Data Exploration Report

INTRODUCTION

FIFA 19 is a football simulation video game developed by EA Vancouver as part of Electronic Arts' FIFA series.
As a fan of football, I am making an exploratory analysis on the FIFA 19 dataset using R. The problem we incur is that the Football dataset contains different features that needs to be analysed and to create a recommended system for the coaches to get quality players within the age and budget limit and play the best possible combination with similar replacement on different formations. This has motivated me to answer the following questions :-

1. Which different skill sets are predominantly required for different positions in football?

2. How are the wages earned and ranking of players from various positional groups (like Strikers, Mid-Field, Defence, Goal-keepers) distributed and which group has the highest value? Does the Age of the player have an impact on overall performance and their wages?

3. What recommendations can we make to the coaches and the team management aiming to enhance their team performance by choosing correct team combinations for different formations based on different positions (as given by Q1) and also provide different options as replacement (as given by Q2)

DATA WRANGLING

Description of the data sources with links if available:-

1. FIFA 19 data has been taken from https://sofifa.com/
2. Detailed attributes for every player registered in the latest edition of FIFA 19 database. Dataset: https://www.kaggle.com/karangadiya/fifa19#data.csv
3. It contains Tabular data with **18208 rows** and **89 columns,** containing player's data with different skill sets and overall rating on scale of 100 for different positions.

Steps in Data Wrangling:

1. The dataset is viewed using excel to note the required parameters like: ['Name', Age', 'Nationality', 'Overall', 'Potential', 'Club', 'Value', 'Wage', 'Preferred Foot', 'Position']

ID	Name	Age	Photo	Nationality	Flag	Overall	Potential	Club
158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	https://cdn.sofifa.org/flags/52.png	94	94	FC Barcelona
20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	https://cdn.sofifa.org/flags/38.png	94	94	Juventus
190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	Brazil	https://cdn.sofifa.org/flags/54.png	92	93	Paris Saint-Germain
193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	https://cdn.sofifa.org/flags/45.png	91	93	Manchester United
192985	K. De Bruyne	27	https://cdn.sofifa.org/players/4/19/192985.png	Belgium	https://cdn.sofifa.org/flags/7.png	91	92	Manchester City

Figure 1: Initial Dataset with 18208 rows and 89 columns

2. In R, open the csv file as: **fifa <- read_csv('data.csv')**

3. In Data Cleaning, we eliminate the variables that are not required for analysis and we reduce the data to **18208 rows** and **45 columns** as:

fifa <- fifa[-c(1,2,5, 7,11, 14,16, 17:21, 23:54)]

Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	Position	Crossing
L. Messi	31	Argentina	94	94	FC Barcelona	€110.5M	€565K	Left	RF	84
Cristiano Ronaldo	33	Portugal	94	94	Juventus	€77M	€405K	Right	ST	84
Neymar Jr	26	Brazil	92	93	Paris Saint-Germain	€118.5M	€290K	Right	LW	79
De Gea	27	Spain	91	93	Manchester United	€72M	€260K	Right	GK	17
K. De Bruyne	27	Belgium	91	92	Manchester City	€102M	€355K	Right	RCM	93

Figure 2: Cleaned Dataset with 18208 rows and 45 columns

4. In Data Transformation, a new column is created as group_position by grouping the positions according to Strikers, Midfielders, Defenders and GoalKeepers and the null values are removed.

Code:

```
filter_cat <- c('ST', 'CF', 'LF', 'LS', 'LW', 'RF', 'RS', 'RW')
fifa$GroupPosition[which(fifa$Position %in% filter_cat)] <- 'Striker'

filter_cat <- c('CAM', 'CDM', 'LCM', 'CM', 'LAM', 'LDM', 'LM', 'RAM', 'RCM', 'RDM', 'RM')
fifa$GroupPosition[which(fifa$Position %in% filter_cat)] <- 'Midfielder'

filter_cat <- c('CB', 'LB', 'LCB', 'LWB', 'RB', 'RCB', 'RWB')
fifa$GroupPosition[which(fifa$Position %in% filter_cat)] <- 'Defender'

fifa$GroupPosition[which(fifa$Position %in% c('GK'))] <- 'GoalKeeper'
```

DATA CHECKING

1. After exploring in the dataset, Value and Wage columns are converted to actual currency values. This is achieved by calling a function that takes a vector as an input and removes the "€" sign from the columns and multiplies it with the appropriate number to convert it into thousand(K) and million(M).

