MONASH University

FIT 5147
DATA EXPLORATION PROJECT

FIFA DATA ANALYSIS

STUDENT NAME

SURBHI SANJAY PESHWE
30060567

Table of Contents

Introduction

Being a football fan, I thought to explore some FIFA data by analysing various trends in the game such as how various factors are correlated with each other and how do they contribute towards the teams' success and players success.
I have used descriptive statistics for the analysis of my objectives. But before that, it is required to filter the data to make it compatible with R, after which then data is cleaned for missing values.
With the above objective in mind, I have chosen a FIFA dataset from the website 'kaggle':

https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset

The dataset consists of 4 different data files:

- Player Attribute Data
- Players Personal Data
- Player Playing Position Data
- Complete dataset.

This dataset contains every players' details who have participated in FIFA, 2018. The dataset has 17981 observation and 70+ attributes which includes attributes such as a players' personal data, players performance and even player position data in the game.

Player Attribute Data.csv: This dataset consists players performance details such as player ID, overall, potential, Aggregation etc.

Player Personal Data.csv: This dataset contains player personal data such as name, nationality, club, age, wage, value etc.

Player Playing Position Data.csv: This dataset contains player positions in game and rating for the allocated positing.

It is a tabular data set which has different types of attributes like spatial and temporal as well as textual attributes.

Motivation

For this report the following questions are answered:

- Based on their age, how are the players distributed with reference to the overall performance of the player?
- What is the relation between age and position of the player?
- For which club does a player have the highest value and wages? Display top 10 clubs.
- Which Nationality has the most players in FIFA?

Data Wrangling:

The original dataset which was obtained from website where the 4 files obtained were: Player Attribute Data, Player Personal Data, Player Playing Position Data, Complete dataset. The complete dataset consists of total combination of the data from the other 3 files, but I have explored the data personally and have combined it. I have used tools such as **Excel, Tableau and R** to get the data formatted and have also used it for visualising.

Excel Usage:

I have used Excel for observing and analysing data at a glimpse. This has also helped in deciding what attributes to consider for the purpose of answering the above questions.

R Usage:

I have used R to explore the data and perform data wrangling on dataset.

```
> #reading data
> PlayerAttributeData <- read.csv("~/Downloads/fifa-18-demo-player-dataset/PlayerAttributeData.csv")
> PlayerPersonalData <- read.csv("~/Downloads/fifa-18-demo-player-dataset/PlayerPersonalData.csv", header=TRUE)
> PlayerPlayingPositionData <- read.csv("~/Downloads/fifa-18-demo-player-dataset/PlayerPlayingPositionData.csv")
```

FIGURE 1: READING THE CSV FILES INTO R

My initial datasets had 17981 rows and different columns based on dataset. For better visualisation, I have omitted a few columns which are not required such as 'unnamed X column' and 'ID' column as these columns are common to all 3 datasets. Finally, my complete dataset consists of 17981 row and 76 columns.

Data	
complete_data	17981 obs. of 76 variables
mydata	17981 obs. of 49 variables
PlayerAttributeData	17981 obs. of 36 variables
PlayerPersonalData	17981 obs. of 15 variables
PlayerPlayingPositionD…	17981 obs. of 29 variables

FIGURE 2: DESCRIPTION OF DATASETS SHOWING THE NUMBER OF OBSERVATIONS AND THE NUMBER OF VARIALES IN EACH OF THEM

Description of the 'Complete Dataset':

The complete dataset has null values, but those where only for the goal keeper player as they stand near goal-post so the position columns are left null. These columns were left untreated as they were not required for my analysis.

Further from my complete dataset, I have used columns such as **Name, Age, Nationality, Overall, Club, Value, Wage and Preferred Position** for my analysis.

After cleaning the whole dataset, now I have selected the above considered columns and made a new dataset for few visualisations using Tableau.

