FIFA19 the Analysis

Daisy Ambaum

28 februari 2019

Introduction

The dataset:

After learning different aspects from R during the course "Introduction to DataScience in R", it is finally time to use all the knowledge on a dataset chosen by me. For this project I chose to use the FIFA19 dataset with the intention of looking if I could find bargains and predict the value of players.

The dataset consists of 18206 players from all different competitions. With this I mean that there is not only 1 competition like La Liga in Spain, or Eredivisie in the Netherlands. From all those 18000+ players I looked for example at their values, wages and potential. But the dataset includes way more: preferred foot, weight, height, nationality etc. Let's just have a quick look what is all included before I will dive into the goal of this project:

```
colnames(fifa)
```

```
##  [1] "ID"                       "Name"
##  [3] "Age"                      "Nationality"
##  [5] "Overall"                  "Potential"
##  [7] "Club"                     "Value"
##  [9] "Value_E"                  "Wage"
## [11] "Wage_E"                   "Preferred Foot"
## [13] "Position"                 "Jersey Number"
## [15] "Height"                   "Weight"
## [17] "Skill Moves"              "Weak Foot"
## [19] "International Reputation" "Crossing"
## [21] "Finishing"                "HeadingAccuracy"
## [23] "ShortPassing"             "Volleys"
## [25] "Dribbling"                "Curve"
## [27] "FKAccuracy"               "LongPassing"
## [29] "BallControl"              "Acceleration"
## [31] "SprintSpeed"              "Agility"
## [33] "Reactions"                "Balance"
## [35] "ShotPower"                "Jumping"
## [37] "Stamina"                  "Strength"
## [39] "LongShots"                "Aggression"
## [41] "Interceptions"            "Positioning"
## [43] "Vision"                   "Penalties"
## [45] "Composure"                "Marking"
## [47] "StandingTackle"           "SlidingTackle"
## [49] "GKDiving"                 "GKHandling"
## [51] "GKKicking"                "GKPositioning"
## [53] "GKReflexes"
```

Before accessing the dataset, I first looked at it using Excel and over there I made the first changes. For example I deleted columns that where redundant in my opinion (e.g. Position and all different scores on the different positions, in this case I chose to only keep the "overall position"). The adjusted dataset is available to the reader.

The goal:

The goal of this report is to take the reader by the hand in the whole process I have gone through, and show the outcome of the process. The goal I set for myself when chosing this dataset is to predict a player's value and see if I can find a player with high overall rating, but who can be "picked up" for fairly nothing. Besides that I want to be able to finds fairly similar players, so I can have a choice in selecting players without having to pay the jackpot (pricewise).

Key steps:

In order to come to the set goal, I preformed serveral steps which can be listed as the following:

1. Preparing and "clean" the data
2. Analyzing the dataset and creating multiple charts for better insights
3. Check different approaches (naive Bayes, Spearman, Pearson)
4. Apply K-Means clustering to the data
5. Find similar players
6. Predict player's Value
7. Look for absolute bargains

This whole list will result in the finding of the right players and supporting the goal of this project. In this report I will make sure the following aspects are outlined:

- Methods & analysis (as stated in point 2 - 4 above)
- Show the results (as stated in point 5 - 7 above)
- Come with a solid conclusion on the goal set: predicting value, bargains and similar players.

So let's dive into the project! NOTE: In this document the commenting from the R-Script has been partly removed to improve readability. For all comments, I would like to refer to the R-Script. In this document accompaning explanation will be done in the text.

Data Analysis –> Analyze & visualize the data

Data preparation

Before I can actually start with the data analysis, I first made sure all the data was actually understood by me. So I first dove into the columns and their representation. As we already saw the columns above, I will not repeat them here. Though there are a few columns that need changing:

- Weight –> this one is in pounds, but being from the Netherlands, I prefer to look at data in kilogram
- Height –> this one is in inches, but same as for weight, in the Netherlands we use centimeters, so I want to change this as well
- Wage & Value –> they are presented as characters, so they need to be converted to numeric as I want to calculate with those.

The way I did that, is show in the script below. Notice that for **weight** it was way easier than it was to covert the **height** of a player. This is due to the fact that **height** has the feet-sign in the middle and there is

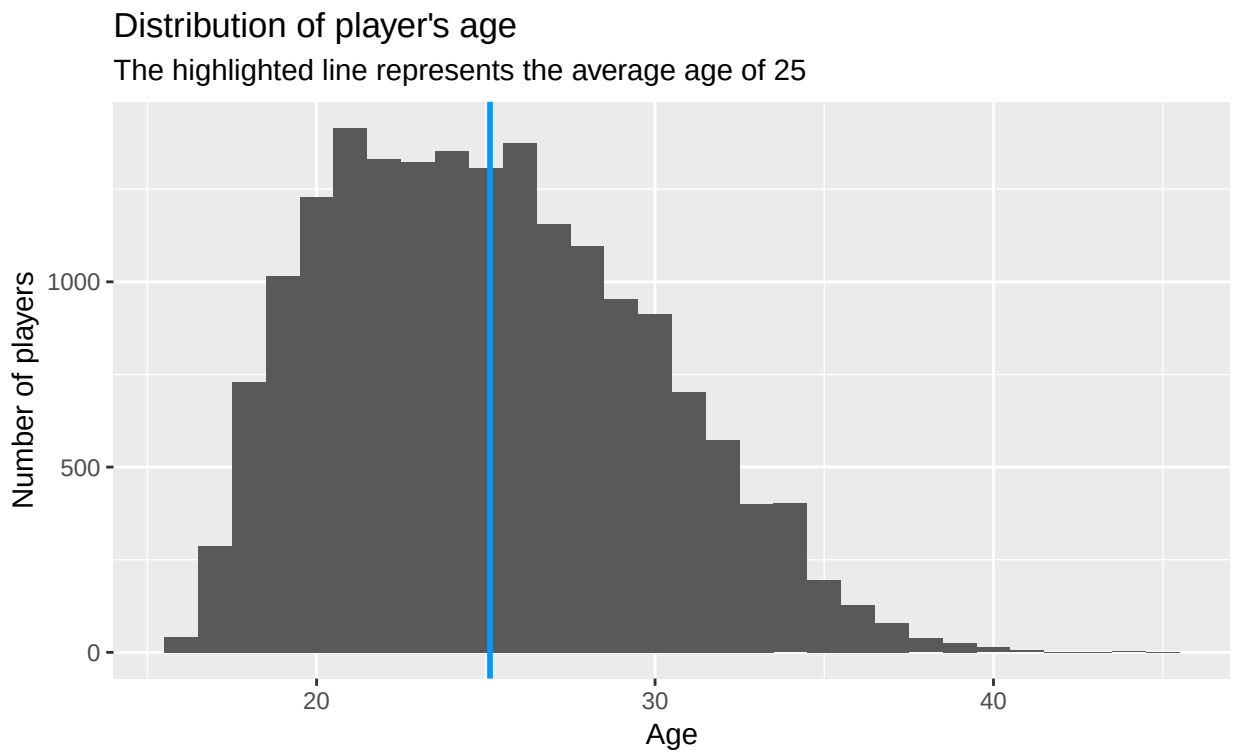
an extra calculation needed to come to only inches (feet have to be converted to inches to be able to convert to total to cm).

```
fifa$Weight <-  str_remove(fifa$Weight, "lbs")

fifa <- as.data.frame(fifa) %>%
  mutate(Value = as.numeric(as.character(Value)),
         Wage = as.numeric(as.character(Wage)),
         Weight = as.numeric(as.character(Weight)))

  # Just in case when some NAs would appear, let's remove them before continueing.
fifa <- fifa[complete.cases(fifa), ]

temp <-  str_split(fifa$Height, "'")

for(i in 1:length(temp)){
  temp[[i]] <- as.numeric(temp[[i]])
}

for (i in 1:length(temp)) {
  temp[[i]] <-  (temp[[i]][1] *12)+temp[[i]][2]
}
temp2 <- as.numeric(unlist(temp))
fifa$Height <- temp2

# By using the "measurement" package, I can easily convert pounds to kg and inches to centimeters witho
fifa$Weight <- conv_unit(fifa$Weight, "lbs", "kg")
fifa$Height <- conv_unit(fifa$Height, "inch", "cm")
```

Now that all the data is correct and presented in the way that I prefer to use them, I can move on to the actual data analysis.

Data anaysis & visualizations

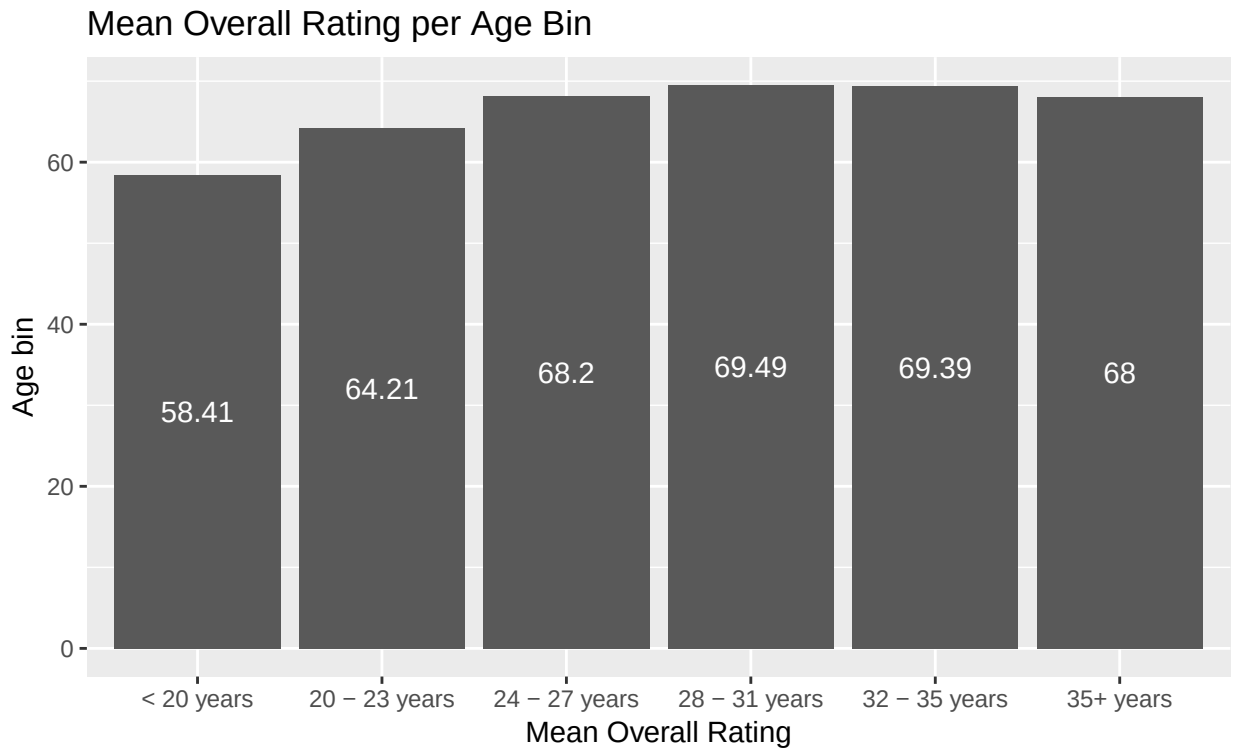
In the first analysis, I would like to stick to the player and the aspect he cannot change anything about: his age. First it is important to know the age distribution as this could influence various of other attributes like value, wage or even skills.

