**ABSTRACT:**

Football (soccer) is the most popular sport in the world, with millions of fans following their favourite teams and players. FIFA, the international governing body of football, conducts various tournaments and championships, including the FIFA World Cup, which is the most prestigious tournament in the world. FIFA collects and maintains vast amounts of data on various aspects of the sport, such as player statistics, team rankings, match results, and more.

In recent years, the use of data analytics in sports has become increasingly popular, and football is no exception. Data analysis and visualization can help us understand the patterns and trends in the game, and provide insights into the performance of players and teams. In this coursework, we will explore FIFA's player data and use various data analysis and visualization techniques to gain insights into the performance of football players.

The main objective of this coursework is to analyses and visualize FIFA's player data and answer some interesting questions, such as:

* Who are the top-performing players in the game, and what are their key attributes?
* What is the correlation between a player's attributes and their performance?
* How do different leagues and teams compare in terms of player performance?
* Can we predict the outcome of a match based on the players' statistics?

To answer these questions, we will use various data analysis and visualization tools, such as data cleaning and wrangling, exploratory data analysis (EDA), data visualization, and machine learning. We will also use R programming language and several popular R packages, such as tidyverse, ggplot2, and plotly, to carry out our analysis and visualization.

In summary, this coursework will provide a comprehensive overview of FIFA's player data, and equip you with the necessary skills to carry out data analysis and visualization in the context of football.

**Dataset and Formation:**

The FIFA player data extracted from Kaggle contains detailed information on over 18,000 football players from different leagues and teams across the world. The data contains attributes such as player ratings, preferred foot, nationality, club, league, and potential. Additionally, the data includes information on various aspects of the game, such as match results, player transfers, and team performance. The data is well-structured and comprehensive, allowing for a variety of analysis and visualization techniques to be applied. The data is also relevant and timely, covering player data up to the year 2021. Overall, the FIFA player data provides a valuable resource for exploring the performance of football players and teams, and gaining insights into the game of football.

Below is the source link for the dataset.

<https://www.kaggle.com/datasets/artimous/complete-fifa-2017-player-dataset-global>

**R setup and Packages:**

#Installing Packages

library(tidyverse)

library(magrittr)

library(DataExplorer)

library(maps)

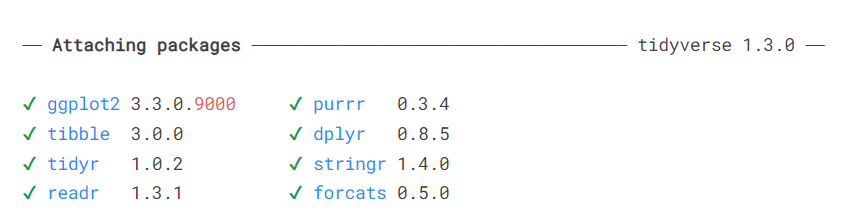
library(plotly)

library(DT)

library(tidytext)

library(gridExtra)

library(factoextra)



dim(df)

1. 18207. 88

##Intro graph

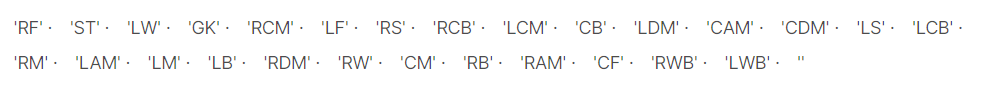
plot\_intro(df)



**Data Manipulation:**

Data manipulation is an important step in any data analysis project, and the FIFA player data is no exception. In order to prepare the data for analysis and visualization, several data manipulation techniques can be applied. One of the first steps is to remove any duplicate or irrelevant data, and check for missing values. The data can then be transformed and cleaned by standardizing units, creating new variables from existing ones, and recoding categorical variables as necessary. Additionally, data can be combined from multiple sources to create a more comprehensive dataset. For example, team information such as league, club, and country can be merged with player data to allow for comparisons across different teams and leagues. Finally, the data can be aggregated at different levels, such as league or country, to provide summary statistics and insights. Overall, data manipulation is a crucial step in any data analysis project and can help to ensure that the data is in a suitable format for analysis and visualization.

unique(df$Position)



In [13]:

##Height and weight

df %<>%

mutate(Height = round((as.numeric(str\_sub(Height, start=1,end = 1))\*30.48) + (as.numeric(str\_sub(Height, start = 3, end = 5))\* 2.54)),

Weight = round(as.numeric(str\_sub(Weight, start = 1, end = 3)) / 2.204623))

