ITU

FINAL PROJECT

Mrinal Mathur, MSSE, ITU

A. PROJECT PLAN

In this project I am going to use the data set provided by Kaggle. The name of the data set is "FIFA 19 Complete player Dataset". This dataset includes 18K FIFA plater with around 90 attributes from the latest FIFA database. This dataset comes under Football Analytics and contains detailed attributes for every player registered in the latest edition of FIFA 19 database.The web scrapping is already done from "https://sofifa.com/".

I analyzed the dataset of players with their attributes likes Age, Nationality, Preferred Foot, Skill Moves, Value etc. The knowledge from the R language course helped me in finding the interesting facts about this dataset.

As I go through the R lectures I will keep on adding the features to the Final Project. The link to kaggle website: https://www.kaggle.com/karangadiya/fifa19

B. PROJECT DEFINITION

SCHEDULE:

Schedule:
1/29: Basic data set import with initial analysis
2/5: Finding appropriate packages to plot graphs
2/12: Advanced Analysis
2/19: Finalizing Phase
2/23: Final Report Completion and Submission

C. PROJECT REPORT

1. Data Import

Below is the information of the imported dataset:

```
> df_fifa19 <- read.csv("Fifa19_data.csv",header = TRUE)
> df_fifa19[1:5,1:6]
> dim(df_fifa19)
> colnames(df_fifa19)
```

There are 89 attributes of the 18207 FIFA players in this dataset. These attributes will be used for performing interesting analysis of the soccer players.

2. Data Analysis

Performing the data analysis for better understanding of the dataset. First step before performing any high level analysis is to check the dataset for any NULL/NA values. The output is not being displayed to keep the report short.

Check if there are any NA"s or NULL values in the dataset using functions like is.na and is.null.

