation rating (1-5) player_traits - Player's special traits pace - Speed rating (1-100) shooting - Shooting rating (1-100) passing - Passing rating (1-100) dribbling - Dribbling rating (1-100) defending - Defending rating (1-100) physic - Physicality rating (1-100)

URL for the dataset https://www.kaggle.com/datasets/avizyt/fifa22-detail-dataset-male-and-female-updated.

```
[1] 1259   26
```

This section was used to to load the data set into the Quarto document, and from now it will be refered as “northAmerica”.Some packages were also loaded.The “dim” function was used to see the dimensions of the data set. However, we decided to not use any other functions as “table” or “head” because they offer a general overview of the data, showing all the variables, which we considered it is not relevant at this point and could cause confusion or distraction.

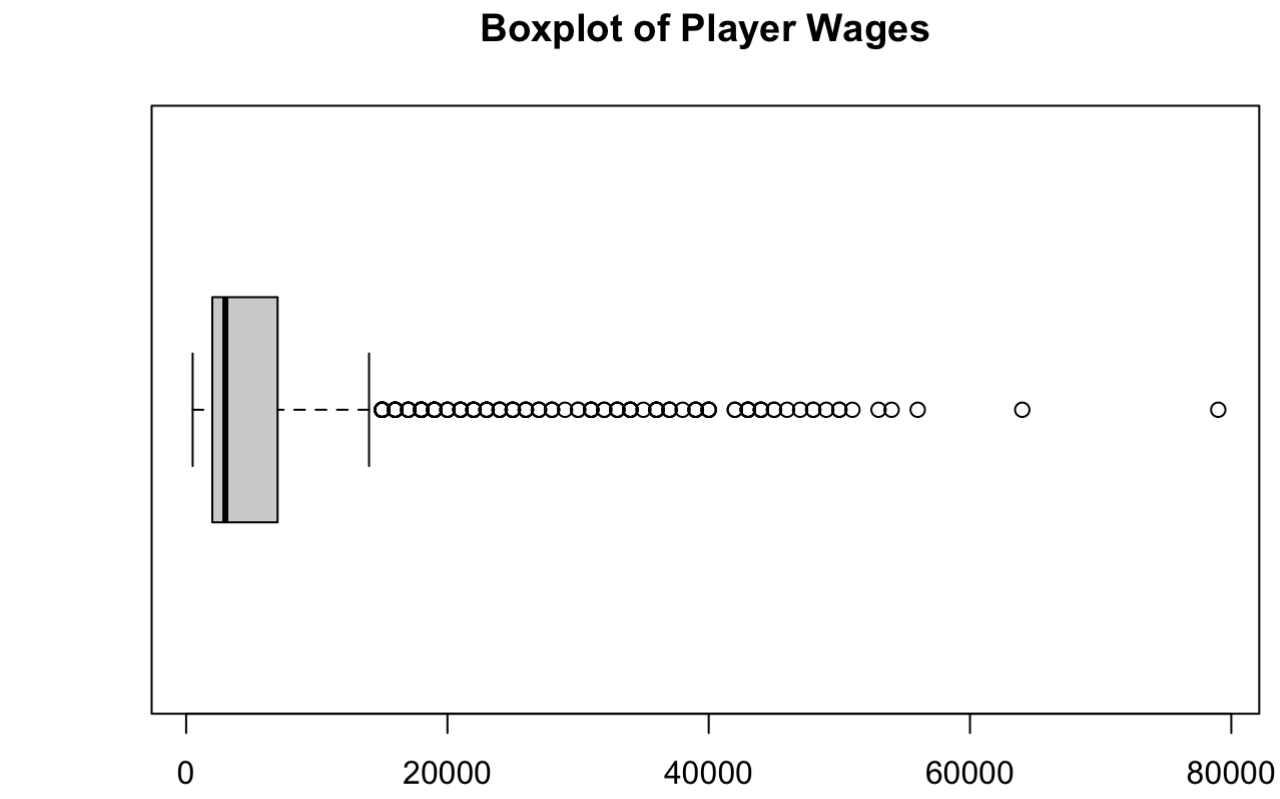
In the following section, we will prepare the data for the future analysis.

