1 Introduction

The purpose of this project was to analyze data to compare FIFA statistics among players across the years 2015 to 2021 and to learn how to use R programming language to clean and analyse data;FIFA is an online soccer game. We pulled a set of data frames from an online data repository, Kaggle. Each data frame corresponds to a yearly FIFA edition, namely editions '15-'21. This means we can observe how players' in-game abilities evolved over time which aids in scouting. Each observation corresponds to its real life soccer player. Every single observation has 106 variables, including player height, weight, date of birth, and other basic information. In addition, there are various player stats that affect the in-game ability of the player (approximately column numbers 34-80). Our group chose this data because we are all avid e-gamers. All of us either watch profession soccer, play soccer, play FIFA onine, or all of the above. We were interested to see the leagues' talent distribution, the total market value of each league, whether a players' preferred foot affects salary, correlation between value and overall score, the probability that a player played for a different country, and analyze specific players.

1. What is the distribution of player ratings by position, club, league, and overall in FIFA 2021?
2. What is the distribution of player value by position, club, league and overall?
3. Which position in FIFA 2021 pays the highest?
4. Does a player's preferred foot affect salary?
5. Do players of a given nationality play within their nation of origin more frequently than migrating to a different country to play football?

2 About Cleaning Data

The kaggle dataset, as we found, includes data from FIFA '15 until FIFA '21. All of them were .csv files, and they were all loaded in R Studio using read.csv(). Each dataset has a different naming convention so cleaning must be done individually.

First, since each dataset contains 106 variables, it was necessary to clean the data by removing the excess columns, columns we decided do not answer our questions, such as sofifa_id, real_face, player_tags, body_type. The dataset contained unwanted ratings for positions that a player has never played in. The deletion part was carried out using the select() function from dplyr library. Secondly, we rename the variable names to be more clear and precise. Columns with long names and the ones with special characters were renamed to shorter versions for user friendliness. Example: "attacking_heading_accuracy" was renamed to just "heading". The process was carried out using the rename()

function from dplyr library. To make interactive plots for the analysis, we used functions like filter() to create subsets of specific rows. Apart from filter(), we also used the data.frame() function to generate required data frames. To extract league wise data, we relied upon the subset() function to copy over each observation to work on our analysis.

Soccer is a global sport and it's mostly played outside the United States, categories like height and weight are listed using the international units. However, some of these were scaled to conversion factors, fifa$weight_kg - fifa$weight_kg * (2.2). Similarly, the player values and wages used euros as their currency, and conversion process could be used to present them in US Dollar valuation. To find a specific player or league, which() was used to find the required indices and manipulate the data from there.

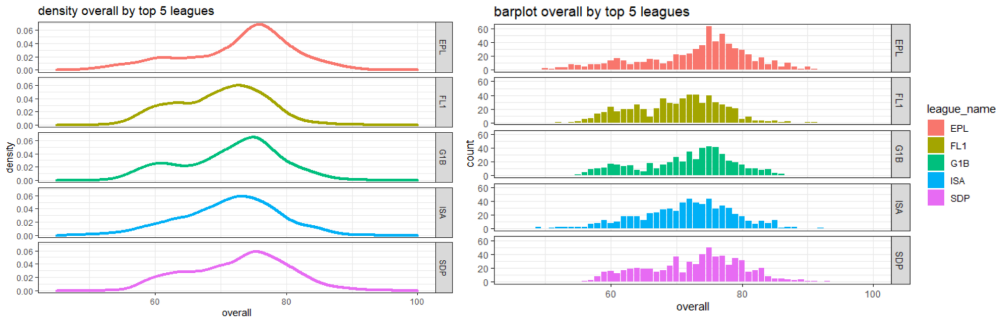
Additionally, for the player evolution graph, we had to use a player's data from the five different FIFA datasets. Since they all contained his name [K. Mbappe], renaming it year wise [K. Mbappe_year] was necessary to avoid confusion so that a subset of each row could be better for readability.

3 Overall Analysis

3.1 Density of Top Five Leagues

The following only pertains to the FIFA 2021 dataset. This plot depicts the density of top 5 leagues in FIFA 2021, namely Spain Primera Division (SDP), English Premier League (EPL), Italian Serie A (ISA), German 1. Bundesliga ($G1B$), and French Ligue 1 ($FL1$). Since these plots are probability density functions, they are non-negative and integrate to be 1. The maximum overall score of 93 belongs to Lionel Messi in Spain Primera Division. The minimum score within these 5 leagues is 49, and two players have this score. The overall score that occurs most often is 75. 238 players have an overall score of 75.

There is a trend of highest density about the median and a smaller high density of players slightly below median. We will call them "big bump" for the bump near the median and "small bump" for the smaller bump. These can be read off from the graphs by observing the two bumps in all graphs but ISA. It may be reasonable to assume that the league with the highest big bump. With this assumption, it appears the English Premier League could be the best league. The same plot is given in discrete form with geom_bar since the variable overall is integer, and could have given more information like frequency instead of probability. The density plot and bar plot look generally the same. Both bumps are observable in the bar plots as well.

