52414: Lab 2

Gil Shiloh and Elky Sandor

June 1, 2021

Lab 2: Visualization Through `ggplot`

Contents:

- Q0. Submission Instructions
- Q1. Basic Statistics (30 pt)
- Q2. Scouting Report (30 pt)
- Q3. Model Building (30 pt)
- Q4. Fix Problematic Plots (10 pt)

Q0.Submission Instructions

This lab will be submitted in pairs using GitHub (if you don't have a pair, please contact us).
Please follow the steps in the GitHub-Classroom Lab 2 (https://classroom.github.com/g/6_Wy5z44) to create your group's Lab 2 repository.
Important: your team's name must be `FamilyName1_Name1_and_FamilyName2_Name2`.
You can collaborate with your partner using the git environment; You can either make commits straight to master, or create individual branches (recommended). However, once done, be sure to merge your branches to master - you will be graded using the most recent *master* version - your last push and merge before the deadline.
Please do not open/review other peoples' repositories - we will be notified by GitHub if you do.

Your final push should include this Rmd file (with your answers) together with the html file that is outputted automatically by knitr when you knit the Rmd. Anything else will be disregarded. In addition, please adhere to the following file format:
`Lab_2_FamilyName1_Name1_and_FamilyName2_Name2.Rmd/html`

Some questions may require data wrangling and manipulation which you need to decide on.
In some graphs you may need to change the graph limits. If you do so, please include the outlier points you have removed in a separate table.

Show numbers in plots/tables using standard digits and not scientific display. That is: 90000000 and not 9e+06.
Round numbers to at most 3 digits after the dot - that is, 9.456 and not 9.45581451044

The required libraries are listed below the instructions. You are allowed to add additional libraries if you want. If you do so, *please explain what libraries you've added, and what is each new library used for*.

Background:

You've been hired as a data analyst at at football (soccer) club. Since this is a small and under-funded club, you will not have access to real-football data, but to data from the football computer game fifa18. Your job is to analyze this dataset and extract meaningful insights from the data in order to help your club make better decisions.

Data File:

You will load and analyze the fifa18 football dataset file called “fifa_data.csv”.
The dataset contains detailed information about each player in the game, including: names, age, nationality, overall ability, estimated potential ability, current club and league, market value, salary (wage), ability at different football skills (also called ‘attributes’, e.g. Ball.control, Sprint.speed …), ability to play at different position in the game (CF, CM, …) and the preferred positions of the player.

Required Libraries:

```
library(ggplot2)
library(dplyr)
library(corrplot)
library(scales)   # needed for formatting y-axis labels to non-scientific type
library(radarchart)
library(tidyr)
library(tidyverse)
library(reshape2) # melt
library(ggthemes)
library(rworldmap) # world map
library(modelr)
library(radarchart) #Spider chart
############################################
library(e1071) #Q1.c -  skewness() and kurtosis()
library(grid) # geom_segment
library(ggrepel)# Use ggrepel::geom_label_repel
library(fmsb)  #Spider chart

options("scipen"=100, "digits"=4)  # avoid scientific display of digits. Take 4 digits.
```

Q1. Basic Univariate Statistics (30 pt)

First, you are requested to load the fifa18 dataset and find and display general information about the players.

a. Make a plot showing the `overall` ability distribution of all players. How skewed is the distributions? does it have fat tails?
   Plot on top of the `overall` distribution a Normal distribution matching its first two moments. Is the distribution described well by a Normal distribution? explain.

b. Make a plot comparing the multiple `overall` ability *distributions* of players according to the `continent` of the players. Describe which continents have especially good/bad players.

c. Make a plot showing the density of players’ `value` distribution.
   Next, make a separate plot showing the density distribution of the *log* of players’ `value` .
   Which of the two visualizations is better? explain.

d. Are the top-10 players with the highest `value` also the top-10 best players in terms of `overall` ability? Show tables for both and compare.
   Who is the best player not in the top-10 valued players?

e. Show a table of the *10 youngest* and *10 oldest* teams in terms of *average* players `age` .

Loading the data:

```
fifa_players <- data.frame(read.csv(url("https://raw.githubusercontent.com/DataScienceHU/Data
AnalysisR_2020/master/data/fifa_data.csv")))
#fifa_players <- data.frame(read.csv("../../../../Datasets/fifa_data.csv"))
# Pre-processing:
for (i in c(3,6,7,10:71)) {
  fifa_players[,i]<-as.numeric((fifa_players[,i]))
}
fifa<-na.omit(fifa_players)
fifa_players <- fifa
fifa_players_info <- fifa[,c(1:11)] # players general info
fifa_players_attribures <- fifa[,c(1,12:45, 6)] # players different skills. Add overall
fifa_players_positions <- fifa[,c(1,46:72,6,7)] # players ability at different positions . Ad
d overall
fifa_players_indicators <- fifa[,c(1,6,7,10,11)] # players general ability
```

PLEASE ADD YOUR SOLUTION BELOW, WITH A CLEAR SEPARATION BETWEEN THE PARTS!

Q2. Scouting Report (30 pt)

You are in charge of the scouting division. The goal of this division is to follow players' `potential` and `overall` ability, and identify undervalued players - that is, players whose current value is lower compared to what would be expected based on their predicted future ability.

a. Plot the *average* `potential` ability by `age` of all players, for players 35 years old or younger

b. Plot the *average difference* between a player's `overall` ability to `potential` ability as a function of `age`, up to age 35. At what ages should we expect to find players for future development based on this graph?

c. We are seeking young ($age \leq 21$) players with high `potential` (> 70). Show a scatter plot of these players comparing their `potential` ability (x-axis) and current `value` (y-axis).
Find the 10 most-undervalued players, i.e. having the lowest `value` compared to their predicted value by `potential` using a simple linear regression model.
Calculate for each of them what is a fair `value` matching their `potential` that you be willing to pay in order to by them to your club and show these 10 players with their name, `age`, `overall` ability, `potential`, actual `value` and fair `value` it a table.

d. Your boss wants to fly abroad to recruit promising players. Use the `rworldmap` package to display the world map and color each country based on the *median* `potential` of players from this nationality.

e. Repeat the above analysis but this time display a world map where each country is colored by the *median ratio* of `potential` to `value` of players. Find an under-valued country you'd recommend to travel to (i.e. a country with cheap players compared to their `potential` average quality).

PLEASE ADD YOUR SOLUTION BELOW, WITH A CLEAR SEPARATION BETWEEN THE PARTS!

Q3. Correlations Analysis (30 pt)

In this question we find and display different skills and their correlations

a. We are interested in finding out which positions are similar in terms of players' performance.
Extract the 26 non-goalkeeper positions (`CAM, CB, ..., ST`). Calculate the correlation between players' ability in each pair of positions and show a heatmap correlation-plot of the correlations' matrix. What three positions have the *least* average correlations with other skills?

We are interested in finding out which skills are similar in terms of players' performance at the position. Extract the 29 skills for non-goalkeeper players (Acceleration, …, Volleys, except 'GK.*' skills). Calculate the correlation between players' ability in each pair of skills and show a heatmap correlation-plot of the correlations' matrix. What two skills seem least correlated with other skills?

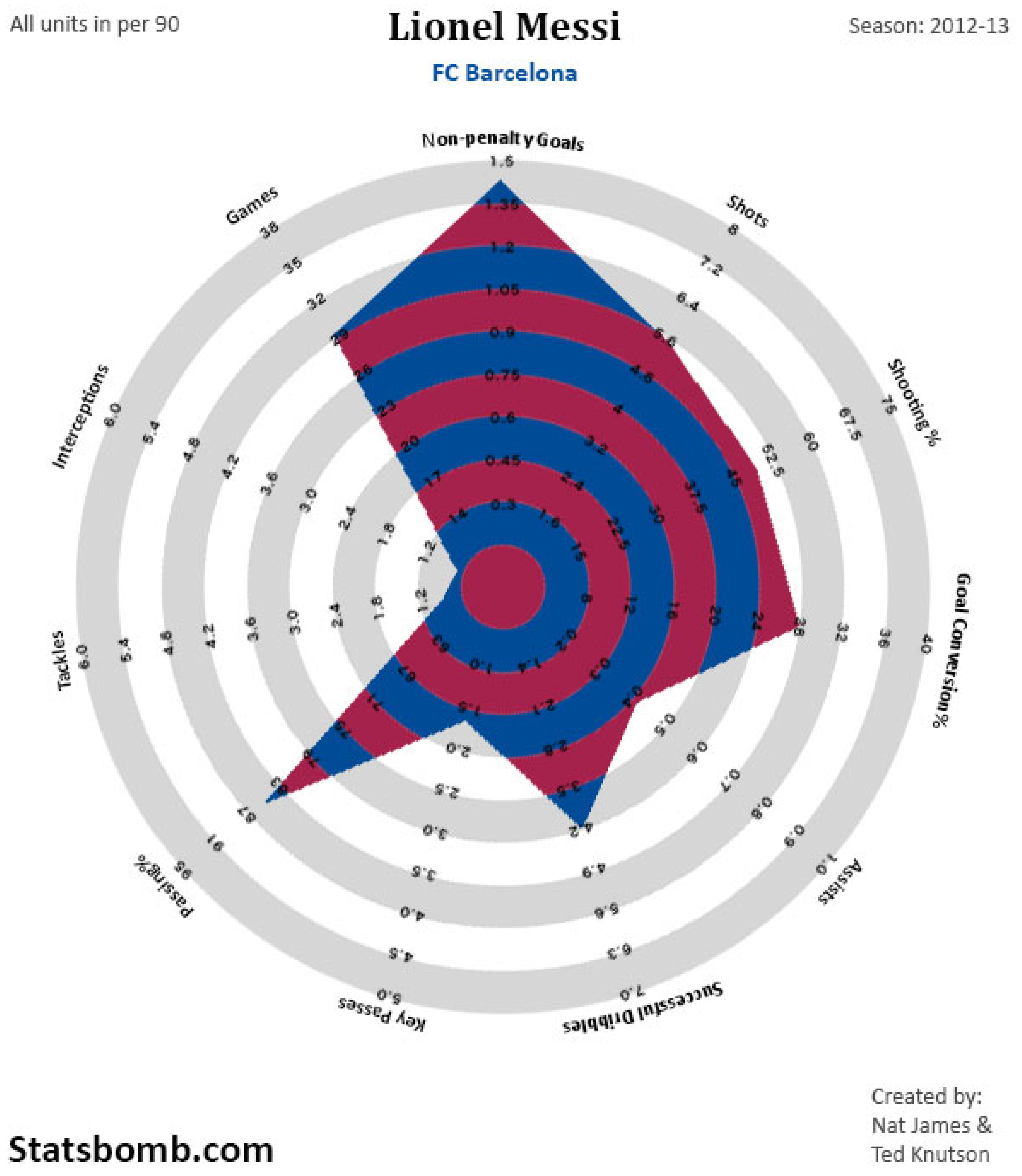
b. Consider the following indicators of players performance: `overall` players' performance, their `potential` , their salary (`wage`) and their market `value` . Show a correlation-plot of players' *34* skill levels (`Acceleration` , …, `Volleys`) vs. these four indicators. Find the *10* skills with the highest *average* correlation with the four inidcators and list them in a table.

c. Build a team of *11 different* players with the following rules:

- For each of the *26* non-goalkeaper positions (*26* from above plus goalkeaper, `GK`), find the player with the best performance at this position.
- Find the goal keaper (`Preffered.Positions` is `GK`) with the best `overall` performance.
- From the players obtained above, find *11 distinct* players maximizing the average `overall` performance of the team, with the constraint that there must be a goalkeaper (preferred position `GK`).
- List the players in a table including their `overall` performance and the team average `overall` score. Next, peak six *different* players of your choice from your team, one of which is the goalkeaper. Using the function `radarchart::chartJSRadar` , graph their abilities (individually for all 6 players) in the top *10* skills according to 3.b in a radar chart (https://en.wikipedia.org/wiki/Radar_chart) (also called 'spider chart') graph. See below an example for such a chart.

d. We are interested in determining how the player's abilities in different positions changes with age. Repeat the analysis of question 2.a., but this time show the *34* different skills
Which skills peak at youngest/oldest ages?

e. Your boss suggests that some players may be currently under-payed compared to their performance, and that we can acquire them by offering them a higher salary (`wage`).
Fit a multiple regression model predicting player's `overall` performance based on their `wage` and `age` . Find the 10 players with the highest difference between their `overall` performance level and the regression model prediction, and list them in a table.



Example of a Spider chart

PLEASE ADD YOUR SOLUTION BELOW, WITH A CLEAR SEPARATION BETWEEN THE PARTS!

