EDA_FIFA22

Soheil Moattar Berenguer

4/6/2022

```
library(data.table) # For Fast Data Loading
library(forcats) # To use the forcats function
library(dplyr)
library(ggplot2)
library(inspectdf) # Automatic Data Exploration Analysis
library(stringi) #For handling strings
library(maps)
library(plyr) #to rename and map levels
#Put path of the directory containing the data files
x <-paste("/Users/soheilmoattarmohammadiberenguer/Desktop/Entrega_Soheil_Moattar_Berenguer",
          sep="")



#Se the specified path as new directory
setwd(x)

source("./Cleaning_Data2.R")

data<- data_eda
```

Most Valuable Teams

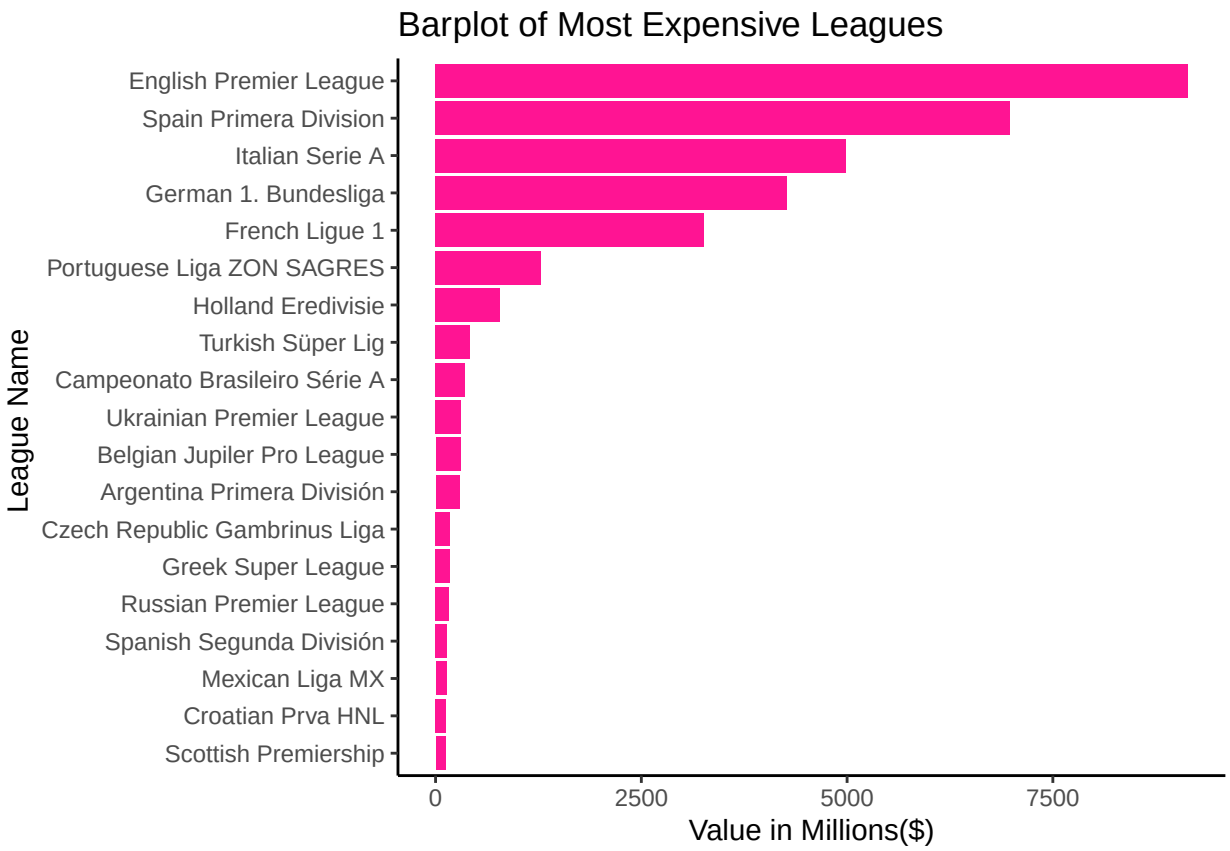
Here we can see a barplot of the most expensive teams per league

```
plot_bar <- function(df) {


    df%>% mutate(league_name=as.factor(league_name))%>%
      ggplot(aes(x = fct_reorder(league_name,value_eur,sum),
                      y = value_eur/1000000)) +
    geom_col(fill = "deeppink")+
    coord_flip()+
    labs(
      title = "Barplot of Most Expensive Leagues", x = "League Name ",
      y = "Value in Millions($)",
    )+
    theme_classic()

}
```

```
plot_bar(data)
```