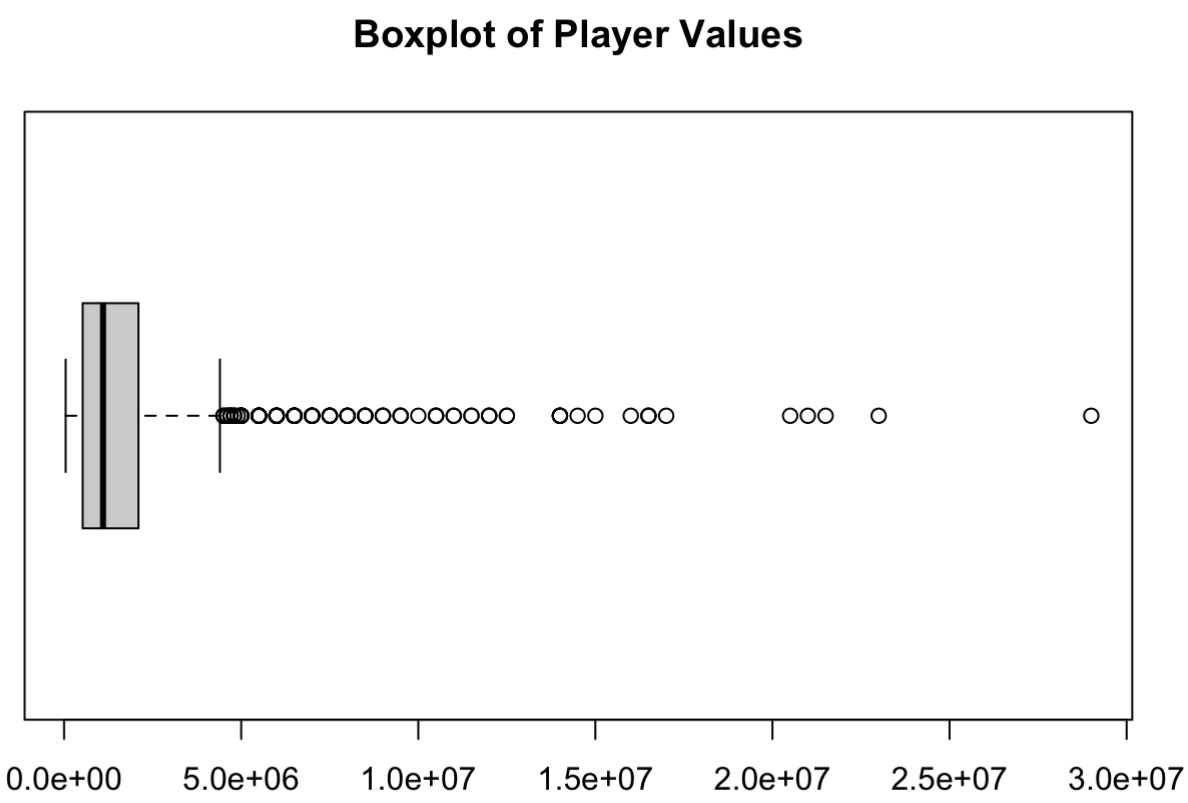
PART B: DATA WRANGLING, MUNGING, AND PREPARATION

```
tibble [1,259 × 26] (S3: tbl_df/tbl/data.frame)
 $ sofifa_id               : num [1:1259] 169416 139068 153244 167664 140233 ...
 $ short_name              : chr [1:1259] "C. Vela" "Nani" "A. Gignac" "G. Higuaín" ...
 $ player_positions        : chr [1:1259] "RW, LW, CAM" "LW, LM" "ST, CF" "ST" ...
 $ overall                 : num [1:1259] 83 81 81 81 80 80 80 80 80 80 ...
 $ potential               : num [1:1259] 83 81 81 81 80 80 80 80 81 80 ...
 $ value_eur               : num [1:1259] 29000000 12000000 11000000 16500000 2300000
16500000 20500000 21000000 16000000 21500000 ...
 $ wage_eur                : num [1:1259] 17000 14000 50000 14000 27000 12000 13000 64000
43000 14000 ...
 $ age                     : num [1:1259] 32 34 35 33 35 32 29 28 30 28 ...
 $ height_cm               : num [1:1259] 177 177 189 186 190 170 170 179 194 170 ...
 $ weight_kg               : num [1:1259] 77 70 85 89 85 69 63 70 83 69 ...
 $ club_name               : chr [1:1259] "Los Angeles FC" "Orlando City Soccer Club"
"Tigres U.A.N.L." "Inter Miami CF" ...
 $ league_name             : chr [1:1259] "USA Major League Soccer" "USA Major League
Soccer" "Mexican Liga MX" "USA Major League Soccer" ...
 $ league_level            : num [1:1259] 1 1 1 1 1 1 1 1 1 1 ...
 $ club_position           : chr [1:1259] "RW" "ST" "RS" "ST" ...
 $ club_joined             : Date[1:1259], format: "2018-01-01" "2019-02-19" ...
 $ nationality_name        : chr [1:1259] "Mexico" "Portugal" "France" "Argentina" ...
 $ preferred_foot          : chr [1:1259] "Left" "Right" "Right" "Right" ...
 $ skill_moves             : num [1:1259] 4 5 3 3 1 4 4 4 1 3 ...
 $ international_reputation: num [1:1259] 3 4 1 4 1 3 2 1 1 2 ...
 $ player_traits           : chr [1:1259] "Leadership, Finesse Shot, Flair, Playmaker
(AI), Outside Foot Shot, Technical Dribbler (AI)" "Injury Prone, Flair, Outside Foot
Shot, Technical Dribbler (AI)" "Leadership, Finesse Shot, Long Shot Taker (AI), Power
Header, Outside Foot Shot" "Leadership, Outside Foot Shot" ...
 $ pace                    : num [1:1259] 80 70 68 58 NA 74 65 79 NA 87 ...
 $ shooting                : num [1:1259] 82 77 83 84 NA 73 75 79 NA 80 ...
 $ passing                 : num [1:1259] 78 81 73 71 NA 83 80 79 NA 61 ...
 $ dribbling               : num [1:1259] 84 85 75 78 NA 81 83 80 NA 79 ...
 $ defending               : num [1:1259] 31 37 41 29 NA 49 52 41 NA 25 ...
 $ physic                  : num [1:1259] 67 68 84 71 NA 74 66 68 NA 67 ...
```

```
[1] 1511
```

sofifa_id	short_name	player_positions
0	0	0
overall	potential	value_eur
0	0	0
wage_eur	age	height_cm
0	0	0
weight_kg	club_name	league_name
0	0	0
league_level	club_position	club_joined
0	0	56
nationality_name	preferred_foot	skill_moves
0	0	0
international_reputation	player_traits	pace
0	603	142
shooting	passing	dribbling
142	142	142
defending	physic	
142	142	






```
# A tibble: 3 × 2
  club_position avg_wage
  <chr>            <dbl>
1 Reserves         2546.
2 Starters         9708.
3 Subs             5736.
```

```
16,001–32,000 32,001–48,000 48,001–64,000    500–16,000 64,001–79,000
           69            43             8          1138             1
```

```
16,001–32,000 32,001–48,000 48,001–64,000    500–16,000 64,001–79,000
           69            43             8          1138             1
```

```
11,613,001–23,181,000 23,181,001–29,000,000      45,000–5,829,000
                   20                      1                  1185
 5,829,001–11,613,000
                   53
```

```
  0   1
961 298
```

A series of actions are performed as part of the data wrangling, munging and preparation process.This is a critical stage since it goes over the cleaning, organization, and translation of raw data into analysis ready format.In order to make our dataset more useful we investigate its structure, manage missing values, transform data types, and new variables. First, I examined the type of data that I was working with and look for any missing values to make sure there is no missing values within the variables we will be working with.After this, we use the box plot function for outliers to be identified.Following this, we transform the wage and value variables that we will be working with through the project into numeric format. Following that, I carefully classified players according to their current roles on their teams as "Starters," "Subs," or "Reserves." At the same time, I divided them into age groups, distinguishing between "Young," "Prime," "Experienced," and "Old," which corresponded to different periods of their careers. In addition, I converted numerical salary and value representations into categorical values, increasing the granularity of our dataset.By categorizing and organizing the data, we lay the foundation for more sophisticated analyses and visualizations that we will be performing later on.

PART C: SUMMARY MEASURES AND DATA VISUALIZATION

Summary Measures

```
   vars    n    mean      sd median trimmed    mad min   max range skew
X1    1 1259 6730.14 9423.07   3000 4484.64 2965.2 500 79000 78500 2.97
   kurtosis     se
X1    10.23 265.57

   vars    n    mean      sd  median trimmed     mad   min     max    range
X1    1 1259 1881219 2625960 1100000 1334688 1000755 45000 2.9e+07 28955000
   skew kurtosis       se
X1 4.23    24.99 74007.42

[1] 0.6089951

[1] 0.1221675

[1] 0.2937364
```

In this section I examined key statistics for my variables Wage and Value to acquire a full knowledge of both variables. Using the describe function I was able to retrieve basic summary metrics such as mean, median, minimum, maximum, and quartiles. This measures will provide me with information about central tendency and dispersion overview of the player's earning and value. Some of the most important insight taken out for the wage variables are that with 9,423.07 as standard deviation coefficient demonstrate a high dispersion of the earning among the players. Also, the mean gives us information about what an average player in North America is earning which is 6,730.14. With an rightward skewness of 2.97,showing a longer right tail and a concentration of individuals with comparatively lower incomes, while a few high earners contribute to the extended tail.This skewness highlights the salary distribution's asymmetry, reflecting the existence of outliers at the top. Furthermore, the kurtosis value of 10.23 indicates a leptokurtic distribution, this means that the dataset contains a higher concentration of extreme values, which could indicate a higher incidence of players with substantially high or low