Usage of R:

After the basic analysis such as summary, dimension, null values, duplicates and invalid values, I have noticed that the Wage and Value column had a value with currency symbol which would be difficult for me while visualising. Snapshot of the dataset where we can see the wage and value column entries with currency symbol which can neither be considered as string or integer.

```
> head(vis_data)
      ID     X              Name Age  Nationality Overall           Club  Value  Wage Preferred.Positions
1  20801     0 Cristiano Ronaldo  32     Portugal      94 Real Madrid CF €95.5M €565K               ST LW
2 158023     1          L. Messi  30    Argentina      93   FC Barcelona  €105M €565K                  RW
3 155862    10      Sergio Ramos  31        Spain      90 Real Madrid CF   €52M €310K                  CB
4 201399   100         M. Icardi  24    Argentina      84          Inter   €42M €105K                  ST
5 208450  1000   Andreas Pereira  21       Brazil      77    Valencia CF   €14M  €86K          CAM CM LM
6 208926 10000       A. Al Enezi  26 Saudi Arabia      65       Al Nassr  €425K   €8K                  GK
```

FIGURE 3: FIRST 6 ROWS THE SHOWING CURRENCY SYMBOL

For visualisation, I have changed the entries into actual currency value by using a function called 'Currency', where I have removed currency symbol and converted the value into thousand and million.

After implementing the above function on wage and value columns, the number have been converted into integers which will be easy while visualising these attributes.

```
> head(vis_data)
      ID     X              Name Age  Nationality Overall           Club    Value   Wage Preferred.Positions
1  20801     0 Cristiano Ronaldo  32     Portugal      94 Real Madrid CF 9.55e+07 565000               ST LW
2 158023     1          L. Messi  30    Argentina      93   FC Barcelona 1.05e+08 565000                  RW
3 155862    10      Sergio Ramos  31        Spain      90 Real Madrid CF 5.20e+07 310000                  CB
4 201399   100         M. Icardi  24    Argentina      84          Inter 4.20e+07 105000                  ST
5 208450  1000   Andreas Pereira  21       Brazil      77    Valencia CF 1.40e+07  86000          CAM CM LM
6 208926 10000       A. Al Enezi  26 Saudi Arabia      65       Al Nassr 4.25e+05   8000                  GK
```

FIGURE 4: AFTER IMPLEMENTATION OF FUNCTION, CURRENCY VALUES ARE CHANGED TO INTEGER VALUES

Later, the 'preferred position' column consists of different position that were: CAM, CB, CDM, CF, CM, GK, LB, LM, LW, LWB, RB, RM, RW, RWB, ST. In a soccer game, the basic positions are divided into goalkeeper, defenders, Midfielders and Forwards. Based on this knowledge of soccer, I have categorised the preferred positions into 4 categories: GK, DEF, MID FWD by adding a new column position.

```
> posi <- as.factor(vis_data$Preferred.Positions)
> levels(posi) <- list(GK  = c("GK"),
+                      DEF = c("LWB", "LB", "CB", "RB", "RWB"),
+                      MID = c("LW","LM","CDM","CM","CAM","RM","RW"),
+                      FWD = c("CF", "ST"))
> vis_data <- mutate(vis_data, Position = posi)
> head(vis_data)
      ID     X              Name Age  Nationality Overall           Club    Value   Wage Preferred.Positions Position
1  20801     0 Cristiano Ronaldo  32     Portugal      94 Real Madrid CF 9.55e+07 565000                  ST      FWD
2 158023     1          L. Messi  30    Argentina      93   FC Barcelona 1.05e+08 565000                  RW      MID
3 155862    10      Sergio Ramos  31        Spain      90 Real Madrid CF 5.20e+07 310000                  CB      DEF
4 201399   100         M. Icardi  24    Argentina      84          Inter 4.20e+07 105000                  ST      FWD
5 208450  1000   Andreas Pereira  21       Brazil      77    Valencia CF 1.40e+07  86000                 CAM      MID
6 208926 10000       A. Al Enezi  26 Saudi Arabia      65       Al Nassr 4.25e+05   8000                  GK       GK
>
```

FIGURE 5: ADDITION OF NEW COLUMN

Finally, I have 17981 observation and 11 variables in the dataset and it can be loaded in Tableau to answer my questions.

Data Checking

I have used R to check null values, duplicates and invalid values.