a. Analysis-1

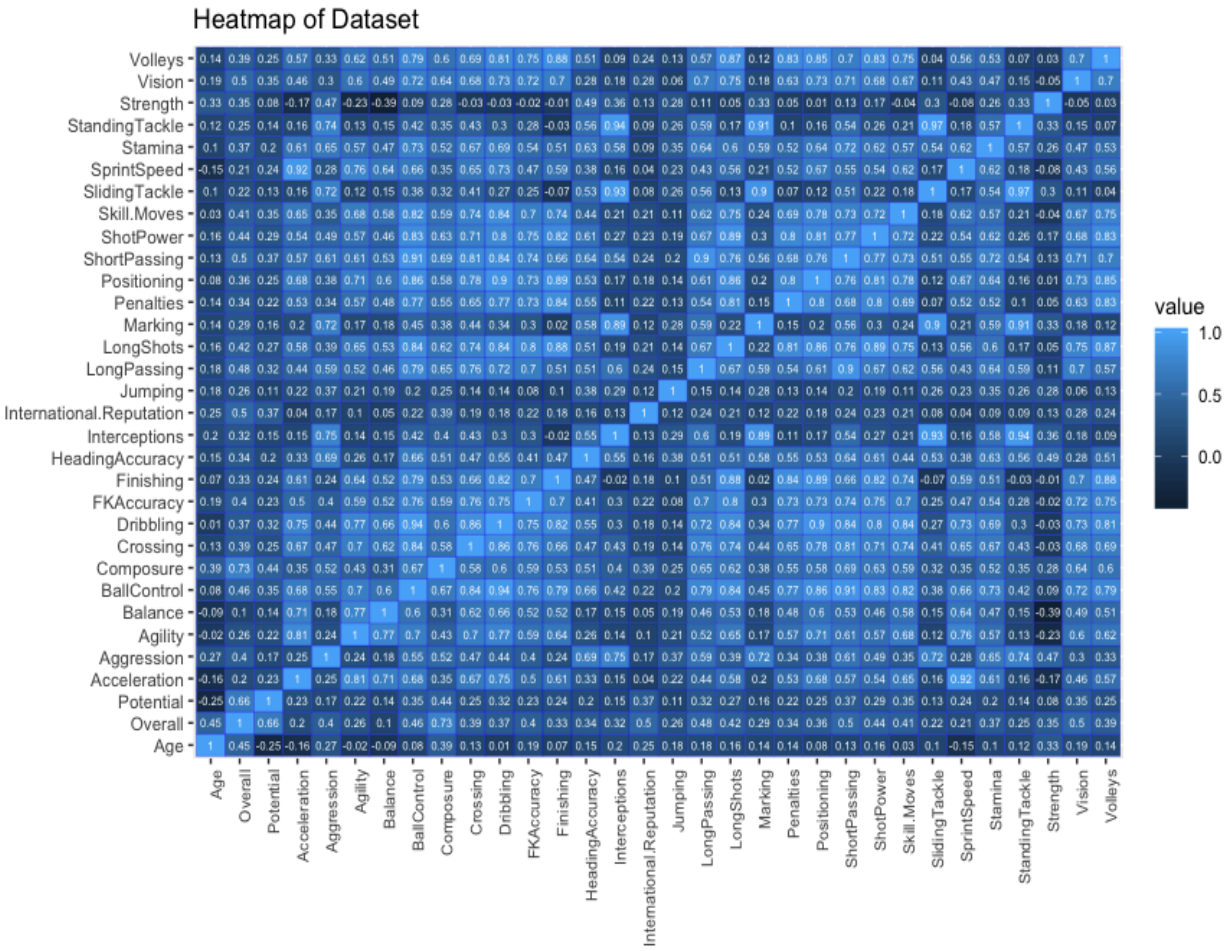
```
> cat("Choosing columns of interest")

Choosing columns of interest

> chosen_col <- c("Name","Age","Nationality","Overall","Potential","Special","Acceleration",
+                 "Aggression","Agility","Balance","BallControl","Body.Type","Composure",
+                 "Crossing","Curve","Club","Dribbling","FKAccuracy","Finishing","GKDiving",
+      "GKHandling","GKKicking","GKPositioning","GKReflexes","HeadingAccuracy",
+      "Interceptions","International.Reputation",
+      "Jersey.Number","Jumping","Joined","LongPassing","LongShots",
+      "Marking","Penalties","Position","Positioning",
+      "Preferred.Foot","Reactions","ShortPassing","ShotPower",
+      "Skill.Moves","SlidingTackle","SprintSpeed","Stamina",
+      "StandingTackle","Strength","Value","Vision","Volleys","Wage","Weak.Foot","Work.Rate"
> df <- df_fifa19[,chosen_col]
> df_cor <- df[,c("Age", "Overall", "Potential", "Acceleration",
+                  "Aggression", "Agility", "Balance", "BallControl",
+                  "Composure", "Crossing","Dribbling", "FKAccuracy", "Finishing",
+                  "HeadingAccuracy", "Interceptions","International.Reputation",
+                  "Jumping", "LongPassing", "LongShots",
+                  "Marking", "Penalties", "Positioning",
+                  "ShortPassing", "ShotPower", "Skill.Moves", "SlidingTackle",
+                  "SprintSpeed", "Stamina", "StandingTackle", "Strength", "Vision",
```

```
+                   "Volleys")]
> cor_mat <- round(cor(df_cor, use = "complete"),2)
> library(reshape2)
> melted_cor <- melt(cor_mat)
> library(ggplot2)
> ggplot(data = melted_cor, aes(x=Var1, y=Var2, color= "blue1",fill =value)) +
+    geom_tile(colour = "Blue") +
+   theme(axis.text.x = element_text(angle = 90, vjust = 1, size = 8, hjust = 1)) +
+   ggtitle("Heatmap of Dataset") + geom_text(aes(label=value),size=2,colour="White") +
+   theme(axis.title.x=element_blank(),axis.title.y=element_blank())
> cat("Below is the correaltion heatmap of important attributes:")
```