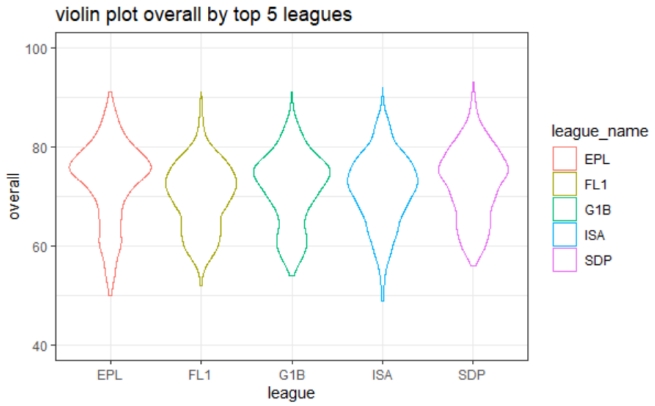


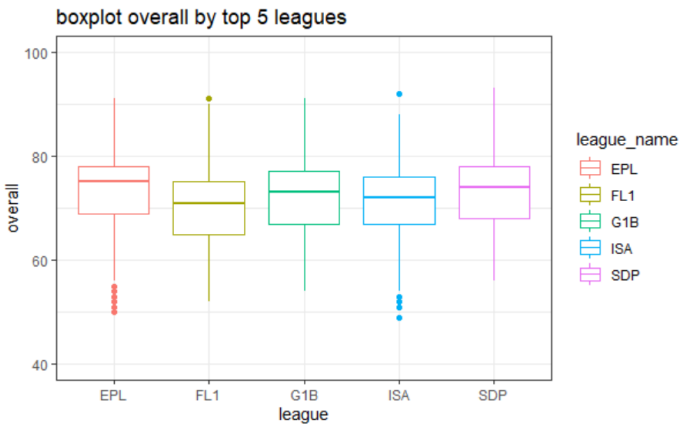
3.2 Plots of Top Five Leagues

Included on slide two, are three graphs of overall score by league. There are about 600 players in each league. In addition, the range of overall score in these leagues are the integers within [49, 93], making only 15 overall scores. The amount of players and the small number of overall scores is why the jitter plot does not work very well. There are too many data points close together to tell anything besides max and min. On the other hand, boxplot and violin plot are good depictions of the data. In both these plots, we can see the spread of data better. With boxplot, we can see outliers and median better than violin. With

the violin plot, we can see the two bumps present in the density graphs, except Italy.

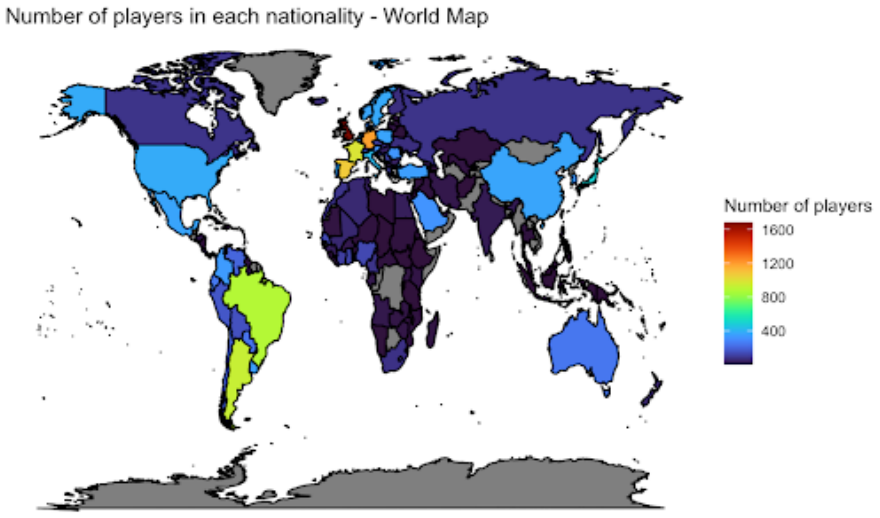
Outliers near the top of the box plot could indicate that there are fewer high scoring players in the league, while outliers near the bottom indicate that there are fewer low scoring players in the league. Looking at the boxplots, we can see that EPL's median is higher than or almost as high as the third quartile of ISA and FL1, that is, half of EPL's players have higher overall scores than FL1 and ISA. In addition, EPL has many outliers below the lower extreme. This suggests that EPL has fewer lower scoring players. Thus, EPL may be the better league. On the other hand, SDP has the highest scoring player, Messi, and the second highest median. This indicates they may be the second best league.





3.3 World Mapping of Player Nationality

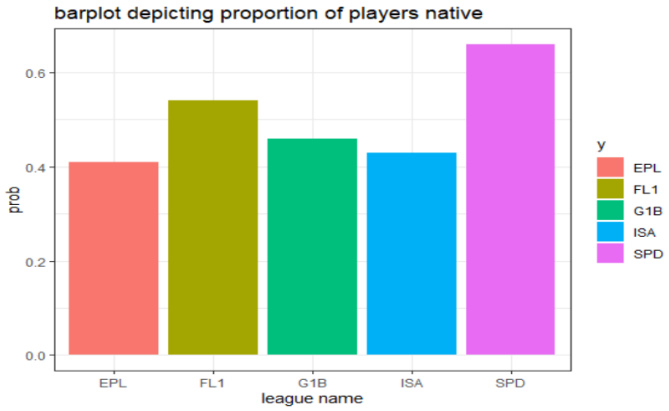
In the world map, we can notice that Asian countries have a low percentage of soccer players while European and South American countries produce the most talent. North America is slowly but steadily improving the number of players [378] produced and making a name for them in the world stage. England, despite being a smaller nation whose national sport is cricket, produced the most players [1685]. Germany in 2017 had a total of 689 players all over the world, and in 2021, Germany has the second highest players in world football with 1,189 players. The number of players in each country equates to how well they have performed in the FIFA world cups with Brazil winning five times, and Germany and Italy tied next with four each.