Q4. Fix Problematic Plots (10 pt)

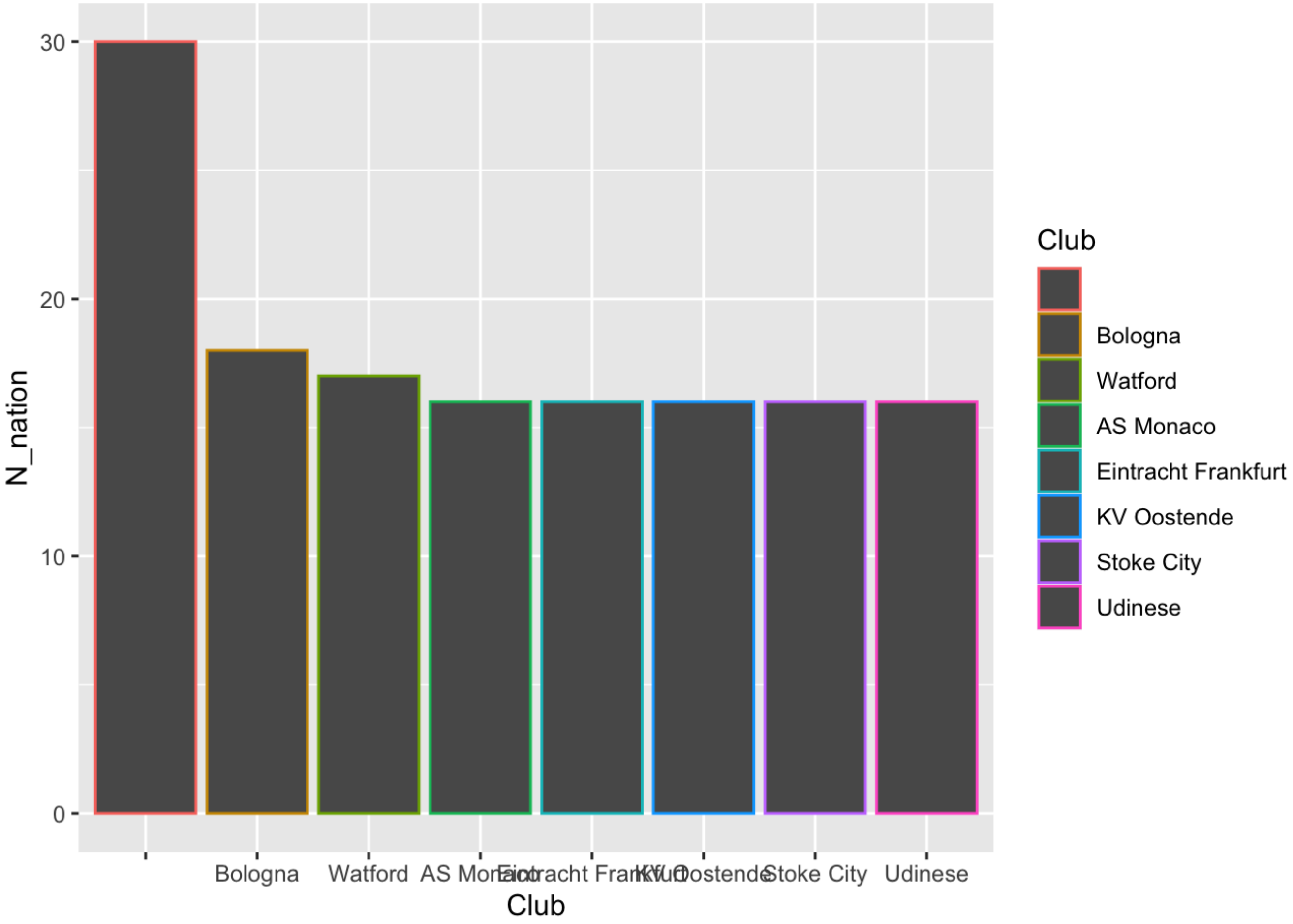
The previous data-analyst of the club was fired for producing poor plots. See below two bar plots that he made including their code.

a. Describe in your own words what did your predecessor try to show in each of the two plots.
b. Find *at least* three *different* problematic issues with his plots, and explain them.
c. Fix the problematic issues above in the code below to generate new, improved plots.
   You will get an additional *bonus* point for finding any additional problem and fixing it.
   (identifying the *same* problem in the two plots counts as *one* problem).

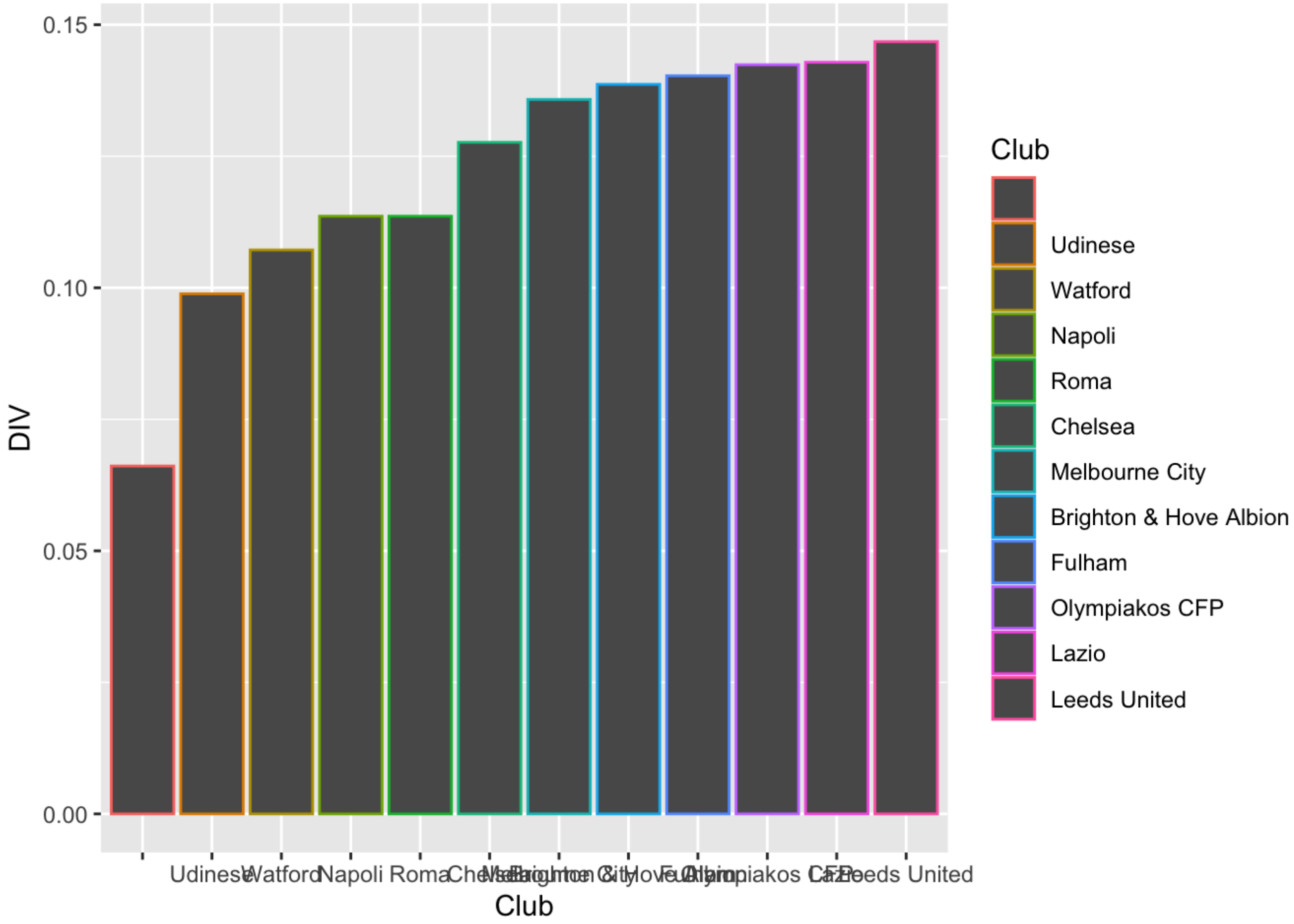
```
# A measure of category's diversity
DIV <- function(category_vec){
  t <- table(category_vec)
  p <- t/sum(t)
  return(sum(p^2))
}

cleaned_data <- fifa_players %>% select(Nationality,Club) %>% na.omit()

number_of_nationality_in_club <- cleaned_data %>% group_by(Club, Nationality) %>% summarise(c
ount = n()) %>% group_by(Club) %>% summarise(N_nation=n()) %>% arrange(desc(N_nation)) %>% mu
tate(Club = factor(Club, level=unique(Club)))
```

```
## `summarise()` has grouped output by 'Club'. You can override using the `.groups` argument.
```

```
DIV_in_club <- cleaned_data %>% group_by(Club) %>% summarise(DIV = DIV(Nationality))%>% arran
ge(DIV)%>% mutate(Club = factor(Club,level=unique(Club)))  # arrange(desc(DIV)) %>%

# Plot number of different nationalities in each club
g <- ggplot(data = number_of_nationality_in_club %>% head(8), aes(x = Club, y = N_nation,colo
r = Club))
g + geom_bar(stat="identity")
```



```
# Plot DIV (diversity?) of different nationalities in each club
g <- ggplot(data = DIV_in_club %>% head(12),aes(x = Club,y = DIV, color = Club))
g <- g + geom_bar(stat="identity")
g
```



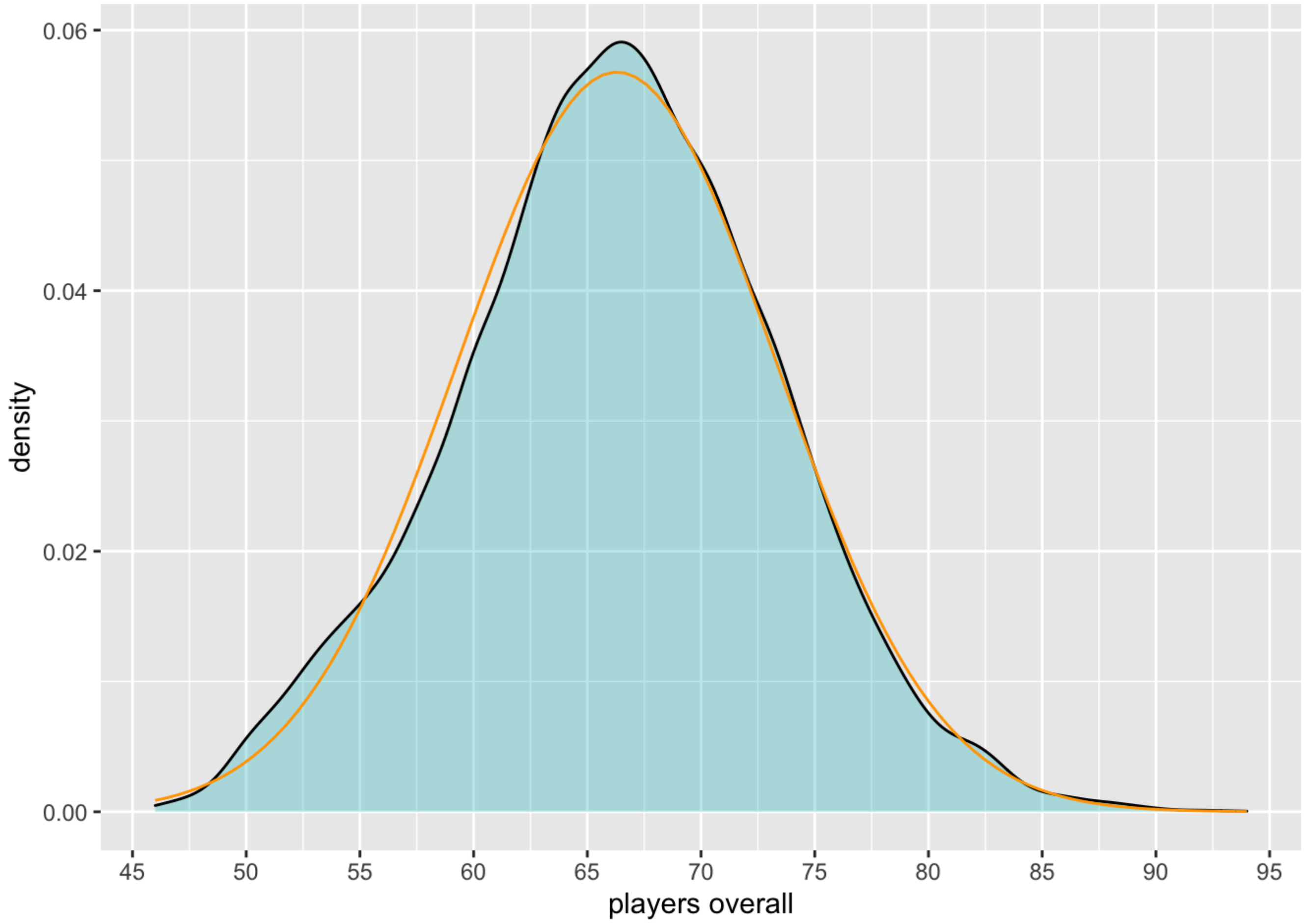
PLEASE ADD YOUR SOLUTION BELOW, WITH A CLEAR SEPARATION BETWEEN THE PARTS!