Checking Null Values:

- There were a few null values for the goal keeper preferred positions. They have been left the same as it makes sense.
- Moreover, these columns were not considered for future exploration

Checking Duplicates:

- As ID column is unique, the main focus for duplicates was done on ID column as it is unique but there were no duplicate IDs.
- There were a few duplicates in the name column and even they are not treated as it is okay to have duplicates in name column since people do have same names.

Checking Invalid Values:

- There were no invalid data in dataset.

```
> colSums(is.na(complete_data))
                X                 ID       Acceleration         Aggression            Agility            Balance
                0                  0                  0                  0                  0                  0
     Ball.control         Composure           Crossing              Curve          Dribbling          Finishing
                0                  0                  0                  0                  0                  0
Free.kick.accuracy         GK.diving       GK.handling        GK.kicking     GK.positioning       GK.reflexes
                0                  0                  0                  0                  0                  0
  Heading.accuracy      Interceptions            Jumping       Long.passing         Long.shots            Marking
                0                  0                  0                  0                  0                  0
         Penalties       Positioning          Reactions      Short.passing        Shot.power     Sliding.tackle
                0                  0                  0                  0                  0                  0
     Sprint.speed            Stamina    Standing.tackle           Strength             Vision            Volleys
                0                  0                  0                  0                  0                  0
       Unnamed..0               Name                Age              Photo        Nationality               Flag
                0                  0                  0                  0                  0                  0
          Overall          Potential               Club          Club.Logo              Value               Wage
                0                  0                  0                  0                  0                  0
          Special                CAM                 CB                CDM                 CF                 CM
                0               2029               2029               2029               2029               2029
              LAM                 LB                LCB                LCM                LDM                 LF
             2029               2029               2029               2029               2029               2029
               LM                 LS                 LW                LWB Preferred.Positions                RAM
             2029               2029               2029               2029                  0               2029
               RB                RCB                RCM                RDM                 RF                 RM
             2029               2029               2029               2029               2029               2029
               RS                 RW                RWB                 ST
             2029               2029               2029               2029
```

FIGURE 6: DATA CHECKING FOR NULL VALUES

Data Exploring

For Data Exploration, I have used the Tableau Plugin to get my answers for each question. Personally, I find Tableau easy to identify patterns and visualise the data and gives better looking graphs.

Q1) Based on their age, how are the players distributed with reference to the overall performance of the player?