3.4 Probability of Player Nationality

This bar plot shows the proportion of players who are native to the country they are playing in. What is the probability that a player is native to the country they are playing in given they play in Spain's league? To determine this, count players in SDP that are native to Spain. This turns out to be 425. Take 425 out of the total players in five leagues, 3092. Then, divide this probability by the probability that the player is from Spain. Thus,

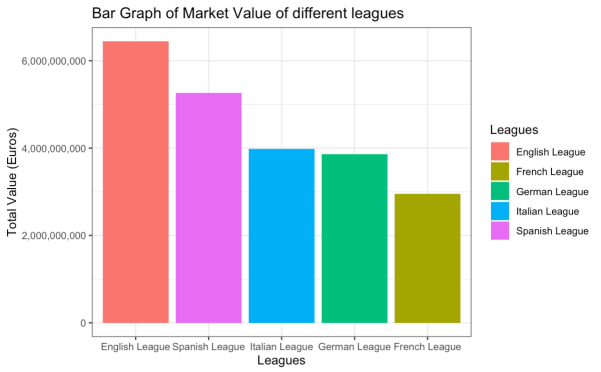
$$\frac{\frac{425}{3092}}{\frac{645}{3092}} = \frac{425}{645} = 0.659$$

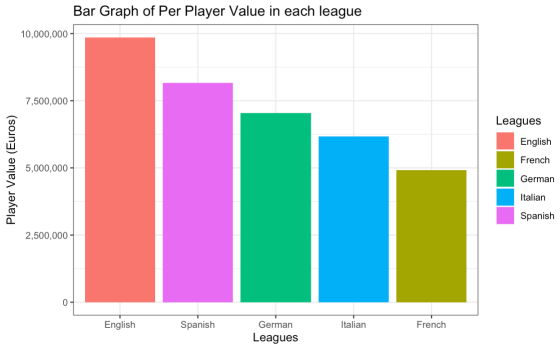


4 Financial Market

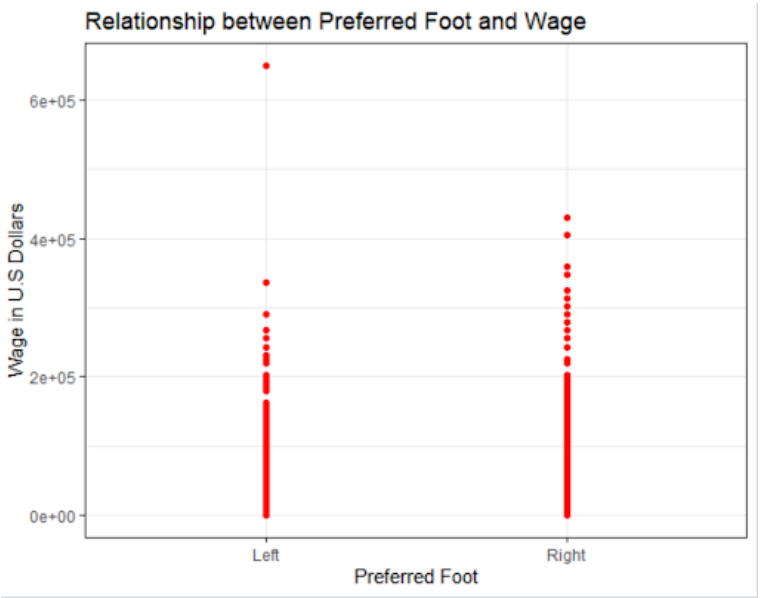
4.1 Total Market Value of Leagues, Per Player Value

The plot shows the total player market value of the top five leagues in descending order. We can clearly see English Premier League leading with over 6.4 billion euros. This plot hints that the English league has more high quality players, and the league is very competitive. That could be partially true in one way, however not accurate. To take our analysis a bit deeper, with more research, we found out that the number of players in each league is not the same despite each league having the same number of teams. English Premier League has the highest number of players [654], followed by Italian Serie A and Spain Primera Division leagues in the next spot [645], and German 1. Bundesliga being the lowest [548]. Most logical approach was to use the total market value and the number of players to calculate "per player value" for each league and then make a new bar plot representing this data. The new bar plot gave us a different analysis and showed why a one-dimensional analysis alone shouldn't be taken as conclusive evidence. From the total market value plot, one would conclude that Italian Serie A is a highly valued league with better quality players, however, the per player value shows us German 1. Bundesliga is in fact the league with higher average value per player. Other spots were similar to the total market values graph.