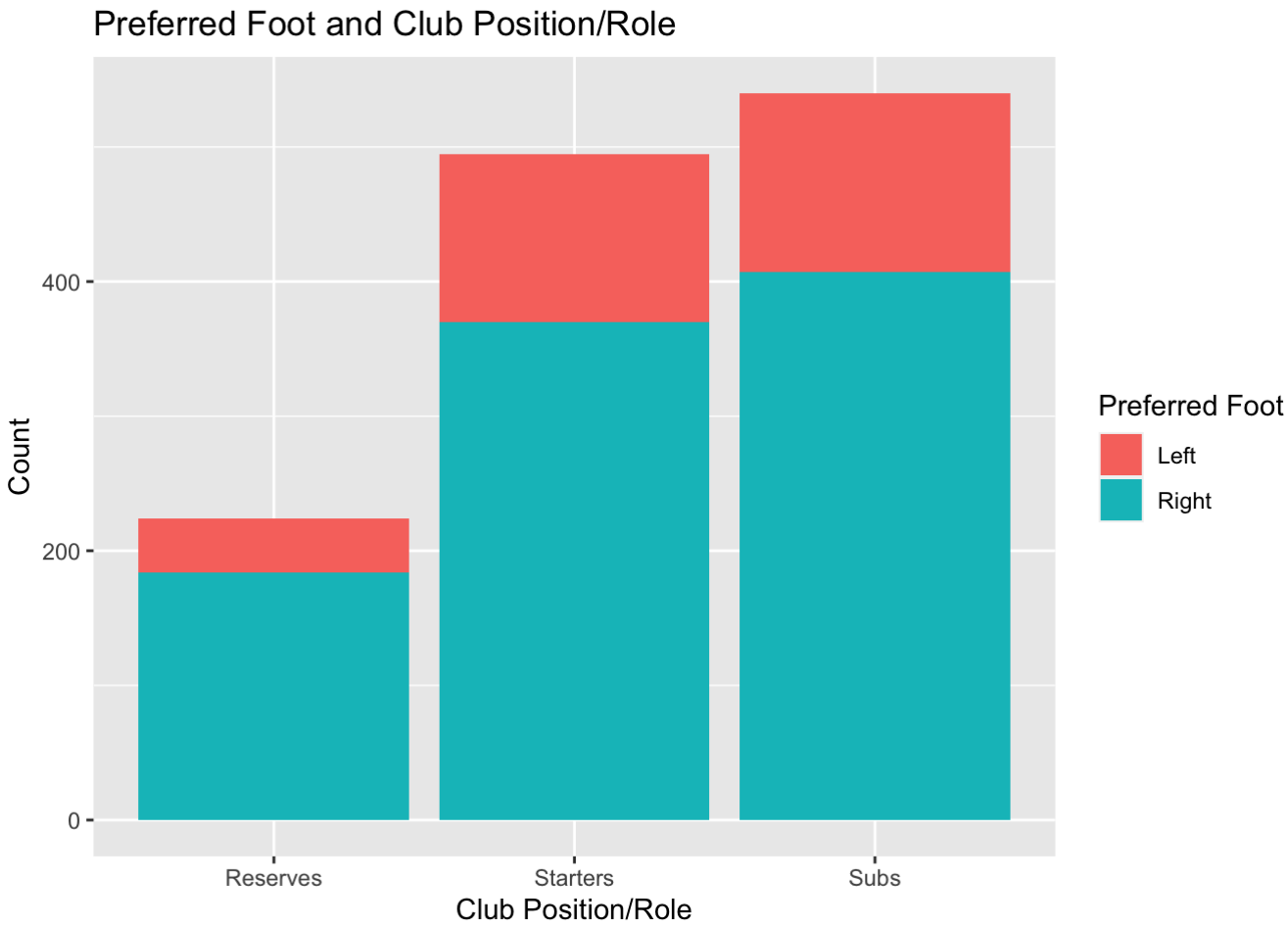
incomes. On the other hand, for the value variable we are able to determine the average worth of players that is roughly 1,881,219, demonstrating a significant financial valuation among this cohort. The standard deviation (sd) of 2,625,960 indicates significant variation in worth, highlighting the different financial profiles of players, with some commanding substantially greater valuations than others.The dataset also has a wide range of financial appraisals, with a minimum value of 45,000 and a maximum value of 29,000,000. The value variable in North American players' financial data from the FIFA dataset has a kurtosis value of 24.99 and a skewness of 4.23, indicating that the distribution has different characteristics. The positive skewness indicates that the distribution is skewed to the right, implying that there are more players with lower financial valuations, but a few players command extremely high valuations, resulting in a longer right tail. Finally we were able to test some correlations,we tested the correlation between Wage and Value where we got a coefficient of 0.6089951 which indicates a moderately strong positive relationship between the two variables. This suggests that as player wages increase, their market values tend to rise as well. Also, we tested how age may affect this two variables,which with a correlation of 0.1221675, this means slight tendency for older players to have higher market values, but the relationship is not very weak. Finally we tested the correlation between wage and age and we got a coefficient of 0.2937364 which suggests that older players tend to have higher wages, however this correlation is not really strong.