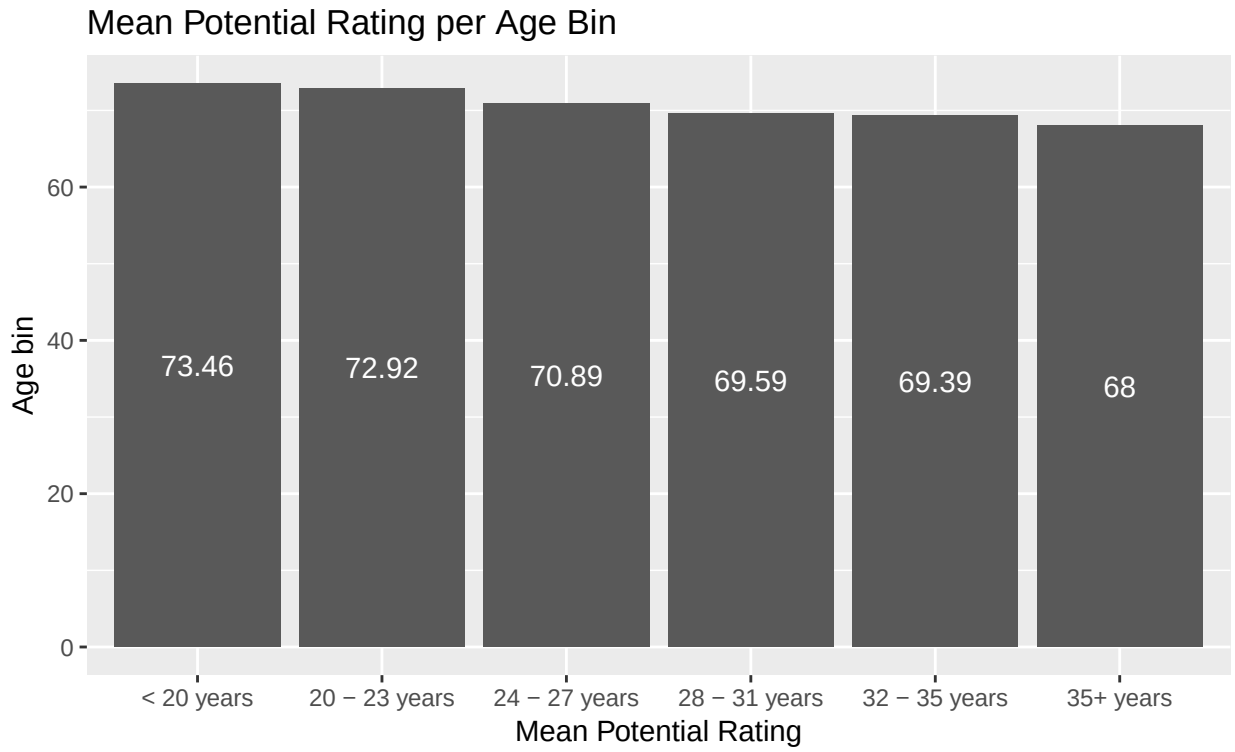


You can see that most players are between the 20 and 30 years old, and after 30 they are rapidly disappearing. This might be because football is a contactsport, and players who are a bit older tend to have longer lasting injuries or they can't compete with the younger talents anymore so they stop playing all together.

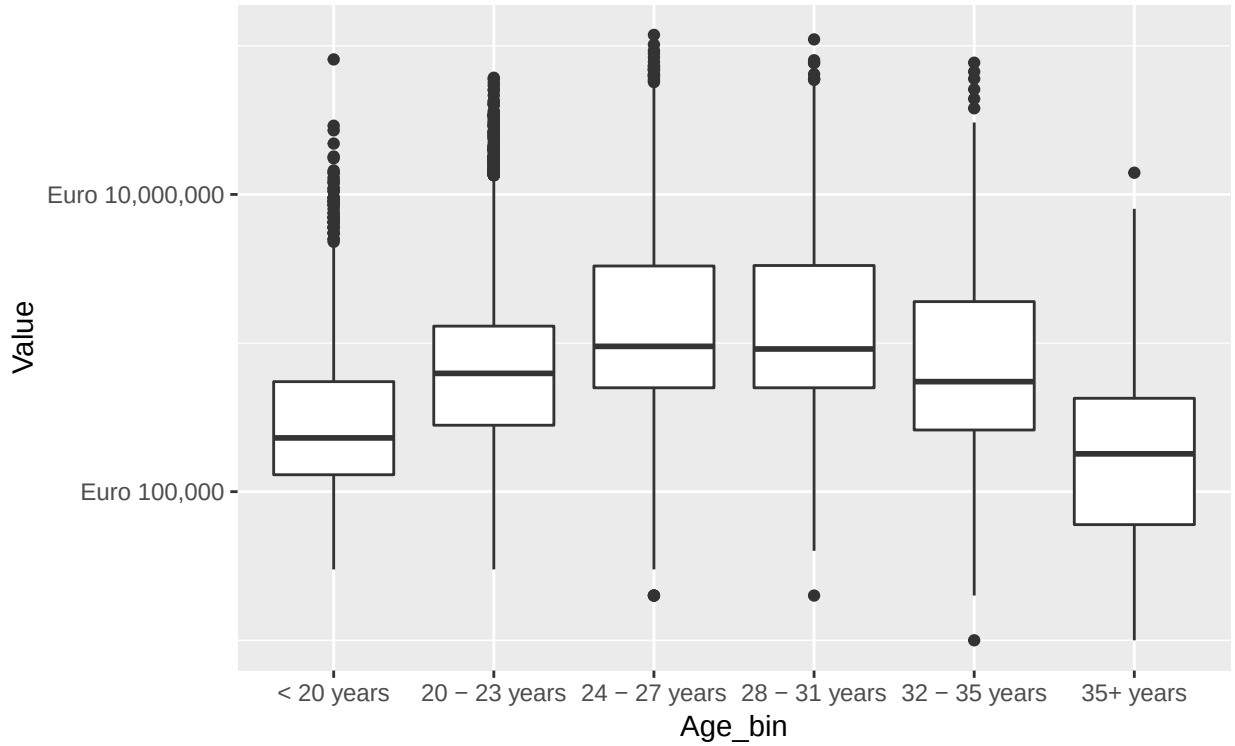
This distribution is important for the next step, as I will create 6 bins of the aging. The next analysis will become unclear when we visualize them per age as there are too many for a clear image. So I will create the following **age_bins**: $<$ 20 years; 20-23 years; 24-27 years; 28-31 years; 32-35 years; 35+ years.

Those age_bins I will use to see the difference in Overall Rating and Potential Rating, as I want to know if it is true that younger players have a higher potential than the older players.





So it is true that younger players have a higher potential than an overall rating. The older players (there are also fewer) already reached their potential. This is very good visable in the age_bin of **32-35 years**. There is no difference between the overall rating and the potential rating. So I assume that a player's value goes hand in hand with that, but let's see if we can prove this with the data



Yes this appears to be true. Though of course there are always outliers, and in this dataset we have many. I did not take them out, as they are of high importance of the market. There is more and more the tendency of overpaying players just to have them come and play for your team (which is also why older players leave for China or the USA). Like Paris Saint Germain (PSG) did with Neymar, making him the most expensive player: *Neymar from Barcelona to PSG for $ 261 million*

So does this mean that, because PSG was willing to pay that much money for Neymar Jr, he is also the most valued player in his age category? Let's just have a quick look at the most valued players per age category:

```
## # A tibble: 6 x 4
## # Groups:   Age_bin [6]
##   Name              Age_bin       Potential     Value
##   <chr>             <chr>             <dbl>     <dbl>
## 1 K. Mbappe         < 20 years         95.0  81000000
## 2 L. Sane           20 - 23 years      92.0  61000000
## 3 Neymar Jr         24 - 27 years      93.0 118500000
## 4 L. Messi          28 - 31 years      94.0 110500000
## 5 Cristiano Ronaldo 32 - 35 years      94.0  77000000
## 6 Z. Ibrahimovic    35+ years          85.0  14000000
```

Yes Neymar Jr is the most valued player in this age category, but is he also the player with the most potential? As it is rather logical that a orice paid for a player is influencing his value, but it cannot influence it's potential can it?

```
## # A tibble: 10 x 4
## # Groups:   Age_bin [6]
##    Name              Age_bin       Potential     Value
##    <chr>             <chr>             <dbl>     <dbl>
##  1 K. Mbappe         < 20 years         95.0  81000000
##  2 L. Sane           20 - 23 years      92.0  61000000
##  3 M. skriniar       20 - 23 years      92.0  46500000
```

```
##  4 Marco Asensio     20 - 23 years      92.0  54000000
##  5 O. Dembele        20 - 23 years      92.0  40000000
##  6 Gabriel Jesus     20 - 23 years      92.0  41000000
##  7 P. Dybala         24 - 27 years      94.0  89000000
##  8 L. Messi          28 - 31 years      94.0 110500000
##  9 Cristiano Ronaldo 32 - 35 years      94.0  77000000
## 10 G. Buffon         35+ years          88.0   4000000
```

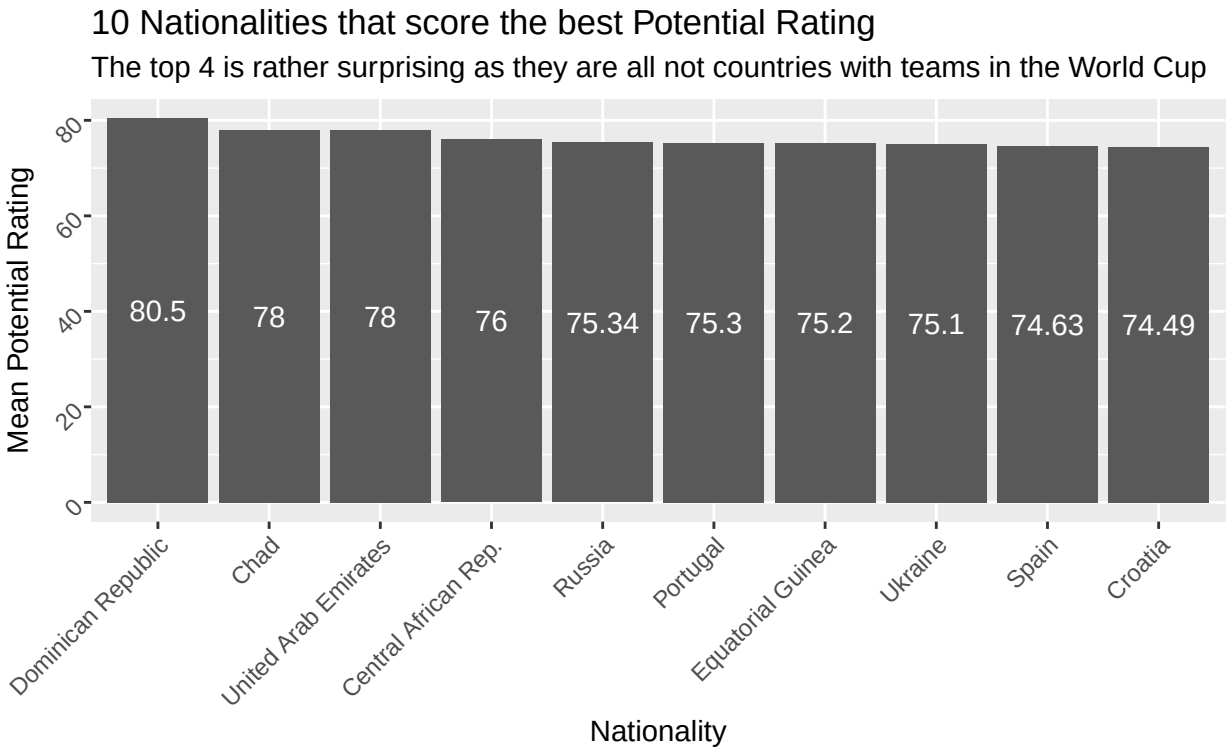
```
## # A tibble: 8 x 4
## # Groups:   Age_bin [6]
##   Name              Age_bin       Overall     Value
##   <chr>             <chr>           <dbl>     <dbl>
## 1 K. Mbappe         < 20 years       88.0  81000000
## 2 L. Sane           20 - 23 years    86.0  61000000
## 3 Bernardo Silva    20 - 23 years    86.0  59500000
## 4 R. Sterling       20 - 23 years    86.0  56500000
## 5 Neymar Jr         24 - 27 years    92.0 118500000
## 6 L. Messi          28 - 31 years    94.0 110500000
## 7 Cristiano Ronaldo 32 - 35 years    94.0  77000000
## 8 G. Buffon         35+ years        88.0   4000000
```