Q1 basic statistic

a

Make a plot showing the overall ability distribution of all players,How skewed is the distributions? does it have fat tails? Plot on top of the overall distribution a Normal distribution matching its first two moments.

```
ggplot(data = fifa_players,aes(x=Overall))+
  geom_density(alpha= .3,fill="#00BFC4")+
  stat_function(fun = dnorm,args = list(mean = mean(fifa_players$Overall),sd =sd(fifa_players
$Overall)),col="orange",lwd=0.5 )+
  scale_x_continuous(name = "players overall",breaks = breaks_width(5))
```



```
skewed_Wage <- skewness(fifa_players$Overall)
skewed_Wage
```

```
## [1] 0.008293
```

```
fat_tail_ind<-kurtosis(fifa_players$Overall)
fat_tail_ind
```

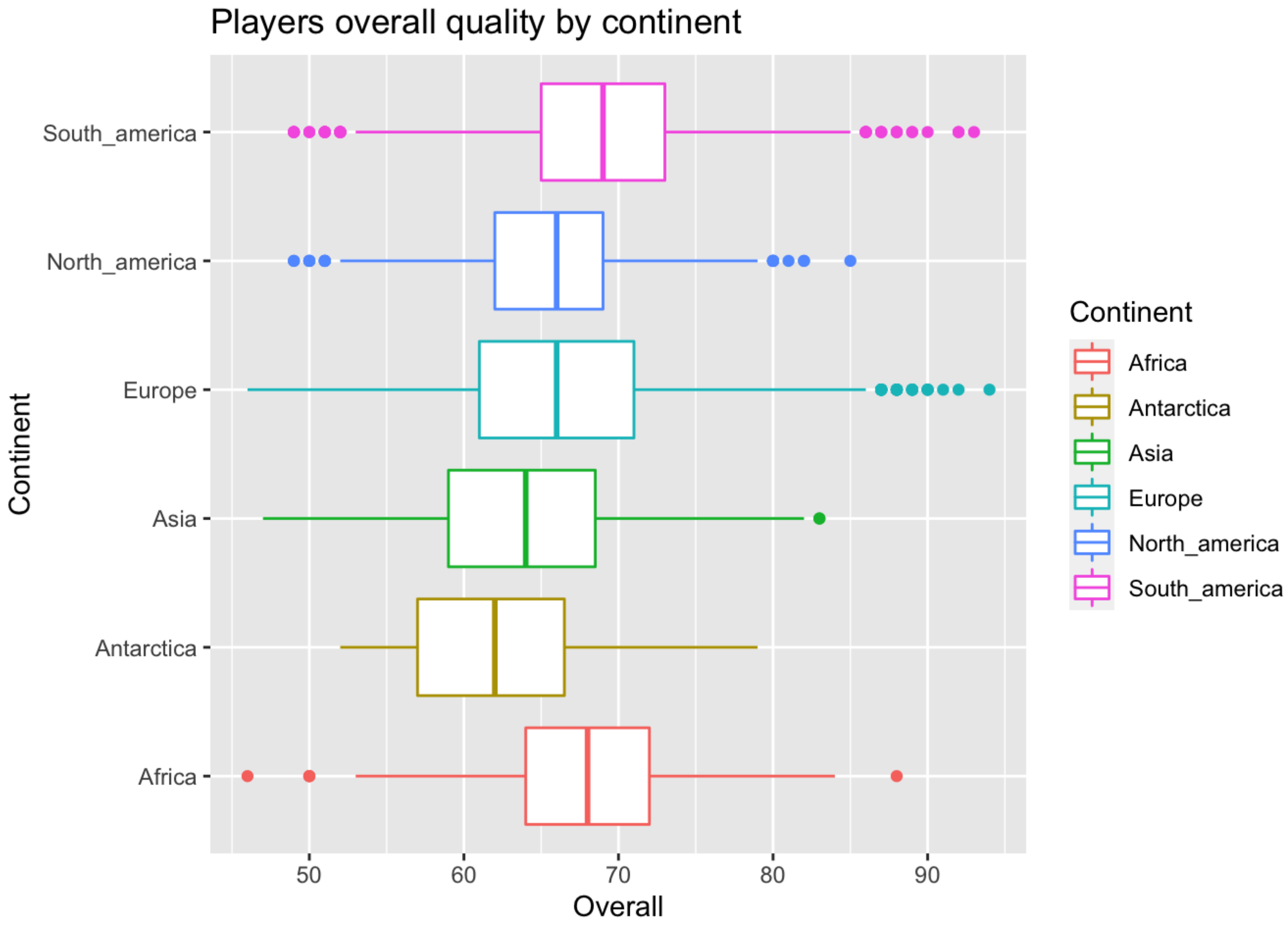
```
## [1] -0.02102
```

The results are that the distribution is 0.008293 skewed and have kurtosis of -0.02102 from the low skewed we can tell that the distribution of the overall is pretty symmetry. and from the low forth moment (compared to normal distribution) we can tell that the tails are thin we can see that the normal distribution above the Overall distribution is almost the same which means that the normal distribution described good the data

b

Make a plot comparing the multiple overall ability distributions of players according to the continent of the players. Describe which continents have especially good/bad players.

```
ggplot(data = fifa_players,aes(x=Overall,y=Continent,color = Continent ))+
  geom_boxplot()+
  labs(title = "Players overall quality by continent")
```

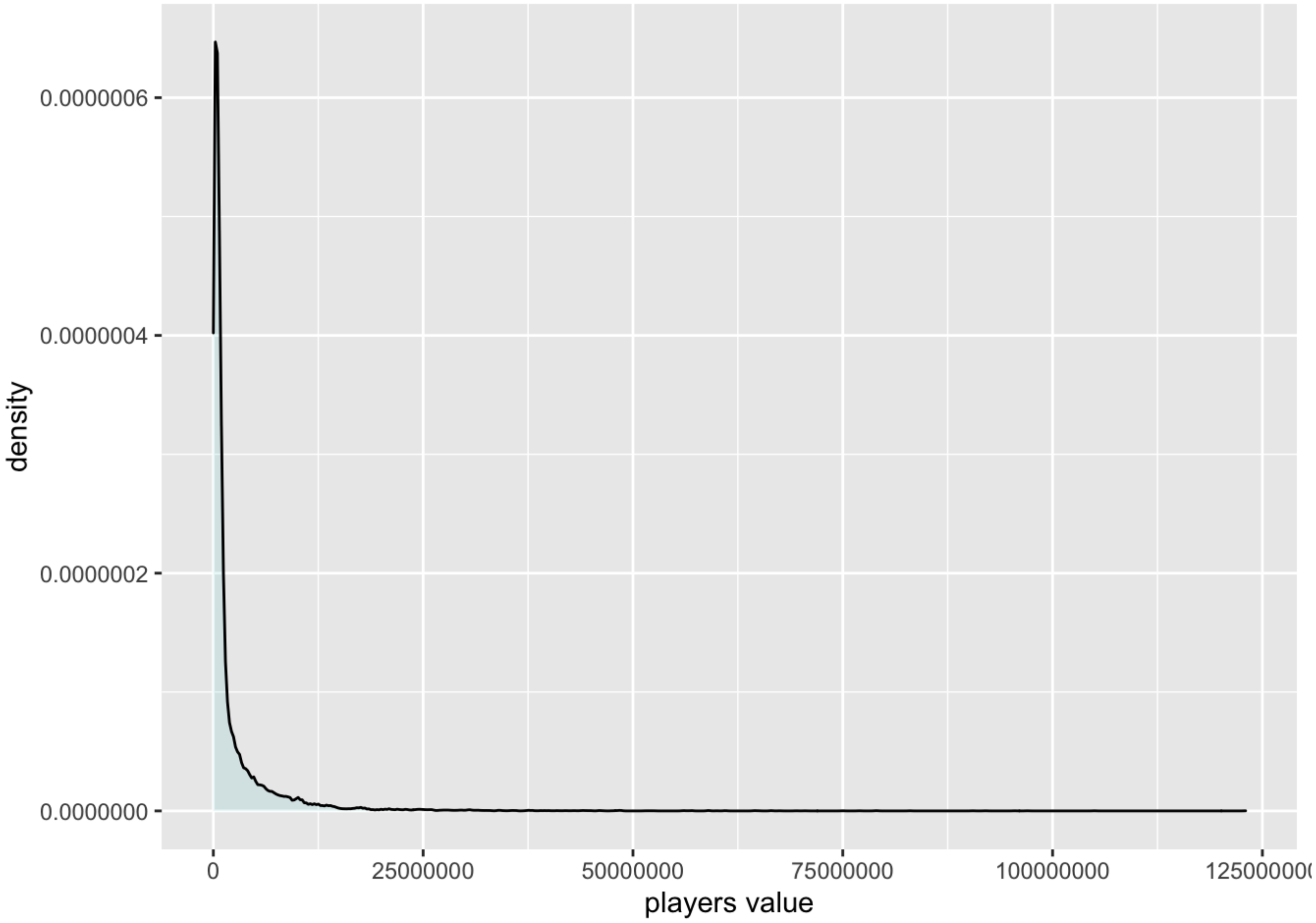


From the graph we can see that Antartica have extremely bad soccer players, on the other side Africa and South America have great players

C

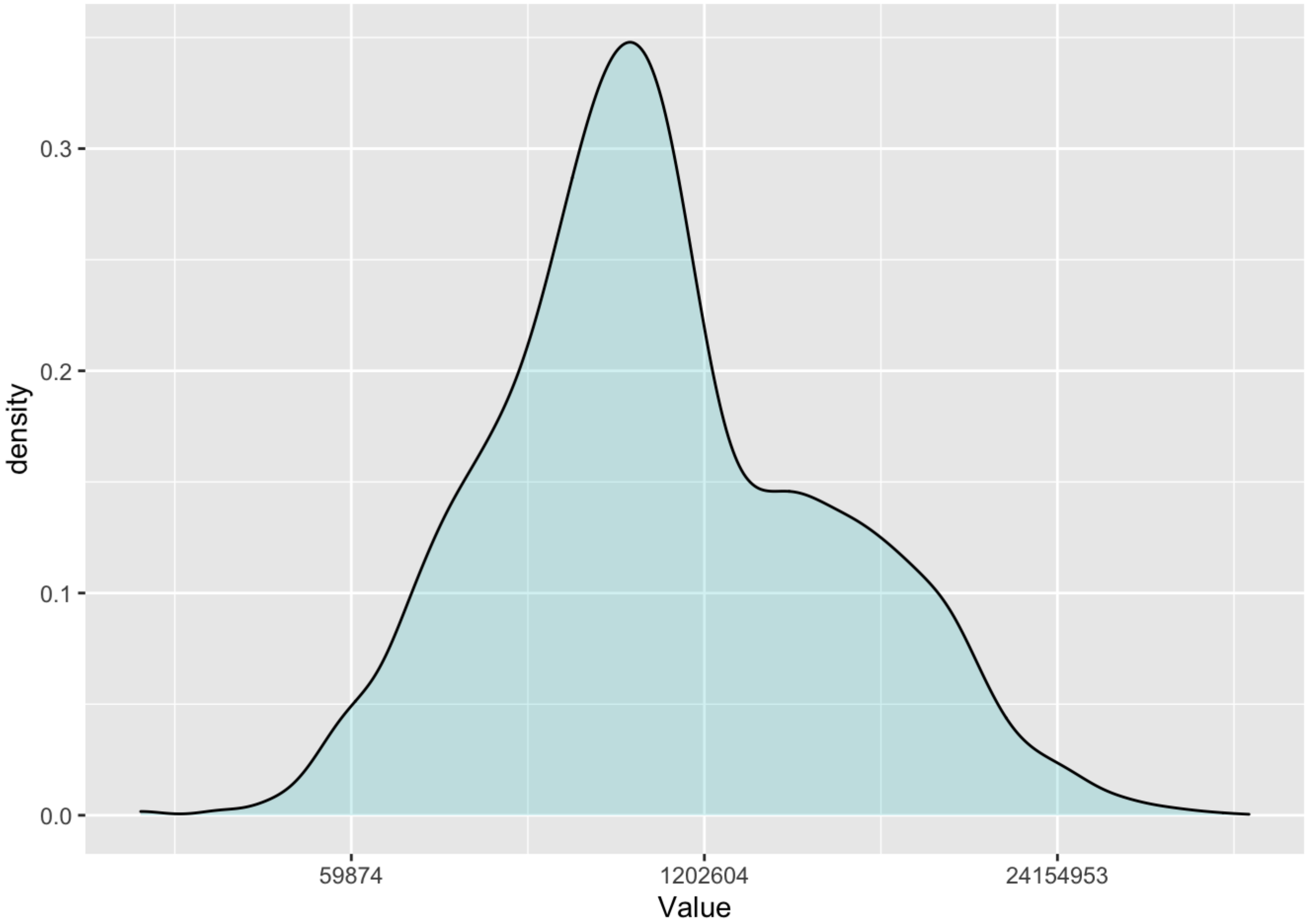
Make a plot showing the density of players' value distribution.

```
ggplot(data = fifa_players,aes(x=Value))+
  geom_density(alpha= .1,fill="#00BFC4")+
  scale_x_continuous(name = "players value")
```



Next, make a separate plot showing the density distribution of the log of players' value.

```
ggplot(data = fifa_players,aes(x=Value))+
  geom_density(alpha= .2,fill="#00BFC4")+
  scale_x_continuous(trans = "log")
```



The second graph (log of value) is a better visualization because in the first graph the values are moving too quickly to understand them in a small graph, on the other hand in the second (log) graph, the scale is the log of the the original values, which reduces very much the differences of the original values. so the data grow in a clearly rate which is easier to see and understand

d

Are the top-10 players with the highest value also the top-10 best players in terms of overall ability? Show tables for both and compare. Who is the best player not in the top-10 valued players?

```
top_10_val <- fifa_players %>% arrange(desc(Value))%>%
  dplyr::select(Name,Overall, Value)%>%
  head(10)
top_10_val
```

```
##                  Name Overall     Value
## 1              Neymar      92 123000000
## 2            L. Messi      93 105000000
## 3           L. Suárez      92  97000000
## 4  Cristiano Ronaldo      94  95500000
## 5      R. Lewandowski      91  92000000
## 6           E. Hazard      90  90500000
## 7        K. De Bruyne      89  83000000
## 8           P. Dybala      88  79000000
## 9            T. Kroos      90  79000000
## 10        G. Higuaín      90  77000000
```

```
top_10_over <- fifa_players %>% arrange(desc(Overall))%>%
  dplyr::select(Name,Overall,Value)%>%
  head(10)
top_10_over
```

```
##                  Name Overall     Value
## 1  Cristiano Ronaldo      94  95500000
## 2           L. Messi      93 105000000
## 3             Neymar      92 123000000
## 4           M. Neuer      92  61000000
## 5          L. Suárez      92  97000000
## 6     R. Lewandowski      91  92000000
## 7          E. Hazard      90  90500000
## 8             De Gea      90  64500000
## 9         G. Higuaín      90  77000000
## 10          T. Kroos      90  79000000
```

```
compare<-anti_join(top_10_over,top_10_val,by = c("Name", "Overall", "Value"))
compare[which.max(compare$Overall),]
```

```
##       Name Overall    Value
## 1 M. Neuer      92 61000000
```