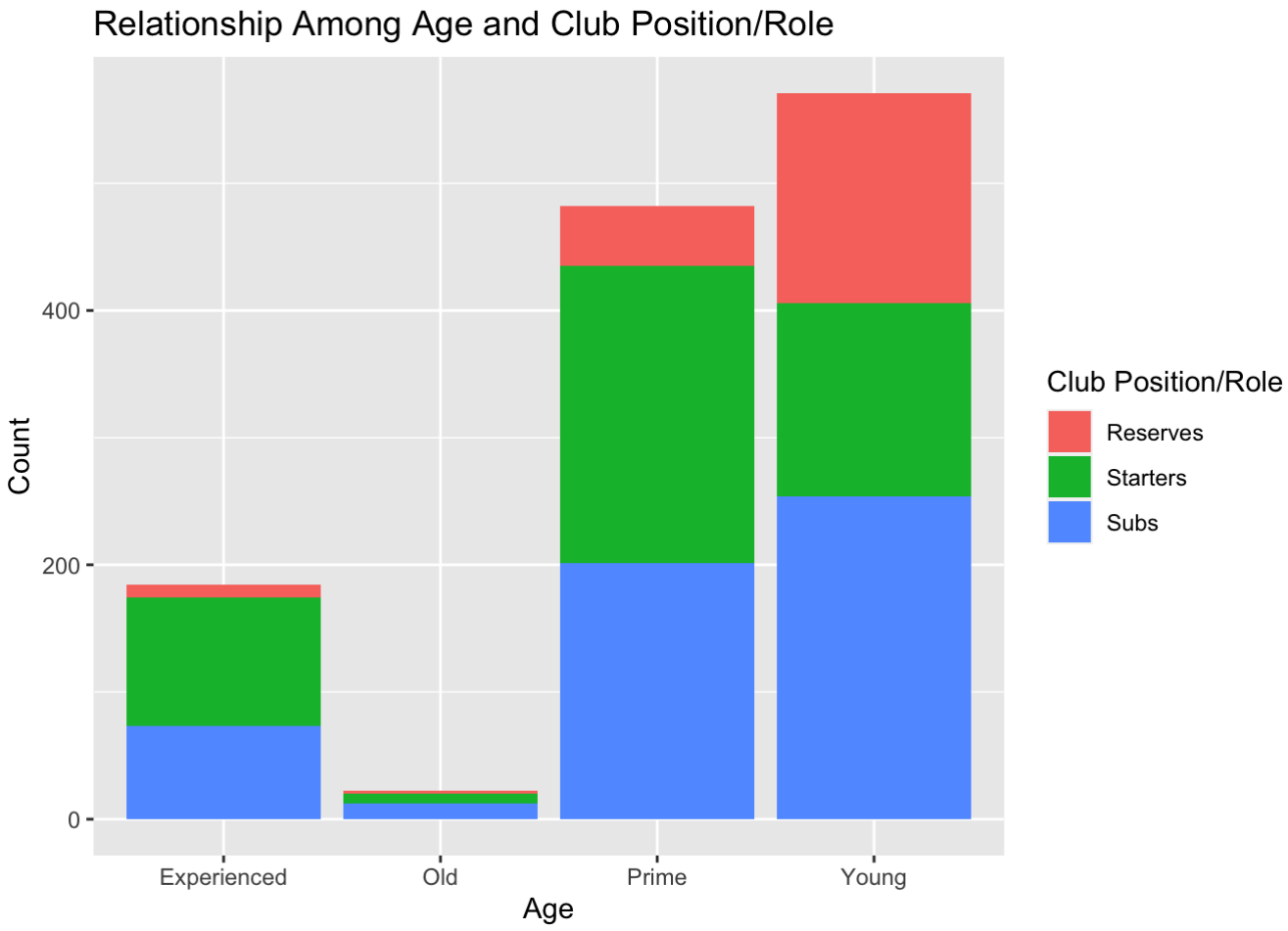
Data Visualization

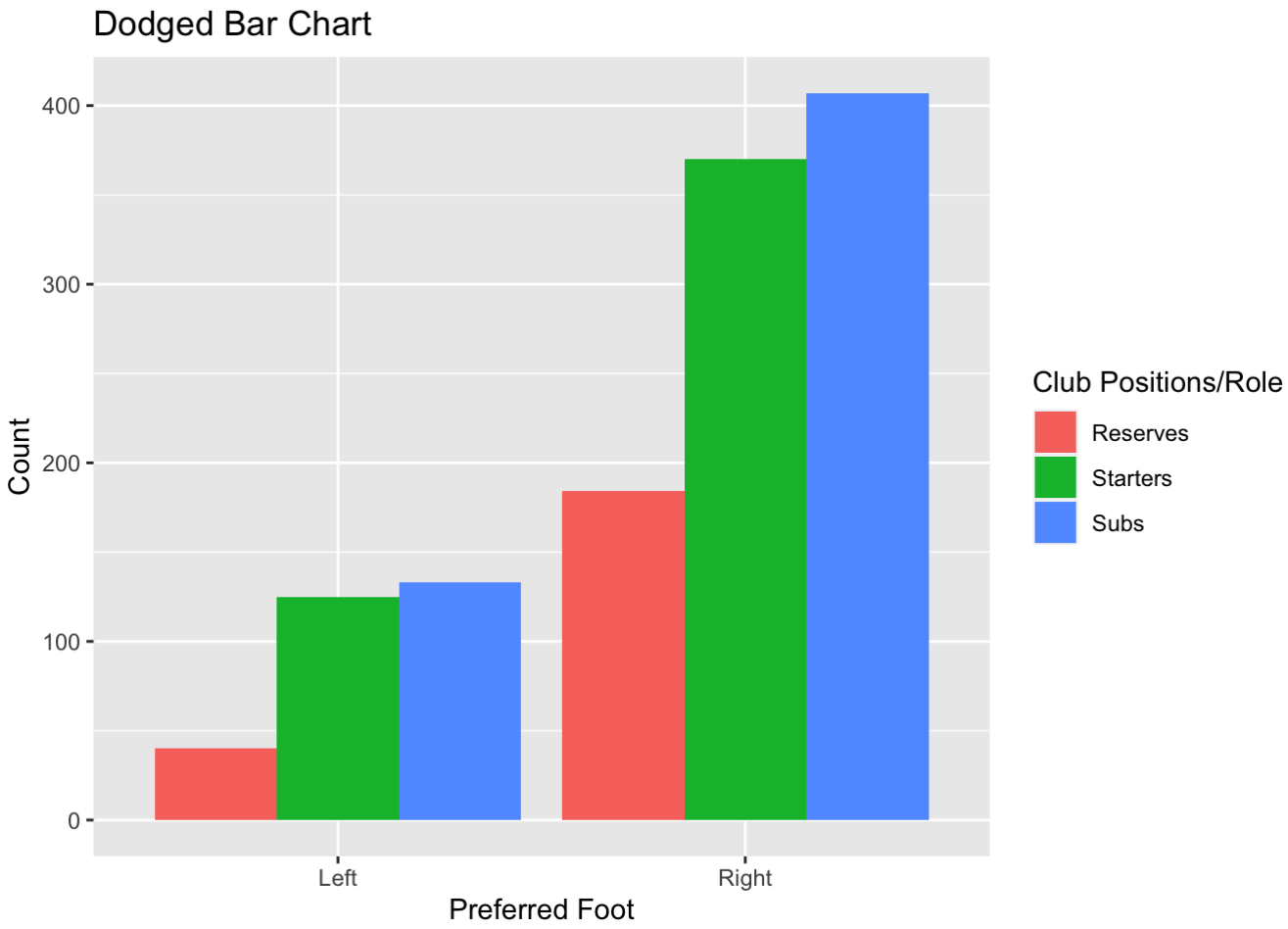
	16,001–32,000	32,001–48,000	48,001–64,000	500–16,000	64,001–79,000
Reserves	2	0	1	221	0
Starters	40	30	6	418	1
Subs	27	13	1	499	0