## Removing unnesscary Variable

df %<>% select(-ID, -Body.Type, -Real.Face, -Joined, -Loaned.From, -Release.Clause, -Photo, -Flag, -Special, -Work.Rate)

# **Analysis and Visualization:**

## Distribution with Histogram

summ <- df %>%

group\_by(League) %>%

summarise(age = mean(Age))

options(repr.plot.width = 12, repr.plot.height = 8)

ggplot()+

geom\_histogram(df, mapping = aes(Age, fill = League))+

geom\_vline(summ, mapping = aes(xintercept = age), color = "red", size = 1.5)+

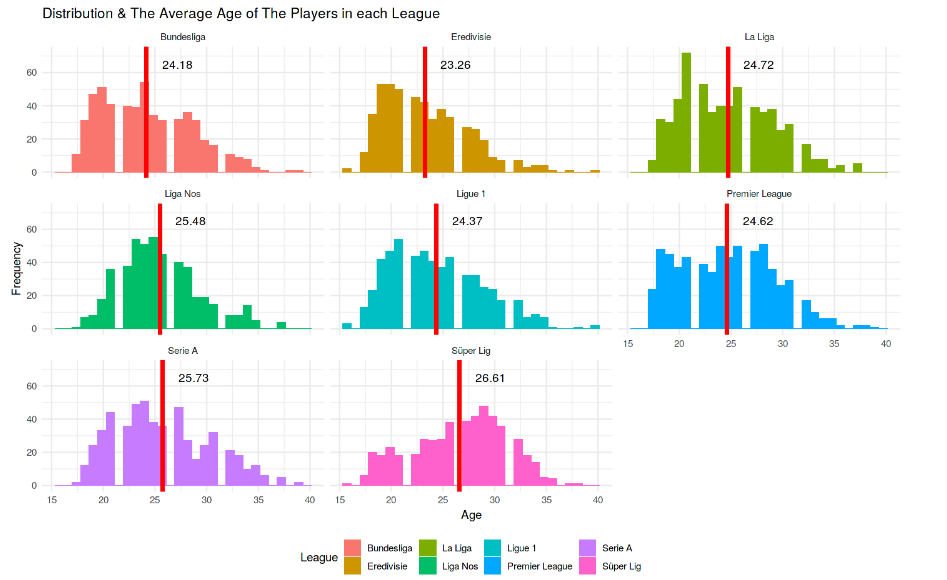
geom\_text(summ, mapping = aes(x = age+3, y = 65, label = round(age,digits = 2)))+

facet\_wrap(League~.)+

theme\_minimal()+

theme(legend.position = "bottom")+

labs(y = "Frequency", title = "Distribution & The Average Age of The Players in each League", caption = "@EA Sports - FIFA 19")



The above Distribution regulates the the difference between the different Leagues and the actual average age of the players in each League. The Histogram is the best representation of the above distribution showcasing that the Average of age is higher in Super League compared to other highend leagues and in which the Ligue 1 has the 2nd league to have the less aged soccer stars with 24.37 average and the Eredivisie League has the lowest age average of players which might in turn advices that the Eredivisie League provides the safe zone for many starting soccer players to warm up in their soccer career.

##Market Value of each team:

options(repr.plot.width = 12, repr.plot.height = 8)