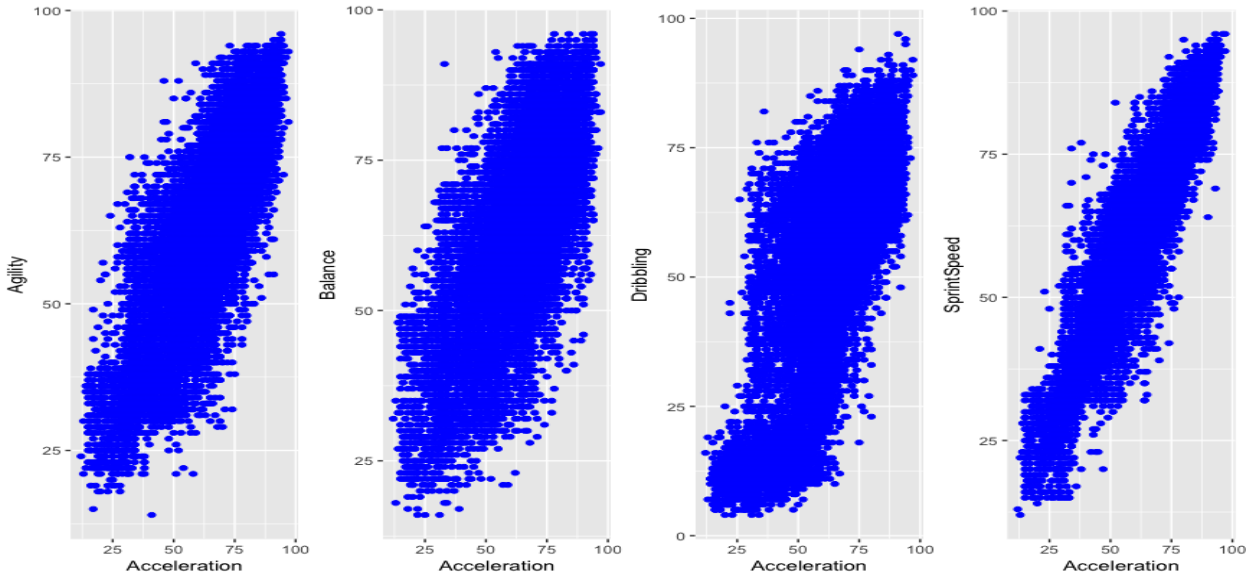
Below is the correaltion heatmap of important attributes:



b. Analysis-2

Using the above heatmap we can see some intersting correlations between different attributes. Here by interesting means the correlation value greater than "0.7". For example, this can be used to plot correlations between Accelerations and other features like : Agility, Balance, Dribbling, SprintSpeed.

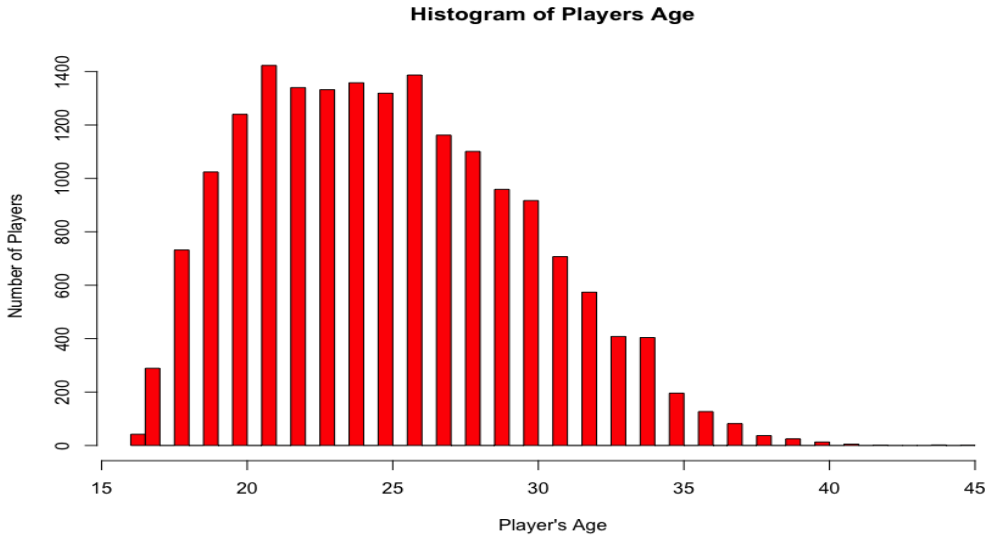
```
> p1 <- ggplot(data=df_fifa19, aes(x=df_fifa19$Acceleration,y=df_fifa19$Agility)) +
+   geom_point(color="blue") + xlab("Acceleration") + ylab("Agility")
> p2 <- ggplot(data=df_fifa19, aes(x=df_fifa19$Acceleration,y=df_fifa19$Balance)) +
+   geom_point(color="blue") + xlab("Acceleration") + ylab("Balance")
> p3 <- ggplot(data=df_fifa19, aes(x=df_fifa19$Acceleration,y=df_fifa19$Dribbling)) +
+   geom_point(color="blue") + xlab("Acceleration") + ylab("Dribbling")
> p4 <- ggplot(data=df_fifa19, aes(x=df_fifa19$Acceleration,y=df_fifa19$SprintSpeed)) +
+   geom_point(color="blue") + xlab("Acceleration") + ylab("SprintSpeed")
> #install.packages("ggpubr")
> library(ggpubr)
> ggarrange(p1,p2,p3,p4,nrow=1,ncol=4)
```