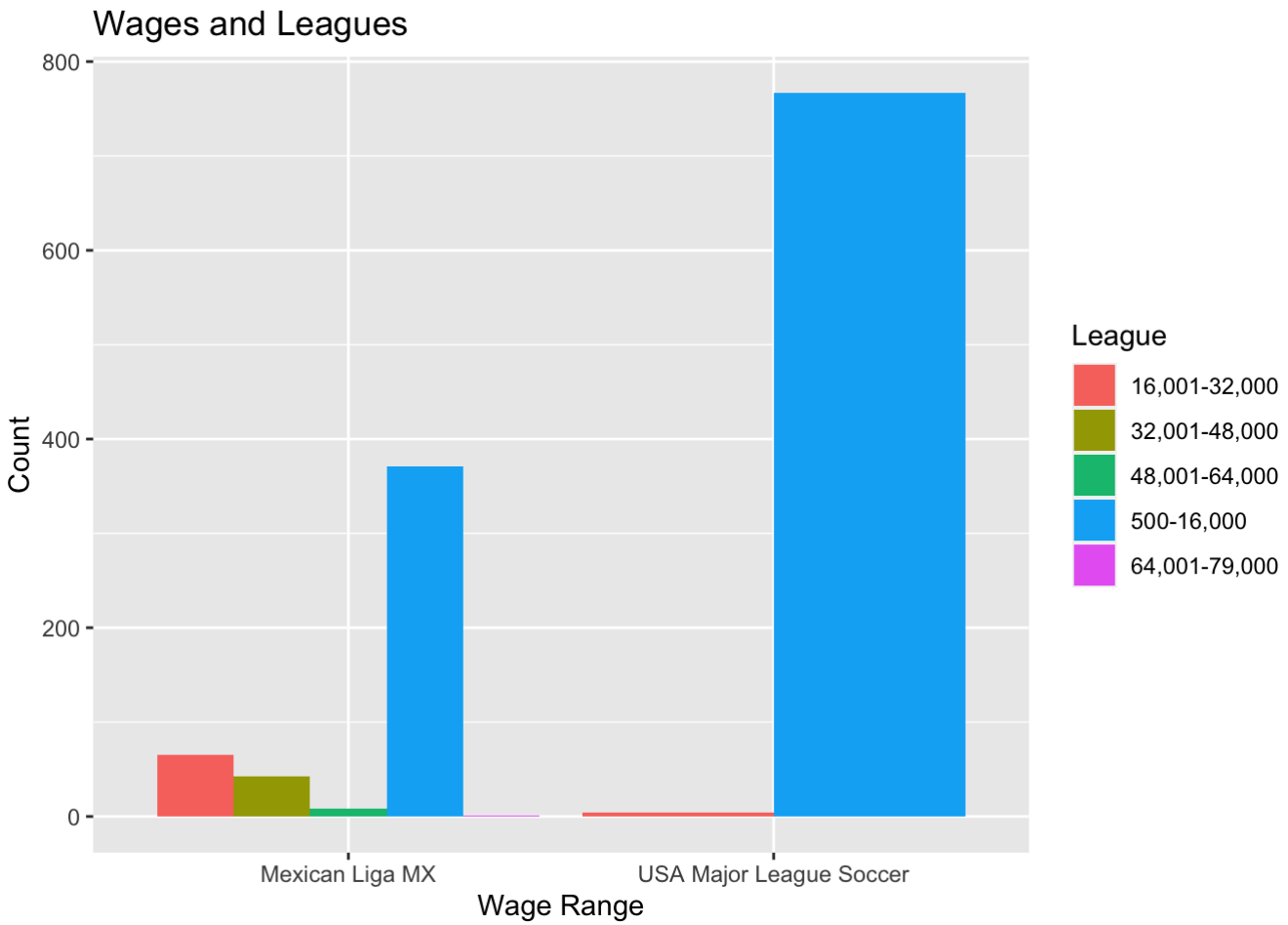
	11,613,001–23,181,000	23,181,001–29,000,000	45,000–5,829,000
Reserves	0	0	222
Starters	19	1	434
Subs	1	0	529

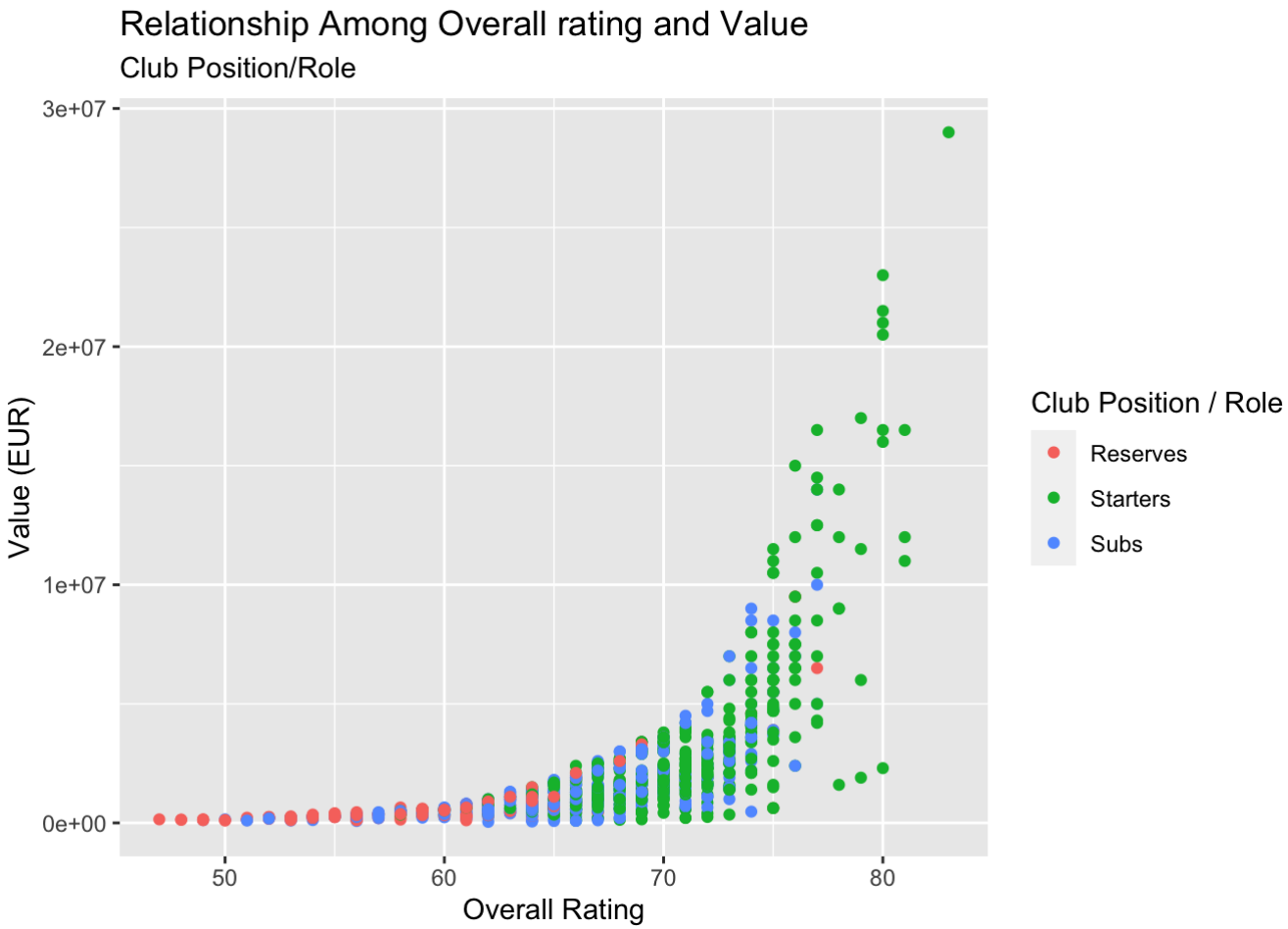
	5,829,001–11,613,000
Reserves	2
Starters	41
Subs	10