We can see that Neuer is the best player not in the top-10 valued players because in soccer the GK is a position that normally is under valued compare to all the others

e

Show a table of the 10 youngest and 10 oldest teams in terms of average players age

```
ave_age_by_team<-aggregate(Age~Club,data = fifa_players,FUN = mean)
youngest_teams<-head(arrange(ave_age_by_team,Age),10)
knitr::kable(youngest_teams, caption = "youngest teams")
```

youngest teams

Club	Age
Sevilla Atlético	19.79
FC Barcelona B	20.38
Werder Bremen II	21.46
LOSC Lille	21.63
PSV	21.88
Crewe Alexandra	21.88
FC Nordsjælland	22.00
Ajax	22.07
KRC Genk	22.08

Club	Age
Barnsley	22.10

```
oldest_teams<-head(arrange(ave_age_by_team,desc(Age)),10)
knitr::kable(oldest_teams, caption = "oldest teams")
```

oldest teams

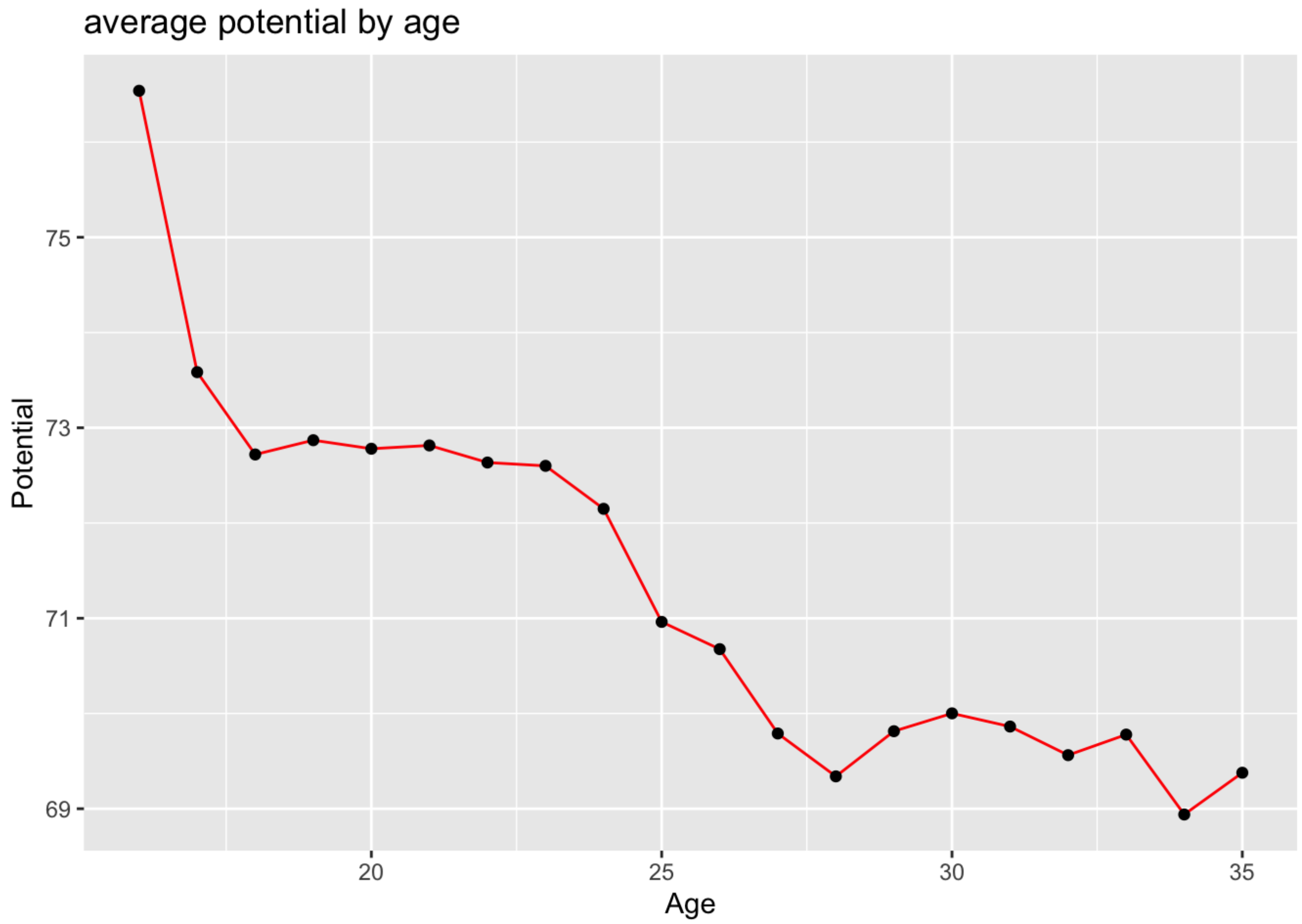
Club	Age
Brisbane Roar	31.00
Newcastle Jets	31.00
FC Seoul	30.75
Western Sydney Wanderers	30.75
Associação Chapecoense de Futebol	30.60
Adelaide United	30.40
Jeonbuk Hyundai Motors	30.33
Clube Atlético Paranaense	30.00
Grêmio Foot-Ball Porto Alegrense	30.00
Sydney FC	30.00

Q2

a

Plot the average potential ability by age of all players, for players 35 years old or younger

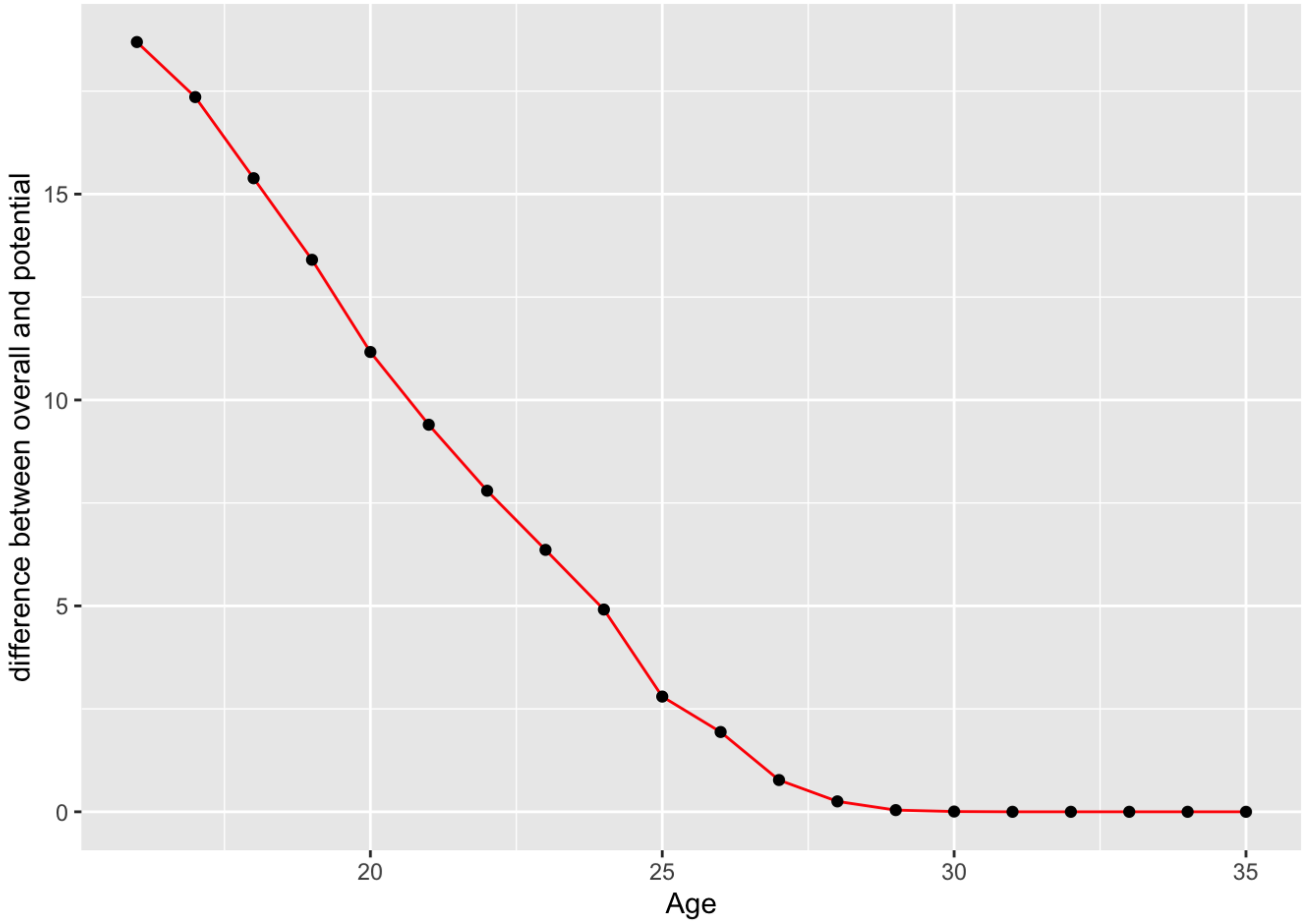
```
under_35<-filter(fifa_players,Age<=35)
potential_by_age<-aggregate(Potential~Age,data = under_35,FUN = mean)
ggplot(potential_by_age,aes(x=Age,y=Potential))+
geom_line(color = "red")+
geom_point()+
scale_x_continuous(name = "Age",breaks = breaks_width(5))+
labs(title = "average potential by age")
```



b

Plot the average difference between a player’s overall ability to potential ability as a function of age, up to age 35

```
Diff<-abs(under_35$Overall-under_35$Potential)
under_35<-mutate(under_35,Diff)
aver_diff<-aggregate(Diff~Age,data = under_35,FUN = mean)
ggplot(aver_diff,aes(x=Age,y=Diff))+
  ylab("difference between overall and potential")+
  geom_line(color="red")+
  geom_point()
```

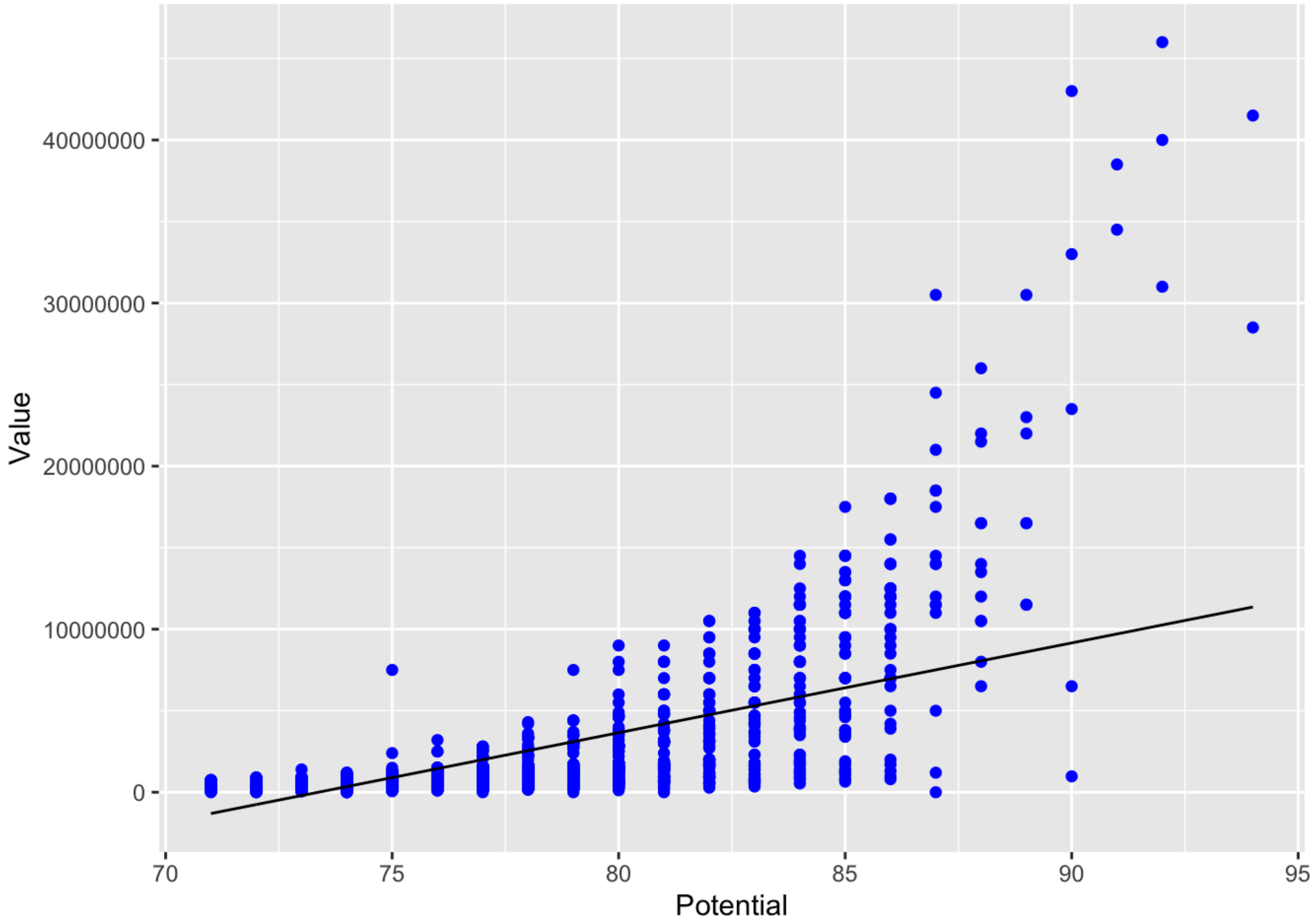


From the graph we can understand that as young as the player it's better for future development until the age of 25 because after that the Difference between the Overall and the Potential becomes under 2.5 so the player overall become more stabilized.