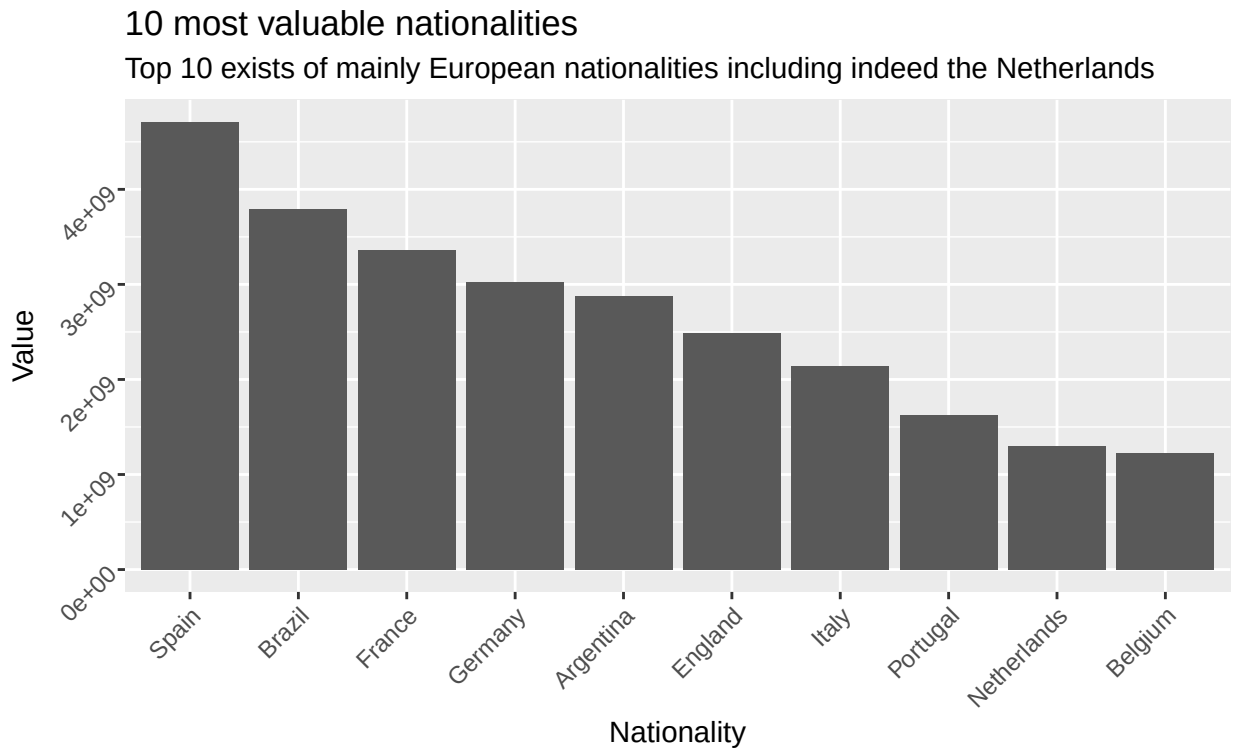
No, in this case Neymar Jr has not the higest potential, but he has still the highest overall rating within his age category. This should mean that **P. Dybala** will have a value higher than Neymar, or is that not how it works. Let's come back to that when we start predicting the player's value, if it is that easy to say.

On more aspect that is interesting to look at, is whether players have reached their potential already. Because if a player cannot grow anymore at age 20, than this is something we want to know before we say he is a "bargain".

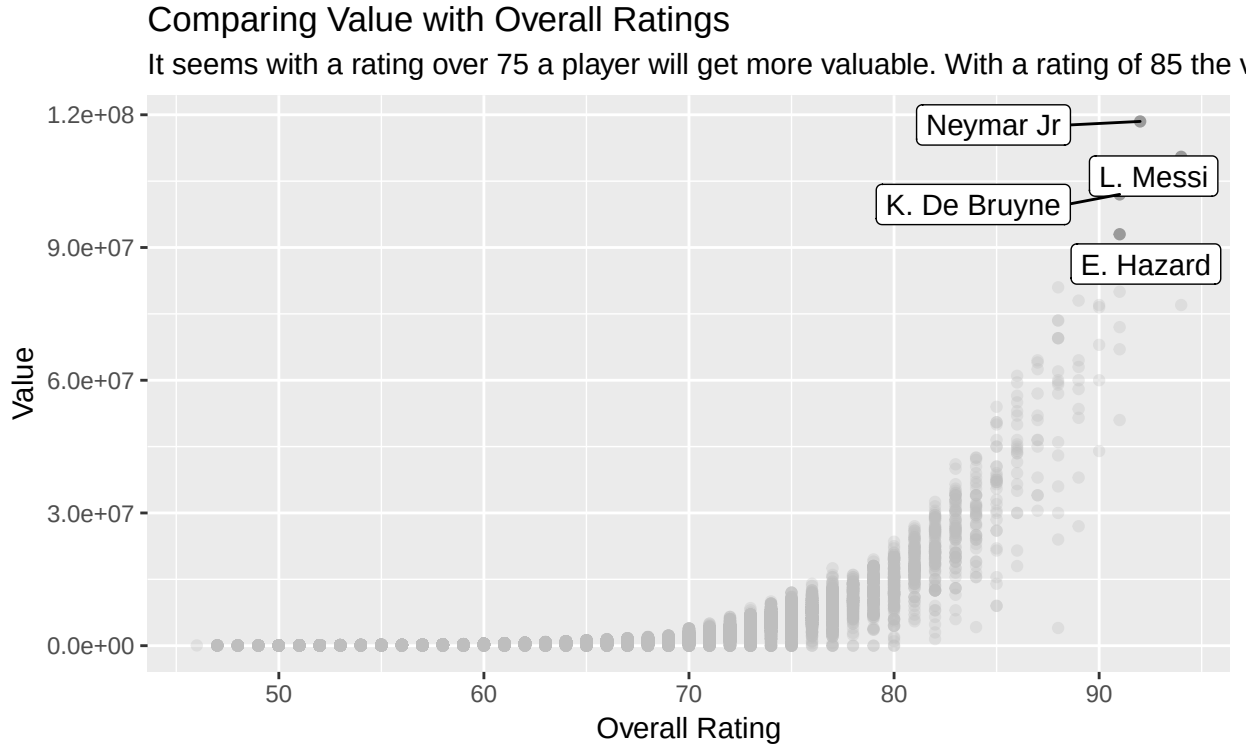
```
##            Name                Club Overall Potential       Age_bin
## 1     K. Mbappe Paris Saint-Germain      88        95    < 20 years
## 2     P. Dybala            Juventus      89        94 24 - 27 years
## 3     Neymar Jr Paris Saint-Germain      92        93 24 - 27 years
## 4        De Gea   Manchester United      91        93 24 - 27 years
## 5  K. De Bruyne     Manchester City      91        92 24 - 27 years
## 6       H. Kane   Tottenham Hotspur      89        91 24 - 27 years
## 7          Isco         Real Madrid      88        91 24 - 27 years
## 8    C. Eriksen   Tottenham Hotspur      88        91 24 - 27 years
## 9  A. Griezmann     Atletico Madrid      89        90 24 - 27 years
## 10     M. Salah           Liverpool      88        89 24 - 27 years
## 11 J. Rodriguez   FC Bayern Munchen      88        89 24 - 27 years
## 12     Coutinho        FC Barcelona      88        89 24 - 27 years
##        Value
## 1   81000000
## 2   89000000
## 3  118500000
## 4   72000000
## 5  102000000
## 6   83500000
## 7   73500000
## 8   73500000
## 9   78000000
## 10  69500000
## 11  69500000
## 12  69500000
```

Than it was last week that some news on the Dutch radio made me want to include an extra part in this analysis. It was stated that Dutch players were preferred to play in Spain. Probably this has something to do with the transfer of Frenkie de Jong going to Barcelona which made me think this news-item was biased. So let's take a look at the nationalities with the highest potential and value and see if indeed the Dutch are so highly valued.