Most Valuable Squads

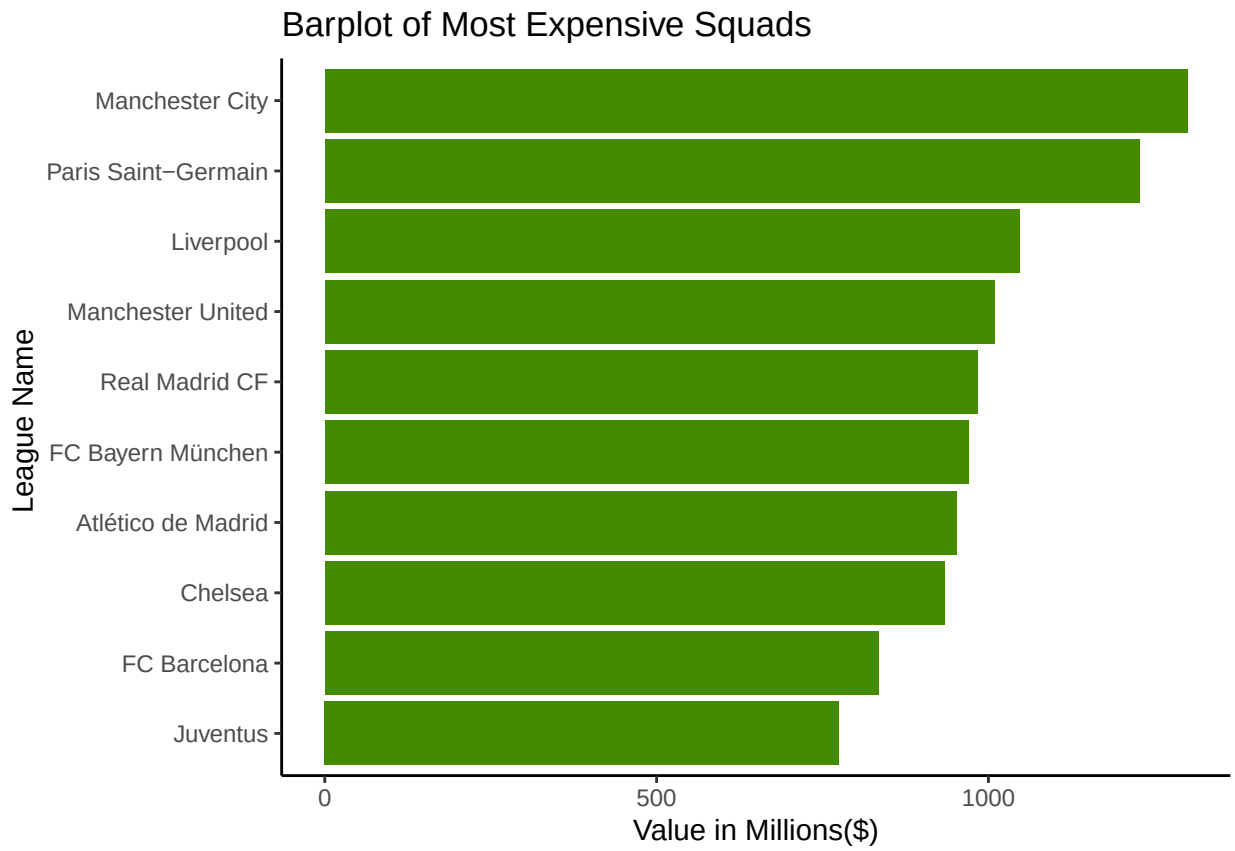
Now we want to see the top 10 most expensive squads.

```
plot_bar <- function(df) {


      df%>% mutate(club_name=as.factor(club_name))%>%
      ggplot(aes(x = fct_reorder(club_name,Sum),
                      y = Sum/1000000)) +
    geom_col(fill = "chartreuse4")+
    coord_flip()+
    labs(
      title = "Barplot of Most Expensive Squads", x = "League Name ",
      y = "Value in Millions($)",
    )+
    theme_classic()

}
```

```
data%>%group_by(club_name)%>%
  dplyr::summarise(Sum=sum(value_eur))%>%
  arrange(desc(Sum))%>%head(10)%>%plot_bar()
```



The two first clubs are petrodollar clubs (Clubes estado) while the only team which in the top 5 which is not either in the Premier League or either backed by a state (Qatar) is Real Madrid.

note: The Premier League teams profit from a better Image Right policies due to their league being the most followed one worldwide.