C

We are seeking young (age≤21) players with high potential (>70). Show a scatter plot of these players comparing their potential ability (x-axis) and current value (y-axis).Find the 10 most-undervalued players, i.e. having the lowest value compared to their predicted value by potential using a simple linear regression model. Calculate for each of them what is a fair value matching their potential that you be willing to pay in order to by them to your club and show these 10 players with their name, age, overall ability, potential, actual value and fair value it a table.

```
good_young<-filter(under_35,Age<=21&Potential>70)
regg_1<-lm(Value~Potential,data = good_young)
predicted_val<-regg_1$fitted.values
ggplot(good_young,aes(x=Potential,y=Value))+
  geom_point(color="blue")+
  geom_line(aes(y=predicted_val))
```



```
preper_under_val<-predicted_val-good_young$Value
index<-which(preper_under_val%in%head(sort(preper_under_val,decreasing = TRUE),10))
under_val_players<-as.vector(good_young$Name[index])
fair_val<-predicted_val[index]
under_val<-good_young$Value[index]
small_under_val<-good_young[index,]
small_under_val<-mutate(small_under_val,fair_val)%>%
  dplyr::select(Name,Age,Overall,Potential,Value,fair_val)
knitr::kable(small_under_val, caption = "top 10 under valued player")
```

top 10 under valued player

	Name	Age	Overall	Potential	Value	fair_val
194	R. Sessegnon	17	67	86	1300000	6953890
218	B. Woodburn	17	65	85	1100000	6403248
240	A. Gomes	16	64	90	975000	9156458
318	M. Edwards	18	65	87	1200000	7504532
413	V. Thill	17	63	85	800000	6403248
581	J. Sancho	17	63	86	800000	6953890
614	C. Früchtl	17	65	86	975000	6953890
643	J. Arp	17	63	85	825000	6403248
1928	E. Abouchabaka	17	62	85	650000	6403248

	Name	Age	Overall	Potential	Value	fair_val
2353	W. Faríñez	19	73	87	0	7504532

As we can see in the question above we need to find the 10 most-undervalued players. which means that we need to take the players with the max difference between their value and their predict value, but only when the predict value is higher then the value(taking the other option will be overvalue players). So as we can see those are the players that we show in our table.

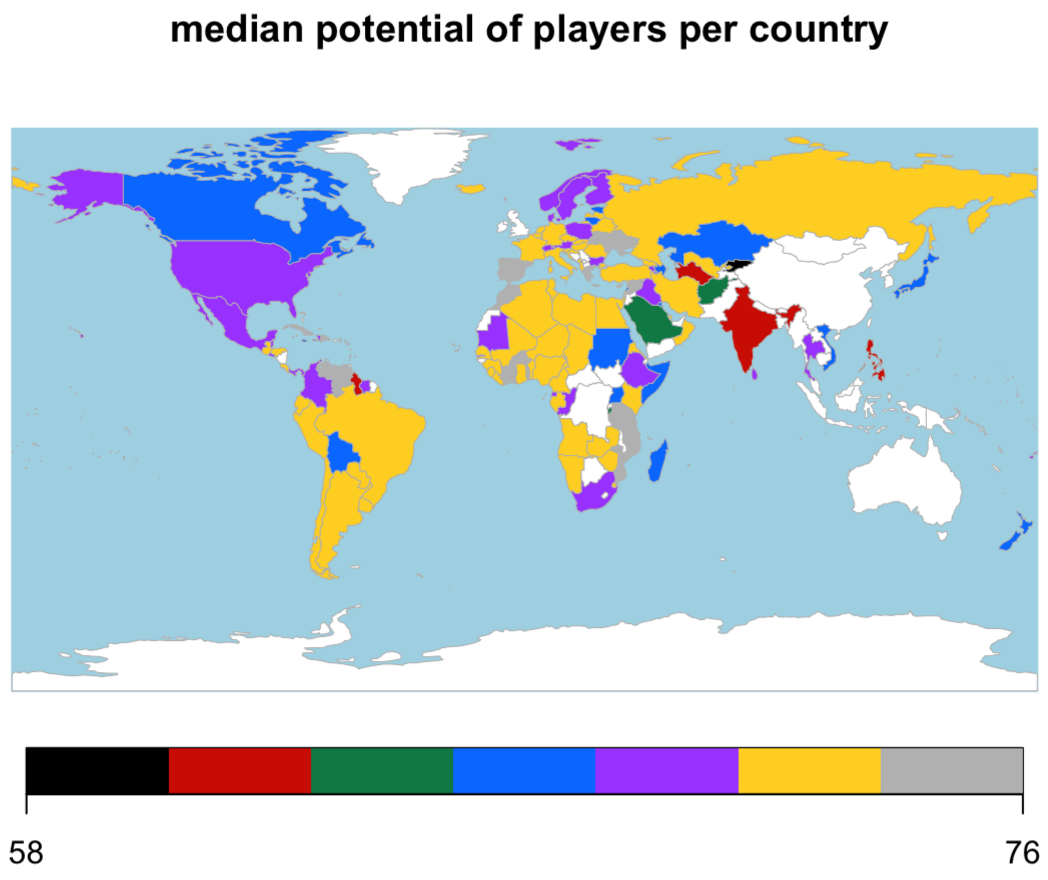
d

Use the rworldmap package to display the world map and color each country based on the median potential of players from this nationality.

```
prep_to_map<-aggregate(Potential~Nationality,data = fifa_players,FUN = median)
mapped_data <- joinCountryData2Map(prep_to_map, joinCode = "NAME", nameJoinColumn = "National
ity")
```

```
## 133 codes from your data successfully matched countries in the map
## 4 codes from your data failed to match with a country code in the map
## 110 codes from the map weren't represented in your data
```

```
theMap <- mapCountryData(mapped_data, nameColumnToPlot="Potential",catMethod ="fixedWidth",co
lourPalette = "palette",
oceanCol = "lightblue", missingCountryCol = "white",mapTitle = " median potential of players
 per country"
,aspect = "variable")
```



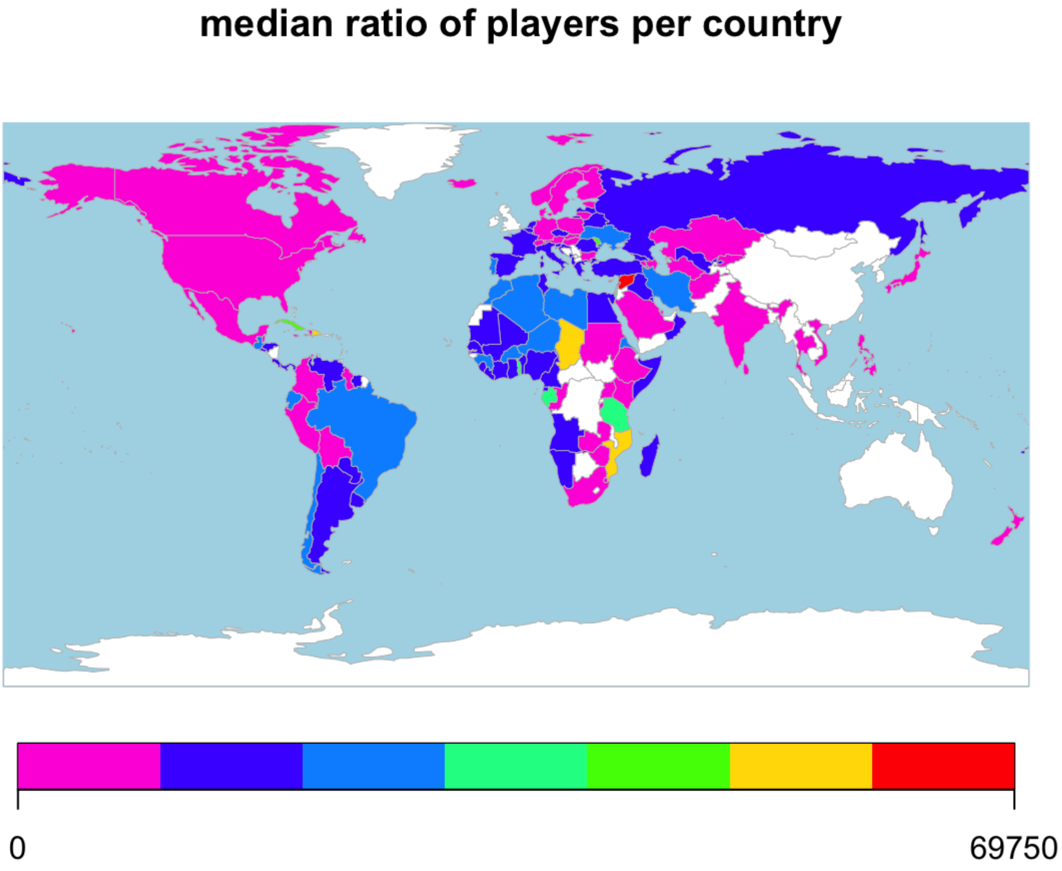
e

Repeat the above analysis but this time display a world map where each country is colored by the median ratio of potential to value of players. Find an under-valued country you'd recommend to travel to (i.e. a country with cheap players compared to their potential average quality).

```
ratio<-fifa_players$Value/fifa_players$Potential
fifa_map<-mutate(fifa_players,ratio)
prep_to_map_2<-aggregate(ratio~Nationality,data = fifa_map,FUN = median)
mapped_data_2 <- joinCountryData2Map(prep_to_map_2, joinCode = "NAME", nameJoinColumn = "Nati
onality")
```

```
## 133 codes from your data successfully matched countries in the map
## 4 codes from your data failed to match with a country code in the map
## 110 codes from the map weren't represented in your data
```

```
theMap_2 <- mapCountryData(mapped_data_2, nameColumnToPlot="ratio",catMethod ="fixedWidth",co
lourPalette = "rainbow",
oceanCol = "lightblue", missingCountryCol = "white",mapTitle = " median ratio of players per
 country"
,aspect = "variable")
```



```
check_1<-aggregate(Potential~Nationality,data = fifa_map,FUN = mean)
check_2<-aggregate(Value~Nationality,data = fifa_map,FUN = mean)
check_3<- full_join(check_2,check_1,by = "Nationality")
check_4<-check_3$Value/check_3$Potential
best_country<-aggregate(ratio~Nationality,data = fifa_map,FUN = mean)
cheap_conti<-top_n(best_country,1,ratio)
knitr::kable(cheap_conti, caption = "recommend country")
```

recommend country

Nationality	ratio
Gabon	93628

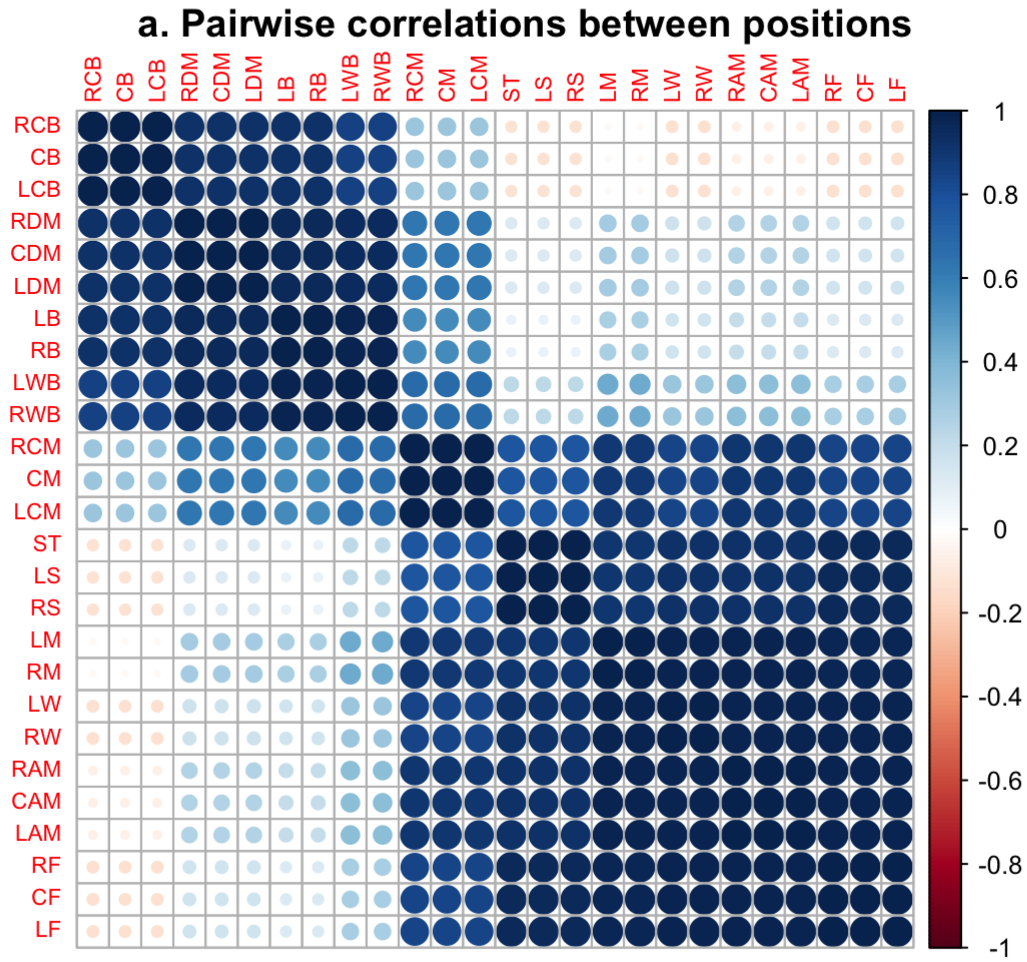
We would recommend to travel to Gabon

Q3

a

We are interested in finding out which positions are similar in terms of players' performance. Extract the 26 non-goalkeeper positions (CAM, CB, …, ST). Calculate the correlation between players' ability in each pair of positions and show a heatmap correlation-plot of the correlations' matrix. What three positions have the least average correlations with other skills?

```
no_gk<-fifa_players_positions[!(fifa_players_positions$Preferred.Positions=="GK "),]
cor_positions<-cor(no_gk[,-c(1,28:30)])
corrplot(cor_positions,order='hclust', tl.cex = 0.7,mar = c(0,0,1,0), title="a. Pairwise corr
elations between positions")
```