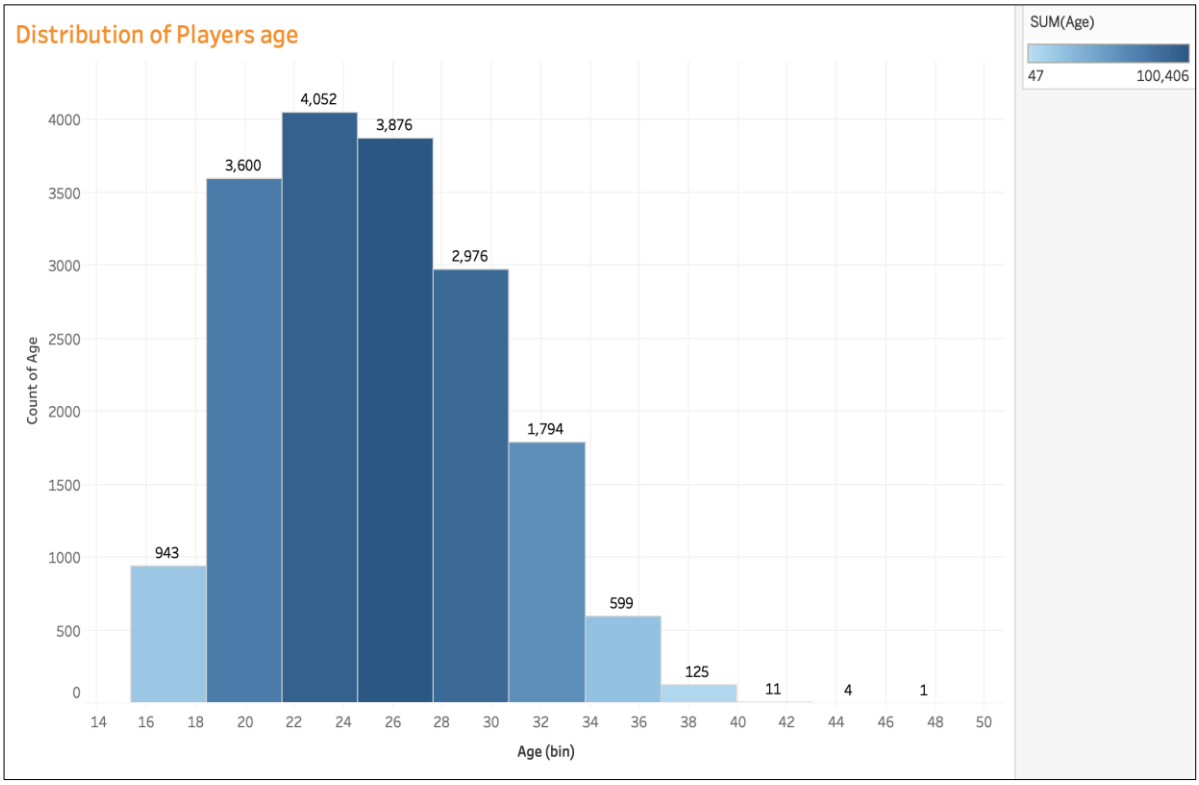

FIGURE 7: DISTRIBUTION OF THE AGE FOR PLAYERS

From the above graph, we can see that most of the players are between 22-26 years old. Usually, this age group is considered to be the best age for athletes as they have learned enough skills and gained experience.

Statistical Inference proves that 'Age' and 'Overall performance' are dependent on each other.

H_0=Age of player and their overall performance are independent of each other
$H_{1=}$ Age of player and their overall performance are dependent on each other.

```
> xt<- xtabs(~vis_data$Age + vis_data$Overall)
> summary(xt)
Call: xtabs(formula = ~vis_data$Age + vis_data$Overall)
Number of cases in table: 17981
Number of factors: 2
Test for independence of all factors:
        Chisq = 10799, df = 1344, p-value = 0
        Chi-squared approximation may be incorrect
```

FIGURE 8: VALIDATION FOR HYPOTHESIS TESTING

Here, since Chisq value= 10799 and p<5%, we reject the null hypothesis at 5% level of significance and conclude that age of player and their overall performance are dependent on each other.

Q2) What is the relation between age and position of the player?