Value	Wage
110500000	565000
77000000	405000
118500000	290000
93000000	340000
80000000	455000
77000000	205000
89000000	205000
83500000	205000
60000000	200000

Code:

```
toNumberCurrency <- function(vector) {
  vector <- as.character(vector)
  vector <- gsub("(€|,)","", vector)
  result <- as.numeric(vector)

  k_positions <- grep("K", vector)
  result[k_positions] <-
as.numeric(gsub("K","",vector[k_positions])) *
1000

  m_positions <- grep("M", vector)
  result[m_positions] <- as.numeric(gsub("M","",
                                vector[m_positions])) *
1000000
  return(result)
}
fifa$Wage <- toNumberCurrency(fifa$Wage)
fifa$Value <- toNumberCurrency(fifa$Value)
```

Figure 3: Value and Wage columns with correct currency value with Code to update

DATA EXPLORATION

1. Which different skill sets are predominantly required for different positions in football?

Different **Bar plots** for each group position is created and the average value of each skill for players in that category is observed. We do this using ggplot, with the X axis as **'Different Skills'** and Y axis as **'Average skill value'.** This is done using a function which takes a data frame of group_position and returns the Top 5 skills.

Code:

```
pos_skills_df <- function(position){
  avg_skills <- lapply(position[1:34], mean,
na.rm = T)
  avg_skills <- unlist(avg_skills)

  new_df <- data.frame(
    positions = names(position[1:34]),
    average_value = avg_skills
  )
  new_df$skills = paste(new_df$positions, "(",
round(new_df$average_value,2) , ")")
  new_df <-
new_df[order(new_df$average_value,
decreasing = TRUE),][1:5,]

  return (new_df)
}
```

Code:

```
ggplot(pos_skills_df(group_position),
aes(fill=reorder(skills, -average_value),
x= reorder(positions, average_value),
y=average_value)) +

   geom_bar(position="stack",
stat="identity") +
  ggtitle("Plotting different skill sets of
strikers") +
  ylab('Average skill value') +
  xlab('Different Skills') +
  theme(legend.title= element_blank())
```

Firstly we check which position has the maximum number of players by plotting the bar graph, with the groupPositions on the x axis and the count of players on the y axis.

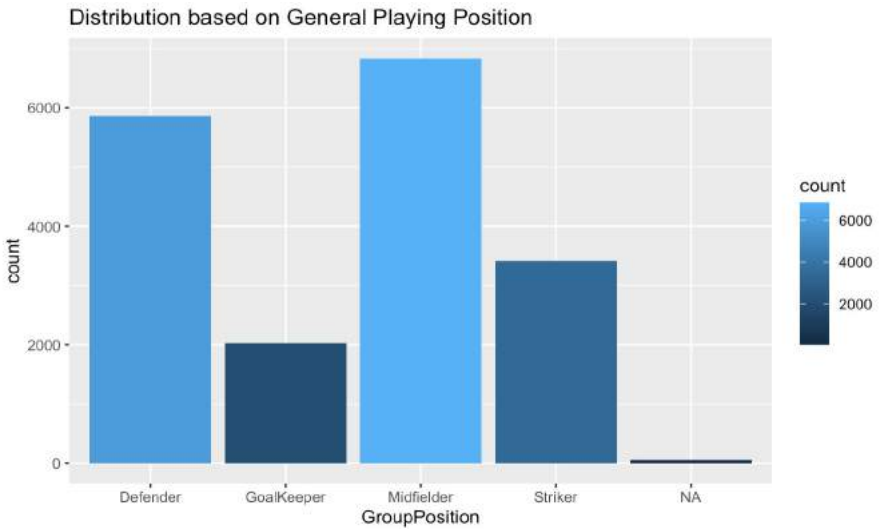

Figure 4: Player distribution on the basis of position

This shows that FIFA 19 dataset has maximum midfielders, followed by the defenders and strikers.

