**Assignment 1. Exploratory Data Analysis**

Keyla Guadalupe Muñoz Guilcapi

Sofía Pérez Pérez

Group 97.

This assignment has been done with the aim of reflecting our knowledge in our first approach to the "Exploratory data analysis" field. Moreover, this work has been a means to achieve understanding, practice and cooperation for us.

For this work we will be using a real Football data set about the games from the Spanish League (First Division) in the Season 2017/2018.

The first thing to do when doing a data analysis is to download and understand the data, so we did it using the code for setting the working directory and the read.csv used to read data like excel documents, i.e documents with extension .csv.

Source Code 1.

Now, we want to visualize the total number of data entries that our data set has

Source Code 2.

When we execute dim(data), we realise there are 380 rows and 64 columns of different variables, and therefore 24320 datapoints that we are going to use during our analysis.

After having understood our data, we are going to start answering questions that we considered curious and interesting, in order to obtain conclusions from our data set.

**Do home teams win more or score more than away teams?**

To approach this question, we have decided to start plotting a table in which we can see the amount of times each Home Team has won at home in the Half Time of the game, and we observe that they do not usually lose at home, they tend to win or draw.

Source Code 3.

A D H

Alaves 5 10 4

Ath Bilbao 6 8 5

Ath Madrid 1 11 7

Barcelona 1 5 13

Betis 5 7 7

Celta 3 8 8

Eibar 9 4 6

Espanol 7 9 3

Getafe 3 11 5

Girona 7 4 8

La Coruna 7 9 3

Las Palmas 8 9 2

Leganes 3 13 3

Levante 5 10 4

Malaga 8 7 4

Real Madrid 1 7 11

Sevilla 3 10 6

Sociedad 3 6 10

Valencia 3 5 11

Villarreal 5 6 8

Next, to have a global understanding of the results for the full time game we arrange all the Half time possible results(H,A,D) and we sum them. After the same thing is done with the Full time results, we can notice that in most of the matches during the first part of the game teams are tied, and it is in the second one when one of them takes advantage in the game. Usually is the home team the one that does that.

Source Code 4.

A D H

93 159 128

Source Code 5.

A D H

115 86 179

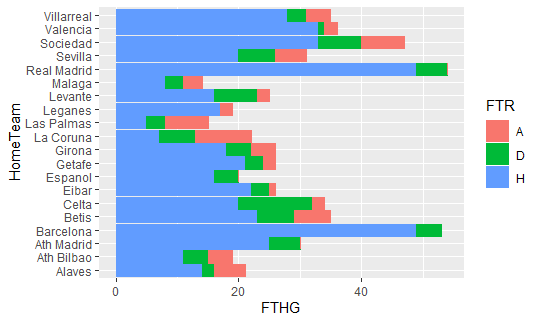
**But do home teams score more than away teams?** Yes, they do.

Source Code 6. Sum of the Full Time Home Team Goals = 588

Source Code 7. Sum of the Full Time Away Team Goals = 436

We are now very close to the answer of our question, and to visualise it better, we can use a histogram. To do so, we use the function library to be able to plot with ggplot.

Source Code 8.



We can see the Home Teams that win the most are Real Madrid, Barcelona, Ath Madrid. Furthermore, the team with more wins is Barcelona,and no Away Team has beated him, only draws. In the same way, the teams that scored more goals were the ones who tended to win more, such as the ones mentioned, but there are few exceptions like La Coruña.

In conclusion, we can affirm that Home Teams tend to win and score more than Away Teams.

Now we are going to deal with the faults committed**: what is the relation between yellow/red cards and faults?**

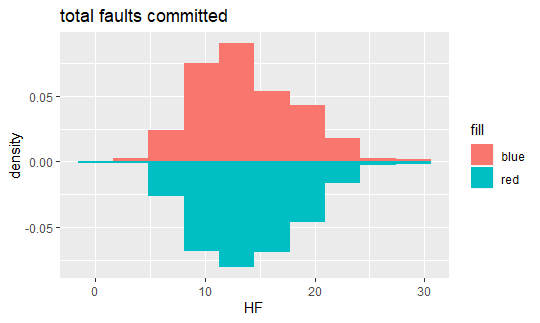
First of all, we study the yellow and red cards, and it seems that there is a huge difference between the total of both of them.