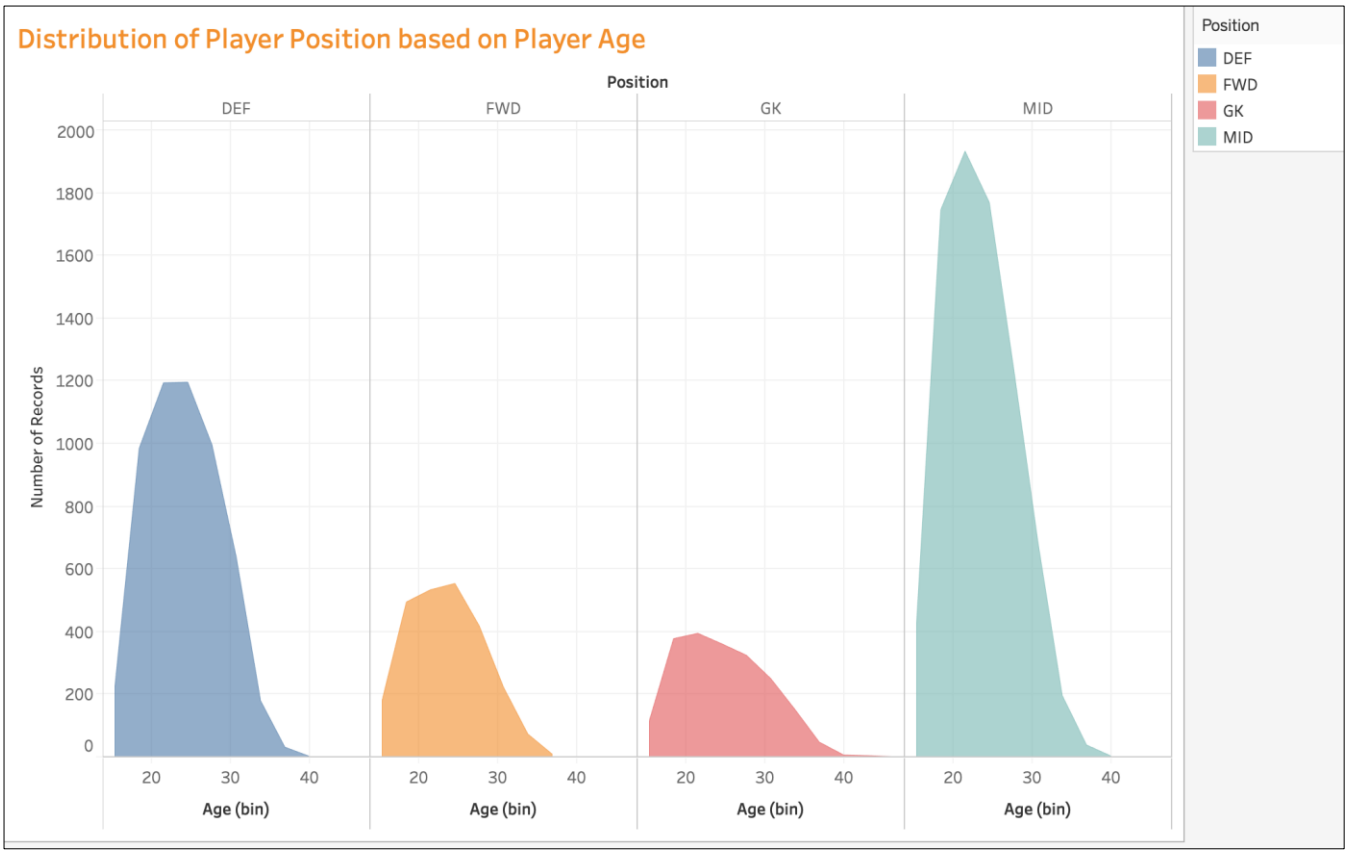

FIGURE 9: DISTRIBUTION OF POSITION BASED ON AGE OF PLAYERS

The above Tableau visualisation gives us the relation between age and position of the players on field. We can see that almost all the positions have most of the players with ages between 21 and 25 years old. One thing can be noticed that when compared to the other positions, there are less number of defenders where players age is above 35 years old.

Along with this, we can also see the effect of wages on different positions of players according to their age. This is shown below in the graph depicting the highest wage for a player with his age for all four positions.