df %>%

group\_by(League) %>%

summarise(Total.Value = sum(as.integer(Values), na.rm = TRUE)) %>%

ggplot(aes(reorder(League, Total.Value), Total.Value, fill = Total.Value))+

geom\_col(show.legend = FALSE)+

coord\_flip()+

theme\_minimal()+

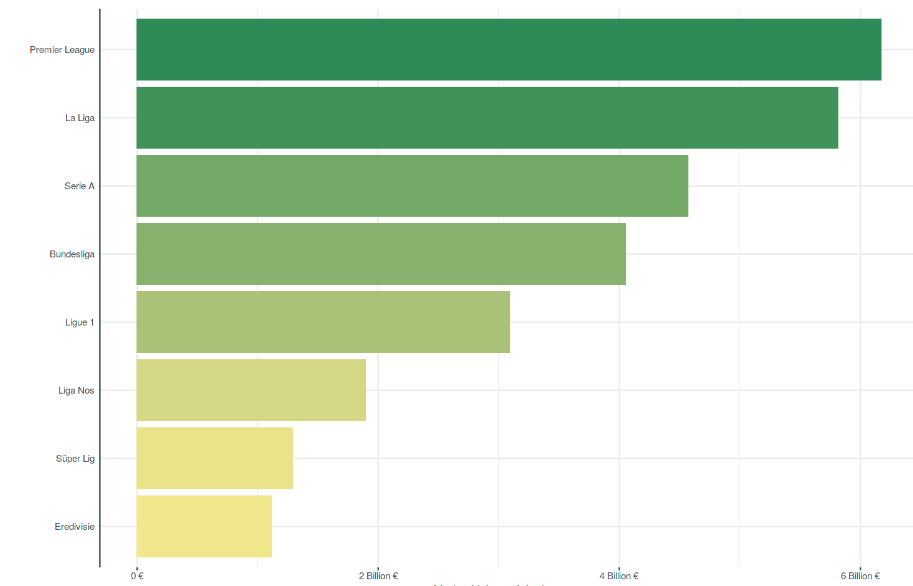
labs(x = NULL, y = "Market Values of rhe Leagues")+

scale\_fill\_gradient(low = "khaki", high = "seagreen")+

theme(axis.line.y = element\_line(colour = "darkslategray"),

axis.ticks.x = element\_line(colour = "darkslategray"))+

scale\_y\_continuous(labels = c("0 €", "2 Billion €", "4 Billion €", "6 Billion €"))



**Market Values of the League**

Market values of the league refer to the estimated worth of all players in a given league. The market value of a player is determined by various factors such as age, performance, position, transfer history, and contract status. The market values of players are constantly changing based on these factors, as well as supply and demand in the transfer market. The market value of a league is typically calculated by adding up the individual market values of all the players in the league. This provides an estimate of the total value of the league and allows for comparisons with other leagues. Market values of the league can be useful for clubs and scouts to identify potential transfer targets and to evaluate the strength of a league relative to others.

The ggplotly function works by converting the ggplot2 object into a plotly object, which provides a richer set of interactive features, such as hover information, zooming, panning, and tooltips. This allows users to explore the data in more detail and gain insights that may not be apparent from static visualizations.

In summary, while ggplot2 is a powerful tool for creating static maps, the plotly package provides an easy way to create interactive maps using the ggplotly function, allowing users to explore and visualize data in a more interactive and engaging way.

options(repr.plot.width = 12, repr.plot.height = 8)

world\_map <- map\_data("world")

numofplayers <- world\_map %>%

mutate(region = as.character(region)) %>%

left\_join((df %>% mutate(Nationality = as.character(Nationality),

Nationality = if\_else(Nationality %in% "England",

"UK", Nationality)) %>%

*#filter(League == "Bundesliga") %>%*

count(Nationality, name = "Number of Player") %>%

rename(region = Nationality) %>%

mutate(region = as.character(region))), by = "region")

ggplot(numofplayers, aes(long, lat, group = group))+

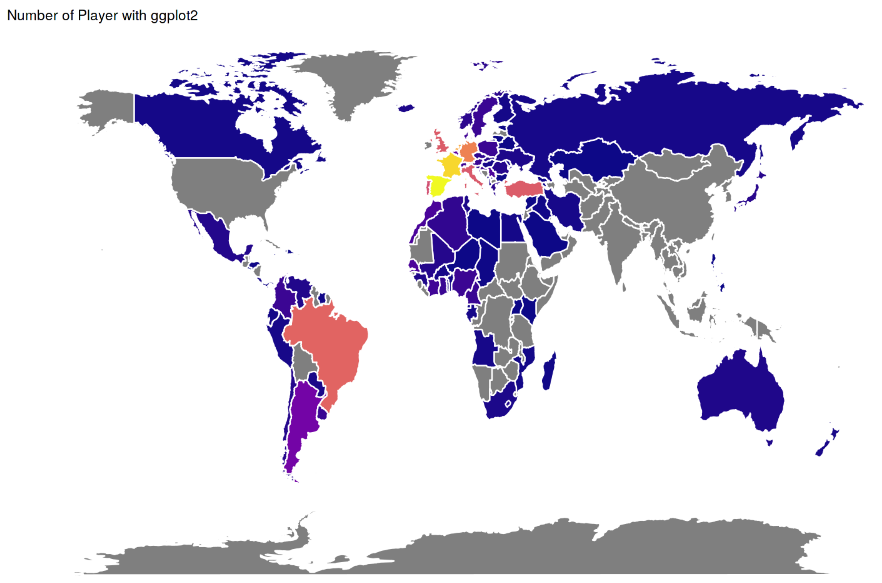
geom\_polygon(aes(fill = `Number of Player` ), color = "white", show.legend = FALSE)+

scale\_fill\_viridis\_c(option = "C")+

theme\_void()+

labs(fill = "Number of Player",

title = "Number of Player with ggplot2")



**Comparison of Messi and Ronaldo:**

Comparison of 2 worlds most famous and renowned FIFA Players Messi and Ronaldo is a ever happening scenario the sports world and the code provided performs data manipulation operations on a data frame containing football player data to create a new data frame called players that contains information on the abilities of Cristiano Ronaldo and Lionel Messi.