4.2 Preferred Foot Salary



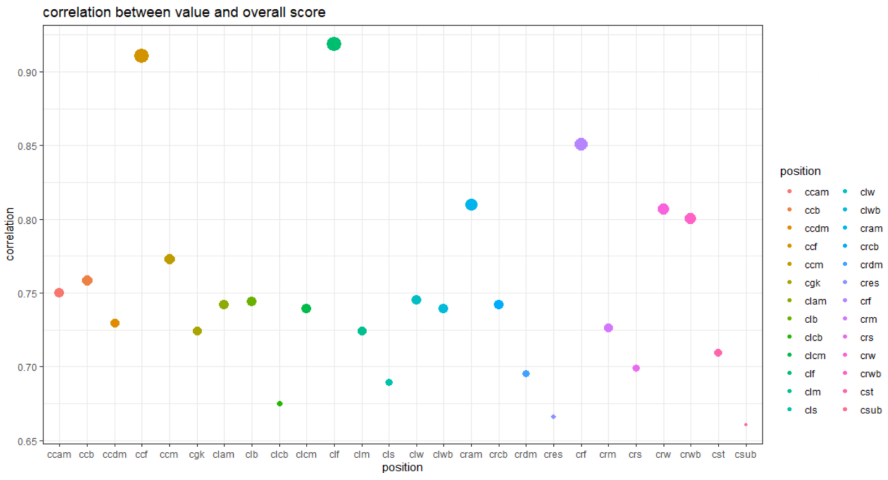
This figure exhibits a scatter plot portraying the relationship between preferred foot and salary in U.S dollars. It reveals that, on average, players who prefer using their right foot tend to earn more than those who prefer to use their left foot. This can be shown by the higher cluster of points on the right side of the graph. This was interesting in the sense that it showed that the foot dominance of a player actually had a noticeable impact on salary. However, what was more fascinating about the plot was that it showed that the highest paid FIFA player of all time preferred his left foot, which goes against the general trend. Given such a high outlier, if one were to compute the average of the salaries of left-footed players and compare it to that of right-footed players, one

would notice that the averages are a lot closer than they should be. Given this information, using a scatter plot that separates each point for visualization is beneficial when trying to answer the question of whether foot preference affects salary on a general level.

4.3 Correlation Between Overall Score, Value

This scatter plot graphs the individual player positions on the x-axis and the correlation between overall score and the market value of a player on the y-axis. The market value of a player is how much money a team is willing to sell a specific player for. There are twenty-six team positions in FIFA. We used abbreviations, and each position in the plot starts with a "c". This "c" means correlation. So, each point represents the correlation between their market value and overall score at a certain team position. Example, "ccam" is Pearson correlation of center attack midfielder, and this correlation is 0.75. The range of the correlations is [0.66, 0.92]. The idea was to determine which positions pay the highest for good players.

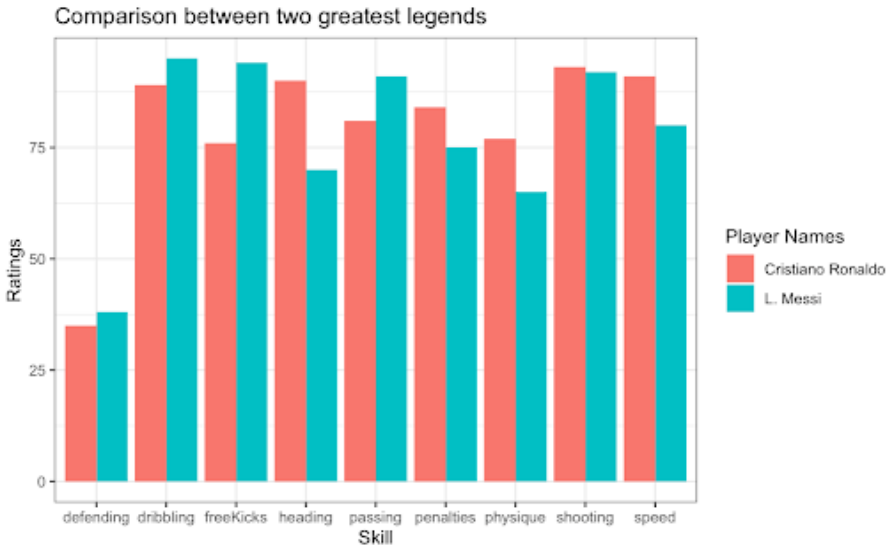
We think that high correlation corresponds to a high overall score player with a high market value. For example, the left center back, LCB position correlation is 0.67, which is lower compared to the other positions. While the left forward, LF, position has a correlation 0.92, which is higher compared to the other positions. We concluded that a good left forward, LF, player is worth more than a good left center back, LCB. The market value of a player does not solely depend on overall score, though obviously good players are worth more. But, as we can see from this graph, team position could affect market value. So, if a team wants to add a center forward, CF, it might have to pay more than if it were to add a left center back, regardless of the players' overall score.



5 Individual Player Analysis

5.1 Greatest of All Time

This figure represents the comparison between two of the greatest athletes, and their fierce rivalry in the year 2021. Cristiano Ronaldo, age 36, still performs at the top and maintains his physical condition like he is 20. On the other hand, Lionel Messi, age 34, is a pure magician and naturally talented . We used the bar plots function from ggplot to differentiate between the players.To get both bar plots simultaneously in one picture, we set parameter position = "dodge". Used the ratings in the y-axis and respective key skills in the x-axis to plot the graph. When comparing, we found out that it was a close contest, as expected. It is evident Messi is better with the ball attributes like dribbling, passing, etc. Ronaldo is more of an athletic person who specializes in off-ball attributes such as speed, physique, and leadership. Ronaldo is a highly determined player, and his work rate [High Work Rate] clearly proves that he works hard for his success. Messi, despite being shorter [170 cm] than most soccer players, he put away his weakness and focused solely on his natural abilities to become the great player he is today.