Where are the players of the top 100 clubs from?

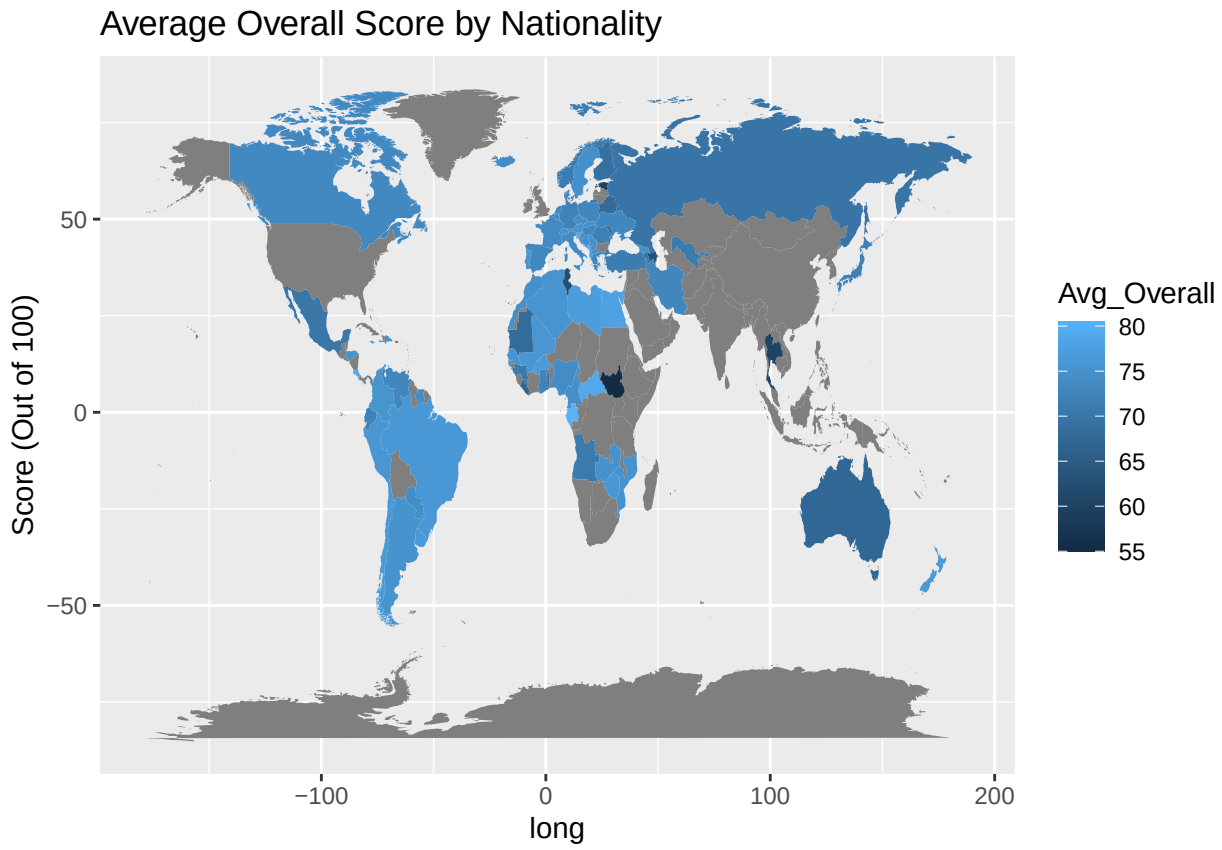
```
overall_data <- data %>%
  group_by(Nationality) %>%
  dplyr::summarise(Avg_Overall = mean(Overall))%>%
  arrange(-Avg_Overall)

worldmap = map_data("world")



merged_data <- merge(x = worldmap, y = overall_data, by.x = "region", by.y = "Nationality", all.x = TRU
```

```
ggplot(data = merged_data, aes(x = long, y = lat, group = group)) +
  geom_polygon(aes(fill = Avg_Overall)) +
   labs(
      title = 'Average Overall Score by Nationality',
      y = "Score (Out of 100)")
```

