SECOND STEP

Miguel Lobo Martin, Borja Ruiz Amantegui, Francesca Sallicati, Antonio MartÃnez PayÃ¡

December 10th 2017

Contents

1 Introduction

The dataset to study in this project is about the virtual football players in FIFA 2018 videogame. In this project we mainly want to establish a model which allows us to classify in Goalkeeper, Defender, Middle fielder and Attacker a possible new observation of a virtual football player. Neverhteless, there is also the qualitative variable Continent that could be studied, for instance, if there are important differences between players that belong to different continents.

The main purpose of the Second Step assignment is carry out a preliminar analysis of our data set. In order to do so, we cleaned the dataset, modifying and erasing some variables that we consider unrelated to the objective of our project.

The outline of this assignment comprises: first, a general plot of the data that could provide us with useful information for the following parts. Second, a descriptive analysis of the data, and last, a brief inter-relational study of the variables.

2 Cleaning data

The first thing to do before starting the analysis was cleaning the data. This part was necessary to refine all the raw data that we got from Internet. In order to do so, we followed different processes.

Substitution: the first data set contained symbols such as the Euro sign or different references such as M (millions in the first data set) or K (1000s). We used the function gsub, which helped us to convert those symbols into manageable data.

Delete: there were many variables that were unnecesary for the purpose of our project. Therefore, we chose to eliminate them. The variables were Photo, Club, Logo, Flag and ID.

Aggregation: we chose to reduce the number of qualitative variables due to the high variety of variables related to the skills that players have. The reduction method that we used is a simple average that will help us to aggregate variables that are closely related to each other. As an ilustrative method:

$$Passing = \frac{Crossing + Long.passing + Short.passing}{3}$$

This aggregation method has been used also to create other variables, such as Pace, Dribbles, Shooting, Defense, Passing, Physical and IQ.

Also, we used the aggregation method to simplify the number of positions in the soccer field that exist. The first categorization contained more than 25 positions, but we thought that by reducting them, we would gain control over the data set. At the end we ended up with four categories: Goalkeeper, Defender, Midlefielder and Attacker.

3 First Detection of Outliers

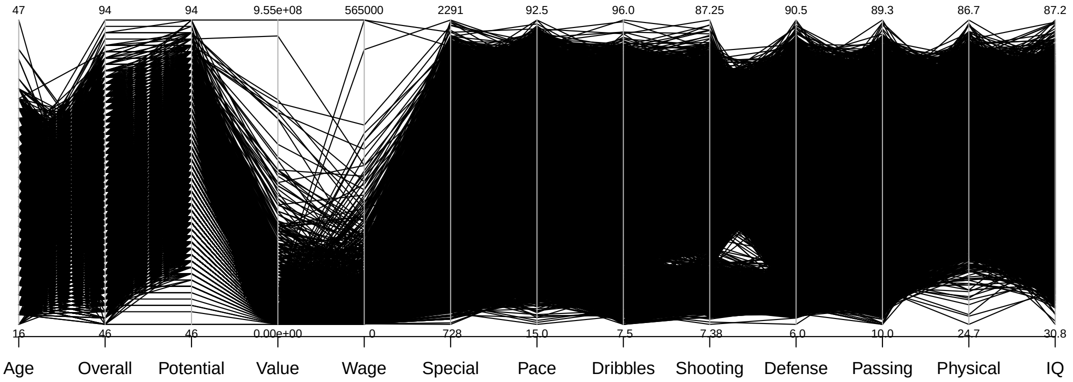
Prior to the core analysis of the data set we wanted to have a well informed idea about where the outliers could lay. In order to identify them we have created another data frame that only contains the quantitative variables. These are: “Age”, “Overall”, “Potential”, “Value”, “Wage”, “Special”, “Pace”, “Dribbles”, “Shooting”, “Defense”, “Passing”, “Physical”, “IQ”.

For this first detection of outliers we used the Parcoord function as well as the Mahalanobis distance (this second one is useful to detect outliers not depending only in a particular variable).

It should be remarked that we are not deleting the outliers, we are only locating them. If it is necessary in the future they will be deleted.

