Predicting Players’ Market Value

Jacopo Belardo

May 8, 2024

**Problem formulation**

As a soccer player and data analytics student, I am very interested in exploring the potential of applying data to the sport. In the past few years, new technologies like GPS trackers and video analysis have helped teams to gather a huge quantity of data that can be used to improve performance. With performance, I refer both to practical execution of the athletes on the field and the overall club in terms of financial results. Most of the time the two elements go hand in hand because clubs perform well on the field by getting valuable players while also maintaining a stable financial situation. However, the soccer market has been out of control in the past few years and clubs are overpaying for players who are not performing as well as they should according to their market value. Therefore, the objective of my analysis is to see if there is actually a relationship between a player’s characteristics and his market value. The main focus of my project is trying to develop a regression model that can predict the value of a player based on his characteristics and statistics. The analysis can produce relevant results to club managers and coaches to better understand the players and what makes them valuable.

**Data description**

The dataset I am using is found on Kaggle (link in the references) and it comprehends data about players in the top 5 European Leagues. The zip file contained data from three seasons from 2017 to 2020. I decided to use only the 2019-20 season because it is the most recent and complete dataset. All the information is collected from Transfermrkt which is a public database containing statistics used by people in the soccer industry to assess performance and value of teams and players.

The dataset about the 2019-20 season initially contained 400 variables, that is why I decided to cut the ones I did not consider relevant. First of all, I removed all the goalkeepers from the dataset and all the statistics related to this kind of players. Goalkeepers are analyzed using a set of statistics (e.g. saves, goals conceded, etc.) that are completely different from the ones used to evaluate field players. At the end of the cutting process, I came up with a 68 columns/ 2434 rows dataset (**SEE APPENDIX 1**). The high number of variables is due to the fact that different statistics might be relevant to different players according to their position on the field. For this reason, creating three different models (defenders, midfielders, forwards) might be ideal to better understand the purpose of my research project.

During the cleaning process, I decided to add two variables that can be useful in my analysis. The first one is “high\_value” and it is a binary variable with “Yes” and “No” factors that tells us whether the player can be considered a high value player. I set the threshold at 12 million euros because it represents the value of the 75th percentile. The second variable that I added is called “age\_group” and it is just a way to transform “age” into a categorical variable. I will use this variable to conduct ANOVA testing and see if there are differences between means across this variable.

Moving to a detailed description of the variables, we can observe from Appendix 1 that the majority is numerical. This structure is expected considering the fact that numbers represent an objective way to evaluate a player’s performance. In terms of categorical variables, it is interesting to use some charts to visualize them and look at the distribution frequencies (**SEE APPENDIX 2**). Starting from the position, we can observe that the minority of the players are forward. This is normal because most of the teams play only with 2/3 forwards in their starting eleven opposed to midfielder and defenders which are usually 3/4 (at times even 5). Moving to the leagues, we mostly have the same number of players in the different competitions. The only relevant difference seems to be between Serie A and Bundesliga, but it is due to the fact that the latter has only 18 teams competing (compared to the usual 20). Another interesting piece of information to visualize is the preferred foot of the players. The graph shows that the majority of the players are righty, and it is a rarity to find players who feel comfortable using both feet equally. Lastly, I created a box plot to show the value of players by position. The first observation we can make is that even if the means are similar and they are valued only few million euros, there are some players who are far above that price (> 100M). We need to keep an eye on these observations because they could heavily impact the analysis as outliers.

To conclude the exploratory analysis of the dataset, it is good to reflect on advantages and disadvantages of using this dataset. Starting with the advantages, the dataset is very detailed, and the presence of many different variables can allow us to get different insights on the soccer market in Europe and predict the value of players. Also, the variety in the numerical variables gives us the chance to analyze separately the players by position. On the contrary, the high number of variables can create confusion in the analysis. Another disadvantage is that the market value of the player might be approximated and give us imprecise outcomes.

**Model Building**

In this I will discuss my general ideas on performing the analysis to create the final model which aims at predicting the market value of a player based on his statistics and characteristics. I will start by reflecting on possible associations between variables and I will proceed with the actual building of the linear regression model which is the focus of the research.

The first step will be to compare means of market value across different variables. The first association that comes to my mind is between position and market value. I believe that forwards and midfielder have a higher value compared to defenders. The first two categories contribute to the success of the teams with key statistics like goals and assists making them more attractive in the eyes of the clubs. The second association that I see is between age groups and market value. Hypothetically, players in the “middle” age group should be more valuable than young and old athletes. Usually, players are in their best shape at 27/28 years old and clubs are willing to spend more money to buy athletes at the peak of their career. The third aspect I want to explore is the difference in market value across the leagues. I am interested in understanding this data because Premier League players are often considered more expensive compared to their peers in other leagues. The reason behind this is that the English championship is considered the most competitive in the world.

After exploring these associations by focusing on the categorical variables I will move forward to develop my linear regression model to predict the market value of players. As I mentioned above, the best way to perform this analysis is by splitting the dataset into three sections according to the position of the players. I will proceed in the following order: defenders, midfielders, and forwards.