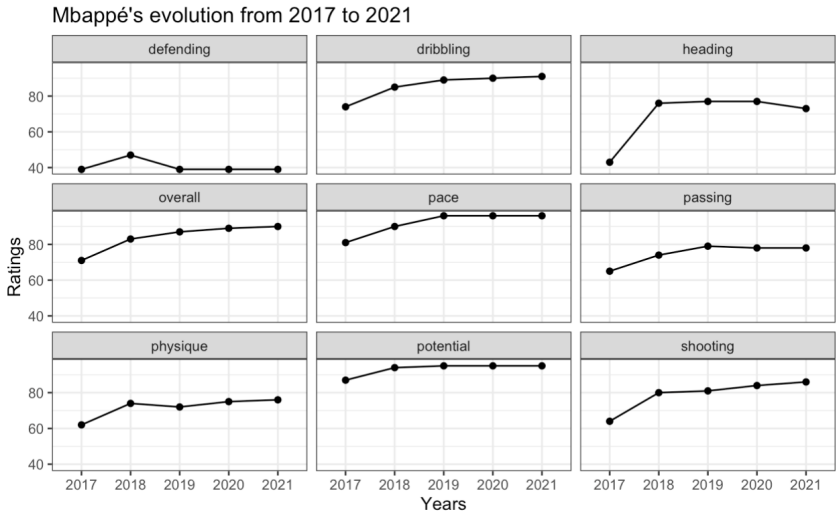


5.2 Trend of Player Ratings

Mbappe is the cover star of the FIFA '21 version. We tracked his growth over the past five seasons with key attribute ratings taken into account. Since he is an offensive star, skills like pace, dribbling, heading, define his value, so we specifically picked out these columns from five different seasons and made

a new dataset. Adding to that, we wanted to show his overall ratings and potential predicted by fifa for each year. To represent this visually, facet_wrap along with line and point plots were used. Years is the x-axis, ratings is the y-axis, and each skill takes over different subplots. Potential ratings for Mbappe have clearly been high since the FIFA '17 edition, showing how much he has impressed the pundits at such a young age [17]. His monstrous growth took place during the transition from FIFA '17 to FIFA '18, where every attribute has a linear growth. A youth product, who was just a fast dribbler, is able to head, dribble, and pass better. This growth was instrumental in defining his market value which had a monumental increase from 3.1 million euros in 2017 to a mighty 105 million euros now.

We do not know for sure the exact reasoning for this drastic growth. To effectively identify the possible reasons, we could potentially track the players who played alongside him during the 2017 and compare them with the ones in the 2018 season. Competitive players tend to push others to a whole new level, or it could also be a result of his hard work and professionalism. Mbappe is just 23 years old now, not even in his peak, is able to compete with most of the top rated footballers with his mind blowing ratings. If he manages to stay fit and improve at the same rate, within a few years, he could be the best player in the world.



6 Conclusion

Through the data frames we obtained from Kaggle, our group was able to explore the leagues' talent distribution, correlation between value and overall

score, probability that a player played for a different country, and an analysis of a specific player's skills over time.

Our group was able to explore the different factors that affected the value of the players. By making a scatter plot comparing the correlation of market value to individual player positions, we saw that the overall score of a player did not have as much of an impact on a player's market value as we thought. Our initial way of thinking was that if a player had a higher overall score, regardless of any other factors, he would have a higher value. However, in the case of the scatter plot we made, it was concluded that the position played itself affected a player's market value to some degree. Another factor we took in the same regards was foot preference. We noticed at first glance from a scatter plot that we made comparing foot preference and salary that right footed players seemed to earn more on average. However, after using R to calculate the exact average salary of left footed players and right footed players, we noticed that there wasn't really much of a difference. We did not expect much of a significant difference regarding salaries of left footed and right footed players since the factor seemed rather random, and the data we found supported our way of thinking. What piqued our interest in the scatter plot was the fact that we saw a significant outlier that displayed the fact that the highest paid FIFA player of all time preferred his left foot. Upon further research for our project, we were able to identify who that player actually was - Messi - which was cool to see how different parts of the project could come together.

Using the data we obtained, our group wanted to see the probability that a player in a league was playing in his native country. From a bar graph that we made we were able to explore this. We concluded that the majority of players in Spain's League were actually native to Spain. As for the other leagues, a big portion of their players were also native to their league's country. This opened our eyes in the sense that we realized that being born and raised in a specific country could be a factor that leagues look at when picking players. This was interesting to see how an unchangeable factor (where someone was born and raised) was one of the potential determining factors when being chosen for a league.

With such a large dataset containing data from years 2015 to 2021, we were able to track a single player's evolution (Mbappe-cover star of FIFA '21) and his journey from a youth product to a global icon. With this information, we made great discoveries. We saw that he made the most improvements in the year 2018. We were able to see how Mbappe grew as a player through the plots we made, which was very cool to see.

There were some things that we wanted to explore but did not have the time. These include body mass index analysis, using additional data, player attributes and grouping similar players, and creating fantasy teams. We wanted to see how BMI could affect a player's performance, market value, and in-game attributes like speed, balance and power. We wanted to compare player attributes so that we could group similar players together. Teams could use this information to

buy players similar to other players if their first picks are already chosen. For instance, we could group together young players and determine the up-and-coming talent. We could use additional data from other sources to analyze how a player's real-life performance over the season (goals/assists) affects in-game abilities, market value each season.

7 Works Cited

- "FIFA 21 Complete Player Dataset." Kaggle.com, www.kaggle.com/stefanoleone992/fifa-21-complete-player-dataset.
- "ESPN." ESPN.com, 2019, www.espn.com/soccer/.
- "A Grammar of Data Manipulation." Tidyverse.org, 2019, dplyr.tidyverse.org/.
- "Soccer World Cup Titles by Country 1930-2018." Statista, www.statista.com/statistics/266464/number-of-world-cup-titles-won-by-country-since-1930/.