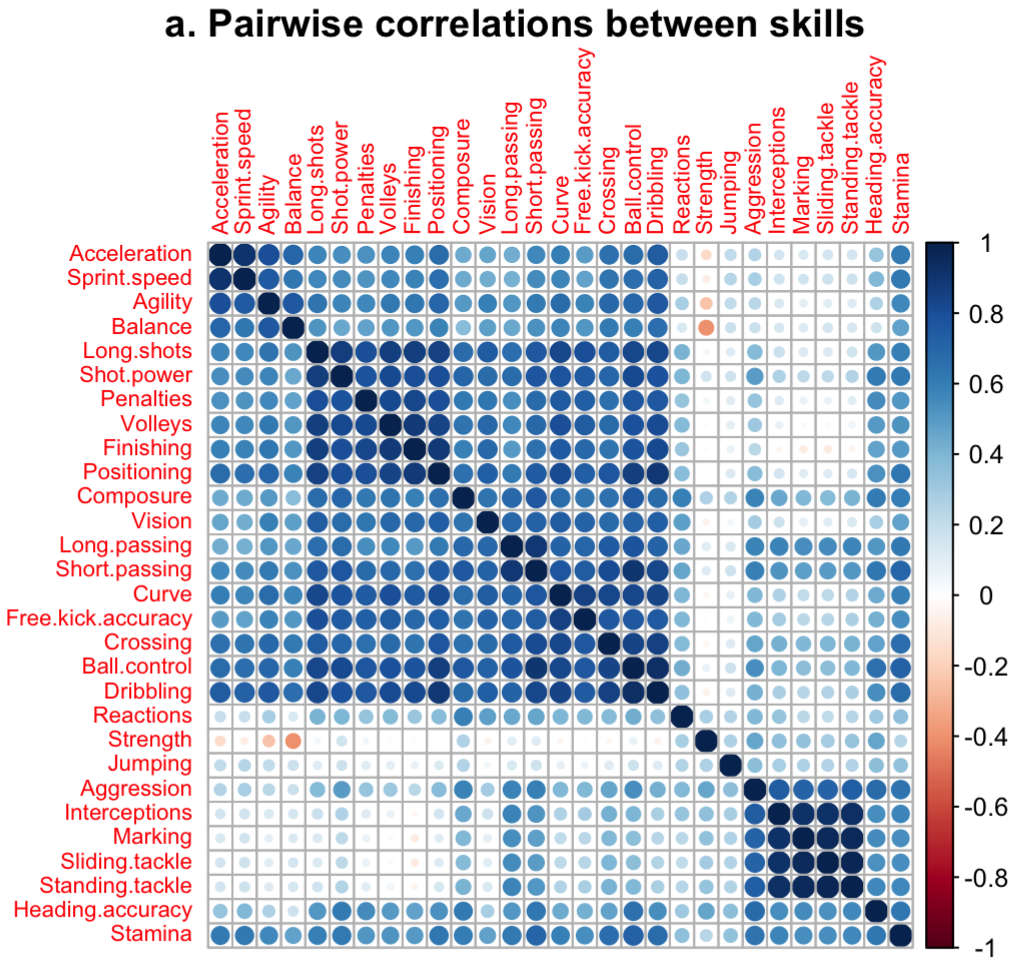
```
mean_of_cor<-rowMeans(cor_positions)
head(sort(mean_of_cor),3)
```

```
##     CB    LCB    RCB
## 0.3459 0.3459 0.3459
```

Three positions that have the least average correlations with other skills are CB, LCB and RCB.

We are interested in finding out which skills are similar in terms of players' performance at the position. Extract the 29 skills for non-goalkeeper players (Acceleration, …, Volleys, except 'GK.' skills). Calculate the correlation between players' ability in each pair of skills and show a heatmap correlation-plot of the correlations' matrix. What two skills seem least correlated with other skills?

```
ready_to_cor_2<-cor(fifa_players_attribures[,c(2:12,18:35)])
corrplot(ready_to_cor_2,order='hclust', tl.cex = 0.7,mar = c(0,0,1,0), title="a. Pairwise cor
relations between skills")
```



```
mean_of_cor_2<-rowMeans(ready_to_cor_2)
head(sort(mean_of_cor_2),2)
```

```
## Strength  Jumping
##   0.1314   0.2198
```

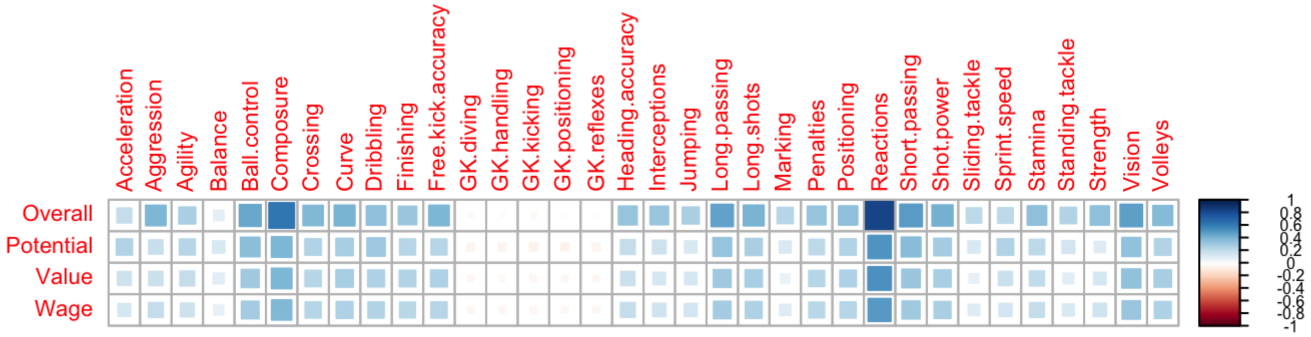
Two skills that are least correlated with other skills are Strength and Jumping.

b

Consider the following indicators of players performance: overall players' performance, their potential, their salary (wage) and their market value. Show a correlation-plot of players' 34 skill levels (Acceleration, …, Volleys) vs. these four indicators. Find the 10 skills with the highest average correlation with the four inidcators and list them in a table.

```
four_indicators<-fifa_players[,c(6,7,10,11)]
skills_dat<-fifa_players_attribures[,c(2:35)]
ready_to_cor_3<-cor(four_indicators,skills_dat)
corrplot(ready_to_cor_3,method= "square",cl.lim = c(-1,1),cl.cex = 0.5,tl.cex = 0.7,mar = c(0
,0,1,0), title="a. Pairwise correlations between skills",is.corr = FALSE)
```

a. Pairwise correlations between skills



```
mean_of_cor_3<-colMeans(ready_to_cor_3)
best_skills<-head(sort(mean_of_cor_3,decreasing = TRUE),10)
best_skills<-as.data.frame(best_skills)
knitr::kable(best_skills, caption = "best skills")
```

best skills

	best_skills
Reactions	0.5994
Composure	0.4576
Short.passing	0.3749
Vision	0.3745
Ball.control	0.3555
Long.passing	0.3502
Shot.power	0.3176
Curve	0.3143
Long.shots	0.3090
Dribbling	0.2994

c

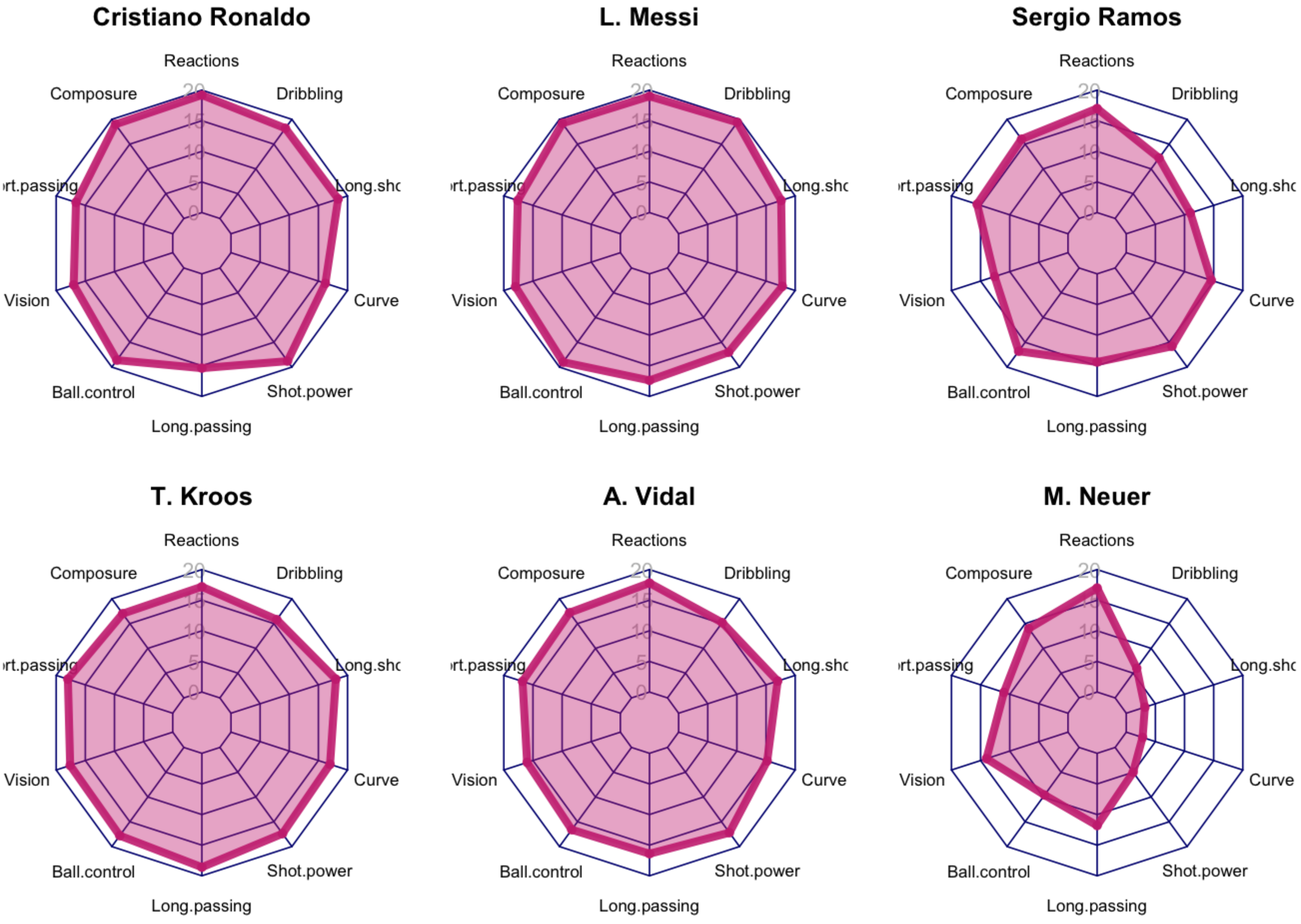
Build a team of 11 different players - For each of the 26 non-goalkeaper positions (26 from above plus goalkeaper, GK), find the player with the best performance at this position.Find the goal keaper (Preffered.Positions is GK) with the best overall performance.From the players obtained above, find 11 distinct players maximizing the average overall performance of the team, with the constraint that there must be a goalkeaper (preferred position GK).List the players in a table including their overall performance and the team average overall score. Next, peak six different players of your choice from your team, one of which is the goalkeaper. Using the function radarchart::chartJSRadar, graph their abilities (individually for all 6 players) in the top 10 skills according to 3.b in a radar chart

```
all_positions <-fifa_players[,c(46:71)]
all_max<-apply(all_positions,2,function(x)  which( x == max(x) ))
index_2<-unlist(all_max)
index_2<-unique(index_2)
best_in_pos<-fifa_players[index_2,]
fifa_players$Preferred.Positions<-as.character(fifa_players$Preferred.Positions)
only_gk<-filter(fifa_players,Preferred.Positions=="GK ")
best_gk<-only_gk[which.max(only_gk$Overall),]
our_team<-rbind(best_in_pos[-which.min(best_in_pos$Overall),],best_gk)
average<-data_frame(Name= "Team average",Overall = mean(our_team$Overall))
our_team<-dplyr::select(our_team,Name,Overall)
final_team<-rbind(our_team,average)
knitr::kable(final_team, caption = "best team")
```

best team

	Name	Overall
17349	L. Messi	93.00
17560	Sergio Ramos	90.00
2439	A. Vidal	87.00
17559	T. Kroos	90.00
3295	Alex Sandro	86.00
17563	Marcelo	87.00
17558	Cristiano Ronaldo	94.00
2441	D. Alaba	86.00
3518	R. Nainggolan	86.00
17356	Jordi Alba	85.00
261	M. Neuer	92.00
1	Team average	88.73

```
chosen_players<-rbind(head(arrange(best_in_pos,desc(Overall)),5),best_gk)
data_to_spider<-dplyr::select(chosen_players,rownames(best_skills))
data_to_spider<- rbind(rep(100,dim(data_to_spider)[2]),rep(0,dim(data_to_spider)[2]),data_to_
spider)
title<-as.character(chosen_players$Name)
par(mar=rep(1,4))
par(mfrow=c(2,3))
for(i in 1:6){
  radarchart(data_to_spider[c(1,2,i+2),],axistype=1, pcol=rgb(0.8,0.2,0.5,0.9),pfcol = rgb(0.
8,0.2,0.5,0.4), plwd=4,plty=1, cgcol="grey",cglty=1, axislabcol="grey", caxislabels=seq(0,20,
5), cglwd=0.8,vlcex=0.8,title=title[i])
}
```