Looking at the variables, there are some factors that are common to the analysis of every position. Nationality, age, height, foot, league, games\_starts, goals, assists, passes\_pct, and CL are variables that I will include in the three initial models. Then, I will dive into more specific statistics. For the defenders, I will use cards\_yellow, cards\_red, passes\_intercepted, passes\_blocked, tackles\_won, interceptions, clearances, fouls, ball\_recoveries, and aerials\_won\_pct. These are all statistics that hypothetically add value to players more involved in the defensive side of the game. The analysis of the midfielders is probably the most complicated because they participate to all the phases of the game, and their defensive and offensive efforts are equally important in the eyes of coaches. The variables that I consider important to the model are cards\_yellow, cards\_red, goals\_assists\_per90, xg, xa, shots\_on\_target\_pct, passes\_completed, assisted\_shots, passes\_into\_final\_third, crosses, passes\_intercepted, tackles\_won, pressures\_regain\_pct, touches, touches\_mid\_3rd, touches\_att\_3rd, dribbles\_completed\_pct, dispossessed, fouls, fouled, and ball\_recoveries. Lastly, in the forwards model I will include goals\_per90, goals\_assists\_per90, xg, xa, xg\_xa\_per90, shots\_on\_target\_pct, shots\_on\_target\_per90, goals\_per\_shot, passes\_into\_final\_third, crosses, touches, touches\_att\_3rd, touches\_att\_pen\_area, dribbles\_completed\_pct, fouled, offsides.

Another insight that I would like to get from my dataset is to predict the participation of the players to the Champions League (CL) based on the statistics. In this case I will try to build a model that includes all the positions and relies on more generic variables. I see “value” as the main factor to predict the dependent variable. The other variables that I will add to the model are nationality, age, height, foot, games\_starts, goals, assists, passes\_pct, goals\_assists\_per90, xg, xa, shots\_on\_target\_pct, passes\_completed, crosses, passes\_intercepted, tackles\_won, pressures\_regain\_pct, touches, dribbles\_completed\_pct, dispossessed, fouls, fouled, ball\_recoveries, aerials\_won\_pct. Participation to the Champions League is an interesting information because the most valuable players usually compete in it. Building a logistic regression model with this objective can help me to predict the impact of the variables on the possibility for a player to compete in the top European competition.

The last analysis that I see meaningful to my research is the building of a machine learning algorithm that can help us to determine if a player can be considered high value or not. Also in this case, I prefer a general approach that can give us a general idea of what are the main characteristics of a player with high market value. Therefore, the variables I will include in the model are the same of the logistic regression with the only difference of replacing “value” with “CL”. It would be meaningless to insert the value in the model because what we are trying to predict derives from it.

In the linear regression approach, I will have the chance to properly evaluate if the metrics I want to consider can be a good predictor of a player’s value with the three different models. Building one model and using all the variables cannot give us good insights into the differences that might be present in the evaluation of an athlete according to his position on the field. The models on the participation to the Champions League and the determination of high value are complementary analysis that allow us to look at different perspectives of the same problem. Being able to predict the market value of a player can be very helpful for clubs to take a data-driven approach in their recruiting strategies. At the managerial level, finding players that can be potentially valuable according to their statistics can be a huge factor in establishing a competitive advantage and positioning themselves ahead of the competitors.

**Analysis methods**

In the model building section I exposed a potential course of action in analyzing the Transfermkt dataset about the 2019/20 season. In this portion of the paper, I will discuss the actual methods and tools that I will use. The analysis will be performed on R Studio because it offers an easy and efficient way to conduct the research and interpret the results. I will proceed in the order previously discussed: associations across variables, linear regression model, logistic regression model, and decision trees/random forests.

The first step will be to compare the mean of market value across different variables like position, age group, and league. In this analysis, the use of the aov function paired with the TukeyHSD will be crucial to achieve our objective. ANOVA tells us if there is a statistically significant difference across the different groups and the Tukey comparison helps us to see what groups are actually differing. ANOVA represents a good way to start or analysis but does not give us any information about the interaction between the variables. For this reason, we need to develop some regression models.

The second step is based on building the three models based on the positions. I will use the lm function and fill it with the variables that I cited in the model building section. I prefer to conduct the analysis relying on a manual approach rather than an automated stepwise because it better fits the purpose of my research question. Also, I carefully examined the variables to pick the ones that I consider most important using my domain knowledge. After building the model, I will run the model diagnostics and check for multicollinearity issues. The strength of a linear regression model is that it provides us with a simple way to understand the direction and strength of the relationship between dependent and independent variables. On the other hand, it assumes the linearity between the variables, and it can be very sensitive to issues with outliers or multicollinearity.

The third step of my analysis is focused on a binary variable; therefore logistic regression is the best model for this purpose. In this case, I will use the glm function to perform the analysis including the generic variables that I consider relevant to the potential participation of a player to the Champions League. Then, I will run diagnostics to evaluate the performance of the model and adjust it accordingly. Overall, logistic regression is a very straight-forward method to handle binary outcomes and understand the class of our observations. However, we still have problems with the linearity assumptions and sensitivity of the model to potential outliers.