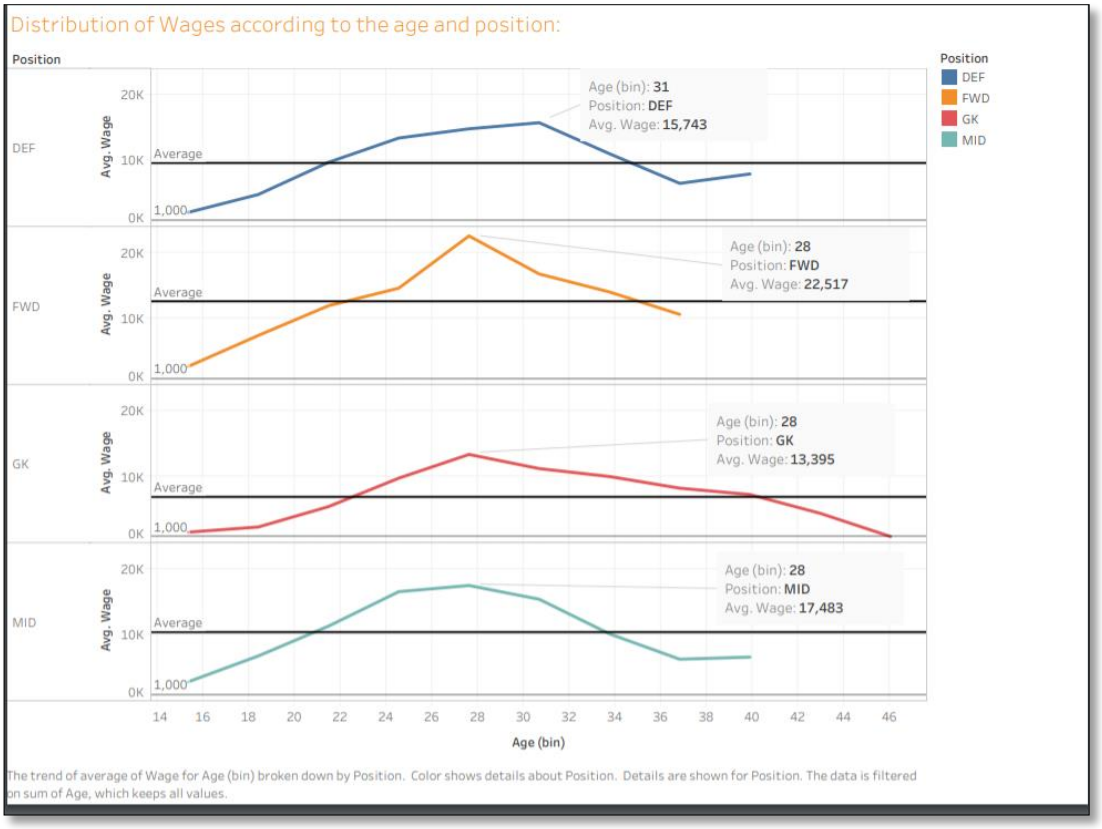

FIGURE 10: DISTRIBUTION OF WAGES ACCORDING TO AGE AND POSITION

From the above graph we can see that for all four positions there are different wages at their highest which have been shown in the boxes above. We can see the following:

- The age corresponding to the highest wage has also been shown for the forward position (FWD) has the highest wage of 22,517 where the forwards age is 28.
- The age corresponding to the highest wage has also been shown for the Mid-fielder (MID) has the highest wage of 17,483 where the forwards age is 28.
- The age corresponding to the highest wage has also been shown for Defender (DEF) has the highest wage of 15,743 where the forwards age is 31.
- The age corresponding to the highest wage has also been shown for Goal Keeper (GK) has the highest wage of 13,395 where the forwards age is 28.

Q3) For which club does a player have the highest value and wages? Display top 10 clubs.

We have already seen the effect of ages of players on their position and the wages that they get. Now, we are interested in the clubs to which these players belong.

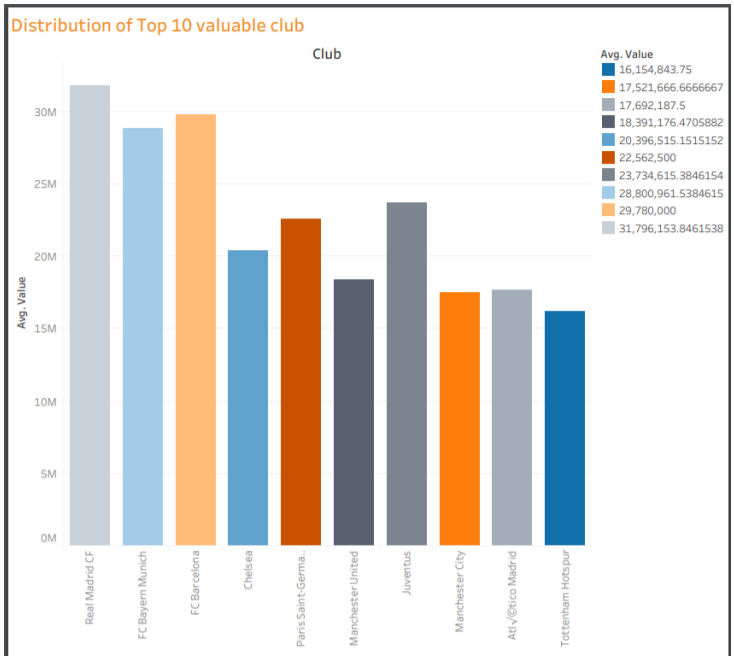

FIGURE 11: DISTRIBUTION OF TOP 10 VALUABLE CLUBS

From above bar charts, it is seen that Real Madrid CF, FC Bayern Munich and FC Barcelona are top 3 clubs who has highest value in market which means these clubs player are the most earning players than the other club players.