8 Code

CODE:

```
# deleting unnecessary columns that are of no use to the analysis
fifa_21 <- fifa_21 %<>%
  select(-sofifa_id, -body_type, -real_face, -player_tags, -ls, -st, -rs, -lw, -lf, -cf, -rf,
         -rw, -lam, -cam, -ram, -lm, -lcm, -cm,
         -rcm, -rm, -lwb, -ldm, -cdm, -rdm, -rwb, -lb, -lcb, -cb, -rcb, -rb)
```

Overall Analysis:

Density overall and Bar plot overall:

```
fifa <- read.csv("players_21.csv")

# getting the indices of each league
spanish <- which(fifa$league_name == "Spain Primera Division")
english <- which(fifa$league_name == "English Premier League")
italian <- which(fifa$league_name == "Italian Serie A")
german <- which(fifa$league_name == "German 1. Bundesliga")
french <- which(fifa$league_name == "French Ligue 1")

# creating a dataframe
df <- fifa[c(spanish, english, italian, german, french), ]
spanish <- which(df$league_name == "Spain Primera Division")
english <- which(df$league_name == "English Premier League")
italian <- which(df$league_name == "Italian Serie A")
german <- which(df$league_name == "German 1. Bundesliga")
french <- which(df$league_name == "French Ligue 1")

#renaming
df$league_name[spanish] <- "SDP"
df$league_name[english] <- "EPL"
df$league_name[italian] <- "ISA"
df$league_name[german] <- "G1B"
df$league_name[french] <- "FL1"

        ggplot(df) +
          geom_density(mapping = aes(x = overall, color = league_name), size = 1.5) +
          xlim(45, 100) +
          labs(title = "density overall by top 5 leagues") +
```

```
  facet_grid(rows = vars(league_name)) +
  theme_bw()

ggplot(df) +
  geom_bar(mapping = aes(x = overall, fill = league_name)) +
  xlim(45, 100) +
  labs(title = "barplot overall by top 5 leagues") +
  facet_grid(rows = vars(league_name)) +
  theme_bw()
```

Jitter plot overall, violin plot overall, box plot overall

```
# box plot
ggplot(data = df) +
  geom_boxplot(mapping = aes(x = league_name, y = overall, color = league_name))
+
  labs(title = "boxplot overall by top 5 leagues", x = "league", y = "overall") +
  ylim(40, 100) +
  theme_bw()

# jitter plot
ggplot(df) +
  geom_jitter(mapping = aes(x = league_name, y = overall, color = league_name)) +
  labs(title = "jitterplot overall by top 5 leagues", x = "league", y = "overall") +
  ylim(40, 100) +
  theme_bw()

# violin plot
ggplot(df) +
  geom_violin(mapping = aes(x = league_name, y = overall, color = league)) +
  labs(title = "violin plot overall by top 5 leagues", x = "league", y = "overall") +
  ylim(40,100) +
  theme_bw()
```

World map analysis:

```
# loading the required libraries to get world map
library(ggplot2)
library(maps)
library(mapdata)
```

```
library(dplyr)
library(mapproj)

#installing viridis to get a color scheme for maps
install.packages("viridis")
library(viridis)

# duplicate data
fifa_21_mapData <- fifa_21

# renaming country names here since maps doesn't recognize them
fifa_21_mapData$nationality[fifa_21_mapData$nationality == "United States"] = "USA"
fifa_21_mapData$nationality[fifa_21_mapData$nationality == "England"] = "UK"
fifa_21_mapData$nationality[fifa_21_mapData$nationality == "China PR"] = "China"
fifa_21_mapData$nationality[fifa_21_mapData$nationality == "Korea Republic"] = "South
Korea"

# creating a variable to load world map and pick them up by nationality
map_world <- map_data("world")
names(map_world)[names(map_world) == 'region'] = 'nationality'

# variable to hold the world map and assign the fifa_21 nationality by counting them
nationality_map <- map_world %>%
  left_join((fifa_21_mapData %>%
             dplyr::count(nationality)), by = "nationality")

# using ggplot and geom_polygon, we plot the nationality numbers into plot
map <- ggplot(nationality_map) +
  geom_polygon(aes(long,lat, group = group, fill = n), color = "black", show.legend = TRUE) +
    ggtitle("Number of players in each nationality - World Map") +
    labs(fill = "Number of players")

# adding color scheme
map + scale_fill_viridis(option = "turbo") +
  theme_void()
```

Migration of players graph:

```
# sorting the data by nationality
sort(table(df$nationality[english]), decreasing = TRUE)
sort(table(df$nationality[spanish]), decreasing = TRUE)
sort(table(df$nationality[french]), decreasing = TRUE)
```

```
sort(table(df$nationality[italian]), decreasing = TRUE)
sort(table(df$nationality[german]), decreasing = TRUE)


x <- c(271/654, 425/645, 325/600, 277/645, 253/548)
y <- c("EPL", "SPD", "FL1", "ISA", "G1B")
z <- data.frame(x, y)
z$x <- round(z$x, 2)
barplot(x, ylim = c(0, 1), main = "prob country of origin", col = c("darkblue", "gold",
"white", "forestgreen", "black"))

# plotting using geom_col
ggplot(z) +
        geom_col(mapping = aes(x = y, y = x, fill = y)) +
        theme_bw() +
       labs(x = "league name", y = "prob") +
        ggtitle("barplot depicting proportion of players native")
```

Salary / Market Value analysis:

Total market value:

```
# loading the required libaries
library(scales)
library(ggplot2)

# to split the commas
options(scipen = 100, big.mark = ",")

# reading in the data csv file to a fifa_21 variable
fifa_21 <- read.csv("players_21.csv")

# getting the indices of the rows that fall under each league category
spanish <- which(fifa_21$league_name == "Spain Primera Division")
english <- which(fifa_21$league_name == "English Premier League")
italian <- which(fifa_21$league_name == "Italian Serie A")
german <- which(fifa_21$league_name == "German 1. Bundesliga")
french <- which(fifa_21$league_name == "French Ligue 1")

# using sum function to add the market value of each league
```

```
sum_spanish <- sum(fifa_21$marketValue[spanish])
sum_english <- sum(fifa_21$marketValue[english])
sum_italian <- sum(fifa_21$marketValue[italian])
sum_german <- sum(fifa_21$marketValue[german])
sum_french <- sum(fifa_21$marketValue[french])

# creating a new column with the sum values
scientific_notation <- c(sum_spanish, sum_english, sum_italian, sum_german, sum_french)


# using data.frame function to create a new dataset with league and sum values
bargraph_marketvalue <- data.frame(name=c("Spanish League", "English League", "Italian
League",  "German League", "French League"),
                                   value = c(sum_spanish, sum_english, sum_italian, sum_german,
sum_french))

# using ggplot to plot a bar graph
legend <- ggplot(bargraph_marketvalue) +
  geom_bar(mapping = aes(x = reorder(name, -value), y=value, fill = name), stat = "identity")
+
  ggtitle("Bar Graph of Market Value of different leagues") +
  labs(x = "Leagues", y = "Total Value (Euros)")  +
    theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(labels = scales::comma)  +
  theme_bw()

legend + labs(fill = "Leagues")



# calculating per-player value since the number of players are not the same in every league
spanish_per_player <- sum_spanish / length(spanish)
english_per_player <- sum_english / length(english)
italian_per_player <- sum_italian / length(italian)
german_per_player <- sum_german / length(german)
french_per_player <- sum_french / length(french)

# creating a new dataset for the new values with leagues
bargraph_per_player <- data.frame(name_per_player = c("Spanish", "English", "Italian",
"German", "French"),
                                  value_per_player = c(spanish_per_player, english_per_player,
italian_per_player, german_per_player, french_per_player))
```

```
# plotting a new barplot for the average value graph
legend <- ggplot(bargraph_per_player) +
  geom_bar(mapping = aes(x = reorder(name_per_player, -value_per_player),
y=value_per_player, fill = name_per_player), stat = "identity") +
  ggtitle("Bar Graph of Per Player Value in each league") +
  labs(x = "Leagues", y = "Player Value (Euros)")  +
    theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(labels = scales::comma)  +
  theme_bw()

legend + labs(fill = "Leagues")
```

Foot-wage graph:

```
        # ggplot and geom_point
        ggplot(data = players) + geom_point(mapping = aes(x = preferred_foot,
                y = wage_eur), color = "red") + labs(x = "Preferred Foot", y = "Wage in U.S
        Dollars",
                title = "Relationship between Preferred Foot and Wage") + theme_bw()
```

Correlation values and overall score:

```
        # variables to store team position
        cam <- which(fifa$team_position == "CAM")

        # use correlation function to obtain the value
        ccam <- cor(fifa$value_eur[cam], fifa$overall[cam])

        # creating a dataframe
        corvect <- c(ccam, cgk, ccb, ccdm, ccf, ccm, clam, clb, clcb, clcm, clf, clm, cls, clw,
        clwb, cram, crcb, crdm, cres, crf, crm, crs, crw, crwb, cst, csub)
        cordf <- data.frame(corvect)
        cordf$position <- c("ccam", "cgk", "ccb", "ccdm", "ccf", "ccm", "clam", "clb", "clcb",
"clcm", "clf", "clm", "cls", "clw", "clwb", "cram", "crcb", "crdm", "cres", "crf", "crm", "crs",
"crw", "crwb", "cst", "csub")

        # using ggplot and geom_point to plot
```

```
        ggplot(cordf) +  geom_point(mapping = aes(x=position, y = corvect, color = position,
size = corvect*1.5)) + labs(y = "correlation") +
                        ggtitle("correlation between value and overall score") +
                        theme_bw()
```

Individual player analysis:

Greatest of all Time:

```
# loading the required libraries
library(magrittr)
library(magrittr)
library(dplyr)
library(tidyr)


# using filer() to separate the Ronaldo-Messi rows
# then selecting the required attributes to compare
# sort them by Skill and Ratings and make a new dataset
data_goats <- fifa_21 %>%
  filter(shortName %in% c("Cristiano Ronaldo", "L. Messi")) %>%
    select(shortName, dribbling, shooting, passing, defending, physique, heading, freeKicks,
speed, penalties) %>%
  gather(Skill, Ratings, dribbling, shooting, passing, defending, physique, heading, freeKicks,
speed, penalties, -shortName)
head(data_goats)



# using position dodge and ggplot, plotting the comparison graph
legend <- ggplot(data = data_goats) +
        geom_bar(mapping = aes(y = Ratings, x= Skill, fill = shortName), stat = "identity",
position = "dodge") +
  ggtitle("Comparison between two greatest legends") +
    theme(plot.title = element_text(hjust = 0.5)) +
  theme_bw()

legend + labs(fill = "Player Names")
```