c. Analysis-3

Histogram of players age can also be plotted:

```
> hist(df_fifa19$Age, col="red", breaks = 15,main="Histogram of Players Age",
+      xlim=c(min(df_fifa19$Age),max(df_fifa19$Age)),xlab="Player's Age",ylab="Number of Pla
>
```



d. Analysis-4

Finding the five eldest and youngest players

```
> eldest <- df_fifa19[order(-df_fifa19$Age),]
> cat("The five oldest players are:")
```

```
The five oldest players are:
```

```
> eldest[1:5,c("Name","Nationality","Age")]
```

```
               Name       Nationality Age
4742       O. Pérez            Mexico  45
17727     T. Warner Trinidad & Tobago  44
18184 K. Pilkington           England  44
10546   S. Narazaki             Japan  42
1121      J. Villar          Paraguay  41
```

```
> youngest <- df_fifa19[order(df_fifa19$Age),]
> cat("The five youngest players are:")
```

```
The five youngest players are:
```

```
> youngest[1:5,c("Name","Nationality","Age")]
```

```
                Name Nationality Age
11458    W. Geubbels      France  16
11733       A. Taoui      France  16
```

```
12497 Pelayo Morilla       Spain  16
12829       Guerrero       Spain  16
13294    H. Massengo      France  16
```

e. Analysis-5

Comparing six best clubs in relation to attribute like Age

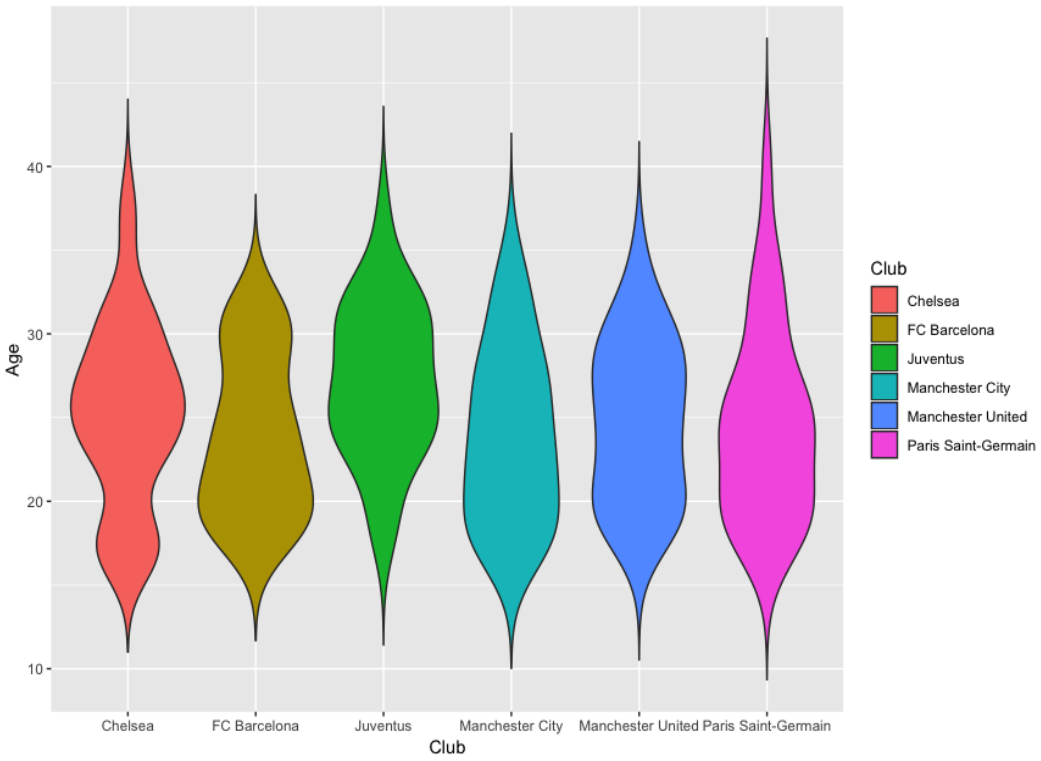
```
> clubs <- df_fifa19[order(-df_fifa19$Overall),]
> best_clubs <- clubs[1:6,c("Club")]
> cat("The six best clubs are:")
```

```
The six best clubs are:
```

```
> best_clubs
```

```
[1] FC Barcelona        Juventus            Paris Saint-Germain
[4] Manchester United   Manchester City     Chelsea
652 Levels:   SSV Jahn Regensburg 1. FC Heidenheim 1846 ... Zag■lębie Sosnowiec
```

```
> df_clubs <- df_fifa19[which(df_fifa19["Club"] %in% best_clubs), c("Age","Club")]
> ggplot(data=df_clubs,aes(x=Club,y=Age)) + ggtitle("Distribution of age in some clubs")+
+   geom_violin(aes(fill=Club),trim = FALSE)
```