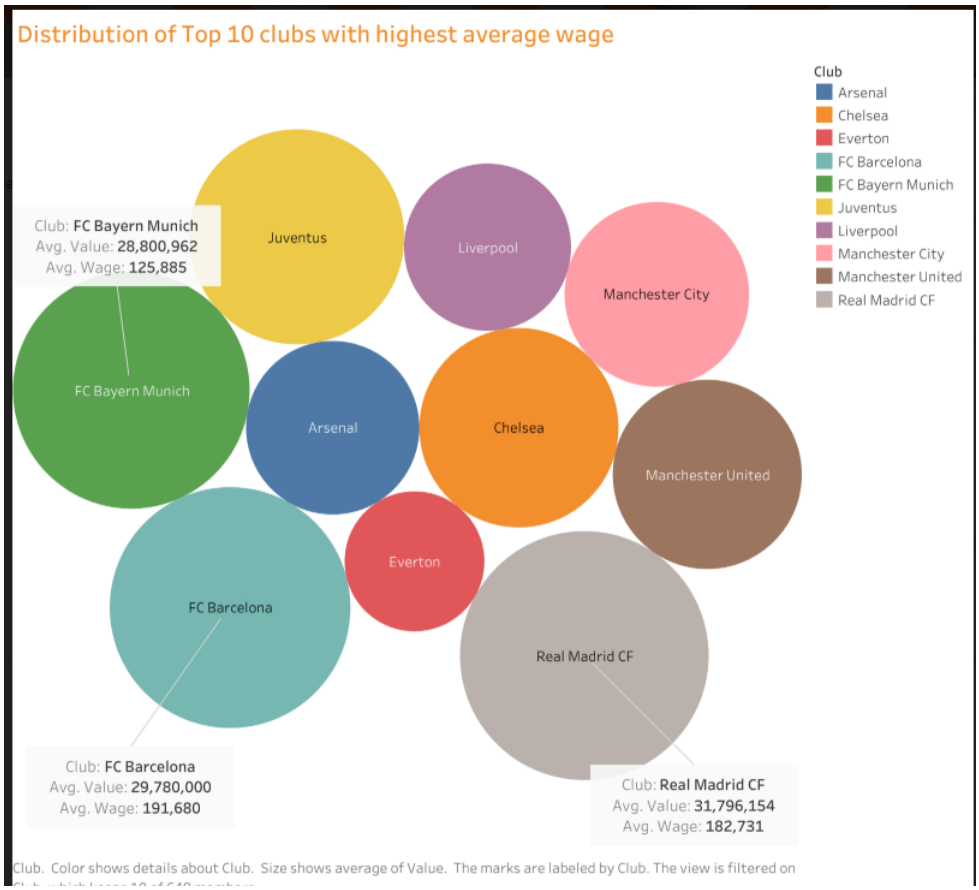

FIGURE 12: DISTRIBUTION OF TOP 10 HIGHEST WAGE CLUBS

The above packed bubbles visualisation depicts that FC Barcelona players have the highest paying wages followed by Real Madrid FC and Bayern Munich.

Q4) Which Nationality has most of the players in FIFA?