The below Insights are critical while creating a recommendation system in Q3:

1. From **3418 strikers**, this visualisation depicts that to be a striker, a player must possess top 5 skills like **SprintSpeed, Acceleration, Agility, Balance and ShotPower.**

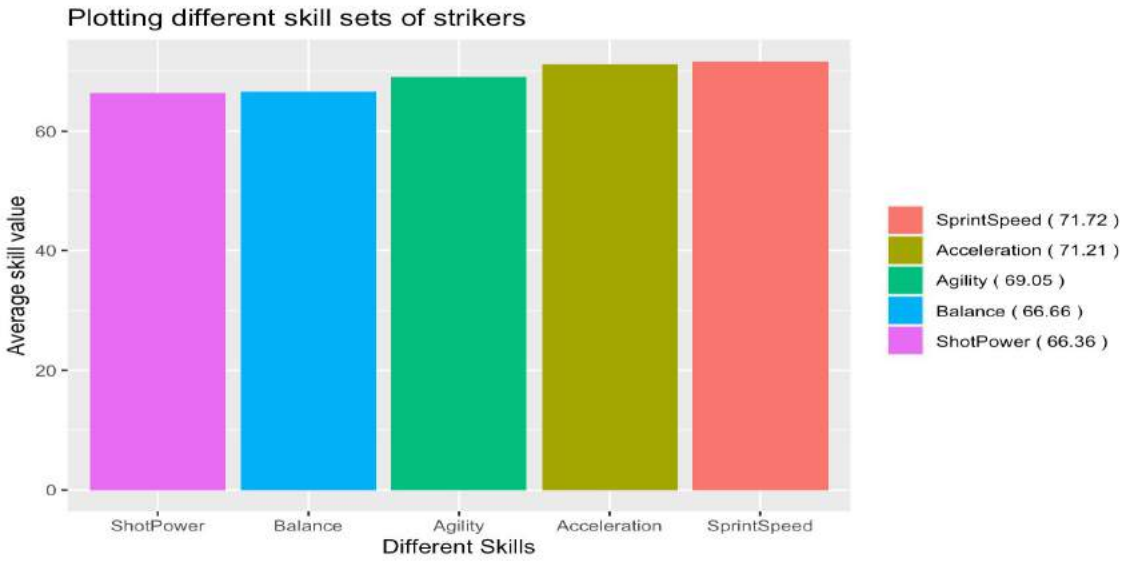

Figure 5: Plot for top 5 Skill sets of strikers

2. From **6838 midfielders**, this visualisation explains that **midfielders** require top skills like **Balance, Agility, Acceleration, SprintSpeed, and Stamina.**

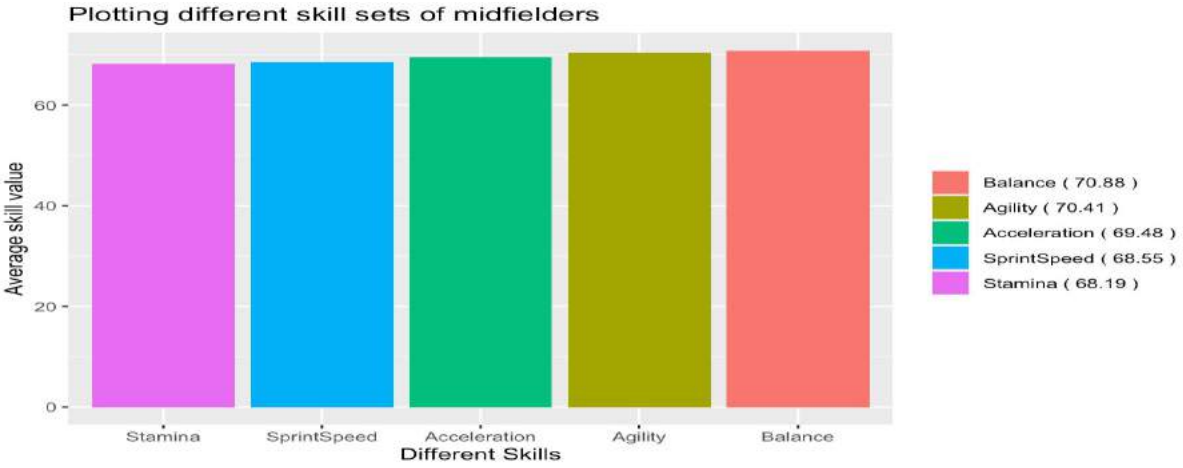

Figure 6: Plot for top 5 Skill sets of midfielders

3. From **5866 defenders**, this visualisation illustrates that **Strength, Jumping, Stamina, Standing Tackle and Aggression** are key skills for **defenders.**