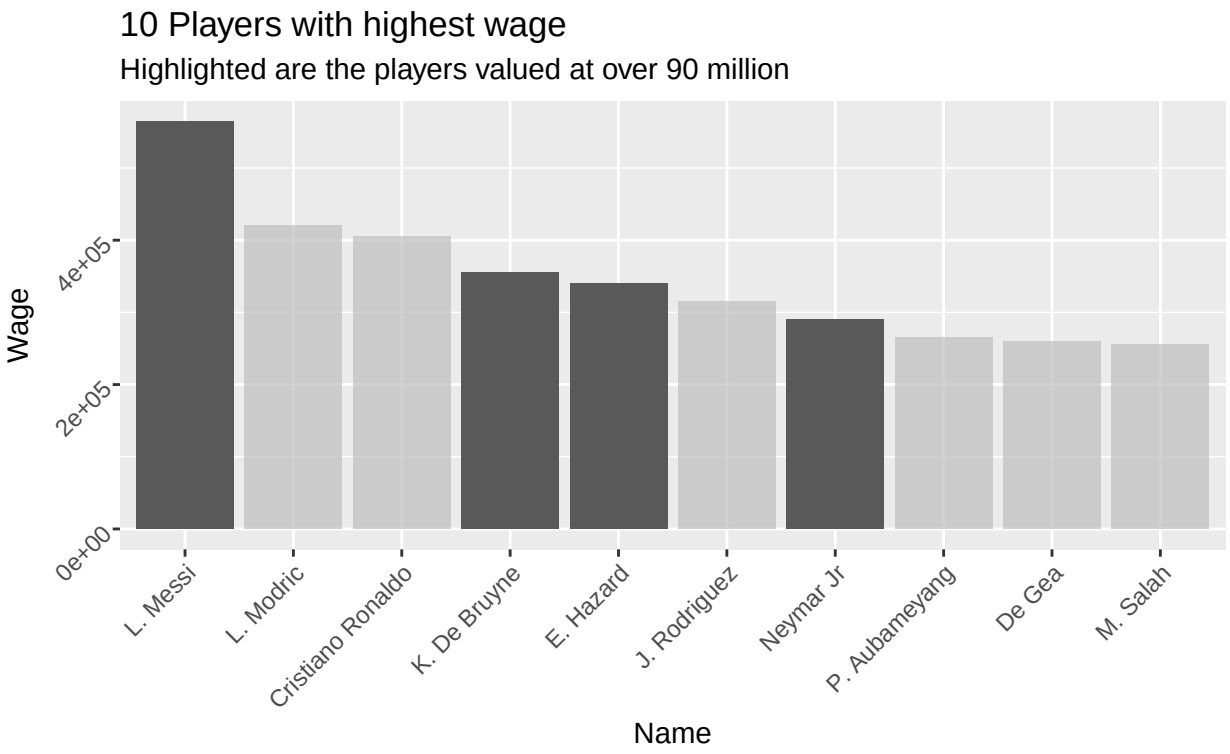


So we now already peaked a little bit at the value of certain players, and it sounds logical that the higher the overall rating of a player is, the higher his value is. I would like to create a chart where I can see the relation between the overall rating and the value of a player to support my thoughts.

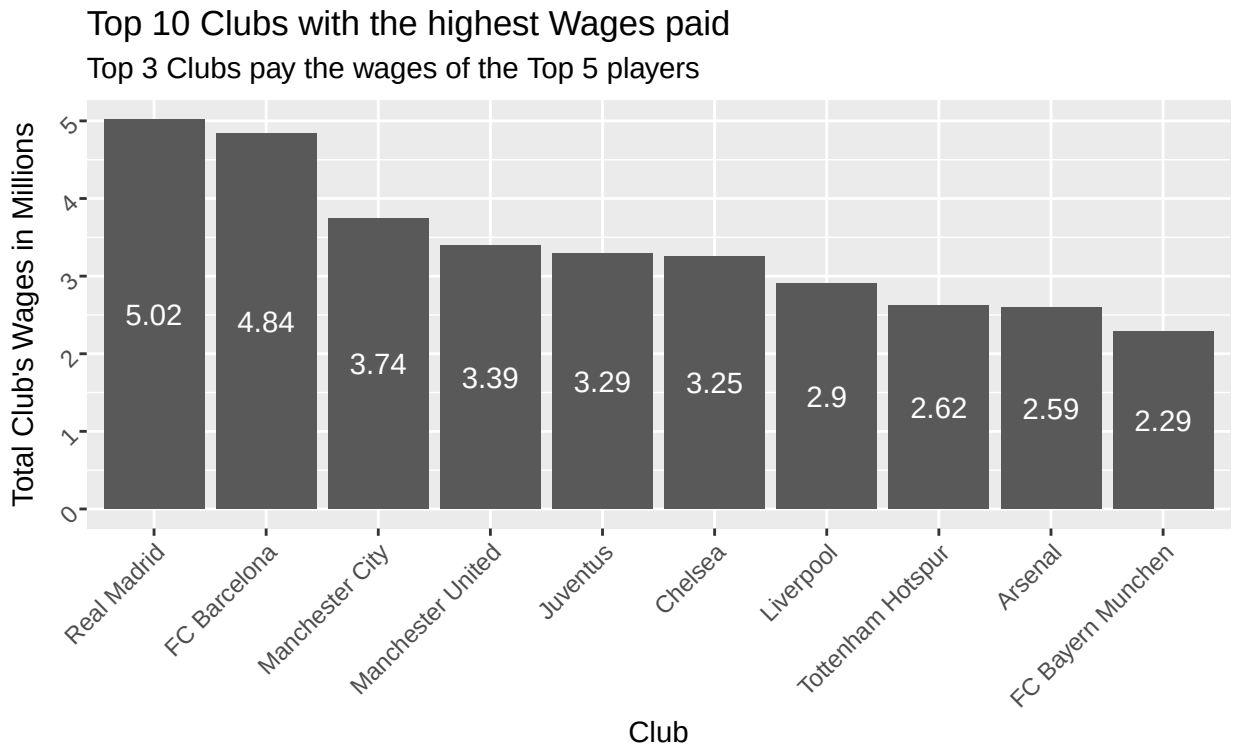


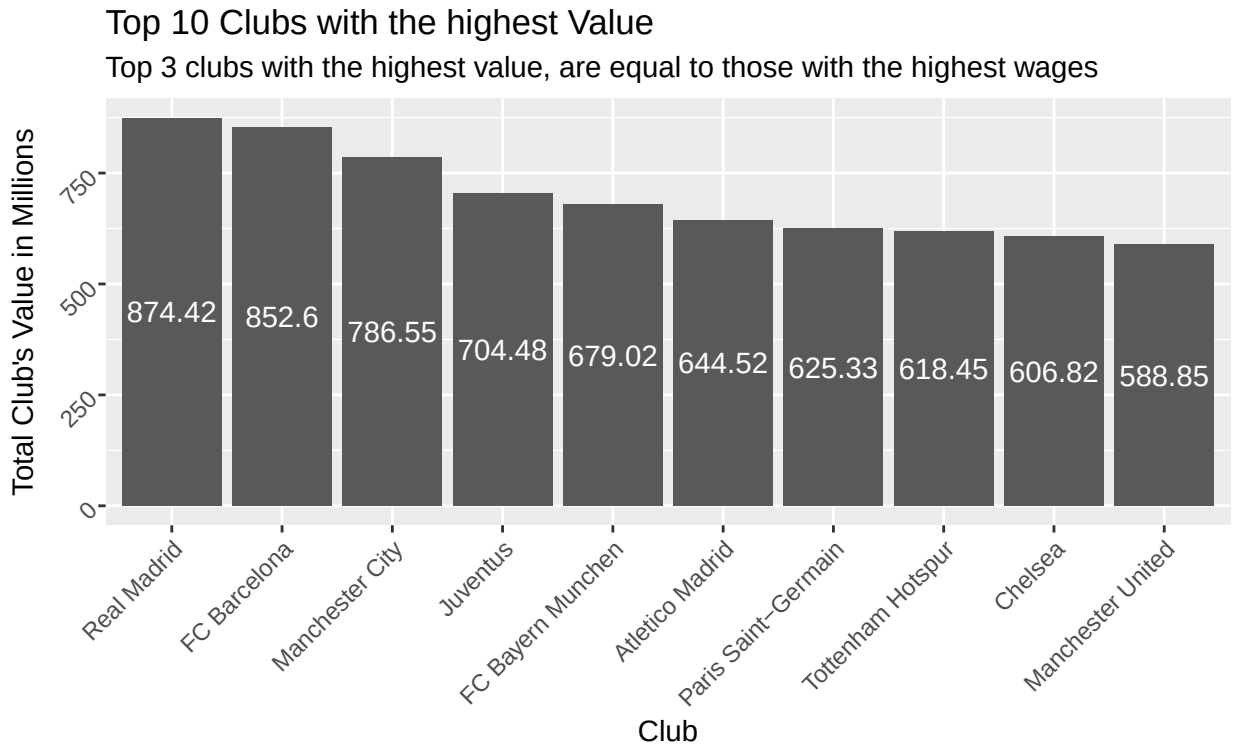
To we can see there is a top 4 of players who have a value over 90.000.000 euro. Let's see if those players are

than also the players who get paid the most as you would expec (or at least I do).



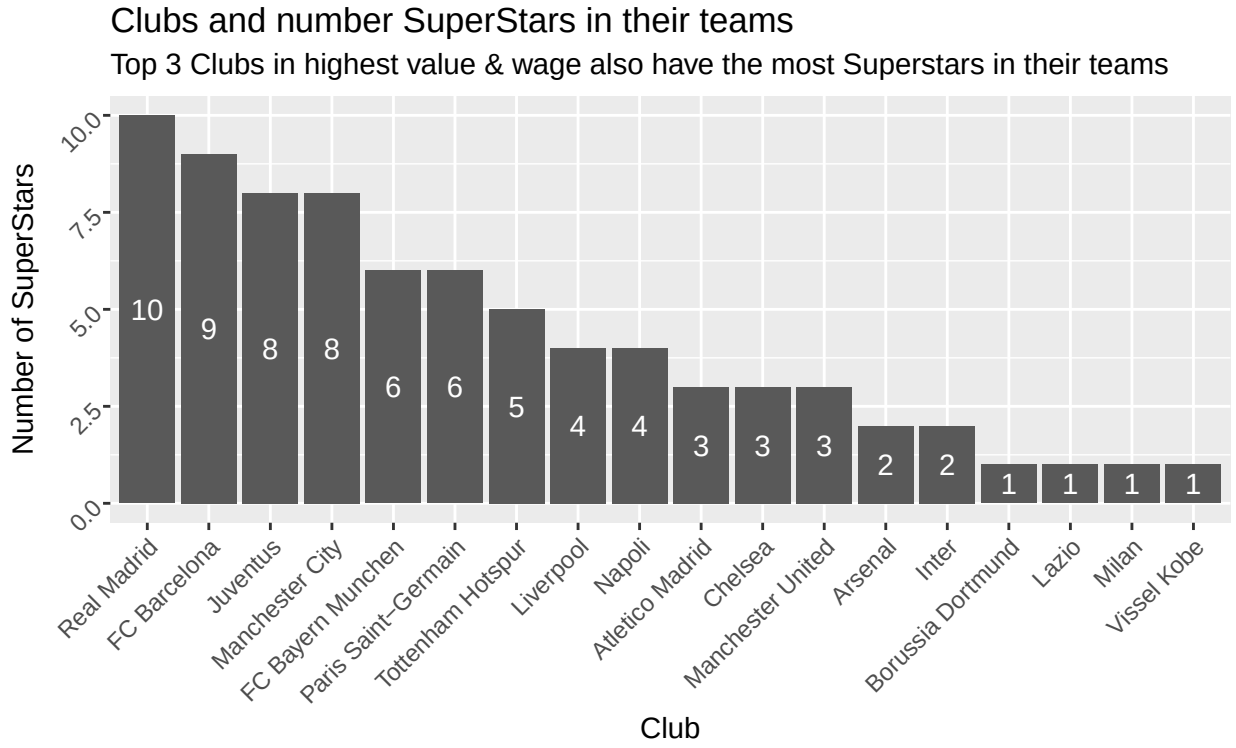
This chart shows that there are players who are paid more even though they are not stated to be the most valuable ones. Wage payment could also be part of the club players play for. Modric and Ronaldo both play for Real Madrid, so let's see if the value and the wage-payments per club give a better insight in why the most valuable players aren't also the best paid.