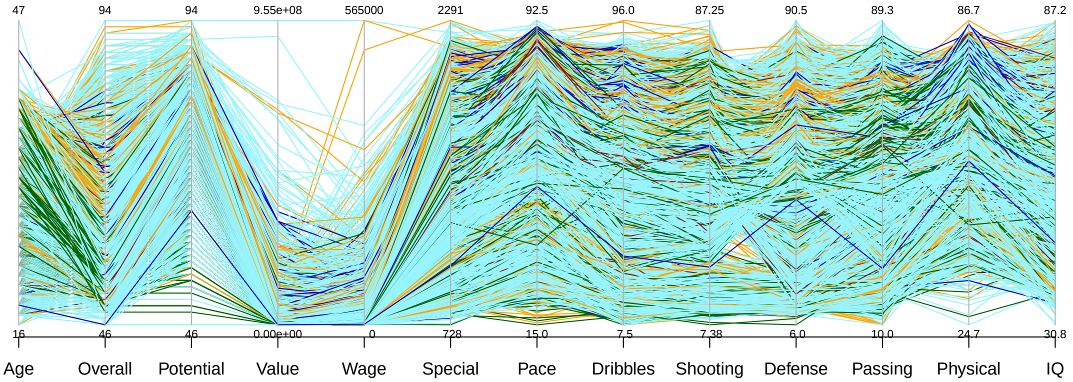
3.1 Parcoords

We basically did three parcoords. The first parcoord graph corresponds to the whole new data frame:



There are some clear outliers in the Wage and Value variable. We will pay attention to other possible outliers in Age, Potential and Physical.

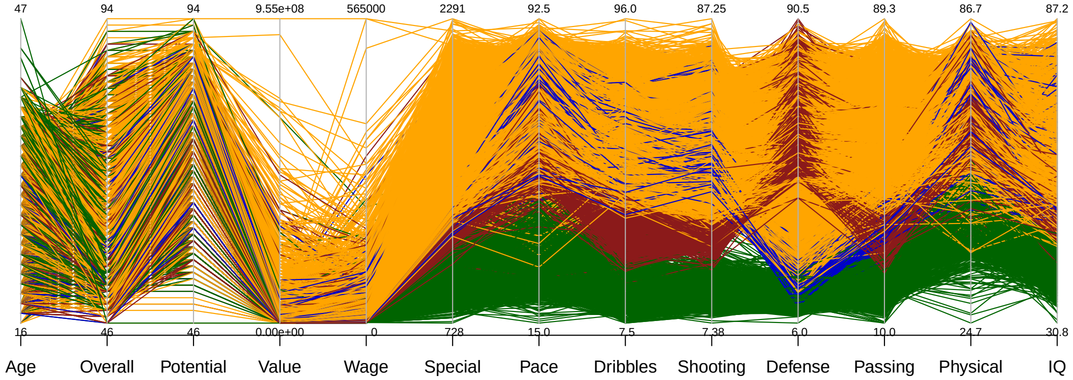
In order to further identify the presence of outliers, we have diferenciated the data corresponded to soccer players based on the continent where they were born. We have considered the following categories: Africa, America, Asia, Europe and Pacific.



Due to complexity of the graph it has not been possible to add a legend. The colors represent: AFRICA (blue), AMERICA (orange), ASIA (green), EUROPE (light blue), PACIFIC (red).

We spotted outliers in: Age (Africa and Europe), Value (America and Europe), Wage (America and Europe).

Moreover, we wanted to clasiffy the dataset based on the preffered position where players normally locate themselves inside the soccer field. The four categories that we used will be also applied during the assignement. Those categories are: Goalkeeper, Defender, Midfielder and Atacker.

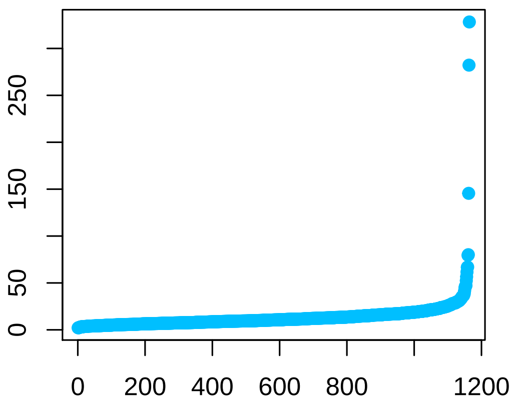


The colors are: Atacker (blue), Defender (red), Goalkeeper (green) and Midfielder (orange). This classification allows us to graphically identify the differences among the different positions. The possible outliers in the Age and Physical variable correspond to the Goalkeeper, while the Wage and Value variable to Midfielder. This result should not be surprising since most of the highest paid players, such as Cristiano Ronaldo or Lionel Messi are classified as Midfielders based on the Eufa classification.