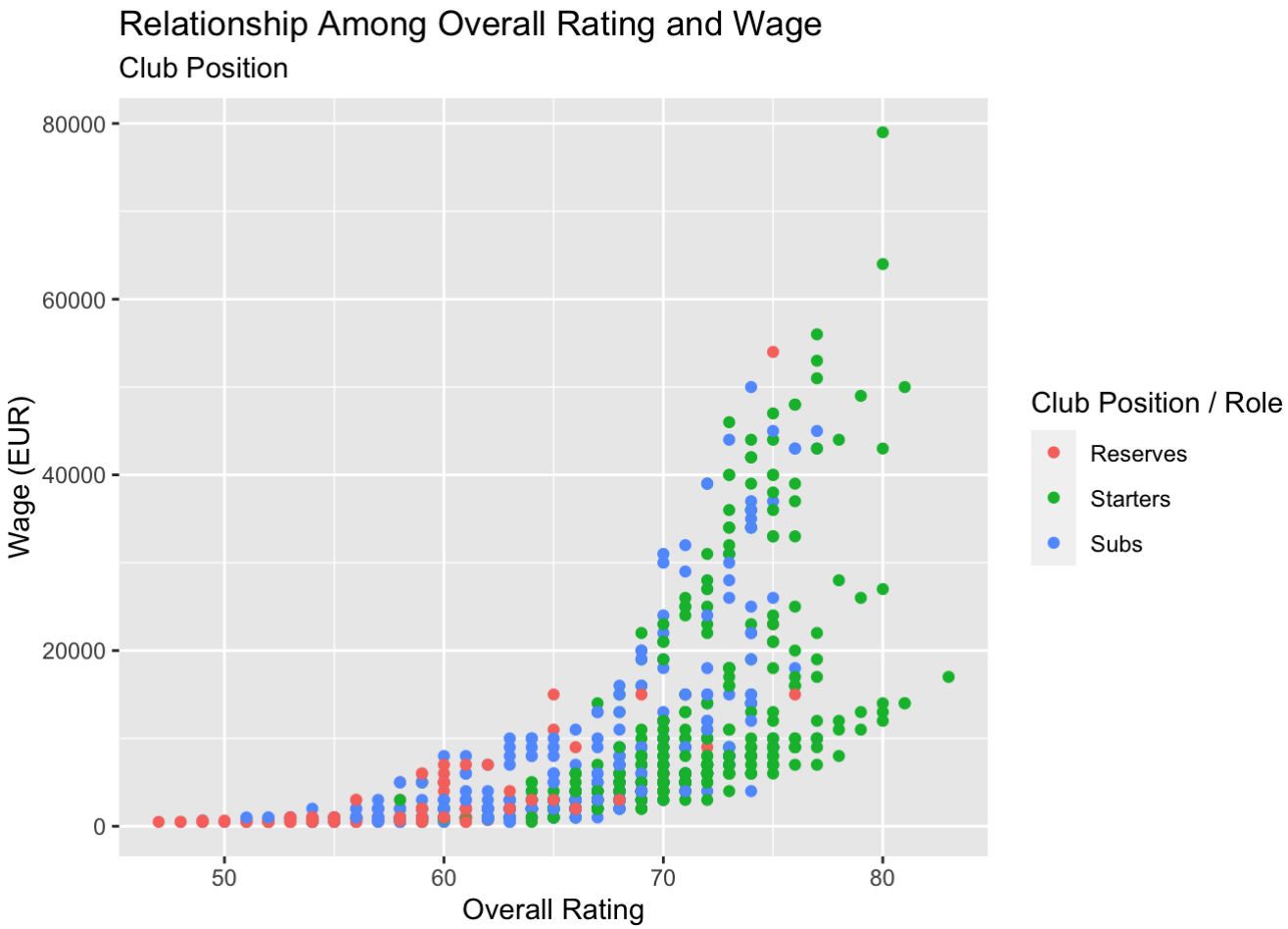

Preferred Foot and Club Position/Role
Count
400
200
0
Reserves
Starters
Subs
Club Position/Role
Preferred Foot
Left
Right


Relationship Among Age and Club Position/Role
Count
400
200
0
Experienced
Old
Prime
Young
Age
Club Position/Role
Reserves
Starters
Subs


Dodged Bar Chart
400
300
200
100
0
Count
Left
Right
Preferred Foot
Club Positions/Role
Reserves
Starters
Subs


Wages and Leagues
800
600
400
200
0
Count
Mexican Liga MX
USA Major League Soccer
Wage Range
League
16,001-32,000
32,001-48,000
48,001-64,000
500-16,000
64,001-79,000


Relationship Among Overall rating and Value
Club Position/Role
3e+07
2e+07
1e+07
0e+00
Value (EUR)
50
60
70
80
Overall Rating
Club Position / Role
Reserves
Starters
Subs