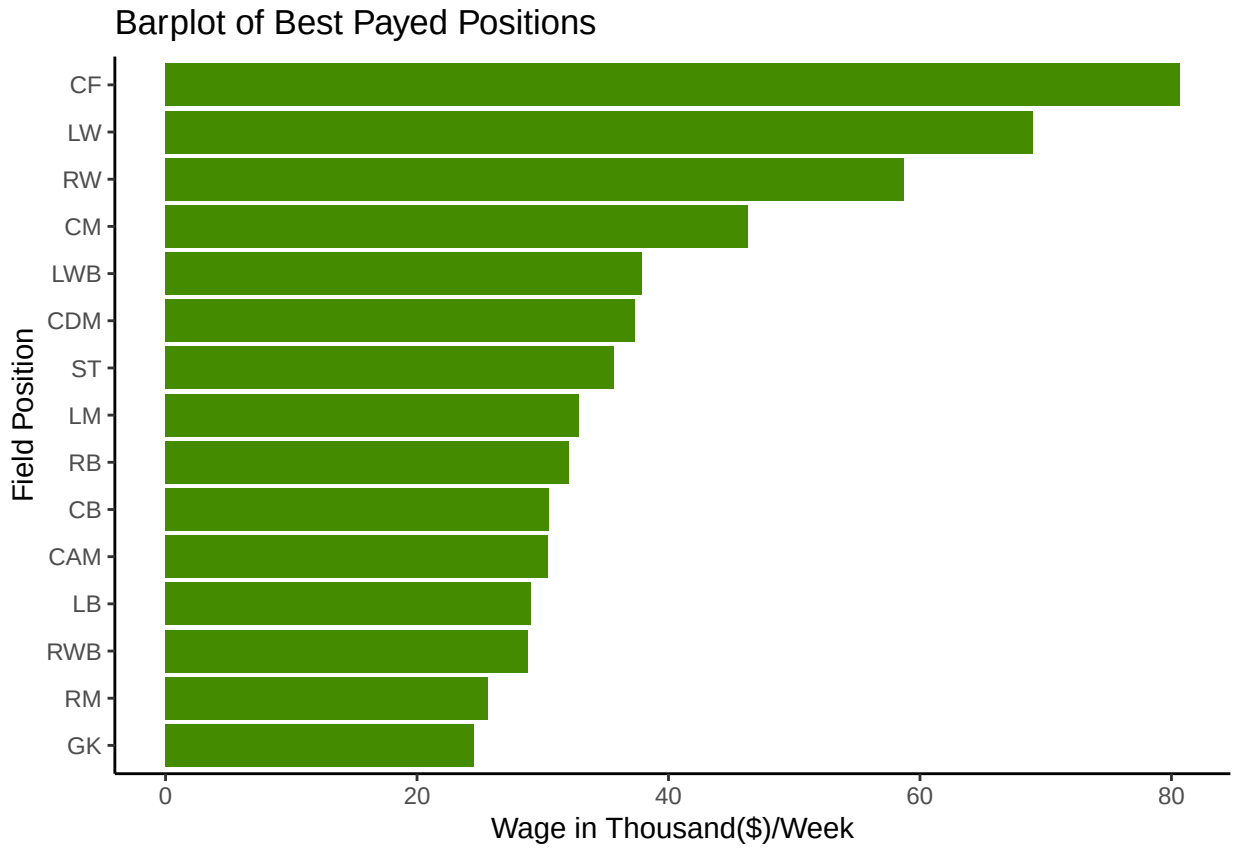


As expected we can see (and expected by the common stereotype) latin and european born players have the highest mean average. In addition it seems like the north-african born players also score high. But we should be cautious about this statement because since we have chosen only players playing in the top 100 clubs it should be noted that there aren't many African players playing in the top 100 clubs and those who do are usually pretty good players, so that leads to an increased average score.