The fourth step of the analysis is the conclusive way to see a different approach to the research question. By using the binary variable “high\_value”, we can have a general idea on how to identify a valuable player. The variables included in the model are discussed in the model building. In this section, I will start by splitting the dataset into training (70%) and testing (30%). I will move forward by developing a decision tree with the rpart function (library(rpart)) using the training data. Then, I will evaluate the performance of our model on the testing data. According to the results, I will proceed with pruning the tree and re-testing. The random forest approach will allow us to have a complete overview of the problem by using multiple trees and predict the outcome more efficiently. In this case the randomForest function (library(randomForest)) will be used to perform the analysis and the model will be tested. In the evaluation of the models, accuracy is the metric we will use to interpret the reliability of the results. The advantages of the random forest approach rely mainly on the absence of a pruning process, the reduction in overfitting issues and the possibility to observe variables’ importance.

**Results and discussion**

In this section we will dive into the specific coding and interpret the outcomes. We will use a significance level of 0.05. We will proceed in the following order: ANOVA tests, linear regression model, logistic regression model, and decision trees/random forest.

ANOVA

The first step consists of performing ANOVA to identify statistical differences in value mean across position, age group and league. Our null hypothesis is that value means do not vary across the categorial variables we will consider. Starting with the position, the p-value of 3.3e-08 makes us reject the null hypothesis. Also, the Tukey comparison show us that there is a statistically significant difference between the means of every position (**SEE APPENDIX 3**). These results support my point that it is better to develop different models for the three positions and the forwards are the category with the highest value due to their important role in terms of goals and assists. Moving to the test on age group, we can reject the null hypothesis with a p-value of 3.53e-11. However, the Tukey comparison shows that the only statistically significant difference is between middle-aged and old players (**SEE APPPENDIX 4**). As I mentioned above, I expected this outcome because middle-aged players are at the peak of their careers and generally a safe choice to invest in. On the contrary, old players can bring experience but clubs are not willing to invest in players with only a few years to compete at a high level. Lastly, we can reject the null hypothesis also in the analysis of leagues (p-value <2e-16). Tukey comparison confirms my initial hypothesis that Premier League players are the most valuable in the world. The average value of Premier League players is significantly higher than the ones of the remaining top 4 European leagues showing an absolute dominance of the English championship. The other leagues are mostly on the same level except for the Ligue One which seems to lag behind La Liga and Bundesliga with a lower average market value (**SEE APPENDIX 5**).

LINEAR REGRESSION

The second step is the analysis of the main purpose of my research. I am trying to predict the value of a player based on his characteristics. As I already mentioned, the best way to create reliable models is by dividing the dataset into three different subsets: defenders, midfielders, and forwards. I will follow this order to interpret the results.

First, I developed the initial models for the defenders in the following way:

lm(formula = value ~ nationality + age + height + foot + league +

games\_starts + goals + assists + passes\_pct + CL + cards\_yellow +

cards\_red + passes\_intercepted + passes\_blocked + tackles\_won +

interceptions + clearances + fouls + ball\_recoveries + aerials\_won\_pct,

data = defenders)

After looking at the summary results with the p-value, I identified some variables that are not statistically significant, and I developed a new model. The latter was not good enough considering the information about r-squared of 0.50 and the issues in the diagnostics plot (**SEE APPENDIX 6**). For this reason, I moved forward with removing the outliers and adding new variables to increase the r squared. After many trials, we can see how the r-squared raised to 0.55 and the diagnostics charts slightly improved. However, there seems to be still problems with the model in terms of linearity, normality and homogeneity (**SEE APPENDIX 7**). There are no multicollinearity issues after checking with the vif function (library(car)). Talking about the independent variables, I was surprised to see that only two specific defending skills (tackles and aerials won) resulted significant in predicting the value of a defender. On the other hand, participation in the Champions League and playing in the Premier League represent two huge factors in the model.

The process for the analysis of the midfielders was very similar. I started with the hypothetical model that I built with my domain knowledge:

lm(formula = value ~ nationality + age + height + foot + league +

games\_starts + goals + assists + passes\_pct + CL + cards\_yellow +

cards\_red + goals\_assists\_per90 + xg + xa + shots\_on\_target\_pct +

passes\_completed + assisted\_shots + passes\_into\_final\_third +

crosses + passes\_intercepted + tackles\_won + pressure\_regain\_pct +

touches + touches\_mid\_3rd + touches\_att\_3rd + dribbles\_completed\_pct +

dispossessed + fouls + fouled + ball\_recoveries, data = midfielders)

Then I removed all the non-statistically significant variables and observed the new model. The latter improved from the initial version but still had variables that need to be removed and other adjustments that could be made (**SEE APPENDIX 8**). Therefore, I opted for the same approach I used for the defenders by removing the outliers and trying to add new variables. In the final model, the variables are all statistically significant and explain the 59% of the change in the dependent variable. Also in this case, no multicollinearity issues are detected. However, we can observe the same problem we had with the defenders in terms of linearity, normality, and homogeneity from the diagnostics chart (**SEE APPENDIX 9**). Reflecting on the variables, we can see that participation in the Champions League and being a Premier League player are still the main two factors in the model. Interesting to observe that goals and assists are not included in the model, but better predictors of the market value of a midfielder are xa (expected assists), and more detailed information like assisted shots and passes into the final third.