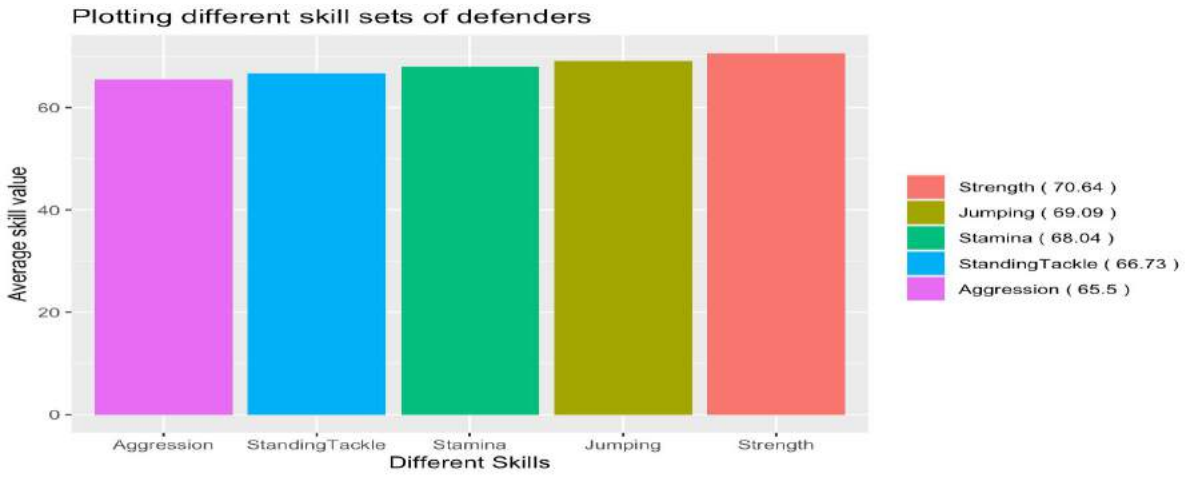

Figure 7: Plot for top 5 Skill sets of defenders

4. From **2025 goalkeepers**, this visualisation outlines that **goalkeepers** need to have skills like **GKReflexes, GKDiving, GKPositioning, GKHandling and GKKicking.**

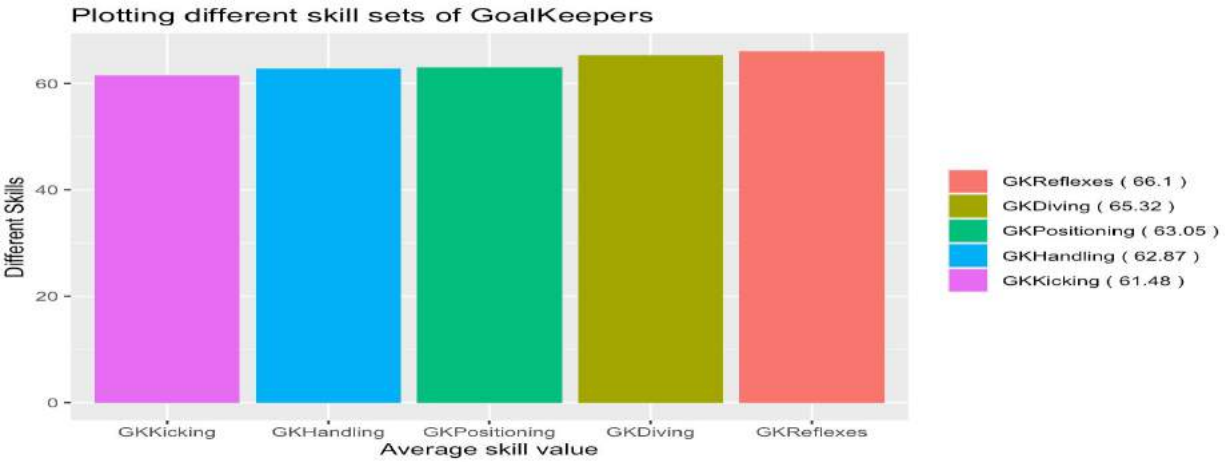

Figure 8: Plot for top 5 Skill sets of goalkeepers

2. How are the wages earned and ranking of players from various positional groups (like Strikers, Mid-Field, Defence, Goal-keepers) distributed and which group has the highest value? Does the Age of the player have an impact on overall performance and their wages?