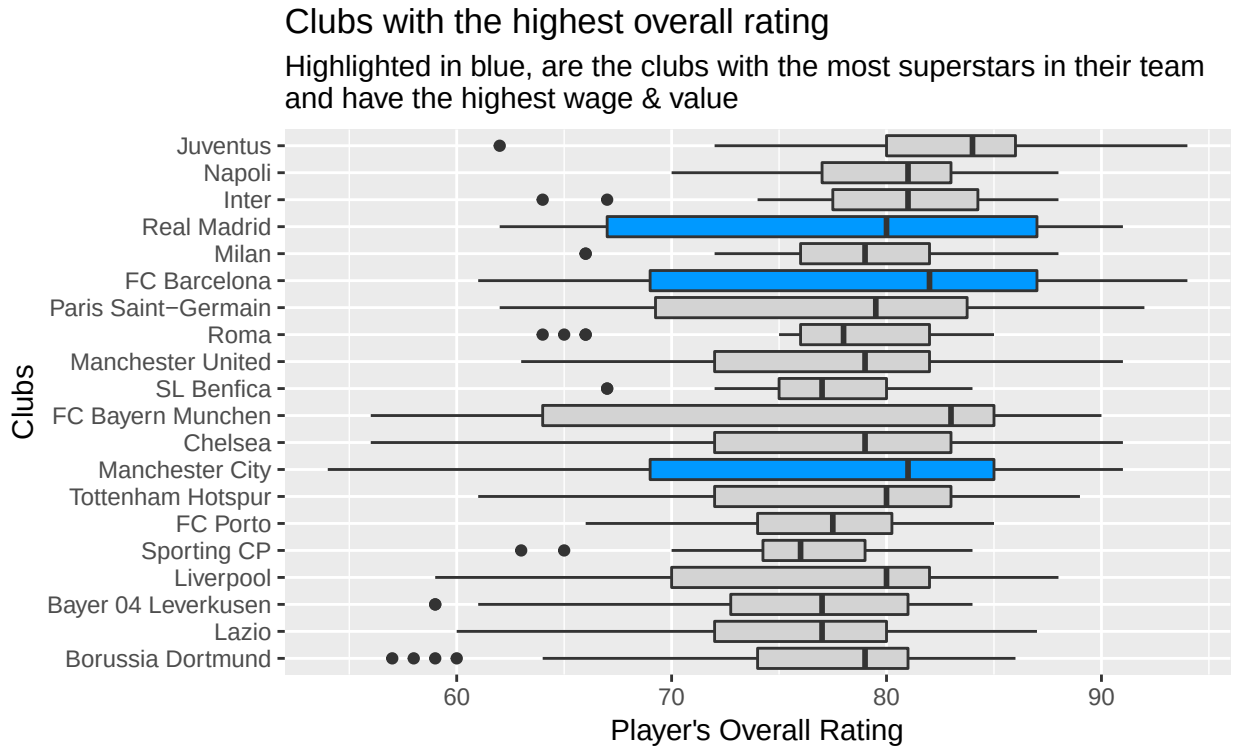


So now we saw that the clubs with the highest wages also have the highest value. Those clubs also paid the players with the highest wages, so would that mean that clubs like that would also have more "very good players"? It would be a logical conclusion, but let's see if the data supports it. From the chart comparing the Overall Rating of Players, with the value a of player the conclusion was drawn that after a rating of 85 the

value of players would "spread" more than with a rating below 85. So let's state that with an Overall Rating over 85 you wouldfall in the category of being a superstar. So let's see if the top 3 Clubs of wage and value, also have the most superstars.



Just to complete the whole analysis, let's include the chart where we see the overall rating of players (wage and value are not taken into account) and see if it is still **Real Madrid**, **FC Barcelona** and **Manchester City** who lead the way.



We can see that it is not the case: the clubs with the most money don't necessarily have the best overall rated teams. You can see that the top 3 Clubs with the best overall rate have a smaller spread in the player's rating. This means that there is less deviation and therefor the overall rating is higher. Manchester City even has players with a rating below 60 as the boxplot shows. This is the reason why it is not even a top 10 Club when looking at the overall rating.

We can analyze even more, but now we have a good base for drawing the correlations and creating the clusters. We know some parts about the players and the clubs which comes in handy during thenext part of the project.

Methods –> Modelling correlations & clusters

Creating the Naive Bayes

Before I will look at the Naive Bayes between Superstar and Value, I will make sure that every player in the dataset is appointed with "Superstar" on "Non Superstar". Just to be sure that all bases are coverted, I include the Unknown for players that don't have an overall rating equal or below 85 or over 85.

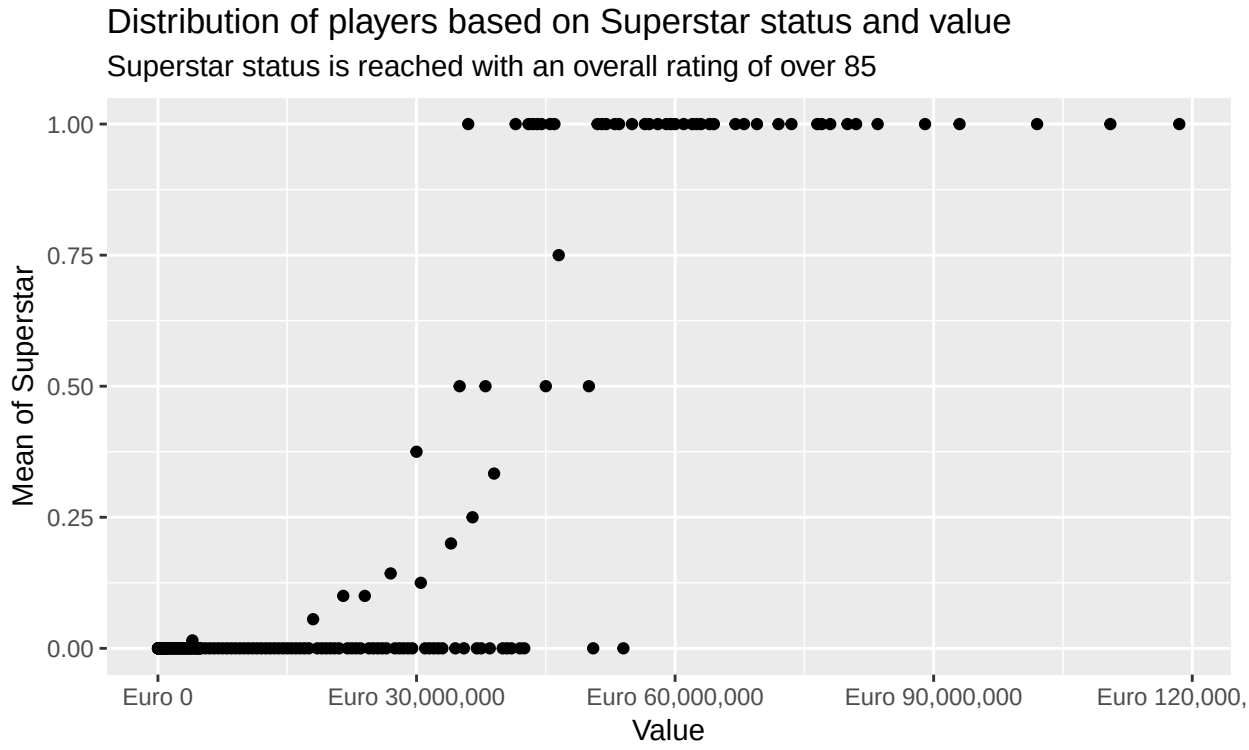
```
for(i in 1:nrow(fifa)){
  super <- fifa[i,]

  if(super$Overall <=85){
      fifa[i, 'Superstar'] <- "No Superstar"
    } else if(super$Overall > 85){
      fifa[i, 'Superstar'] <- "Superstar"
    }else {
      fifa[i, 'Superstar'] <- "Unknown"
    }
}
```

I will now only use the value of "superstar" to hope I can predict the value of a player. This is indeed rather naive, but I small insight will make sure I know how to proceed further on. So let's first create the data I need to be able to predict the value:

```
y <-  fifa$Value
set.seed(2)
test_index <- createDataPartition(y, time = 1, p = 0.5, list = FALSE)
train_set <- fifa %>% slice(-test_index)
test_set <- fifa %>% slice(test_index)

params <- train_set %>%
  group_by(Superstar) %>%
  summarize(avg = mean(Value), sd = sd(Value))

pi <-  train_set %>%
  summarize(pi = mean(Superstar == "Superstar")) %>%
  .$pi

f0 <-  dnorm(test_set$Value, params$avg[2], params$sd[2])
f1 <-  dnorm(test_set$Value, params$avg[1], params$sd[1])

p_hat_bayes <- f1*pi / (f1 *pi + f0*(1 - pi))

head(p_hat_bayes)
```

```
## [1] 1.066039e-146  1.740687e-54  2.531050e-77  4.981532e-35  4.841885e-18
## [6]  1.863617e-09
```

We already saw in the previous chart that there is no normal distribution in the value of the players. So estimating the value of a player only based on the status of being a Superstar or not (or unknown ofcourse) is already stated to be an impossible task. From the code above, we can see that the naive Bayes is realy small, which confirms this. But let's look at the visualization of this and see if there is really a regressionline to be determined.



There is no obvious line to see, though it can be seen that the more you are a superstar, the higher you value is. Though even for superstars it can be seen that the value varies a lot (between 40 million and a 120 million). So I can say that only using the superstar status as a way to determine the value of a players has proven not successfull.