Lastly, we can look at the model for the forwards. The procedure involves the same steps used for the other positions. The initial model is:

lm(formula = value ~ nationality + age + height + foot + league +

games\_starts + goals + assists + passes\_pct + CL + goals\_per90 +

goals\_assists\_per90 + xg + xa + xg\_xa\_per90 + shots\_on\_target\_pct +

shots\_on\_target\_per90 + goals\_per\_shot + passes\_into\_final\_third +

crosses + touches\_att\_3rd + touches\_att\_pen\_area + dribbles\_completed\_pc + fouled + offsides, data = forwards)

As for the previous analysis, I removed the non-statistically significant variables, and I came out with a revised model (**SEE APPENDIX 10**). To improve the model, I proceeded with the elimination of the outliers and the addition of different variables. As a result, the model could explain 64% of the variation in the dependent variable, after been adjusted for multicollinearity problems. The diagnostics charts prove that this model also has issues in terms of linearity, normality, and homogeneity (**SEE APPENDIX 11**). The main takeaways from the analysis are that CL and League (Premier League specifically) confirm themselves as the main predictors and most influential factors to determine the market value of a player. In addition, I was surprised to observe that none of the shooting statistics are included in the model. On the contrary, game starts, passes percentage, expected assists, and touches in the attacking penalty area are all relevant variables. Considering these results, we might assume that clubs are looking for forwards who like to be involved in the game and not only shooters or pure goals scorers.

LOGISTIC REGRESSION

After the exploration of the linear regression models to predict the market value of players according to their position, we can proceed with an accessory analysis on the participation to the Champions League. In performing this kind of analysis with a regression model, I split the dataset into training and testing to observe the goodness of the predictions. The initial model that I proposed is based on the following variables:

glm(formula = CL ~ value + age + height + foot + games\_starts +

goals + assists + passes\_pct + goals\_assists\_per90 + xg +

xa + shots\_on\_target\_pct + passes\_completed + crosses + passes\_intercepted + tackles\_won + pressure\_regain\_pct + touches + dribbles\_completed\_pct +

dispossessed + fouls + fouled + ball\_recoveries + aerials\_won\_pct,

family = "binomial", data = train.data)

This model was not performing well and I decided to take another approach by building another one with all the variables (excluded player, nationality, squad, league and others that I considered inappropriate). The next step consisted of removing the variables with p-value greater than 0.05 and adjusting for multicollinearity issues. The final model included many statistics that I did not consider extremely relevant to affect the probability of a player participating in the Champions League. As we can see from the table, the model was pretty accurate at predicting the probability of a player not participating in the competition. The real problem of the model lies in the prediction of the actual participants (**SEE APPENDIX 12**) With this analysis, I came to the conclusions that the dataset is missing important information like team performance that can be an interesting factor to observe in answering this question.

DECISION TREE / RANDOM FOREST

The last analysis I want to conduct is based on the use of decision trees and random forests. The objective is to have a general idea of the characteristics that determine a high value player. The starting point of the coding was always the splitting between testing and training data. The initial model I used to build the decision tree is the following:

rpart(formula = high\_value ~ CL + age + height + foot + games\_starts +

goals + assists + passes\_pct + goals\_assists\_per90 + xg + xa + shots\_on\_target\_pct + passes\_completed + crosses + passes\_intercepted + tackles\_won + pressure\_regain\_pct + touches + dribbles\_completed\_pct + dispossessed + fouls + fouled + ball\_recoveries + aerials\_won\_pct,

data = train.data, cp = 1e-04)

I decided to set cp = 0.0001 to build a deep tree and explore its performance. The outcome was a tree that used 19 variables and could predict correctly 80.7% of the high value players in the testing data (**SEE APPENDIX 13**). The performance was not terrible, but I saw room for improvement through a pruning of the tree. After the pruning, the decision tree used only 7 variables, was much easier to visualize, and improved the performance to 81.6% (**SEE APPENDIX 14**). Even if the models uses only few variables, they are very representative to evaluate a player. Starting from the participation to the Champions League, moving to expected goals, age, and statistics on the passing skills of a player. These are generally good predictors for a player’s value regardless of his position and the decision tree model confirms this theory.

Lastly, we can use a random forest to create an algorithm that can help us to predict if a player has a high market value or not. Considering the vast amount of variables, I carefully selected a group of factors that I believed to be relevant to my analysis and could summarize the rest of the dataset. Then, I created a subset of the soccer dataset with the variables: "high\_value", "age", "height", "foot", "games\_starts", "goals", "assists", "passes\_pct", "goals\_assists\_per90", "xg", "xa", "shots\_on\_target\_pct", "passes\_completed", "crosses", "passes\_intercepted", "tackles\_won", "pressure\_regain\_pct", "touches", "dribbles\_completed\_pct", "dispossessed", "fouls", "fouled", "ball\_recoveries", "aerials\_won\_pct", "league", "CL". I used this dataset to create the random forest and I obtained a model with an 84.4% accuracy (**SEE APPENDIX 15**). Also in this case, the model performs very well in predicting the “no” but not in predicting the “yes”. The random forest does a good job in showing the importance of the factors in the model and we can observe that passes\_completed, CL, age, and touches are always the main variables in predicting the value of a player.