To analyze the wage earned by the players, we create wage_brackets as: 0–100k, 100k-200k, 200k-300k, 300k-400k, 400k-500k, 500k+ where all the earnings are in the currency of “Euro”.

GroupPosition	wageBrackets
Striker	500k+
Striker	400k-500k
Striker	200k-300k
GoalKeeper	200k-300k
Midfielder	300k-400k
Striker	300k-400k
Midfielder	400k-500k
Striker	400k-500k
Defender	300k-400k
GoalKeeper	0-100k
Striker	200k-300k
Midfielder	300k-400k

Code:

```
wage_breaks <- c(0, 100000, 200000, 300000, 400000,
500000, Inf)
wage_labels <- c("0-100k", "100k-200k", "200k-300k",
"300k-400k", "400k-500k", "500k+")
wageBrackets <- cut(x=fifa$Wage,
breaks=wage_breaks,labels=wage_labels, include.lowest =
TRUE)
fifa <- mutate(fifa, wageBrackets)
```

Figure 9: Creating wageBracket column from the column Wages

We check the wage distribution for each group position by mapping the wages from 100K – 500K+, and observe the findings.

Code:

```
gw1 <- filter(fifa, Wage > 100000)
g1 <- ggplot(gw1, aes(Position)) +
geom_bar(aes(fill=wageBrackets)) +
ggtitle("Position based on Wage (100K- 500K+)")
```

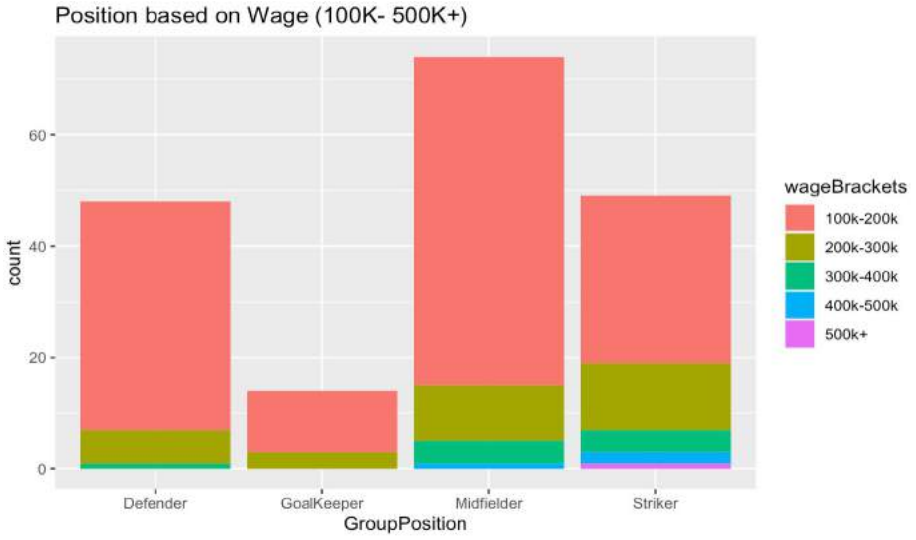

Figure 9: Plot of Wage distribution and GroupPostion

We can further analyse the trend to find out, at which position players are paid well. This can seens by jitter plot where GroupPosition of the players is on the x axis and the wage brackets on the y axis:

Code:

```
g2 <- ggplot(fifa, aes(GroupPosition,wageBrackets)) +
  geom_jitter(color="dark green") +
  geom_point(color="darkgreen",size = 1.2) +
  ggtitle("Position based on Wage (100K-500K+)")
```