Cristiano Ronaldo

L. Messi

Sergio Ramos

T. Kroos

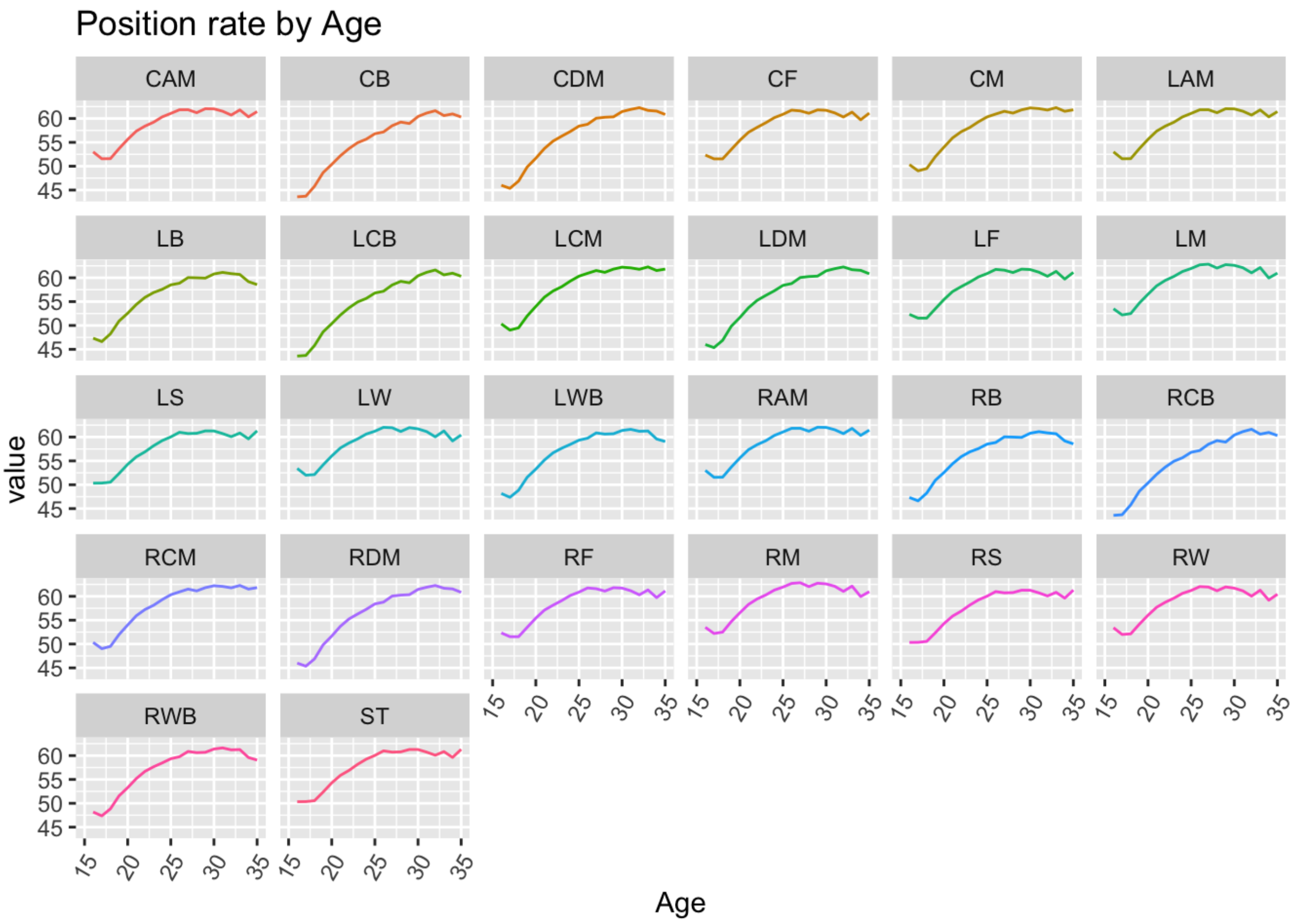
A. Vidal

M. Neuer

d

Repeat the analysis of question 2.a., but this time show the 34 different skills

```
no_gk_under_35<-under_35[!(under_35$Preferred.Positions=="GK "),]
check_pos<-no_gk_under_35[,c(3,46:71)]
mean_check<-aggregate(.~Age,data = check_pos,FUN = mean)
trans_check<-t(mean_check)
vec_to_insert<-c(colnames(under_35[,c(46:71)]))
trans_check<-as.data.frame(trans_check)
colnames(trans_check) <- trans_check[1,]
trans_check<-trans_check[-1,]
trans_check<-mutate(trans_check, position = vec_to_insert)
skills_to_P<-melt(trans_check,id.vars = "position")
colnames(skills_to_P)[2]<-"Age"
skills_to_P$Age<-as.numeric(as.character(skills_to_P$Age))
ggplot(data = skills_to_P,aes(x=Age,y=value,group=position))+
geom_line(aes(color=position),show.legend = FALSE)+
theme(axis.text.x = element_text(angle = 60, hjust = 1))+
facet_wrap(~position)+
 xlim(15,35)+
  labs(title = "Position rate by Age")
```



As we can see there’s no one position who peak at the youngest age, it can be explained by the lack of experience We noticed that the defense players peak at the age of 30, the center players peak at the age of 35 and the attacking players peak at the age of 25.

e

Fit a multiple regression model predicting player’s overall performance based on their wage and age. Find the 10 players with the highest difference between their overall performance level and the regression model prediction, and list them in a table.

```
regg_2<-lm(Overall~Wage+Age,data = fifa_players)
differnce_2<-regg_2$residuals
index_3<-which(differnce_2%in%head(sort(differnce_2,decreasing = TRUE),10))
top_10_diff<-fifa_players[index_3,]%>% dplyr::select(Name)
top_10_diff<-mutate(top_10_diff,Differnce = differnce_2[index_3])%>%
  arrange(desc(Differnce))
knitr::kable(top_10_diff, caption = "players with highest difference")
```

players with highest difference

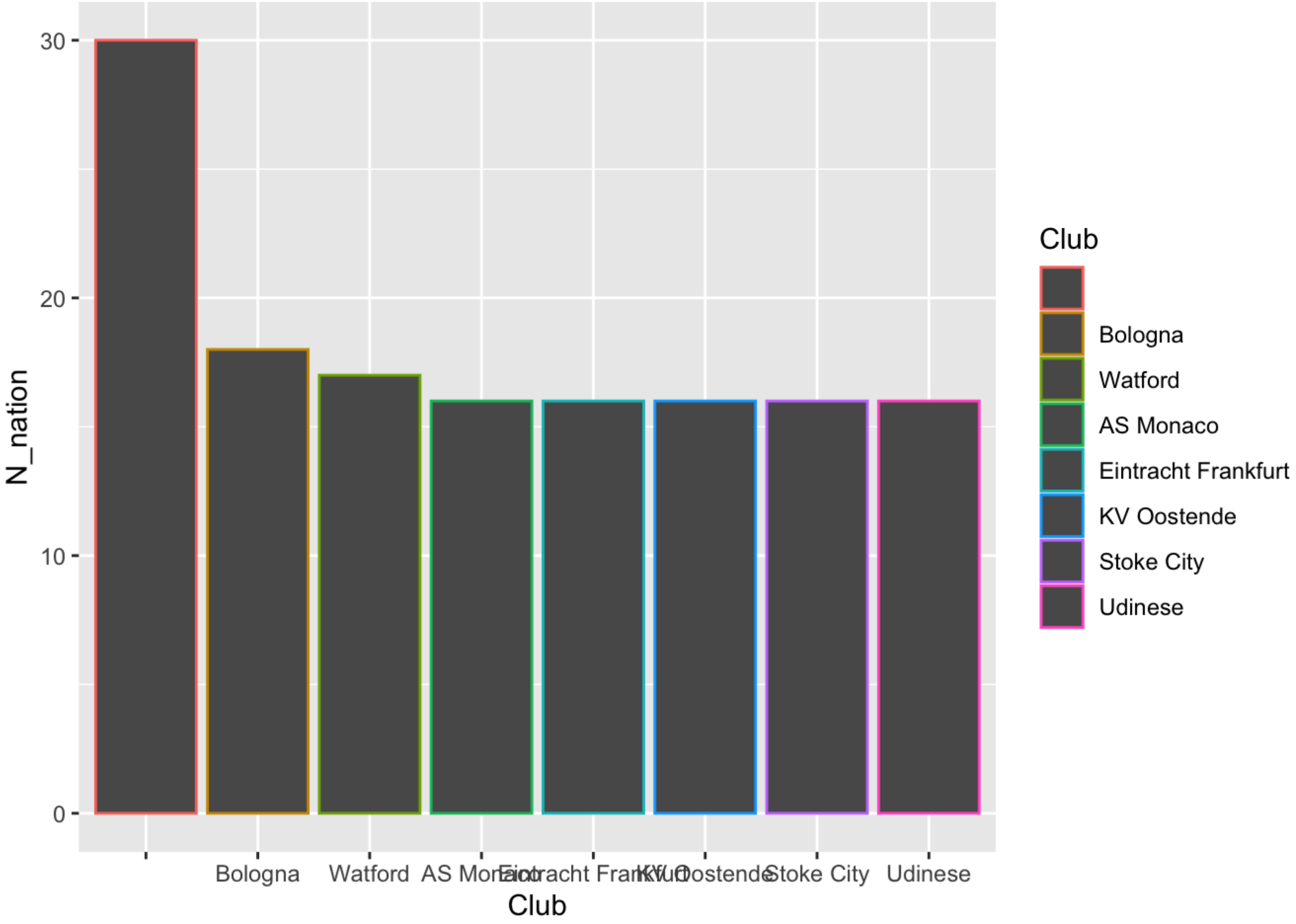
Name	Differnce
Oscar	18.67
K. Mbappé	17.72
Adrien Silva	16.94
Sergio Rico	16.25
Fred	16.08
G. Donnarumma	16.08
Grimaldo	16.05
A. Witsel	15.94
Danilo Pereira	15.78
K. Dolberg	15.52