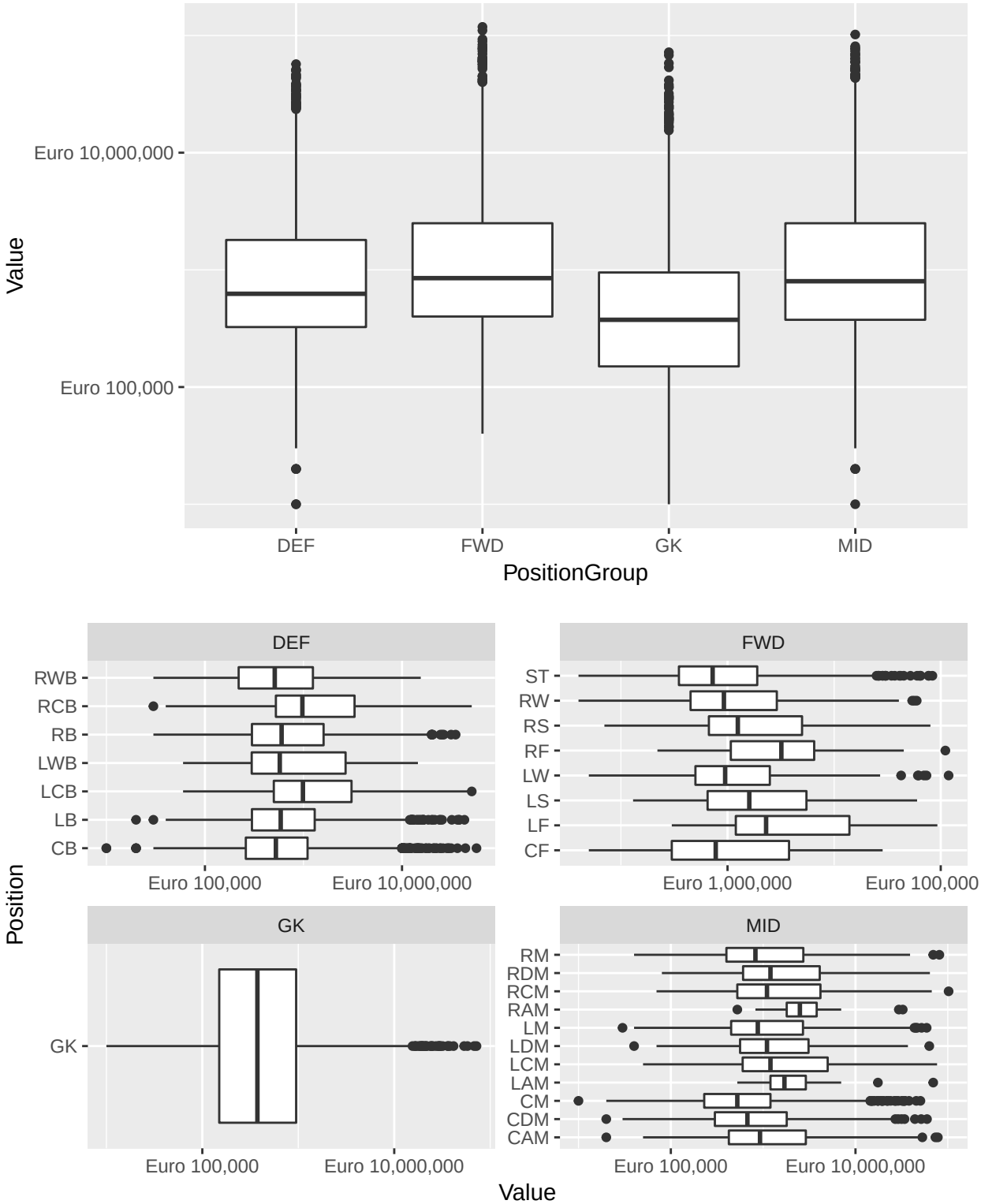
Influence of the player's position

So let's look at another aspect which could influence the value of a player: his position. I will create 4 categories of positions: forwards (FWD), midfielders (MID), defenders (DEF) and goalkeepers (GK). I will assign those groups as they are also used in the award ceremonies of FIFA when awarding "best player" like when *Navas, Ramos and Modric voted best in their positions*.

```
positions <-  unique(fifa$Position)

gk <- "GK"
def <- positions[str_detect(positions, "B$")]
mid <- positions[str_detect(positions, "M$")]
fw1 <- positions[str_detect(positions, "F$")]
fw2 <- positions[str_detect(positions, "S$")]
fw3 <- positions[str_detect(positions, "T$")]
fw4 <- positions[str_detect(positions, "W$")]
fwd <- c(fw1, fw2, fw3, fw4)

fifa <- fifa %>%
  mutate(PositionGroup = ifelse(Position %in% gk, "GK",
                                ifelse(Position %in% def, "DEF",
                                       ifelse(Position %in% mid, "MID",
                                              ifelse(Position %in% fwd, "FWD", "Unknown")))))
```

So now let's have a look at the distribution of value regarding the positions. First I will look at the distribution

within the positiongroups I created earlier, and than I will look at the distribution of the position within these positiongroups. LEt's see what we can conclude from that.



From the first chart we can see that forwards and midfielders have the highest value. When looking at the most valuable position within those groups we can see it is **right forward** (RF) and **left forward**

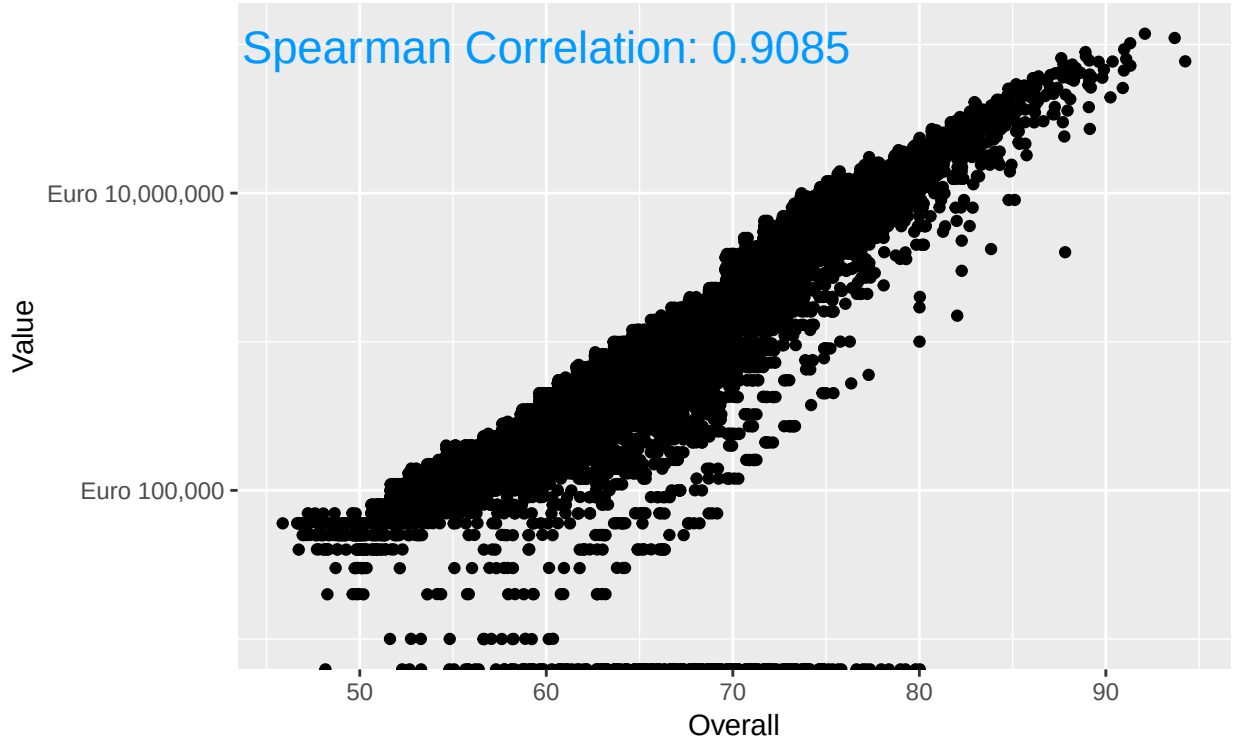
(LF) in the Forwards group who are most valuable. In the group of midfielders it is the **Right Attacking Midfielder** (RAM) or the **Left Attacking Midfielder** (LAM) who are the most valuable. This could be explained by the fact that the game of football has the attacking strategy of "going on the flanks and finish in the middle". So players who are strong in attacks and play on the flanks would actually earn more according to the above charts. Let's check this with looking at the most valuable players and their positions.

```
##           Name Position PositionGroup     Value
## 1    Neymar Jr       LW           FWD 118500000
## 2     L. Messi       RF           FWD 110500000
## 3 K. De Bruyne      RCM           MID 102000000
## 4    E. Hazard       LF           FWD  93000000
```

The players indeed belong to the positiongroups of Fowards (3 out of 4) and midfielders (1 out of 4). Though when looking at their exact position, only Neymar and Hazard belong to the most valuable positions. So apparently the position of a player does not have as high of an impact as assumed.

Correlations between difference aspects

To know what actually does influence the value of a player, it is time to look at the correlations between all different aspects which can be taken into account. But before we dive into it, it is important to first determine **IF** there actually is a correlation. The overall rating is based on all different "smaller" ratings like Accuracy and Speed. So let's take the Overall Rating and see if there is a correlation with the Value of a player.



I used a Spearman correlation in this case, as there are many large outliers who influence the outcome. As is stated in the chart, the correlation is very high with a score of 0.9085. BUt this is only based on the Overall rating, which as we already stated consists of multiple smaller ratings. So lets check with the Spearman as well as with the Pearson method where the correlations exsist. In this we will not take the goalkeepers into account, as they have a different set of qualities needed.

```
##         Feature  Spearman   Pearson
## 1         Value 0.9163082 0.6353774
```

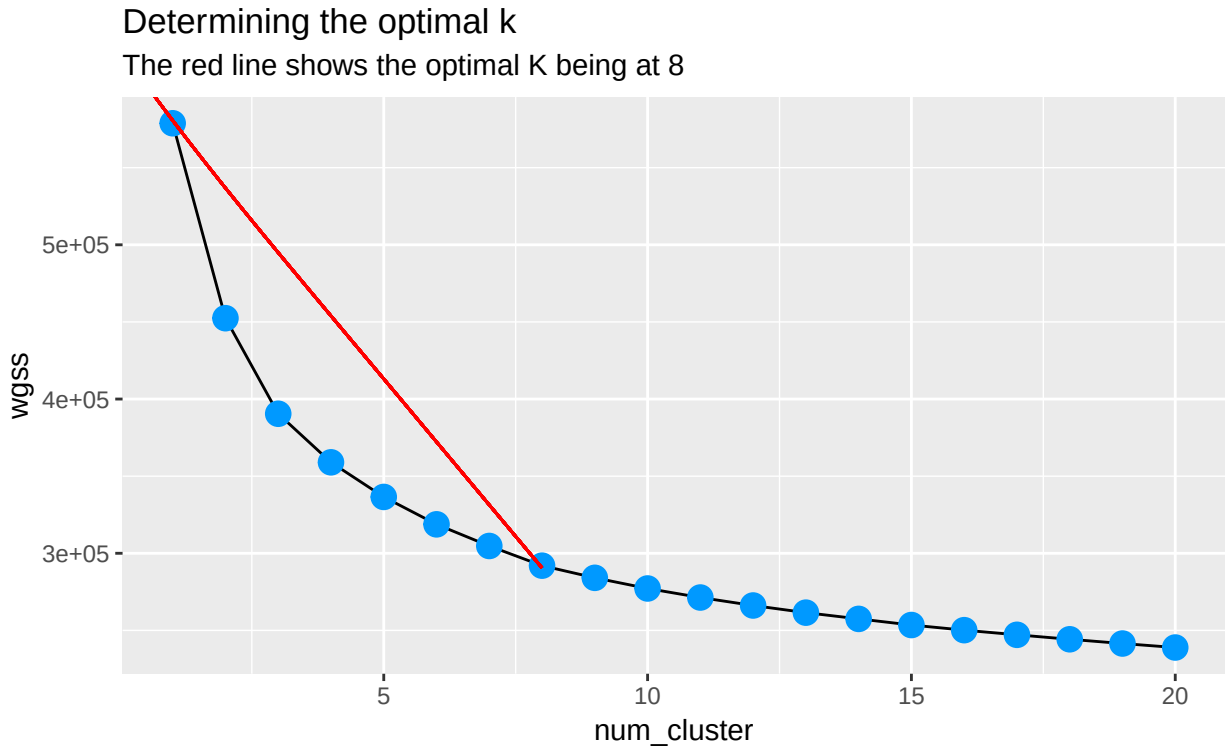
```
## 2     Reactions 0.8429790 0.8477147
## 3     Composure 0.7924707 0.8013021
## 4          Wage 0.7790184 0.5762863
## 5   BallControl 0.7323579 0.7174420
## 6  ShortPassing 0.7203556 0.7224494
## 7     Potential 0.6137577 0.6495115
## 8     ShotPower 0.5922090 0.5623670
## 9   LongPassing 0.5870439 0.5844998
## 10    Dribbling 0.5703460 0.5159061
```

We see that the highest correlation with the Overall rating is the value, followed by reaction and Composure. The difference between the outcome of the Spearman and Pearson method is also shown. So it is shown that there are multiple strong correlations between the overall rating and several other aspects of the players skillset. Before we use this correlations to predict" the value of the players, we first will cluster them.