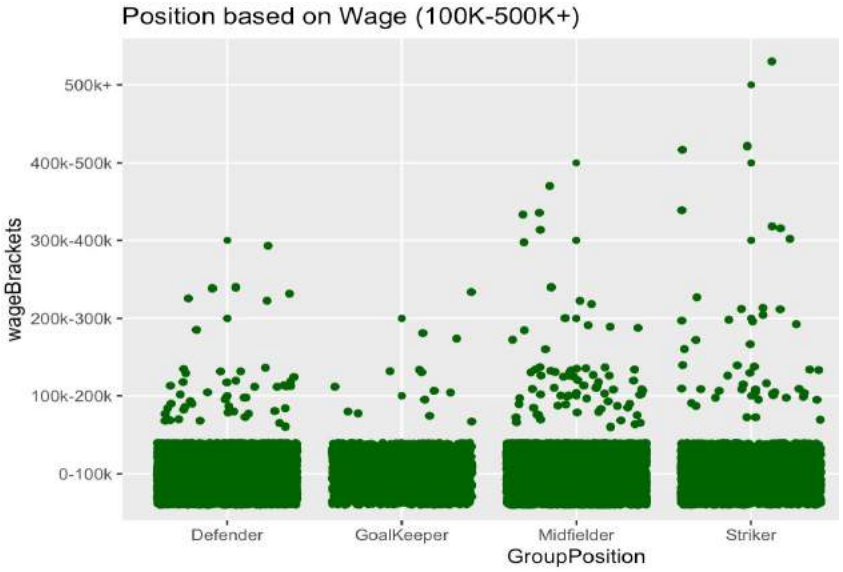

Figure 10: Plot of Wage distribution(100K - 500K+) and GroupPostion

While providing the player recommendation, we must check whether Age is factor impacting the overall performance of the player. This is done by checking the relation between plotting age on the x axis and the overall rating of a player on the y axis.

Code:

```
g_plot <- ggplot(fifa,aes(x=Age,y=Overall)) +
  geom_point(color="dark green",size = 1.2) +
  geom_smooth(colour="red") +
  ggtitle("Age vs Overall")
```

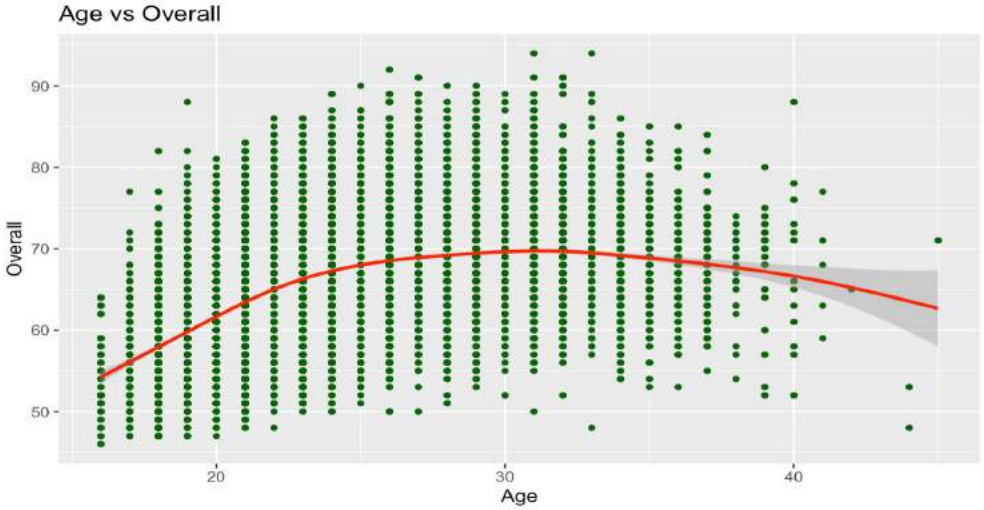

Figure 11: Plot of Age vs Overall performance of the players

From the above graphs we can infer that :-

1. Generally, 0-100K is the maximum amount of wage distribution among all Positions, followed by 100-200K.
2. **Strikers are highly paid amongst the rest of the players.**
3. With few jitters appearing above 500K for strikers, they can even earn more than 500K+.
4. Midfielders can earn up to 500K.
5. Defenders can earn a up to 300K-400K.
6. Goalkeepers can earn majorly of upto 200K-300K.
7. We noticed that the player age between 15 to 45 years and as the age increases the overall permanence also increases.
8. An interesting insight can be drawn that the **overall rating is at the optimum till the age of 30, and after that it tends to decrease as the player grows old**.

3. What recommendations can we make to the coaches and the team management aiming to enhance their team performance by choosing correct team combinations for different formations based on different positions (as given by Q1) and also provide different options as replacement (as given by Q2)

For creating a recommendation system we will take into account the insights drawn from Q1 and Q2. We consider the **formation of 4-3-3 and for different positions,** we will provide best players along with their alternative depending upon the position they play. We will also like to consider **Age = 30 years** as a factor for optimum performance along with the **Wages** they earn. We will plot a radar graph of players with top skill sets (give from Q1).