**Recommendations**

I started my project with the idea of trying to better understand the application of data in soccer to provide some advice to clubs that are trying to figure out a player’s value. The results that I obtained with the different analysis prove that data can give interesting insights into the sports world but do not tell the full story. There is a “magic” element in sports that cannot be summarized with any statistics, and it is something that we should take into account when analyzing players. Therefore, managers and coaches should always remember to pair the use of data with an accurate observation of players by watching games.

By combining the results of my different analysis, we can conclude that the profile of the most valuable player is middle aged, participates in the Champions League, plays in Premier League, is very involved in the game (high number of touches), and has good passing skills. This information can be very useful to clubs in investing their money and planning the work in their academy. The results of my analysis prove that soccer is changing and is not a game where experts value the goal and the finalization of the play. Thanks to the interpretation of the game by the most famous coaches, soccer is becoming a sport where everyone is involved in the play and must possess good passing skills to compete at the highest level. Also, factors like the participation to the Champions League and the Premier League help us to complete the story. To become valuable, players must compete against the best and just accumulating a bunch of statistics is meaningless if the championship is not competitive. Therefore, every club must invest all their resources with the objective of finishing their season among the top 4 and qualify for the Champions League if they want to increase the value of their players. Also, Serie A, Bundesliga, Ligue One and La Liga should set the objective of imitating the competitive style of the Premier League to avoid a monopolization of the European market by the latter.

**Limitations**

Starting with the dataset, the main limitations that I found was the absence of team performance. Soccer is a team sport, and it is impossible to discern individual performance from team performance. Probably, having a dataset with more information about the teams could have helped me to come up with more relevant results. Also, the dataset does not include the unquantifiable measures that I defined “magic” in the previous section and play a huge role in a player’s valuation. Another limitation of the dataset is that it cannot be used to draw conclusions about the current soccer market. The information dates back to 2019/20 and the soccer world might have been influenced by many external factors. Lastly, the dataset is limited to the Top 5 European competitions, thus it does not provide a general overview of the soccer market in the world.

In terms of methods, linear regression and logistic regression are not good enough to capture the complex relationship between a player’s statistics and his market value. The assumption of linearity which is necessary for these kinds of analysis is not respected and the use of transformation techniques can misrepresent the data. Looking at decision trees and random forests, they limit the problems that we see with simple regression models, but they still have some issues. Decision trees represent an easy method to visualize data, but it is hard to find the right balance between complexity and performance to create a simple but efficient model. On the other hand, random forests can be very complex to explain and are more sensitive to outliers. Also, for these last two methods I decided to use a comprehensive approach and analyze the whole dataset without splitting into categories according to players’ position. A more specific analysis could facilitate the creation of better performing models.

In conclusion, a bigger dataset and more complex techniques are required to perform a complete analysis of the soccer market and provide advices applicable outside of the academic context.

**References**

[Soccer players values and their statistics (kaggle.com)](https://www.kaggle.com/datasets/kriegsmaschine/soccer-players-values-and-their-statistics?select=transfermarkt_fbref_201920.csv)

**APPENDIX**

1.