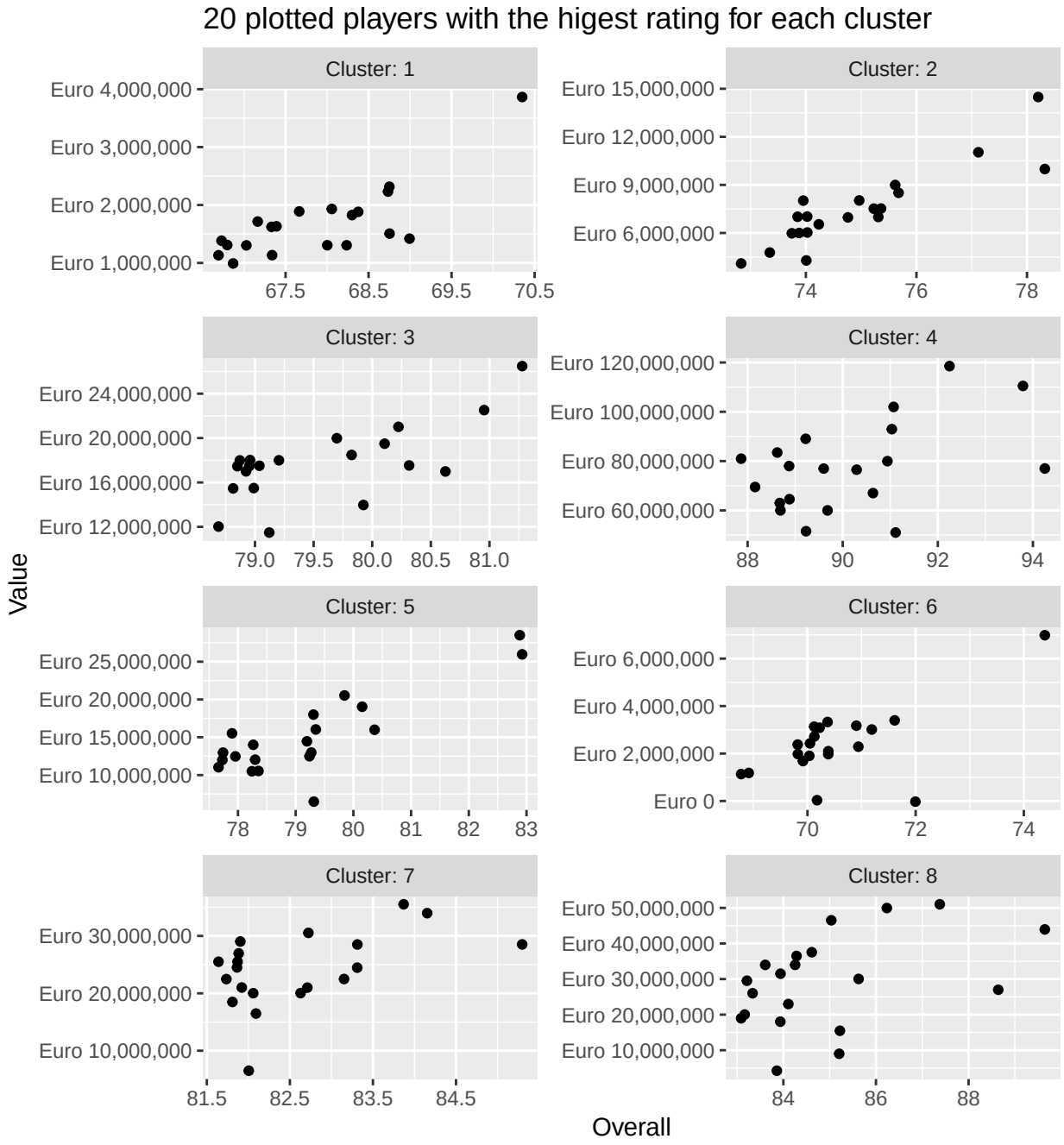
Clustering the players with K-Means

In order to find similar players, I have to make sure that there is a way I can compare them to one another. To do so, I will cluster them using K-Means. First I will filter out the numeric columns that I don't want to use in my comparison. This will be columns like player's ID and Jersey Number, but also the Wage and Value I will not take into account. This is because I want to compare players who have the same skills, not the same value. Also in this case I will filter out all skills that apply to goalkeepers as I will not be comparing goalkeepers with one another.

In this first chart I will determine what the optimal K is. After that I will use this number to actually define the clusters.



The optimal K is 8 as the red line points out. So I will define 8 clusters where the players will be assigned to. First I will create a cluster without the goalkeepers and the players without a position. Than I will cand their overall rating. All this will result in the clustering below.



The fourth cluster has the highest rating and the highest value, let's just quickly check who is in it

```
##                Name                Club Cluster PositionGroup Overall
## 1           L. Messi        FC Barcelona       4           FWD      94
## 2  Cristiano Ronaldo            Juventus       4           FWD      94
## 3          Neymar Jr Paris Saint-Germain       4           FWD      92
## 4       K. De Bruyne     Manchester City       4           MID      91
## 5          E. Hazard             Chelsea       4           FWD      91
## 6          L. Modric         Real Madrid       4           MID      91
## 7          L. Suarez        FC Barcelona       4           FWD      91
## 8       Sergio Ramos         Real Madrid       4           DEF      91
## 9     R. Lewandowski   FC Bayern Munchen       4           FWD      90
```

```
## 10          T. Kroos         Real Madrid       4           MID      90
## 11       David Silva     Manchester City       4           MID      90
## 12          N. Kante            Chelsea       4           MID      89
## 13         P. Dybala           Juventus       4           FWD      89
## 14           H. Kane   Tottenham Hotspur       4           FWD      89
## 15      A. Griezmann     Atletico Madrid       4           MID      89
## 16   Sergio Busquets        FC Barcelona       4           MID      89
## 17         E. Cavani Paris Saint-Germain       4           FWD      89
## 18         S. Aguero     Manchester City       4           FWD      89
## 19         K. Mbappe Paris Saint-Germain       4           MID      88
## 20          M. Salah           Liverpool       4           MID      88
```

And again in this case, we also see that only 1 out of the 20 selected is not a forward or a midfielder. But also the defender has a overall rating over 90. So now it is time to look if we can find similar players, predict the value and look if we can find "bargains".

Result-set

Finding similar players

First part we set as a goal is finding players that are fairly similar to one another. In order to find this out, I need to write a function doing so. This function will look at the value and the cluster and the value of a certain player. I will be able to set the number of results coming out of it (when there are that many, just take that into account) and what the fraction of a difference might be.

```
similar_players <- function(player, num_results, return_fraction){

  cluster_filter <- cluster_analysis$Cluster[cluster_analysis$Name == player]
  player_value <- cluster_analysis$Value[cluster_analysis$Name == player]

  cluster_analysis %>%
    filter(Cluster == cluster_filter,
           Value >= (player_value * (1 - return_fraction)) & Value <= player_value * (1 + return_fracti
    head(num_results)
}
```

Now let's use this funcion to see if there are any players comparable with the Top4 most valuable players: L. Messi, Neymar Jr, K. De Bruyne, E. Hazard and Cristiano Ronaldo as het had the second best overall rating in the 4th Cluster we saw before. So actually we would look at the top 5 players from the fourth cluster: the highest raking, highest value cluster.

```
similar_players("L. Messi", 10, .05)
```

```
##       Name         Club Age PositionGroup Overall Potential Cluster
## 1 L. Messi FC Barcelona  31           FWD      94        94       4
##       Value   Wage Superstar
## 1 110500000 565000 Superstar
```

```
similar_players("Cristiano Ronaldo", 10, .05)
```

```
##                Name             Club Age PositionGroup Overall Potential
## 1 Cristiano Ronaldo         Juventus  33           FWD      94        94
## 2         L. Suarez      FC Barcelona  31           FWD      91        91
## 3    R. Lewandowski FC Bayern Munchen  29           FWD      90        90
## 4          T. Kroos       Real Madrid  28           MID      90        90
## 5      A. Griezmann   Atletico Madrid  27           MID      89        90
```

```
## 6              Isco       Real Madrid  26           FWD      88        91
## 7        C. Eriksen Tottenham Hotspur  26           MID      88        91
##   Cluster    Value   Wage Superstar
## 1       4 77000000 405000 Superstar
## 2       4 80000000 455000 Superstar
## 3       4 77000000 205000 Superstar
## 4       4 76500000 355000 Superstar
## 5       4 78000000 145000 Superstar
## 6       4 73500000 315000 Superstar
## 7       4 73500000 205000 Superstar
```

```
similar_players("Neymar Jr", 10, .05)
```

```
##        Name               Club Age PositionGroup Overall Potential
## 1 Neymar Jr Paris Saint-Germain  26           FWD      92        93
##   Cluster     Value   Wage Superstar
## 1       4 118500000 290000 Superstar
```

```
similar_players("K. De Bruyne", 10, .05)
```

```
##           Name            Club Age PositionGroup Overall Potential Cluster
## 1 K. De Bruyne Manchester City  27           MID      91        92       4
##      Value   Wage Superstar
## 1 1.02e+08 355000 Superstar
```

```
similar_players("E. Hazard", 10, .05)
```

```
##        Name     Club Age PositionGroup Overall Potential Cluster   Value
## 1 E. Hazard  Chelsea  27           FWD      91        91       4 9.3e+07
## 2 P. Dybala Juventus  24           FWD      89        94       4 8.9e+07
##     Wage Superstar
## 1 340000 Superstar
## 2 205000 Superstar
```

You can see that as we take the value of a player into account, the top most valuable players (Messi, Neymar and de Bruyne) don't have any similar players. They are so different from one another already that they aren't even similar to each other. But Ronaldo is less valuable, so you can see that he does have some similar players.

So we can determine similar players depending on their cluster and value. The Cluster is used as we made sure the comparable players are already combined into this one cluster. We will see more of that when we start looking for the bargains, but first: let's get to the prediction of a player's value as it seems to be a rather important point and a red line thoughout this whole analysis.

Predicting a player's value

Wouldn't it be the best thing: buying something and knowing the value of it would increase drastically? Yeah it would and even with football players it is the case. They are basically very valuable assets of a Club. So for a club to gain any profit, they would like to be able an predict what their players will be worth one day. But is it really that simple? Can we just find a linear curve and assign a value to it? Or is there more to it?