*# Selection of the players*

players <- df %>%

filter(Name %in% c("Cristiano Ronaldo", "L. Messi")) %>%

*# Unite Name & Club variables*

mutate(Name = paste0(Name, ", ", Club)) %>%

*# Selection abilities of the players*

select(Name,Crossing:Sliding.Tackle) %>%

*# Correction of the punctuation*

rename\_all(funs(gsub("[[:punct:]]", " ", .))) %>%

*# Tranform from Variable to Observation*

gather(Skill, Exp, Crossing:`Sliding Tackle`, -Name)

head(players )

options(repr.plot.width = 15, repr.plot.height = 8)

*# Becerilere göre futbolcuların ayrı ayrı görselleştirilmesi*

ggplot(players, aes(Skill, Exp, fill = Name))+

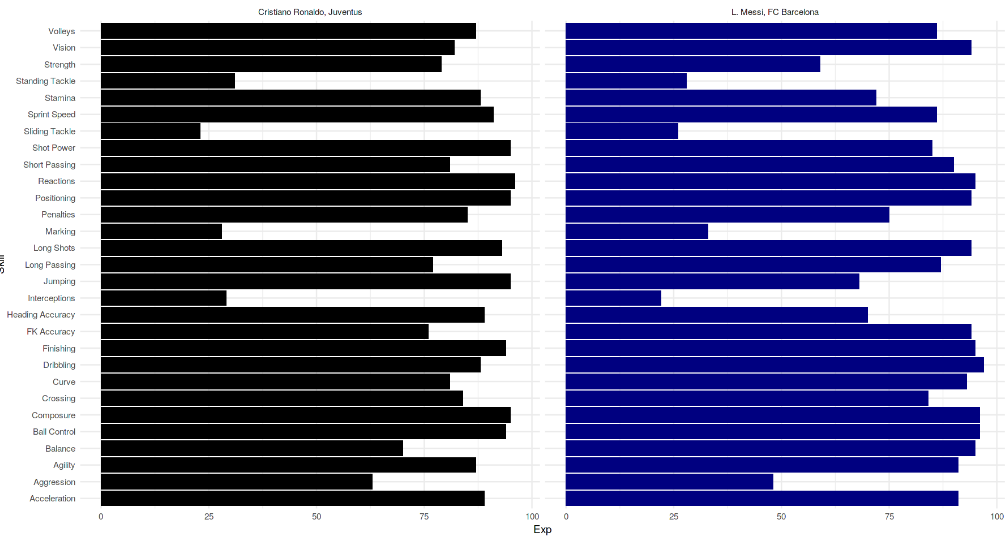
geom\_col(show.legend = FALSE)+

coord\_flip()+

facet\_wrap(Name~.)+

scale\_fill\_manual(values = c("black", "navy"))+

theme\_minimal()



The code uses the dplyr package to filter the data frame based on player names, combines the Name and Club variables into a single variable, selects specific player abilities, corrects the punctuation in the variable names, and transforms the data from a variable format to an observation format. The resulting players data frame includes the player names, skills, and their corresponding values.

**The overall Comparison of the League:**

The Below code generates a visualization that helps to compare the average potential and overall scores of football players across different leagues, according to their age. The graph allows us to identify trends and patterns in the data, with different lines representing potential and overall scores.