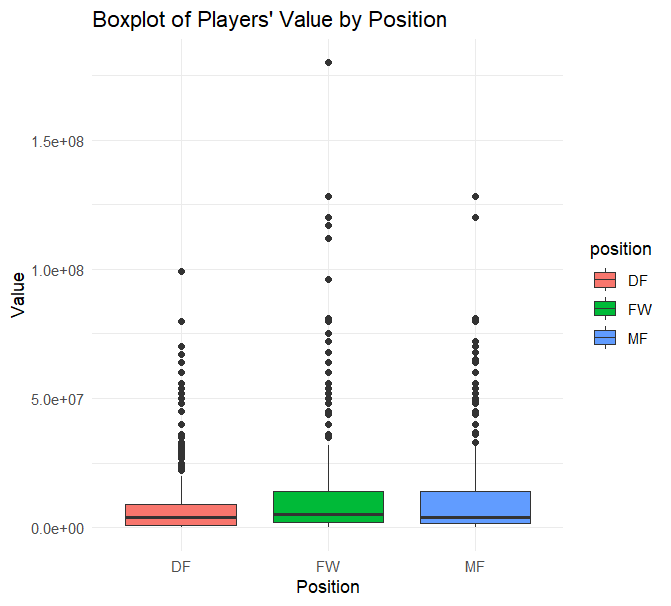
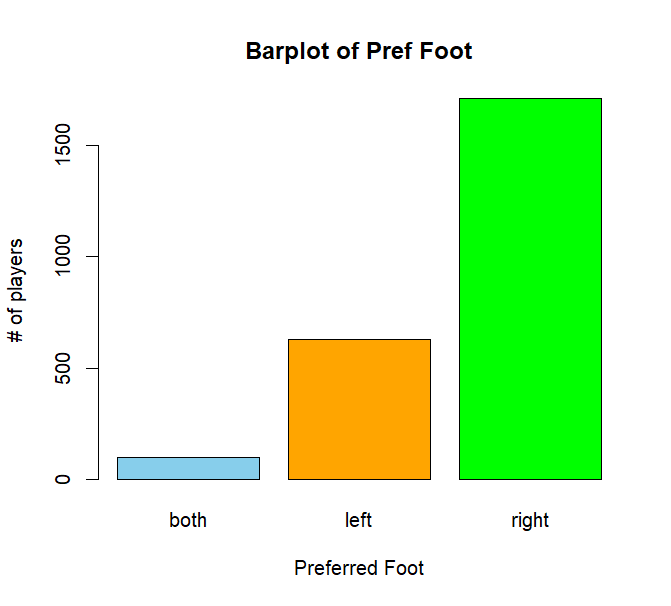
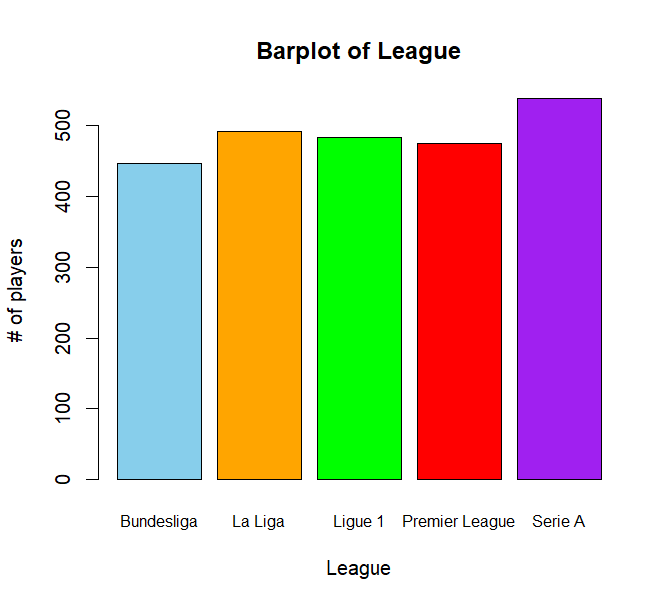
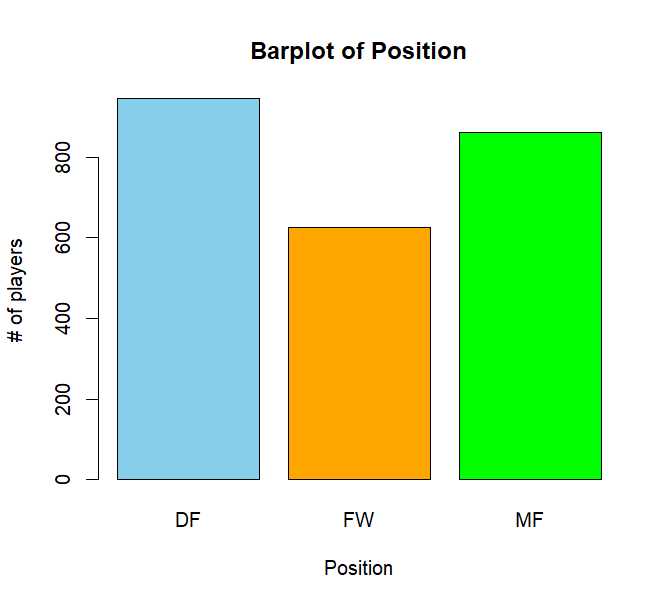
A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

2.



3.

A screenshot of a computer

Description automatically generated

4.

A screenshot of a computer

Description automatically generated

5.

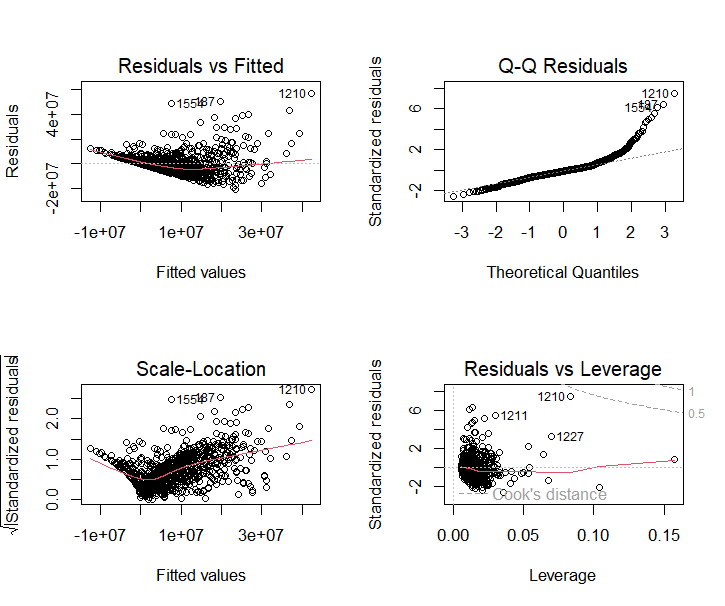
A screenshot of a computer

Description automatically generated

6.