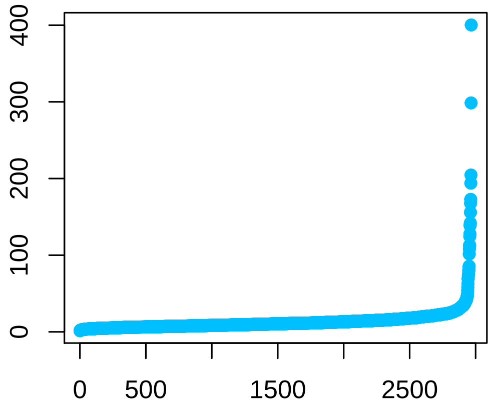
3.2 Mahalanobis distances

In order to further identify outliers, we used the Mahalanobis distance. Assuming normality in our data, the Mahalanobis Distance is a good measure of what we can consider as an outlier since it will measure the distance of each individual to the vector mean. Observing the 4 graphs below, one per category: Goalkeeper, Defender, Middle Fielder and Attacker, we can notice in every graph that we have some outliers.

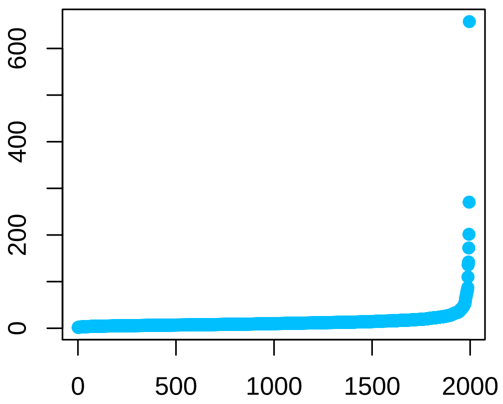
In this first assignment we are going to keep in our data these outliers, but we will take into consideration their existance for the future development of the study.



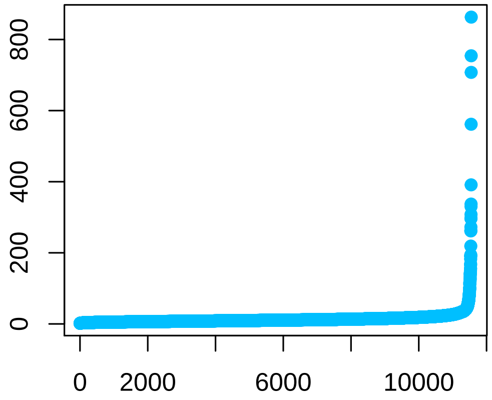
Mahalanobis Attackers



Mahalanobis Defenders



Mahalanobis Goalkeepers



Mahalanobis Midfielders

Maximun value for mahalanobis Attacker = 328.2288

Maximun value for mahalanobis Defender = 400.3001

Maximun value for mahalanobis Goalkeeper = 657.5702

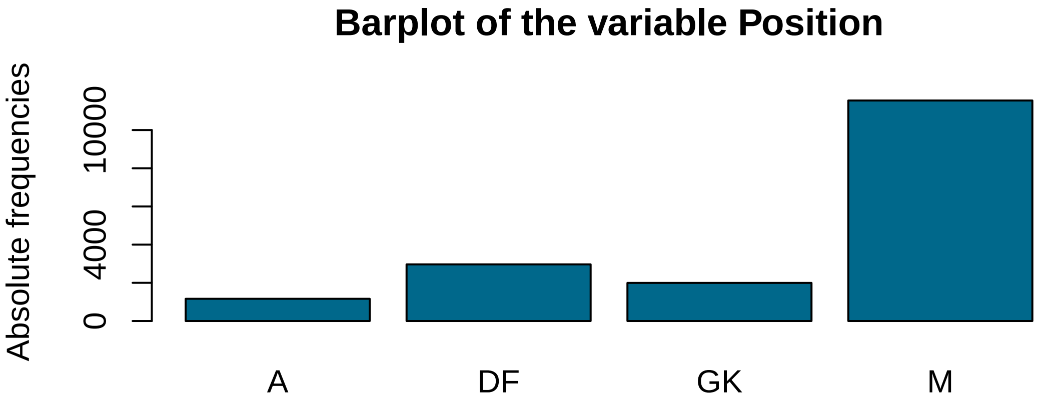
Maximun value for mahalanobis Midfielder = 862.9238

In the graphs avobe we can easily see the outliers. The maximun value for the mahalanobis distance corresponds to the Midfielder. This is something that we expected, since the best players in the world (Cristiano and Messi) are included in this category.