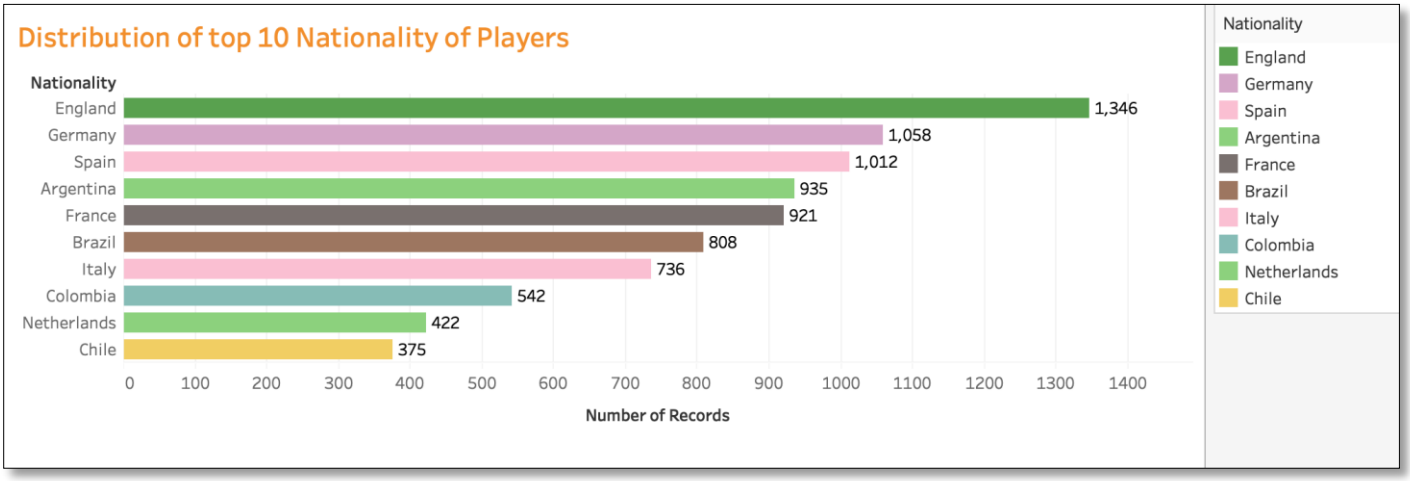

FIGURE 13: DISTRIBUTION OF PLAYERS NATIONALITY

From the above horizontal bar graph, we can see that England is at the topmost position with the highest number of players, i.e. 1346 followed by Germany and Spain. At the 10th position, Chile has 375 players.

Conclusion

- From the above project, one can get the knowledge of how various factors that affect different attributes.
- Most of the players in the game are of age 20-25 years old and the it is proven that players overall performance is based on their age and there are many various factors such as skill, experience etc which are also important.
- When looking at the payment of wages for players at different positions, it is seen that the goal keeper is paid the least whereas the forwards are paid the most, regardless to their age.
- Based on the players performance, the club's value in market has also grown. It has been seen that Real Madrid, Fc Barcelona and FC Bayern Munich are the top 3 clubs whose players have the highest market value and the highest paid wages when compared to other clubs.
- People from England have more players in FIFA followed by Germany and Spain, but it is not necessary that top players belong to these nationalities.

Reflection

From this project, I have learnt functions in R and various ways to clean and analyse data in R. After doing in depth research about visualisations using Tableau, the above answers have been provided to make them easily understandable and visually appealing.

Formatting data was a challenging part, but I have understood that for visualising purpose basic step is to analyse the data properly and based on analysis, format the data as any analysis is dependable if there are no missing values in the dataset.

With regards to the ages of players, it is seen that it is not the primary constraint when considering the most wages paid. Wages are reflected by the position a player has on the field, where a forward is the most paid and a goal keeper is the least paid.

Considering the data, which plot needs to be plotted is also a task, for better presentation and this I have learned from this project.

References

Convert currencies with commas into numeric. (n.d.). Retrieved from Stack Overflow: https://stackoverflow.com/questions/31944103/convert-currency-with-commas-into-numeric

HOW PLAYER RATING IS CALCULATED? IT IS A TOTAL MESS. (n.d.). Retrieved from EA sports: https://fifaforums.easports.com/en/discussion/277545/how-player-rating-is-calculated-it-is-a-total-mess

Loo, E. d. (2013). *An introduction to data cleaning with R.* Statistics Netherlands.

Vijay Kotu, B. D. (2015). Data Exploration. *Data Exploration*. Retrieved from Science Direct: https://www.sciencedirect.com/topics/computer-science/data-exploration