A screenshot of a computer

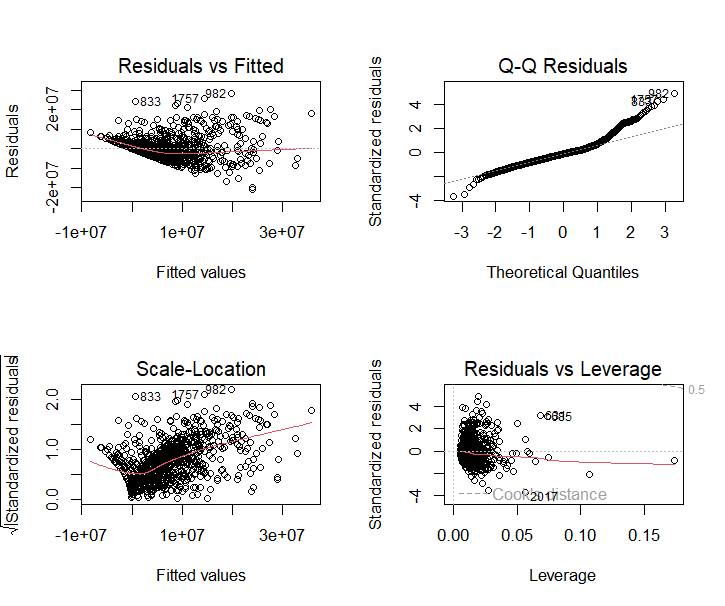
Description automatically generated



7.

A screenshot of a computer

Description automatically generated



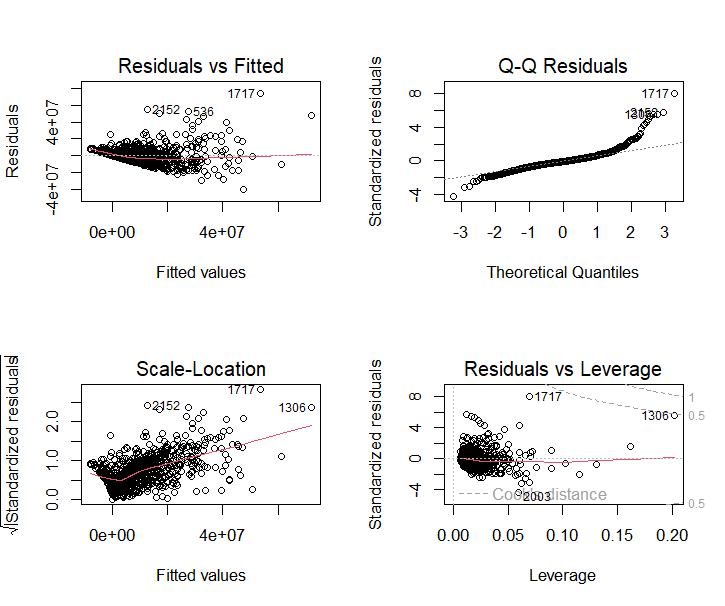
8.

A screenshot of a computer

Description automatically generated

A white background with black numbers

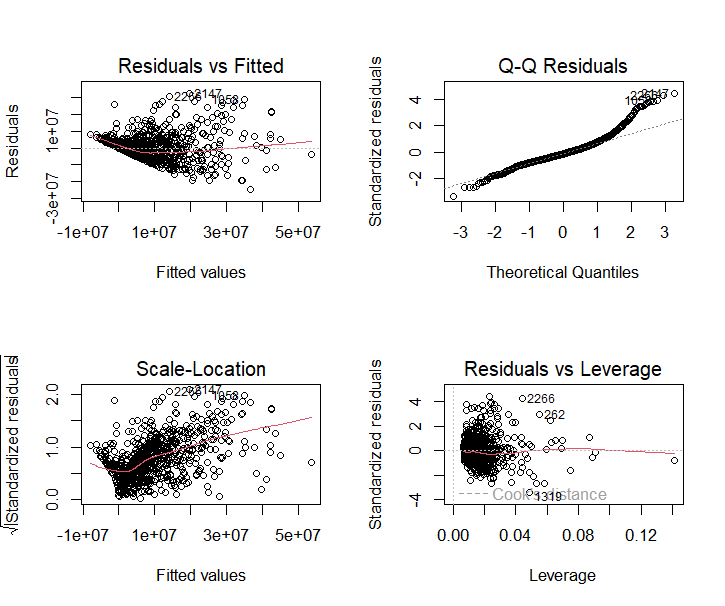
Description automatically generated



9.

A screenshot of a computer

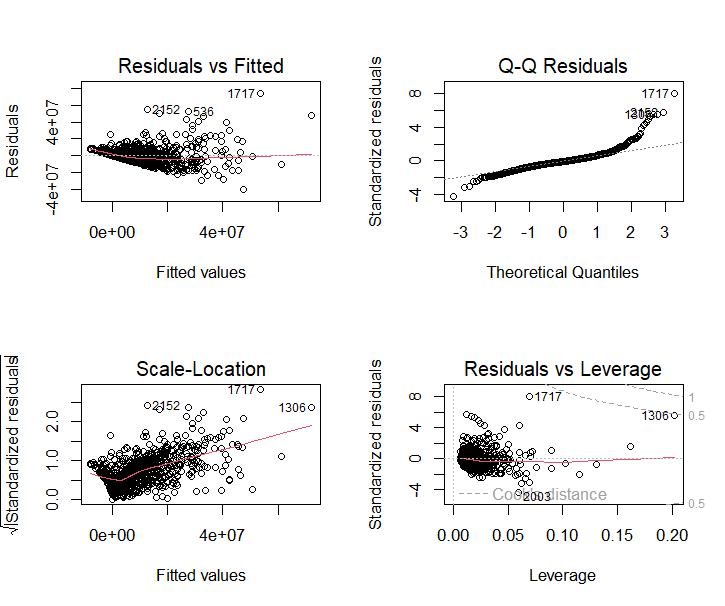
Description automatically generated



10.