4 Descriptive Analysis of the Variables

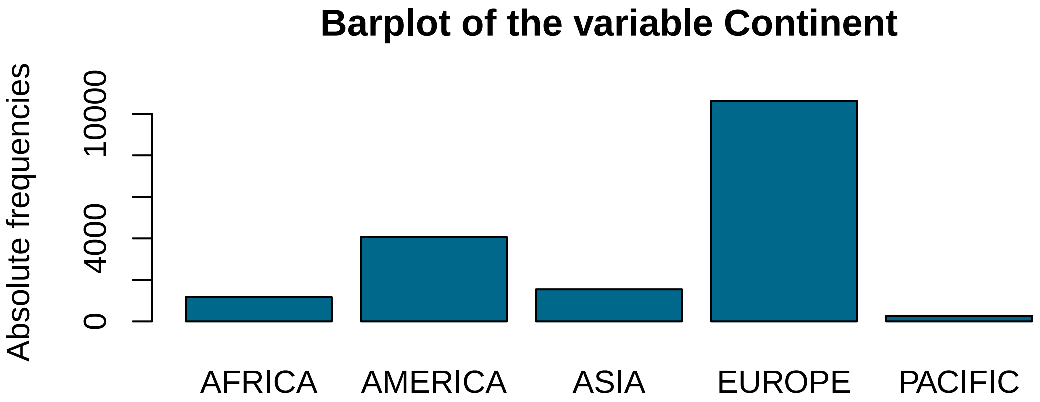
4.1 Position

That is the main variable, the one that we want to predict in the future. Tha players are classified following the Uefa criterias. Is remarkable the large number of players that classified as Midfielder.



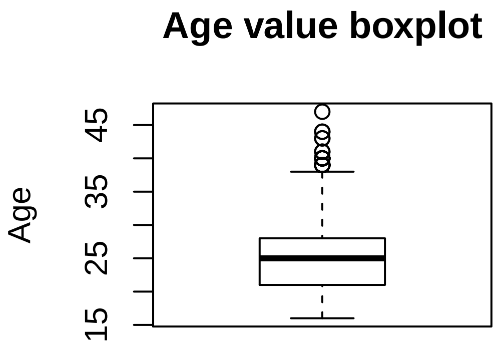
4.2 Continent

That is the other qualitative variable we are going to use in our study. Is also remarkable the large number of european players.



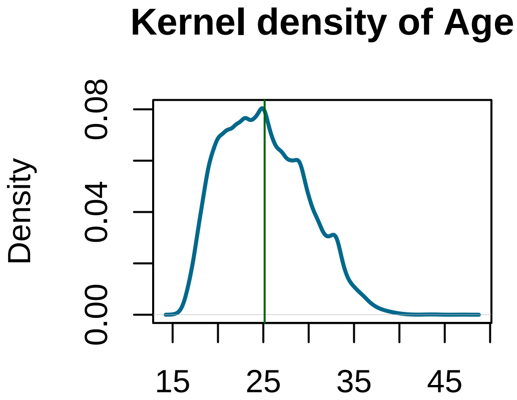
4.3 Age

The first variable is Age. The plots that we showed in the last part suggested the presence of outliers. The boxplot will help us identifying those points:



As we can see there are some outliers. The boxplot yields an upper bound value of 38.5. For this variable the quartiles are:

Min	1st Qu.	Median	3rd Qu.
16.00	21.00	25.00	28.00



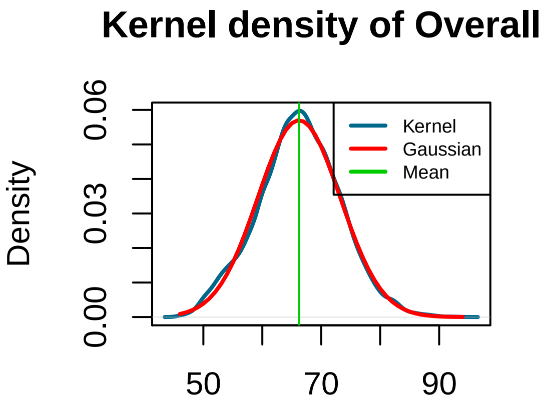
We have used the Kernel density estimation method to depict the distribution of the Age variable. The green line corresponds to the average value, that is showed below.