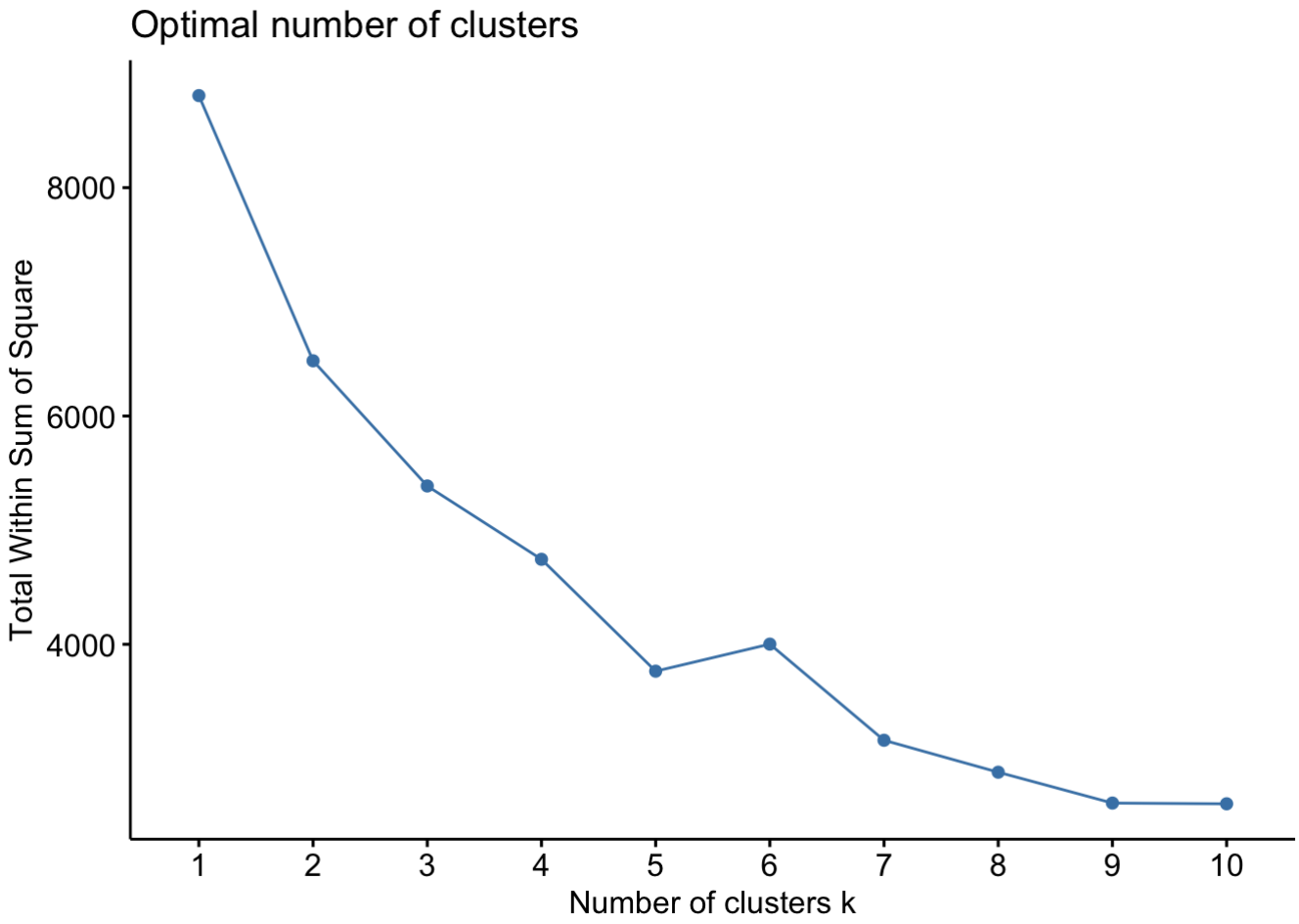
Data visualization is critical for gaining useful insights and comprehending complex data sets.As a result, visual representations such as graphs have been introduced in this part to improve the clarity of linkages and outcomes, ensuring a more accessible interpretation of the data.I've included a contingency table with two categorical, a stacked bar chart, a dodging bar chart, and a scatter plot demonstrating relationships between two numeric variables with one categorical component in this section. From the contingency table we are able to see a trend where players in the "Starter" category exhibit a higher likelihood of having elevated values and wages. This underscores the impactful relationship between a player's role on the team and their financial standing, emphasizing the influence of team role on both player value and wage. Moreover, from the first stackbar chart that relates variables preffered food and player's position, it is clear that players' favorite foot does not have a high propensity to influence their position or role within the squad. After looking at the second stackbar that provides with the relationship between Age and Club positio we can see how clubs and leagues in the North America area strategically invest in young talent, as evidenced by the significant representation in the "Young" group compared to other age groups. Furthermore, the region has a substantial number of players in their "Prime," indicating a concentration of individuals at the pinnacle of their abilities. The data also shows a high incidence of "Starters" among the players, indicating a dedication to fielding experienced and quality players. Also, the "Young" category contains a higher proportion of "Subs" and "Reserves," supporting the idea that young players are in a developing period, gradually adjusting to the demands of professional football. The scatter plot chart demonstrates a clear pattern indicating that players with higher overall ratings are more likely to be designated as "Starters" on their respective teams. Furthermore, a direct association between total rating and pay develops, meaning that players

with higher ratings fetch greater wages. This insight emphasizes the importance of player performance, as represented in their overall rating, in defining their place on the team as well as their wages.Finally,the Dodged Bar Chart, which compares player earnings across leagues, gives information about the distribution of player wages throughout each league. Notably, both leagues include a significant proportion of players earning lesser earnings, which corresponds to the data reported in the summary measures section. This depiction emphasizes the assumption that a considerable part of players in both leagues are paid in the lower wage bands, emphasizing the importance of further research into the reasons influencing these salary distributions.