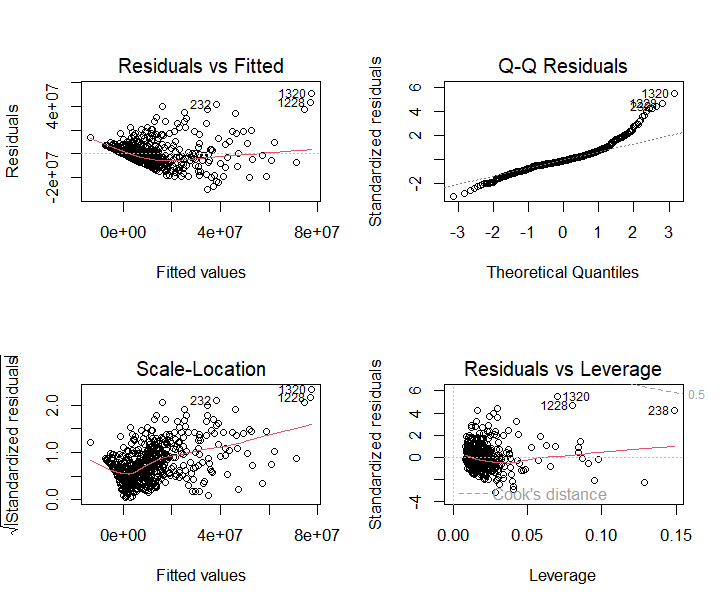
A screenshot of a computer

Description automatically generated

  
11.

A screenshot of a computer

Description automatically generated



12.

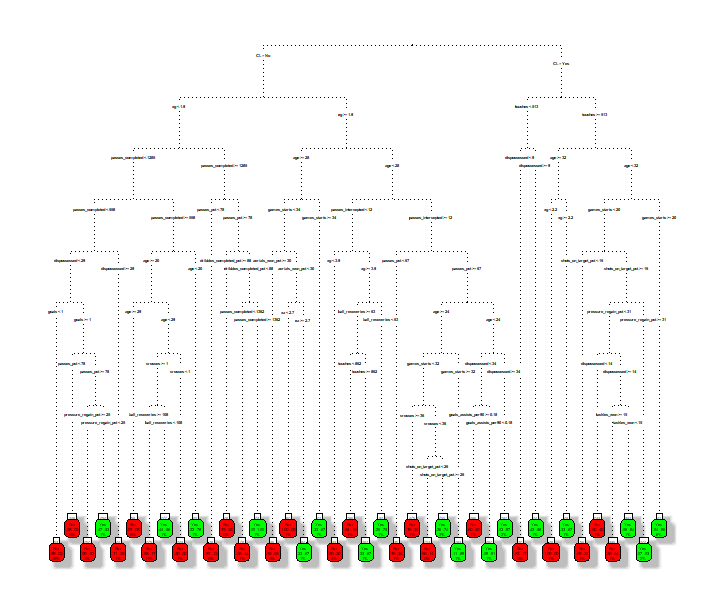
A screenshot of a computer program

Description automatically generated

13.

A screenshot of a computer

Description automatically generated



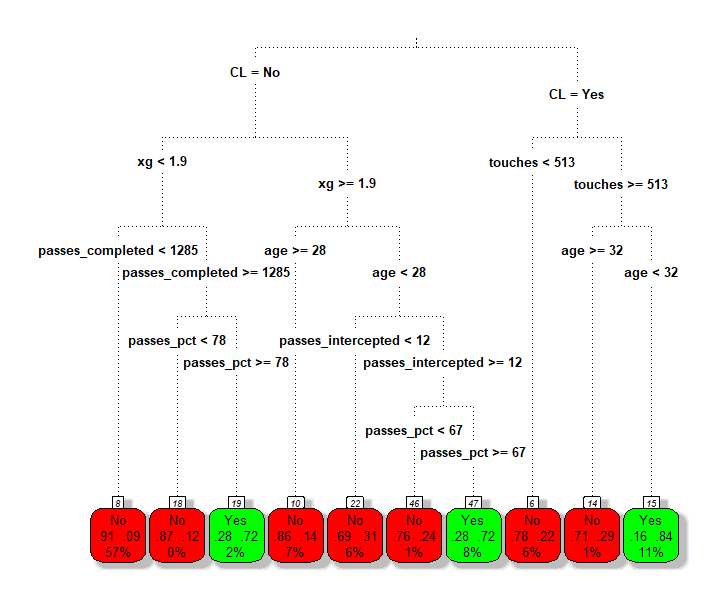
A computer code with blue text

Description automatically generated

14.

A screenshot of a computer error

Description automatically generated



A computer code with blue text

Description automatically generated

15.

A computer screen shot of a computer code

Description automatically generated