Mean	Standard Devation
25.14454	4.614272

Mean	Standard Devation

4.4 Overall

This is the variable in which Eufa gives a final score for each player. We do not know how this value is given for each player, but we could assume that it could take into account certain parameters that are measured.

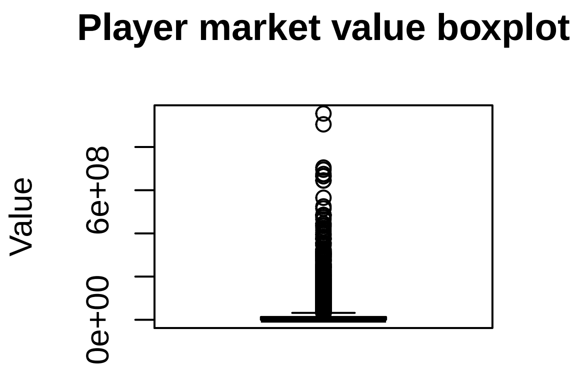


In the previous graph we plotted the Kernel and the Gaussian distribution function just as a matter of comparison.

Mean	Standard Devation
66.22545	6.998794

4.5 Value

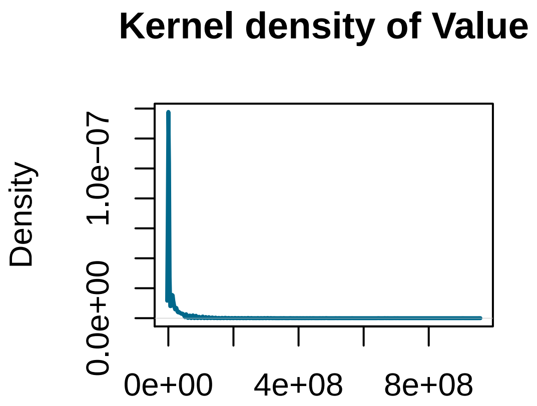
Value reflects the money a player costs in the current market. The Value variable was at a first glance recognized as a variable with multiple outliers. The following graph confirms what we expected:



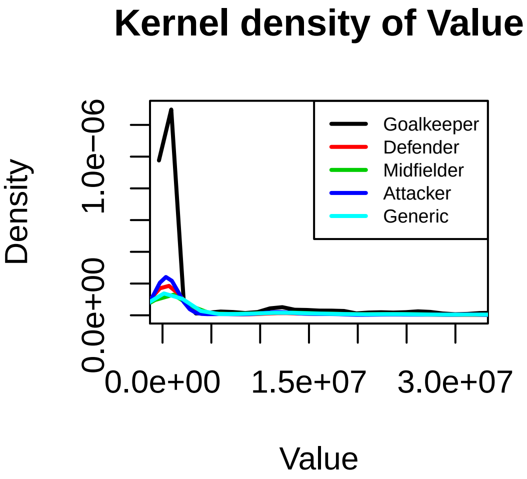
Based on the summary of the variable, we calculated the upperbound.

Min	1st Qu.	Median	Mean	3rd Qu
0	300000	675000	13480000	13000000

The upperbound value is 32050000. Therefore, all the players' value that is above this number could be considered as an outlier.

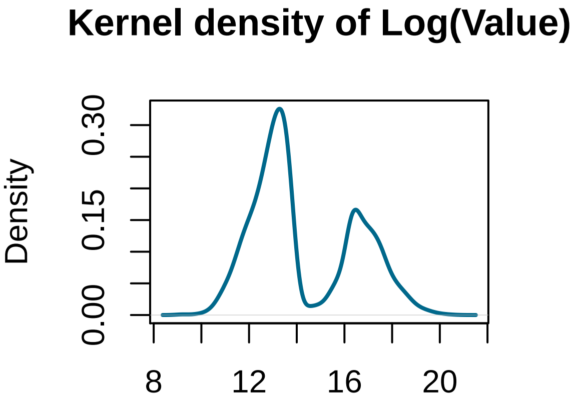


Due to the presence of the outliers the previous graph does not provide with useful information. Therefore, we created a kernel density function for each of the four players' categories.