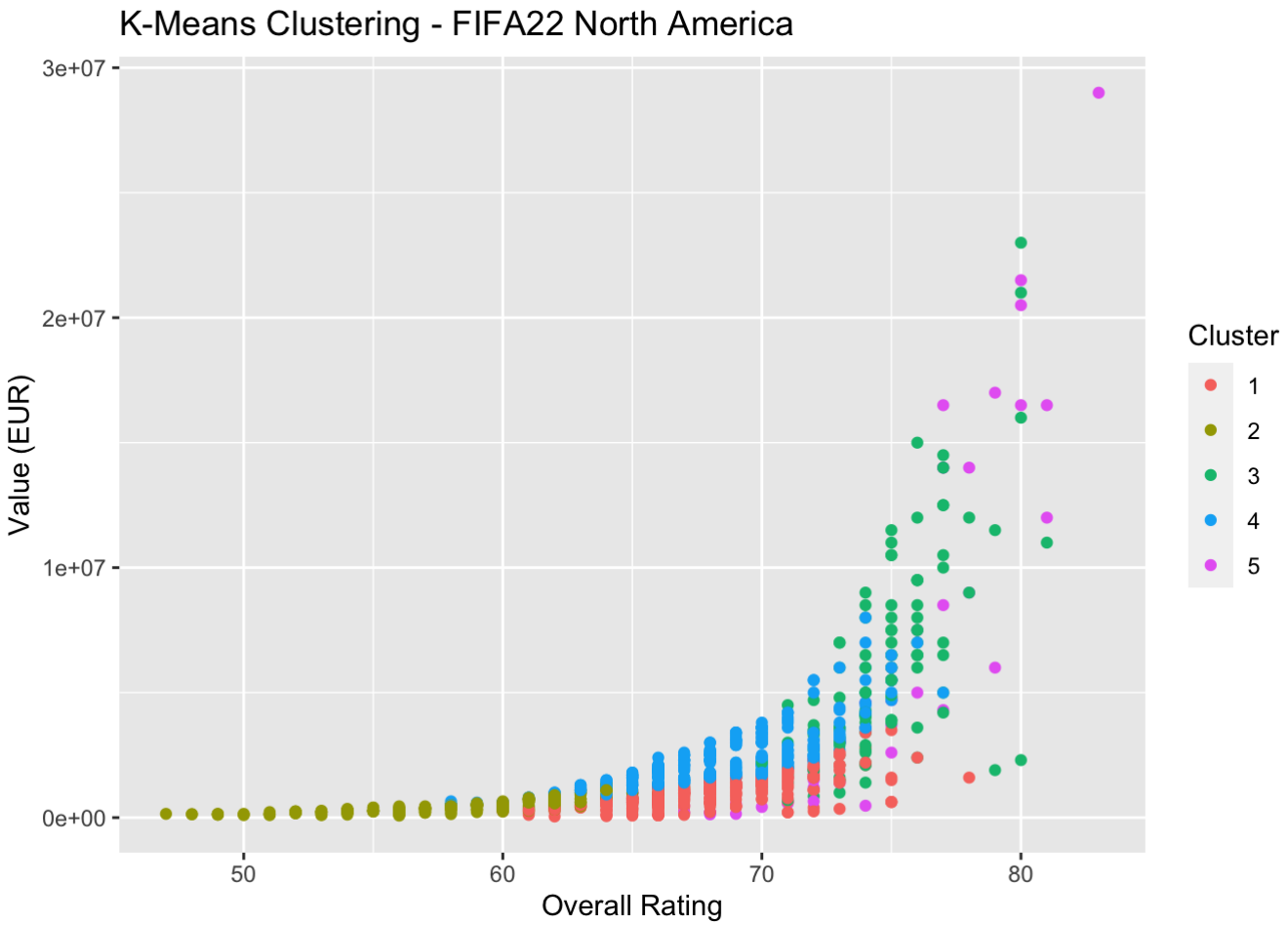
PART D: ANALYTIC MODELS

Clustering

We will now use a clever way to classify individuals based on their skills, value, and more using the clustering method. Our investigation identified five clusters, each of which told a different player tale.




```
cluster_assignments
  1   2   3   4   5
380 365 115 351  48
```



After we used a clustering approach to examine several player factors such as overall rating, potential, value, wage, age, skill, and international reputation in our research. We discovered that the ideal number of clusters was 5 through experimentation. When the data was clustered, distinct player profiles emerged. Cluster 1 players had good overall ratings but very low market valuations, indicating seasoned players nearing the end of their careers. Cluster 2 included players with low overall ratings and value, implying duties as substitutes or reserves, which are common in the early stages of a player's career.Cluster 3 had players with good overall ratings and moderate-low market valuations, suggesting that they were undervalued for their ability level. Cluster 4 players had moderately high overall ratings and moderately low market prices, indicating skilled players with modest valuations. Finally, Cluster 5 featured players with strong overall ratings and market prices who are likely to start for their respective teams.

3. Regression

We aim to identify the key elements impacting a player's market worth in North American football. Therefore we are going to use a regression model to investigate the relationship between a player's market value and how is impacted by critical variables such as overall rating, international reputation, wage, potential, and age.

```
Call:
lm(formula = value_eur ~ overall + international_reputation +
    wage_eur + potential + age, data = northAmerica)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-5621778  -546521  -137907   353997 18269188 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -1.235e+07  9.097e+05 -13.576   <2e-16 ***
overall                   2.262e+05  1.705e+04  13.269   <2e-16 ***
international_reputation  2.619e+06  1.736e+05  15.087   <2e-16 ***
wage_eur                  1.026e+02  6.000e+00  17.101   <2e-16 ***
potential                 1.718e+04  1.726e+04   0.995     0.32    
age                      -2.102e+05  1.989e+04 -10.566   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1582000 on 1253 degrees of freedom
Multiple R-squared:  0.6385,    Adjusted R-squared:  0.6371 
F-statistic: 442.6 on 5 and 1253 DF,  p-value: < 2.2e-16
```

The regression model investigates the association between a player's market value (value_eur) and a variety of variables such as overall rating, international reputation, wage, potential, and age. The coefficients give details on the effect of each variable. For example, as evidenced by the significance of the coefficient levels (p-values 0.05), a player's overall rating, international reputation, wage, and age all have a substantial influence on their market value. The model's goodness of fit is confirmed by a high Multiple R-squared value of 0.6385, indicating that the included variables explain nearly 64% of the variability in market value. The F-statistic is also significant, supporting the overall relevance of the regression model Therefore, we can conclude that this is a good model.

PART E: ETHICAL IMPLICATIONS

While certain factors such as player positions, skills, and physical attributes are derived from public knowledge and are included in our current analysis, other personal details such as age, nationality, height, weight, and wage are also considered public knowledge within the FIFA dataset. Nonetheless, we underline the importance of maintaining and handling this sensitive information with extreme caution and in accordance with the ethical principles set in the Data Science Code of Professional Conduct. Respecting players' private rights and adhering to legal and ethical requirements remain important to our commitment, even when dealing with publicly available information.This approach not only ensures the quality and confidence of our data analysis, but it also strengthens our commitment to ensuring the responsible and ethical use of player information in sports analytics Maintaining this ethical attitude is critical not just for the credibility of our findings, but also for sustaining FIFA's commitment to fair and ethical practices in sports analytics. As an analyst, my primary goal is to provide significant insights that help FIFA achieve its goals, however, it is also critical to emphasize that these insights should not be abused to the harm of athletes or leagues. We seek to ensure that the information created by our studies serves the soccer community as a whole without causing harm to anyone by prioritizing transparency, accountability, and responsible data use.

CONCLUSIONS

Finally, this study has shown to be a significant tool for understanding factors that determine player worth and wages in the world of North American football. We discovered detailed insights into the diverse dynamics of the football industry in this region by methodically evaluating criteria such as overall rating, international reputation, wage, potential, role in the club and age. The findings not only improve our understanding of the factors that influence player market value, but they also provide a solid framework for strategic decision-making among clubs and the FIFA organization. The importance of this study goes beyond football analytics; it provides useful data for clubs and leagues trying to optimize their squad compositions, negotiate contracts successfully, and plan for the future. This research equips decision-makers to navigate the competitive landscape of North American football with better precision and foresight by diving into the complicated interplay of elements determining player worth.

REFERENCES

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2022). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.4.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

OpenAI. (2022). ChatGPT (version 3.5) [Computer software]. Retrieved from https://openai.com.

END