Q4

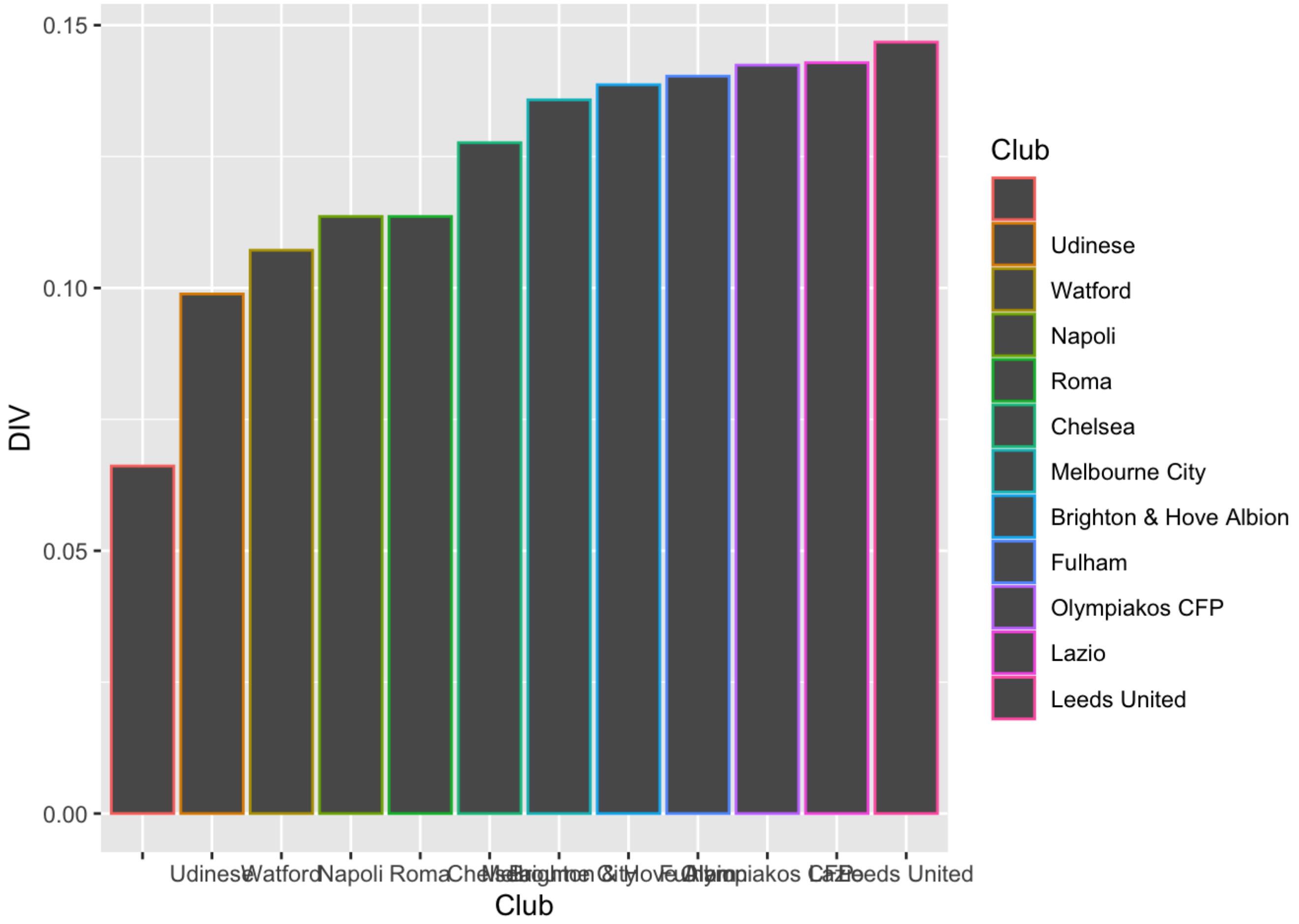
```
# A measure of category's diversity
DIV <- function(category_vec){
  t <- table(category_vec)
  p <- t/sum(t)
  return(sum(p^2))
}

cleaned_data <- fifa_players %>% dplyr::select(Nationality,Club) %>% na.omit()

number_of_nationality_in_club <- cleaned_data %>% group_by(Club, Nationality) %>% summarise(c
ount = n()) %>% group_by(Club) %>% summarise(N_nation=n()) %>% arrange(desc(N_nation)) %>% mu
tate(Club = factor(Club, level=unique(Club)))
```

```
## `summarise()` has grouped output by 'Club'. You can override using the `.groups` argument.
```

```
DIV_in_club <- cleaned_data %>% group_by(Club) %>% summarise(DIV = DIV(Nationality))%>% arran
ge(DIV)%>% mutate(Club = factor(Club,level=unique(Club)))  # arrange(desc(DIV)) %>%

# Plot number of different nationalities in each club
g <- ggplot(data = number_of_nationality_in_club %>% head(8), aes(x = Club, y = N_nation,colo
r = Club))
g + geom_bar(stat="identity")
```



```
# Plot DIV (diversity?) of different nationalities in each club
g <- ggplot(data = DIV_in_club %>% head(12),aes(x = Club,y = DIV, color = Club))
g <- g + geom_bar(stat="identity")
g
```



a

The goal of the both plots is to show which team is the most diverse in terms of players nationality. The first plot represent the highest 8 clubs with players from different number of nations. And The second plot shows the Worst 12 clubs proportion. This measure takes into account the proportions of players from each nation, not just the overall number of nation.

b

The problems that we found:

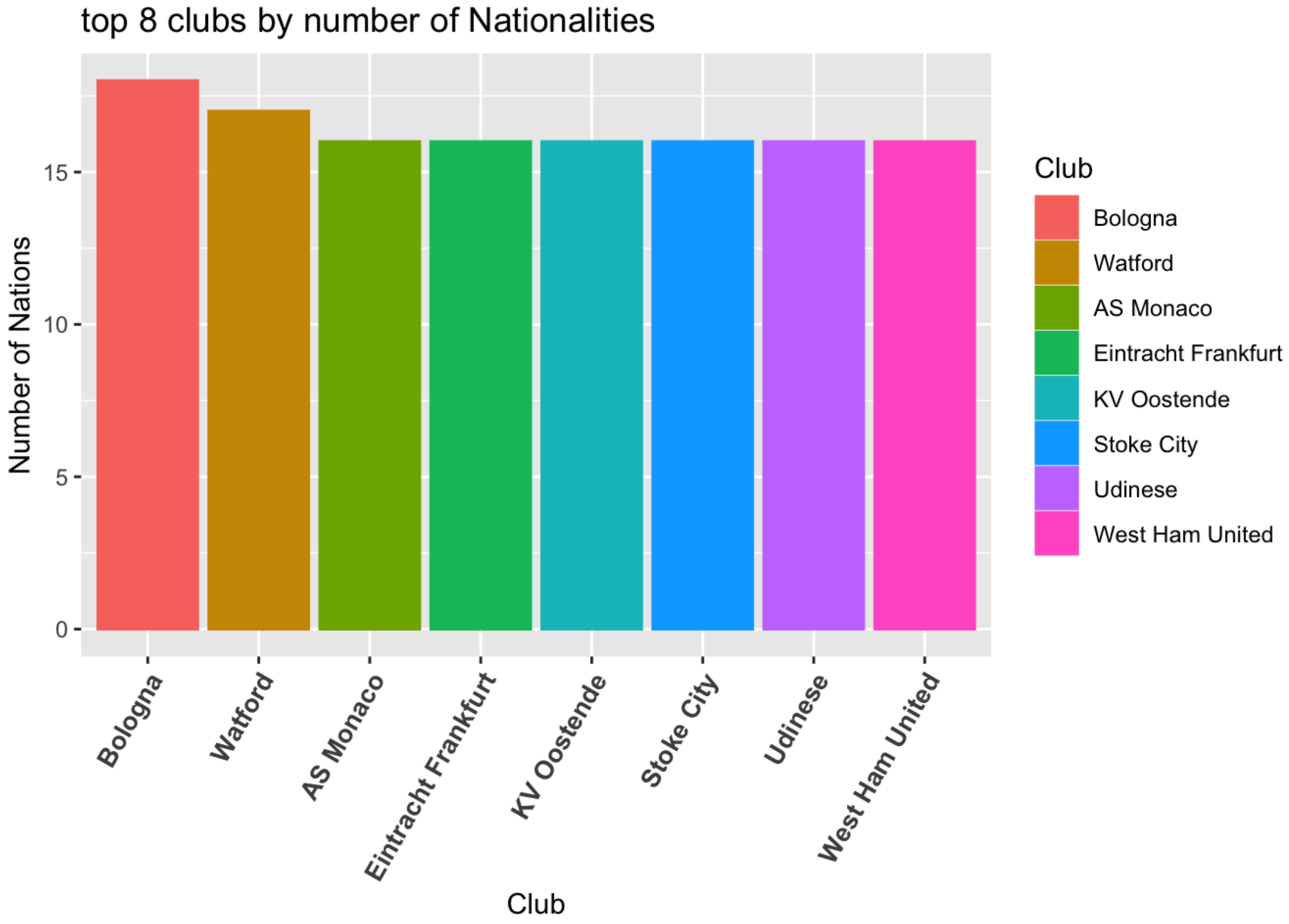
1-The first column is NA and those values should be removed. 2-There’s no title. 3-The names of the clubs is stepping on the others names so we can’t read them. 4-The filling of each column is with the same color, and it makes it hard to understand. 5-The names of the axes can be defined better

c

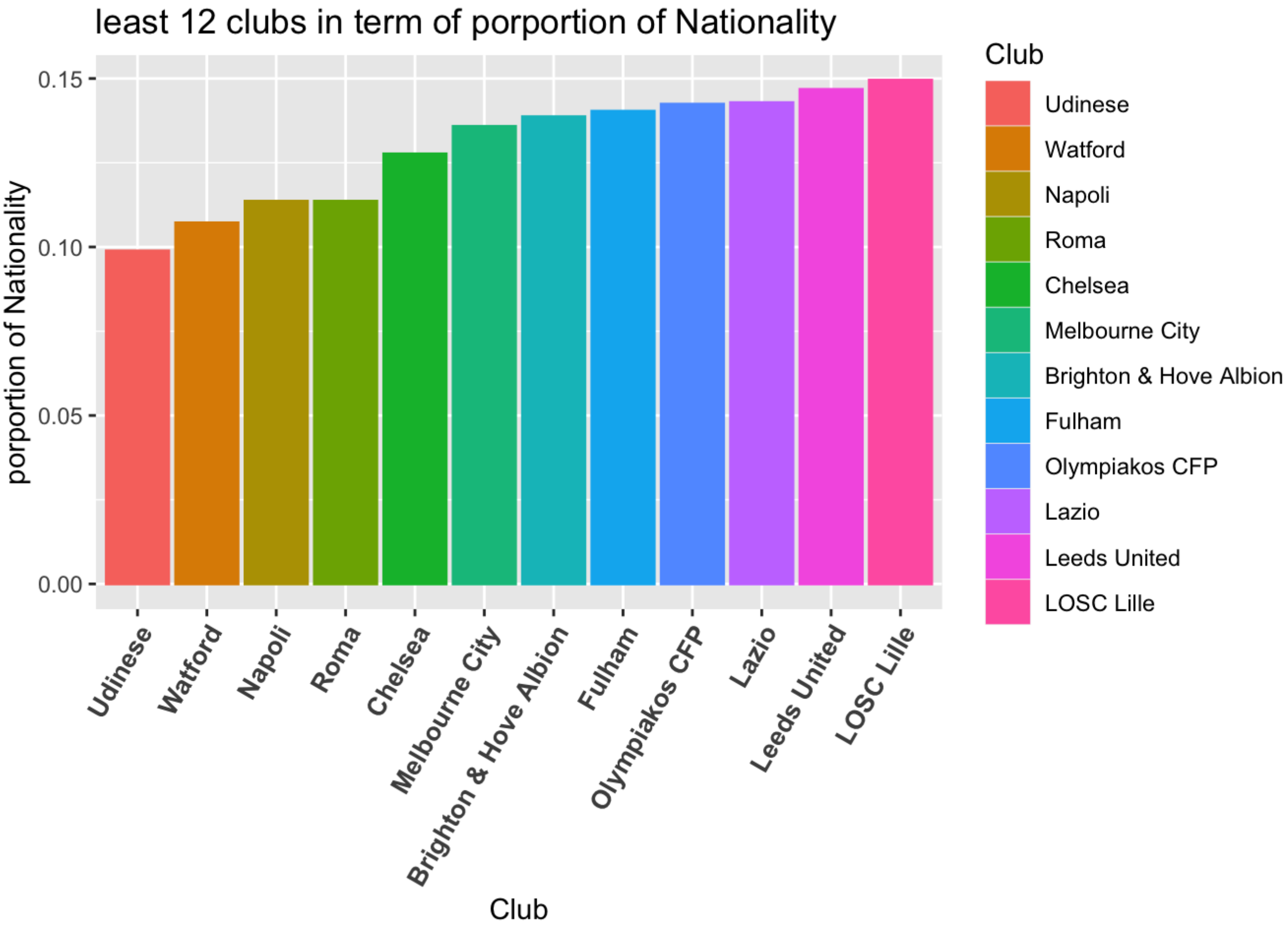
```
number_of_nationality_in_club <- cleaned_data %>% group_by(Club,Nationality) %>% summarise(co
unt = n()) %>% group_by(Club) %>% summarise(N_nation=n()) %>% filter(Club!="") %>% arrange(de
sc(N_nation)) %>% mutate(Club = factor(Club,level=unique(Club)))
```

```
## `summarise()` has grouped output by 'Club'. You can override using the `.groups` argument.
```

```
DIV_in_club <- cleaned_data %>% group_by(Club) %>% summarise(DIV = DIV(Nationality))%>%  filt
er(Club!="") %>% arrange(DIV)%>% mutate(Club = factor(Club,level=unique(Club)))
g <- ggplot(data = number_of_nationality_in_club %>% head(8),aes(x = Club,y = N_nation,color
 = Club, fill = Club))
g + geom_bar(stat="identity") +
theme(axis.text.x = element_text(face = "bold", size = 10, angle = 60, hjust = 1)) + ylab('Nu
mber of Nations') +
labs(title="top 8 clubs by number of Nationalities")
```



```
g <- ggplot(data = DIV_in_club %>% head(12),aes(x = Club, y = DIV, color = Club, fill = Clu
b))
g + geom_bar(stat="identity") + theme(axis.text.x = element_text(face = "bold", size = 10, an
gle = 60, hjust = 1)) + ylab('porportion of Nationality') +
labs(title="least 12 clubs in term of porportion of Nationality")
```


least 12 clubs in term of porportion of Nationality
porportion of Nationality
0.15
0.10
0.05
0.00
Club
Udinese
Watford
Napoli
Roma
Chelsea
Melbourne City
Brighton & Hove Albion
Fulham
Olympiakos CFP
Lazio
Leeds United
LOSC Lille
Club
Udinese
Watford
Napoli
Roma
Chelsea
Melbourne City
Brighton & Hove Albion
Fulham
Olympiakos CFP
Lazio
Leeds United
LOSC Lille