It is possible to extract useful information from the previous graph. The most important point is that Goalkeeprs' value is, on average, much smaller than the other players'. However, we can not infer any further information about the other three categories since the density function is much alike.

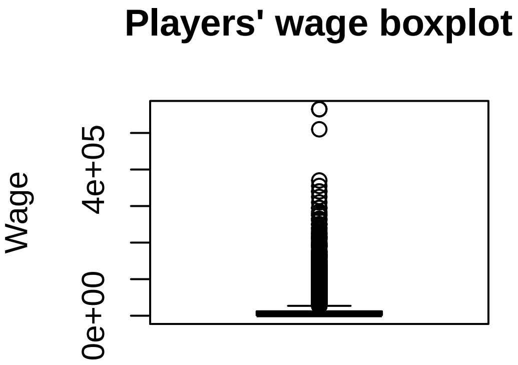
Mean	Standard Devation
13480394	37355377



Applying a logarithm we can easily notice the behaviour of the variable.

4.6 Wage

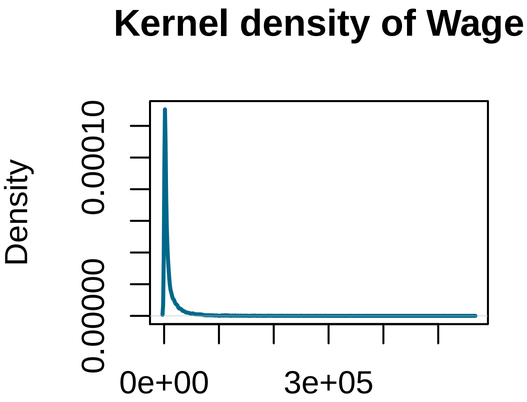
Something very similar to the last case happens to the Wage variable.

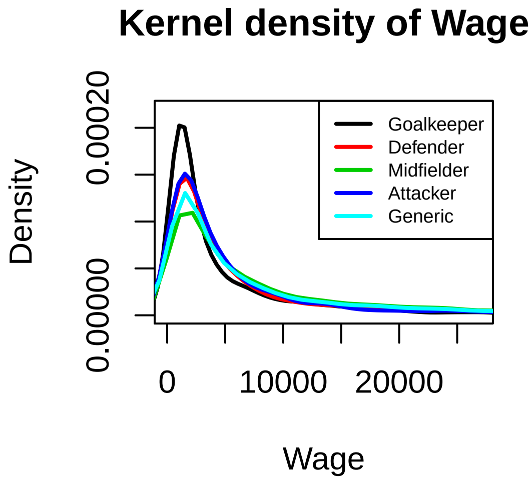


Min	1st Qu.	Median	Mean	3rd Qu
0	1750	4000	11530	12000

The boxplot shows an important quantity of outliers. This situation arises from the fact that most of the players' wage is in the lower part of the graph. However, there a relatively small number of players that are employed by powerful soccer club and get very high pay checks.

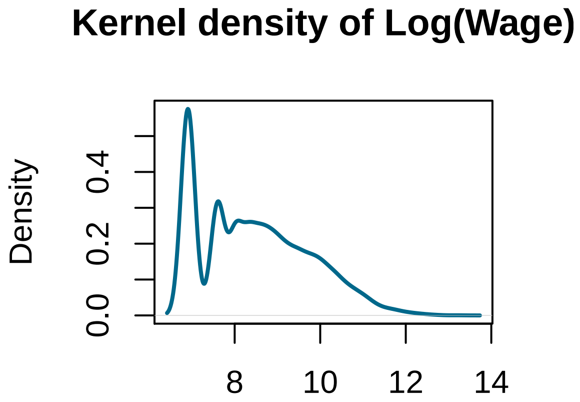
The upperbound in this case is 27000.





Once again, we separated the data into four different categories, that correspond to the four different locations on the soccer field. As we can see in the graph, the flatter density line corresponds to Midfielder, which indicates a higher average Wage.