First I looked for the skills with a correlation with **Value** and from that I only took the ones with a value higher than 0.30 (columnnames are shown below). Those columns I applied to the linear model with the summary shown below.

```
##  [1] "Overall"                 "Potential"
##  [3] "Value"                   "Wage"
##  [5] "Skill Moves"             "International Reputation"
```

```
##  [7] "ShortPassing"             "LongPassing"
##  [9] "BallControl"              "Reactions"
## [11] "Vision"                   "Composure"
```

```
##
## Call:
## lm(formula = Value ~ Overall + Potential + Wage + `Skill Moves` +
##     `International Reputation` + ShortPassing + LongPassing +
##     BallControl + Reactions + Vision + Composure, data = train,
##     na.action = na.omit)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -19989694   -756760    -21742    684350  46027237
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.589e+07  3.546e+05 -44.816  < 2e-16 ***
## Overall                     6.213e+04  8.728e+03   7.118 1.16e-12 ***
## Potential                   1.364e+05  5.572e+03  24.482  < 2e-16 ***
## Wage                        1.659e+02  1.602e+00 103.552  < 2e-16 ***
## `Skill Moves`               2.316e+05  3.735e+04   6.200 5.85e-10 ***
## `International Reputation`  1.537e+06  8.694e+04  17.673  < 2e-16 ***
## ShortPassing                1.685e+03  5.833e+03   0.289 0.772728
## LongPassing                -6.839e+03  3.827e+03  -1.787 0.073950 .
## BallControl                -5.435e+02  3.888e+03  -0.140 0.888818
## Reactions                   1.361e+04  5.421e+03   2.511 0.012062 *
## Vision                      1.525e+04  2.770e+03   5.504 3.79e-08 ***
## Composure                  -1.331e+04  3.997e+03  -3.330 0.000871 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2569000 on 10652 degrees of freedom
## Multiple R-squared:  0.794,  Adjusted R-squared:  0.7938
## F-statistic:  3732 on 11 and 10652 DF,  p-value: < 2.2e-16
```

Based on this information, I will create the information I need to be able to predict the value of a player. I will create a data frame with a new value: predicted value. THis value will be based on a variance of 20%. When my predicted value will fall into this range, a new column (Accurate) will show a *Yes*, if not this column will show a *No*. Let's see if the top 10 players will have a correct predicted value.

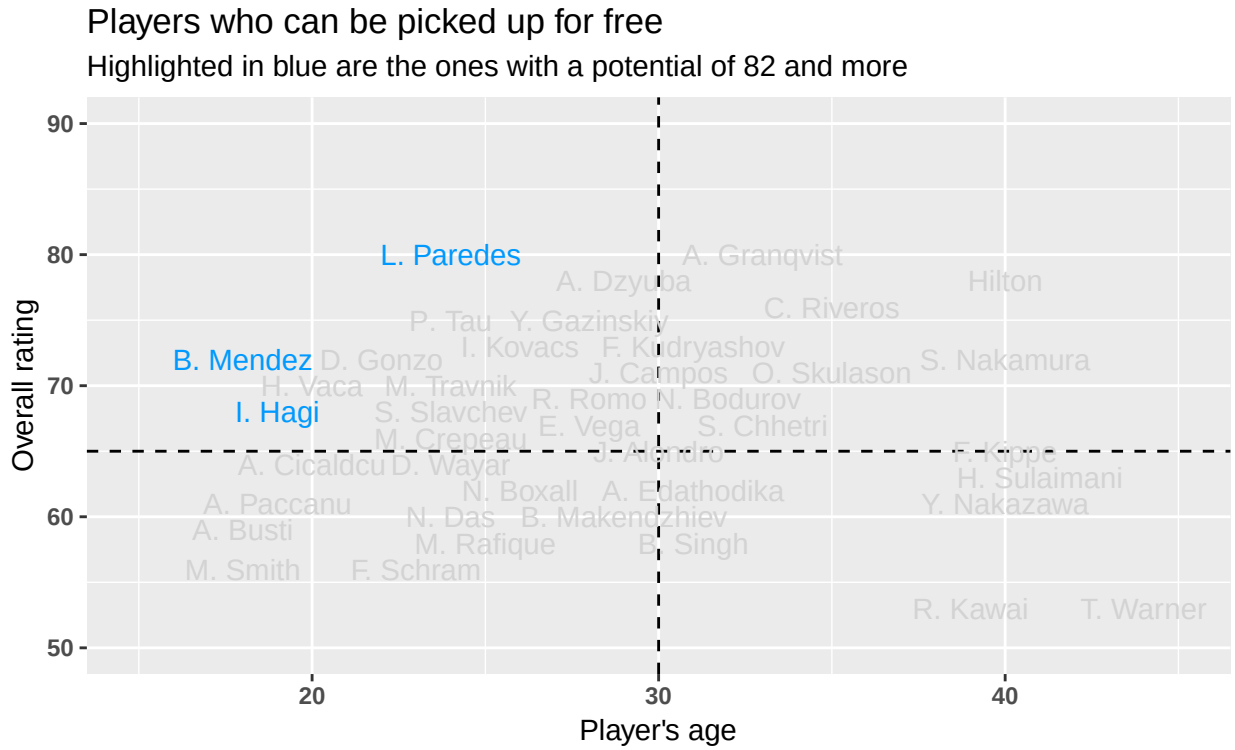
```
##
##   No  Yes
## 6204 1227
```

```
##               Name     Value Predicted.Value Difference Accurate
## 1        Neymar Jr 118500000        60379553   58120447       No
## 2           De Gea  72000000        53034990   18965010       No
## 3     Sergio Ramos  51000000        72652082  -21652082       No
## 4   R. Lewandowski  77000000        43964471   33035529       No
## 5         T. Kroos  76500000        68581511    7918489      Yes
## 6         D. Godin  44000000        28261023   15738977       No
## 7      David Silva  60000000        57237564    2762436      Yes
## 8          H. Kane  83500000        42171428   41328572       No
## 9  Sergio Busquets  51500000        61738760  -10238760      Yes
## 10       E. Cavani  60000000        42855885   17144115       No
```

The table shows that with this formula, I was able to predict 1227 values within the range of the previously set variance. This means that we only predicted about 19.7% of the players correctly. The higher I will set the variance/range, the higher the amount of accurate predictions will be. Though this is not the way a value should be predicted. So we will stick to the 20% and in the conclusion I will elaborate more on why it is so difficult in predicting a players value.

Players available for a bargain

Let's first define what I see as a bargain: A bargain is something that can be picked up for free. Ofcourse not everything that is for free, is something you actually want. I am only interested in players with a potential rating over 82. This means they are not classified as a superstar(86 and over), but they are likely to be players that can support the superstars. In addition: a team with players overall rating of 82 could be way more successful than a team with a very wide range in their overall rating like we saw with Manchester City earlier.



There are 3 players highlighted, who all have the age under 25. This is something which is rather obvious, as we previously saw that young players have a higher average potential than the older squad. This is because they are still developing their skills and ones they are over 30, they probably already reached their potential.

I would want to point out the issues we could face when looking for the bargains in the conclusionsection, as there are of course some aspects we need to take into acocunt when just looking at 1 dataset.

Conclusion

The conclusion is the last and one of the most important parts of this project. As a start I would like to draw conclusions on the method and analysis as those have outcomes as well. Followed by that I want to reflect on the previously set goal and the conclusions on that.

Conclusion on methods

As a first method we had the **Naive Bayes**. On this we can conclude the following:

- Only using a Naive Bayes appraoch on Superstar vs Value will not give you the opportunity to predict the value of the players. There is more information needed.

- A correlation can be seen between in the chart, as the higher the value, the closer a player is to being a superstar. Though only using "superstar status" as point to define a correlation and with that having the ability to predict the value of a player will not be sufficient.

A second method was using **the position** of the player to determine the players value. From this we can conclude the following:

- Players who play the position of Forward or Midfielders have the highest value
- Players who play on the flanks tend to be more valuable than centrum players. Especially Right Forwards, Left Forwards, Right Attacking Midfielders and Left Attacking Midfielders.
- Top 4 players all 4 play as Forwards or Midfielders, which confirms the fact that players on this position are the most valuable. So the position of a player DOES have a large influence on the players value, and should therefore be taken into account when obtaining the goal set.

The next aspect is looking at the **correlations**. For this, I first looked at the most obvious correlation we also already visually saw in the *Naive Bayes*: Value vs Overall Rating

- There is a very strong correlation between a player's Value and his overall rating: 0.9085
- Besides a high correlation with the value, there is also a correlation between the overall rating and Reactions, Composure and Wage. When determining the "similar players" this is an important aspect to also take into account. And besides that, those aspects are also likely to influence the players value. Therefor they will be used in the linear model in the result part.

The last part of this concluding section is the conclusion on the **clustering**. For this I used K-Means and I determined the K with a model.

- The optimal K is set at 8, so there will be 8 clusters created to cluster all players.
- To keep the clusters clear, only the first 20 players are plotted in the clusters. This shows that the fourth cluster has the highest value and the highest rating as well. All 4 Top players are plotted in this cluster and there we can conclude that they can be compared to one another.
- The clustering also shows that in fact only 1 in 20 top players of the fourth cluster is a defender so we can concluded that the position of a player does influence the cluster and with that the value it has.

Based on these conclusions, a conclusion on the resultset can be drawn and with that a conclusion on the goal.

Goal reflection & Conclusions

As a next part I would like to look at the goal we set at the beginning of this report: being able to predict player's value, find similar players and look for bargains. When we take a look at the result-section we already see that we managed to do all three, though there are some conclusions which can be drawn from the outcome:

- Not all players have players that are similar to them as the **similarity** is based on the value and positiongroup. This means that the top 4 players have barely any similar players as their value is extra-ordinarily high. When the value drops, more similarities appear which is shown with the similarities to Cristiano Ronaldo.

- A player's **value** can be **predicted** up till a certain level. Only 19.7% of the predictions were accurate. Value of a player is not only influenced by the skills of the player which were available in this dataset but there are way more aspects that can influence the value of a player:

1. The budget of the Club willing to buy the player,
2. The willingness of the player to move to a certain Club,
3. The willingness of the old Club to let the player leave to another Club,
4. The competition of other players the Club wants to buy,
5. etc.

This information is not available for now, but I would advice to include such information when aiming for a better and more detailed prediction. Though we always have to take into account that it is business and prices rise when options are limited and vice versa.

- There is something like "first row for nothing" and there are players who can be picked up as a **bargain**. But when looking at the players who have an acceptable level, they are all without a Club. This means that they don't develop their skills. With that they might be injured or banned from playing (let's assume this is not the case though). So we can conclude that there is something like a "bargain", though there are more things to consider when taking players like that into account. Also extra knowledge would come in handy in this case.

Though in all 3 cases we managed to come to an outcome and with that we can conclude that the project as a whole was a success.

Overall conclusion & advice

Overall we can conclude that based on just 1 dataset it is difficult to predict values that are not only influenced by the aspects within the dataset. Though as FIFA19 is a game, and outside influences will not be part of the game and the structure it is created on, the conclusions can be stated to be correct.

For "real life" examples I would advice the parties to have multiple sources available with different information and a longer trackrecord of the players itself. In this case it could have been done by combining datasets of FIFA18, FIFA17 etc. Though as the game will not be influenced by the statistics from the passed, they have not been taken into account in this projec.t

So the conclusion and advice I would like to give myself and everybody else in this field is: *Don't rely on just any data, but use your mind to find out if you need more... Numbers are a lot and very helpful, but clear insight is everything.*