Source Code 9. Total sum of the yellow cards = 1906

Source Code 10. Total sum of the red cards = 72

After that, we use the ggplot function to plot a mirror histogram showing the faults committed by the home teams and an away teams.

Source Code 11.



As we can see there are not huge differences between the number of faults committed by a home team or an away team.

Source Code 12. Home team Faults = 13.73

Source Code 13. Away team Faults = 13.95

Taking the mean, we also can observe that the average amount of faults committed is 13.73 for the home teams and 13.95 for the away team, since they tend to be more offensive.

Now we are going to see if the faults committed did involve a card or not.

Source Code 14. Total sum of cards = 1906

Source Code 15. Total sum of faults = 10521

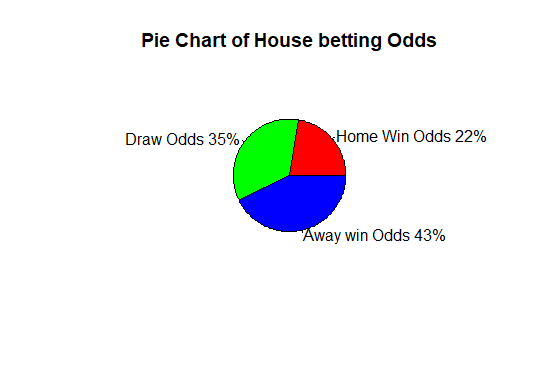
Most of the faults committed did not involve a card.

To conclude with this part, we have answered our question, since there is a very little relation between cards (red and yellow) and faults.

Our next question is: **What is the result most betted by the bookmakers?**

The aim of this question is to know if Betting Houses bet more for Home Win, Away Win or Draw. To do so, we plot a PieChart showing the mean of the bets from 7 different Bookmakers (Betting Houses).

Source Code 16.



As we can notice from our graph generally, houses tend to predict that the away team will win. However, according to the answer of our first question the team that wins more often is the Home Team. Therefore, we can conclude that Bookmakers win more money than bettors, because people who place a bet (generally on away teams as we can see in the pie chart) lose.

Finally, we will conclude with our last question: **Do the betting statistics change in the closing odds?**

To answer this question, we will use the data from the Pinnacle Betting House

PSCH = Pinnacle closing home win odds

PSCD = Pinnacle closing draw odds

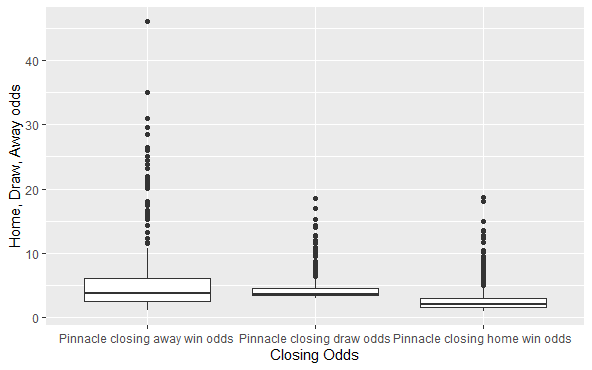
PSCA = Pinnacle closing away win odds

As we have discovered before, betting houses tend to put a higher bet in the Away Teams. Nonetheless, as it is said that closing odds tend to be more accurate, we want to observe if these closing odds present any change from the normal odds.

As we may have heard the closing odds are based on how the market has reacted to all available information, statistics and news about the team that has been received in the time between the original odds being released and the match start.

We will use a boxplot in order to examine the closing odds according to the shape of the distribution, and its variability

Source Code 17.



As we can observe, there are many outliers, which means that the values vary a lot going from 3-4 (which is the mean aproximattely) to higher values like for example 46 (aprox.) for the “closing away win odds”.

In conclusion we have noticed, that closing odds do not vary from the normal odds, because the higher values still being the ones from the Away Win Odds.

Finally, to sum up we definitely can affirm that this assignment has been very useful for us, because we have developed and enhance our abilities at using RProgram and doing Exploratory Data Analysis. Furthermore, we can see that we can take very useful information from a real data set if we work with the right tools. We know there is a lot more to learn and we would still practicing and learning.