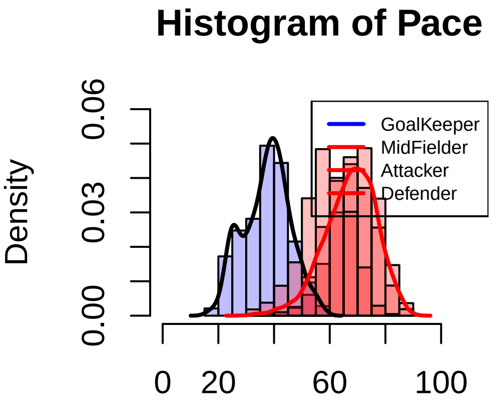
Mean	Standard Devation
11532.19	23156.03



Applying a logarithm we can easily notice the behaviour of the variable.

4.7 Pace

In this variable we have separated the distribution based on the preffered position on the field. This approach is based on the findings of the first plot of the assignment. This plot suggested a big gap in players' skills based on their position.

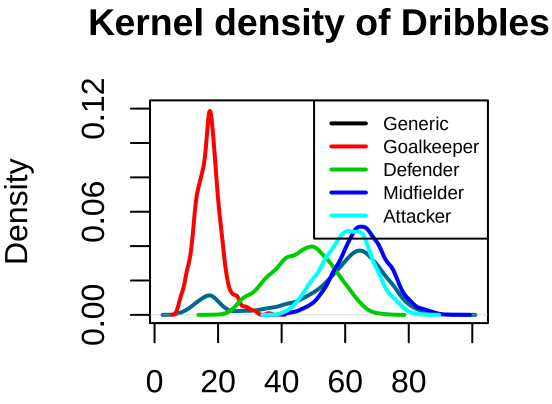


The histogram that is showed above indicates an approximation of the distribution of four division of the Pace variable. As it was expected Goalkeepers normally have a worse ability than other players that get in contact with the ball more often.

The two Kernel density functions correspond to the Goalkeeper and all-non-Goalkeeper positions, highlighting that way the skill difference between these two "collectives".

4.8 Dribbles

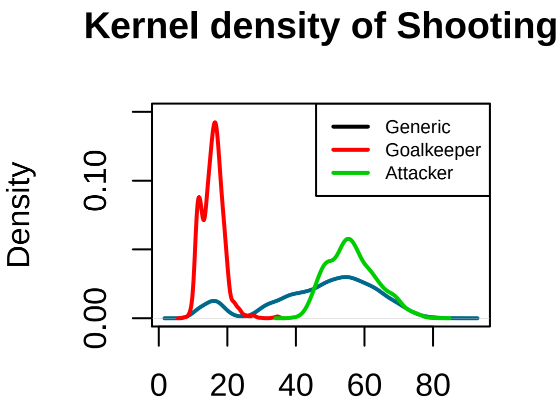
This variable measures the ability of dribbling in the player.



Mean	Standard Devation
56.56333	17.60328

4.9 Shooting

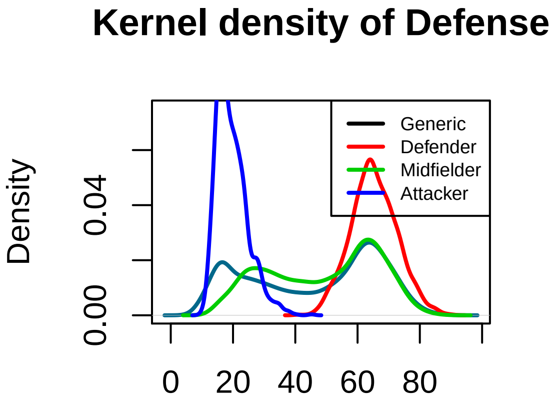
This variable measures the aspects related with the shoot of a player.



Mean	Standard Devation
47.87082	15.64134

4.10 Defense

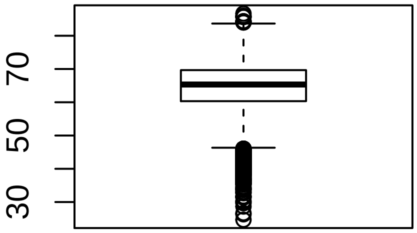
This variable measures the how much a player can defend in the team.



Mean	Standard Devation
45.92149	20.95529

4.11 Physical

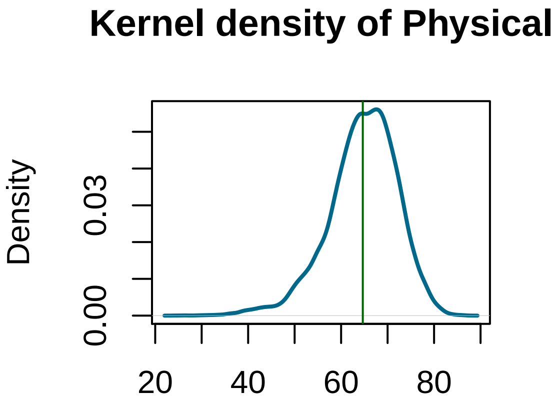
This variable measures the physical capacity of a player. In the Physical variable we spotted some possible outliers when we first analyzed the data.