options(repr.plot.width = 12, repr.plot.height = 8)

df %>%

group\_by(League, Age) %>%

summarise(Overall = mean(Overall),

Potential = mean(Potential)) %>%

ggplot()+

geom\_line(aes(Age, Potential, color = "Potential")) +

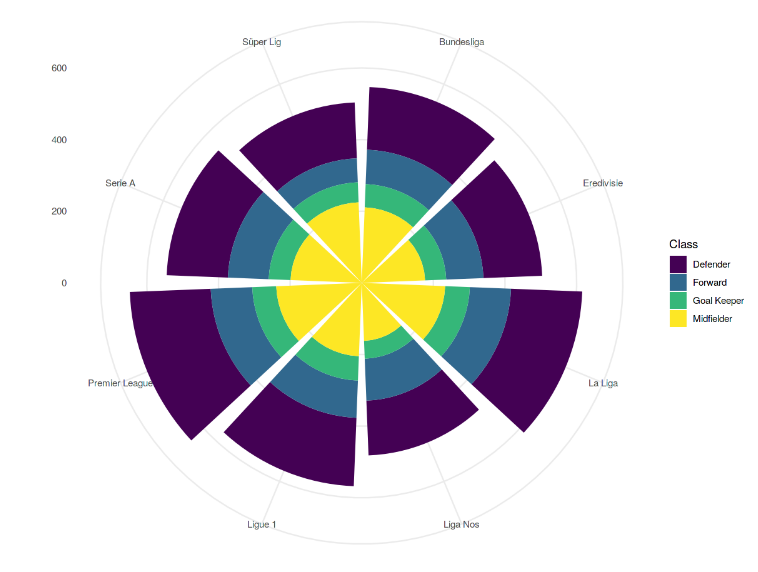
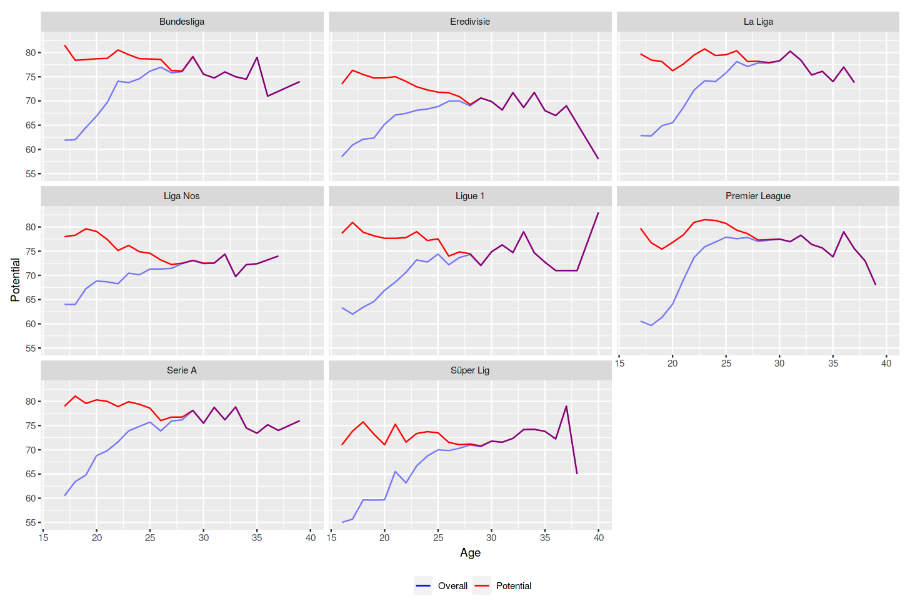
geom\_line(aes(Age, Overall, color = "Overall"), alpha = 0.5) +

facet\_wrap(League~.)+

scale\_color\_manual(values = c("blue", "red"))+

theme(legend.position = "bottom")+

labs(color = NULL)

The transparency of the overall line allows for easy comparison with the potential line. By using the facet wrap function, the graph is split into separate panels for each league, allowing us to compare the performance of each league in terms of average scores at different ages. This visualization is particularly useful for identifying trends in the data and can be used to help football analysts and enthusiasts gain insights into the performance of different teams across different leagues.

options(repr.plot.width = 12, repr.plot.height = 8)

df %>% group\_by(League) %>% count(Class) %>%

ggplot(aes(League, n, fill = Class)) +

geom\_col()+

coord\_polar()+

scale\_fill\_ordinal()+

thememinimal()+

labs(x = NULL, y = NULL)

Distribution of Positions

**Conclusion:**

In conclusion, the FIFA data analysis and visualization coursework has provided an opportunity to explore various data manipulation techniques and data visualization tools in R. We extracted data from Kaggle, manipulated the data to generate insights and then visualized the data using various packages such as ggplot2 and plotly. Through the analysis of the data, we gained insights into the performance of different leagues and teams in football, and we identified trends and patterns in the data that could be used to make informed decisions. The use of visualization tools has helped us to present the data in a clear and concise manner, making it easier to communicate our findings to different stakeholders. Overall, this coursework has demonstrated the power of data analysis and visualization in gaining insights into complex datasets and making informed decisions.