Player evolution growth:

```
# reading in fifa_17 dataset
fifa_17 <- read.csv("players_17.csv")

# renaming the columns to shorter names
fifa_17 <- fifa_17 %<>%
  rename(
    "finishing" = attacking_finishing,
    "heading" = attacking_heading_accuracy,
    "shortPassing" = attacking_short_passing,
    "freeKicks" = skill_fk_accuracy,
    "penalties" = mentality_penalties,
    "positioning" = mentality_positioning,
    "shortName" = short_name,
    "physique" = physic,
    "marketValue" = value_eur,
     "weeklyWage" = wage_eur
  )

# renaming mbappe's name to year wise
fifa_17$shortName[fifa_17$shortName == "K. Mbappe Lottin"] <- "K. Mbappé_17"

# filtering and selecting the mbappe row with required skills
mbappe_17 <-
  filter(fifa_17, shortName == "K. Mbappé_17") %>%
    select(shortName, overall, potential, dribbling, shooting, passing, defending, physique,
heading, pace)


# reading in fifa 18 dataset
fifa_18 <- read.csv("players_18.csv")

# renaming based off year
 fifa_18$shortName[fifa_18$shortName == "K. Mbappé"] <- "K. Mbappé_18"

# renaming columns for 18 version
```

```
fifa_18 <- fifa_18 %<>%
  rename (
    "finishing" = attacking_finishing,
    "heading" = attacking_heading_accuracy,
    "shortPassing" = attacking_short_passing,
    "freeKicks" = skill_fk_accuracy,
    "penalties" = mentality_penalties,
    "positioning" = mentality_positioning,
    "shortName" = short_name,
    "physique" = physic,
    "marketValue" = value_eur,
     "weeklyWage" = wage_eur
  )

# filtering out data as fifa_18
mbappe_18 <-
  filter(fifa_18, shortName == "K. Mbappé_18") %>%
    select(shortName, overall, potential, dribbling, shooting, passing, defending, physique,
heading, pace)

# reading in fifa 19
fifa_19 <- read.csv("players_19.csv")

fifa_19 <- fifa_19 %<>%
  rename (
    "finishing" = attacking_finishing,
    "heading" = attacking_heading_accuracy,
    "shortPassing" = attacking_short_passing,
    "freeKicks" = skill_fk_accuracy,
    "penalties" = mentality_penalties,
    "positioning" = mentality_positioning,
    "shortName" = short_name,
    "physique" = physic,
    "marketValue" = value_eur,
     "weeklyWage" = wage_eur
  )

# renaming name
 fifa_19$shortName[fifa_19$shortName == "K. Mbappé"] <- "K. Mbappé_19"


# filtering out the data
mbappe_19 <-
```

```
  filter(fifa_19, shortName == "K. Mbappé_19") %>%
    select(shortName, overall, potential, dribbling, shooting, passing, defending, physique,
heading, pace)

# reading in fifa 20

fifa_20 <- read.csv("players_20.csv")

# renaming columns
fifa_20 <- fifa_20 %<>%
  rename (
    "finishing" = attacking_finishing,
    "heading" = attacking_heading_accuracy,
    "shortPassing" = attacking_short_passing,
    "freeKicks" = skill_fk_accuracy,
    "penalties" = mentality_penalties,
    "positioning" = mentality_positioning,
    "shortName" = short_name,
    "physique" = physic,
    "marketValue" = value_eur,
    "weeklyWage" = wage_eur
  )

# renaming the name
 fifa_20$shortName[fifa_20$shortName == "K. Mbappé"] <- "K. Mbappé_20"

mbappe_20 <-
  filter(fifa_20, shortName == "K. Mbappé_20") %>%
    select(shortName, overall, potential, dribbling, shooting, passing, defending, physique,
heading, pace)

# loading the fifa_21 data
fifa_21 <- read.csv("players_21.csv")

# renaming each columns
fifa_21 <- fifa_21 %<>%
  rename (
    "finishing" = attacking_finishing,
    "heading" = attacking_heading_accuracy,
    "shortPassing" = attacking_short_passing,
    "freeKicks" = skill_fk_accuracy,
    "penalties" = mentality_penalties,
    "positioning" = mentality_positioning,
```

```
    "shortName" = short_name,
    "physique" = physic,
    "marketValue" = value_eur,
    "weeklyWage" = wage_eur
  )


# renaming his name
 fifa_21$shortName[fifa_21$shortName == "K. Mbappé"] <- "K. Mbappé_21"


# filtering out data for new dataset
mbappe_21 <-
  filter(fifa_21, shortName == "K. Mbappé_21") %>%
    select(shortName, overall, potential, dribbling, shooting, passing, defending, physique,
heading, pace)


# creating a year column
year <- c("2017", "2018","2019", "2020", "2021")

# adding it to existing dataset
mbappe_career <- rbind(mbappe_17, mbappe_18, mbappe_19, mbappe_20, mbappe_21)
mbappe_career <- cbind(mbappe_career, year)


# skill as another column
skill <- c("dribbling", "overall", "potential", "shooting", "passing", "defending", "physique",
"heading", "freeKicks", "pace",
          "penalties")


# loading required library
library(tidyr)


# pivot_longer will create a data set while also allowing us to lengthen the data
pivot_longer(
  mbappe_career, !c(shortName, year),
  names_to = 'skill', values_to = 'rating'
) %>%
ggplot(aes(x = year, y = rating)) %+%
  geom_point() %+%
```

```
geom_line(group=1) %+%
facet_wrap(~skill) %+%
ggtitle("Mbappé's evolution from 2017 to 2021") +
labs(x = "Years", y = "Ratings") +
theme_bw()
```