The boxplot shows multiple outliers that are in both sides.

Min	1st Qu.	Median	Mean	3rd Qu
24.67	60.33	65.33	64.65	69.67

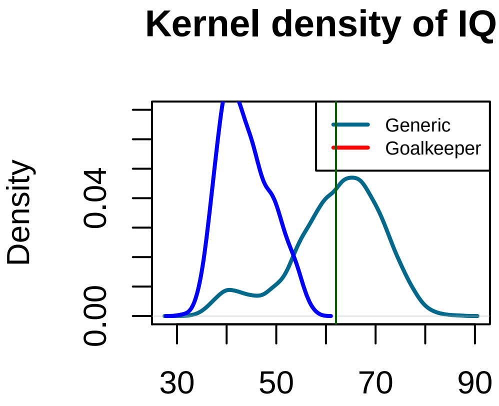
The lower bound is 55.66, while the upper bound is 83.68.



Mean	Standard Devation
64.64745	7.444996

4.12 IQ

The variable IQ measures how well a player adapts himself to a change in the soccer game.

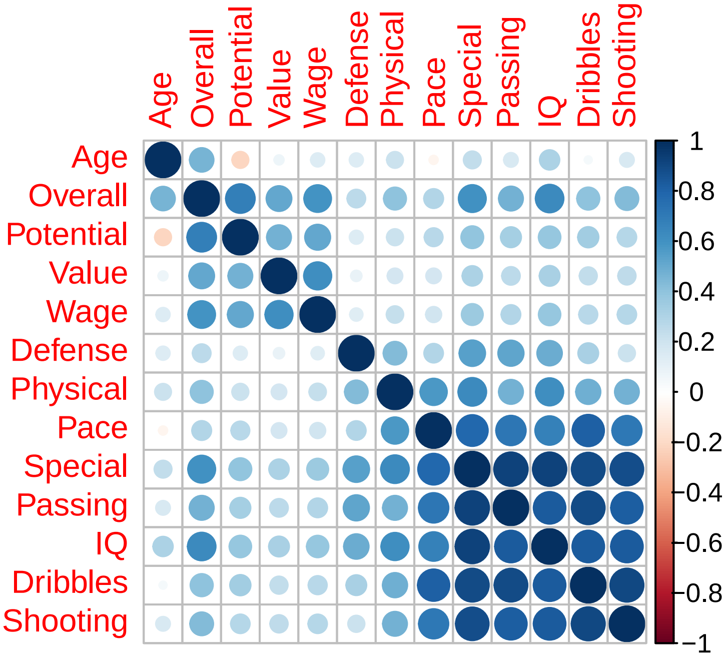


We can see a big difference in the IQ that Goalkeepers have relatively to the other players.

Mean	Standard Devation
62.0292	9.352232

5 Correlations in the data

To study the correlations in the data, we simply calculated the correlation matrix and interpreted it. For its easy visualization is useful the corrplot function in R, obtaining the following plot:



There are some remarkable things from the graph above:

1) Overall is correlated with Potential, which seems reasonable because potential is an important variable that measures how good a player could be in the future. Besides it is correlated with the variable Special (we could think that the variable Special is one of the most important ones). This variable is correlated with the IQ too (IQ is the one that is most correlated with). We could think that this fact is quite curious. Why IQ?

2) Overall is also correlated (but not so much) with the variables Value and Wage, which is logical because of the market behaviour. Also due to market reasons the variables Value and Wage are correlated with the variable Potential (the better carrer expectation a player has the more you are going to pay for him). And of course, Value and Wage are correlated between them.

3) We can highly appreciate a set of correlations between the variables Pace, Passing, Special, IQ, Dribbles and Shooting. For the future study of our dataset, we should take into account that we have a correlation matrix with high correlations among the variables.

Another point to take into account is that the Covariance Matrix Determinant is 2.974204e+42. That is so large, being possible that affects to some calculations, for instance the determinant of its inverse will be close to cero.