Highest Earning Positions

```
data%>%group_by(`Best Position`)%>%
  dplyr::summarise(Avg_wage = mean(wage_eur))%>%
  arrange(-Avg_wage)%>%mutate(`Best Position`=as.factor(`Best Position`))%>%
      ggplot(aes(x = fct_reorder(`Best Position`,Avg_wage),
                 y = Avg_wage/1000)) +
    geom_col(fill = "chartreuse4")+
    coord_flip()+
    labs(
      title = "Barplot of Best Payed Positions", x = "Field Position",
```

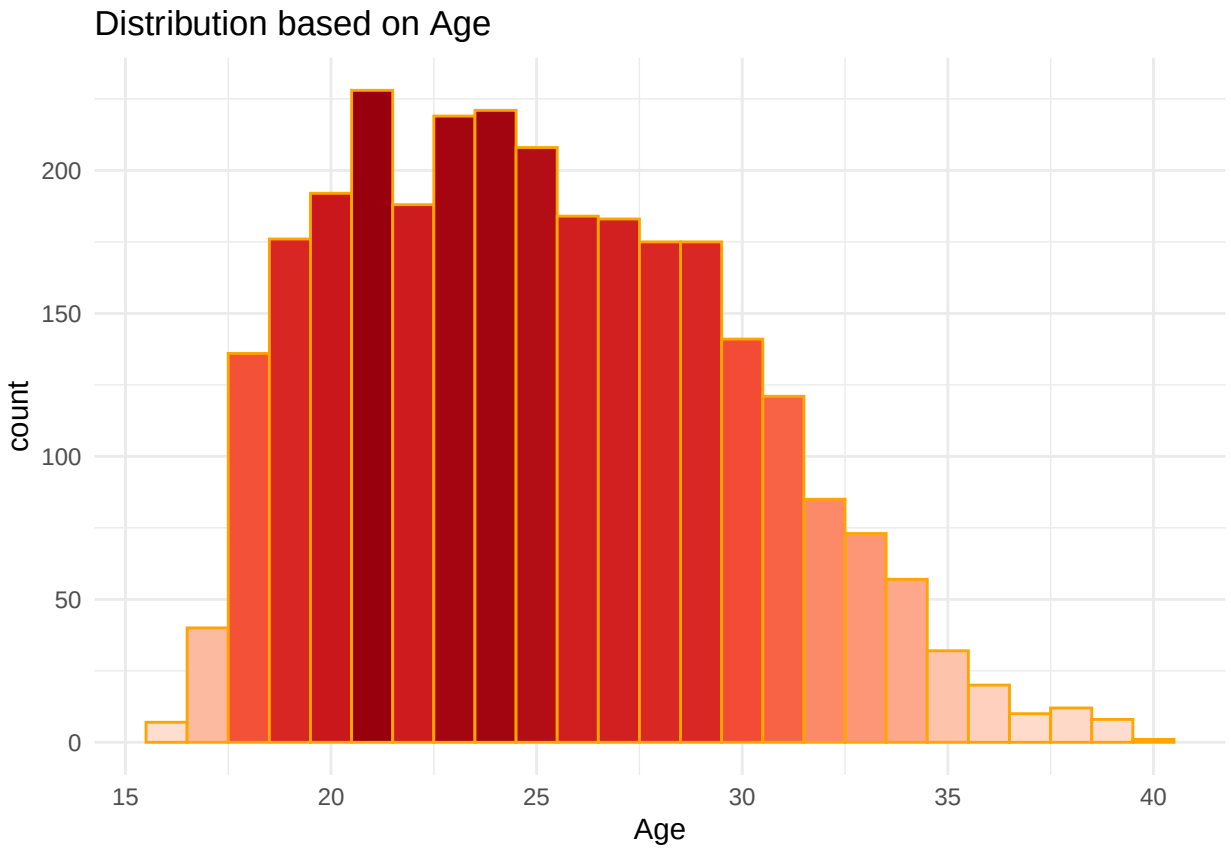
```
    y = "Wage in Thousand($)/Week",
  )+
  theme_classic()
```



CFs(Central Forward) along the LW and RW (Left Wing and Right Wing) are usually the top goal scorers and clubs tend to pay them more to keep them happy.

Also it can be seen that becoming a goalkeeper will not make you as rich as forward playing players.

```
g_age <- ggplot(data = data, aes(Age))
g_age +
  geom_histogram(binwidth = 1, col = "orange", aes(fill = ..count..)) +
  scale_fill_distiller(palette = "Reds", direction = 1) +
  ggtitle("Distribution based on Age") +
  theme_minimal() +
  theme(legend.position = 'none')
```



As expected, after 35 the number of active players decreases a lot. Curiously the Age distribution has almost a bell shaped although it is skewed more towards the young ages.