Figure 12: Football formation of 4-3-3 with their positions

1. For **Defender**, wage bracket is 400K+:

Figure 13: Position as LB

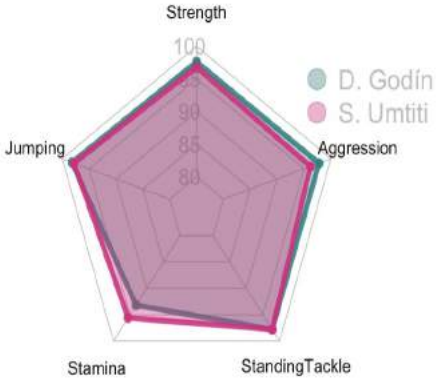

Figure 14: Position as CB

Figure 15: Position as RB

2. For **Midfielders**, wage bracket is 500K

Figure 16: Position as CM

Figure 17: Position as DM

3. For **Strikers**, wage bracket is 500K+

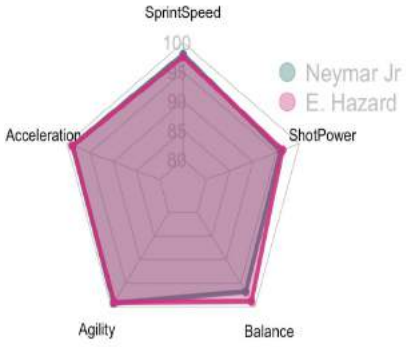

Figure 18: Position as LW

Figure 19: Position as ST

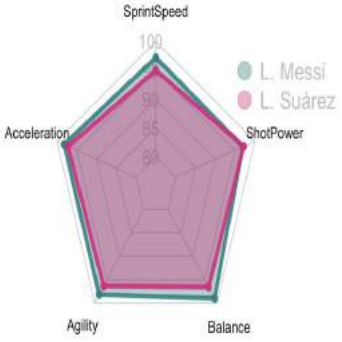

Figure 20: Position as RW

4. For **Goalkeepers**, wage bracket is 300K

Figure 21: Position as GK

From the above plots, it can be inferred that since most of the graphs are overlapping, they provide appropriate alternative options.

CONCLUSIONS

In Football, players from different positions possess different sets of skills. For a Striker, skills like SprintSpeed, Acceleration, Agility, Balance and ShotPower play a key role. In the Midfield position, players need to have strong Balance, Agility, Acceleration, SprintSpeed, and Stamina whereas for defence position, skills like Strength, Jumping, Stamina, Standing Tackle and Aggression impact their performance.

The players in football are handsomely paid with 0-100K is the maximum amount of wage distribution among all Positions, followed by 100-200K. The wages for Strikers is comparatively higher than the players at other positions. Age definitely affects the overall rating of a player, as age increases the overall rating of a player tends to increase. The overall rating is at the optimum till the age of 30, and after that it tends to decrease as the player ages.

Recommendation systems can be made for coaches and the team management which are aiming to enhance their team performance by choosing correct team combinations for different formations based on different positions and also provide different options as replacement.

REFLECTION

I learned a great deal of significant things while exploring this project. For me it was a combination of applying my knowledge of data science with my passion of knowing more insights about the game of Football. This taught me how analysis can be done to know the obvious and hidden facts about the data through visualisation which can be understood by non technical people as well. Data checking was a bit tough as the dataset contains a large number of rows and columns which made manual checking impossible. But, with the help of R, I was able to finish the data wrangling, checking, exploration part.

If anything that I would have done differently, it would be visualising my plots in more varied form. I would have made recommendations based on training the model rather than by exploring. I would have taken into consideration other factors also.

My understanding expanded over the FIFA dataset, and helped me answer my inquiries in the simplest way. Generally, it was a fun and knowledgeable undertaking.

BIBLIOGRAPHY

1. https://sofifa.com/
2. https://www.kaggle.com/karangadiya/fifa19#data.csv
3. https://www.r-graph-gallery.com/
4. https://ggplot2.tidyverse.org/reference/geom_point.html
5. https://www.r-graph-gallery.com/143-spider-chart-with-saveral-individuals.html