f. Analysis-6

Positioning Statistics:

```
> cat("The 10 best players per position:")
```

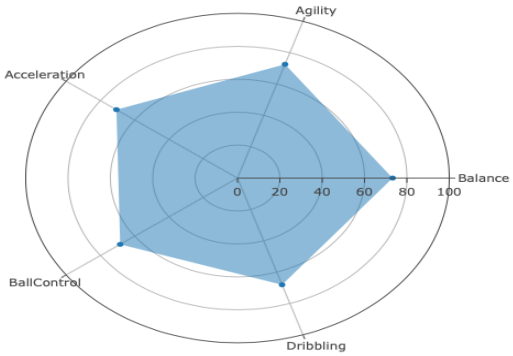
```
The 10 best players per position:
```

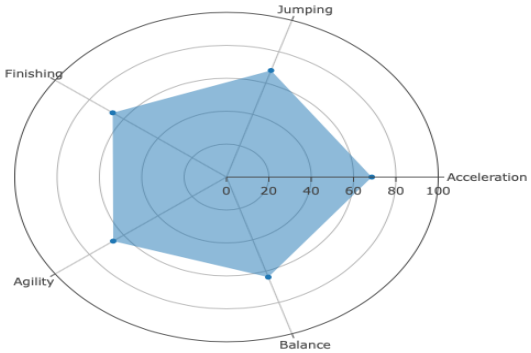
```
> best_player <- df_fifa19[order(-df_fifa19$Overall),]
> best_player[1:10,c("Name","Position")]
```

```
                Name Position
1           L. Messi       RF
2  Cristiano Ronaldo       ST
3          Neymar Jr       LW
4             De Gea       GK
5       K. De Bruyne      RCM
6          E. Hazard       LF
7          L. Modrić      RCM
8          L. Suárez       RS
9       Sergio Ramos      RCB
10           J. Oblak       GK
```

```
> feature_CAM <- c("Balance","Agility","Acceleration","BallControl","Dribbling")
> feature_ST <- c("Acceleration","Jumping","Finishing","Agility","Balance")
> r_cam <- apply(df_fifa19[which(df_fifa19[,"Position"] == "CAM"), feature_CAM],2,mean)
> r_st <- apply(df_fifa19[which(df_fifa19[,"Position"] == "ST"), feature_ST],2,mean)
> #install.packages("plotly")
> library(plotly)
> p1 <- plot_ly( type = 'scatterpolar', r = r_cam, theta = feature_CAM, fill = 'toself') %>%
+       layout(polar = list(radialaxis = list(visible = T,range = c(0,100))),
+       showlegend = F)
> p2 <- plot_ly( type = 'scatterpolar', r = r_st, theta = feature_ST, fill = 'toself') %>%
+       layout(polar = list(radialaxis = list(visible = T,range = c(0,100))),
+       showlegend = F)
> cat("Below is the Radix plot for CAM and ST position which is pretty cool, as it is
+     quite similar to graphs shown in FIFA PS4 game:")
```

```
Below is the Radix plot for CAM and ST position which is pretty cool, as it is
    quite similar to graphs shown in FIFA PS4 game:
```





g. Analysis-7

Left-footed vs Right-Footed Comparison

```
> cat("Top right-footed players:")
```

```
Top right-footed players:
```

```
> df_right <- df_fifa19[which(df_fifa19[,"Preferred.Foot"] == "Right"),]
> head(df_right[,c("Name","Overall")])
```

```
               Name Overall
2 Cristiano Ronaldo      94
3         Neymar Jr      92
4            De Gea      91
5      K. De Bruyne      91
6         E. Hazard      91
7         L. Modrić      91
```

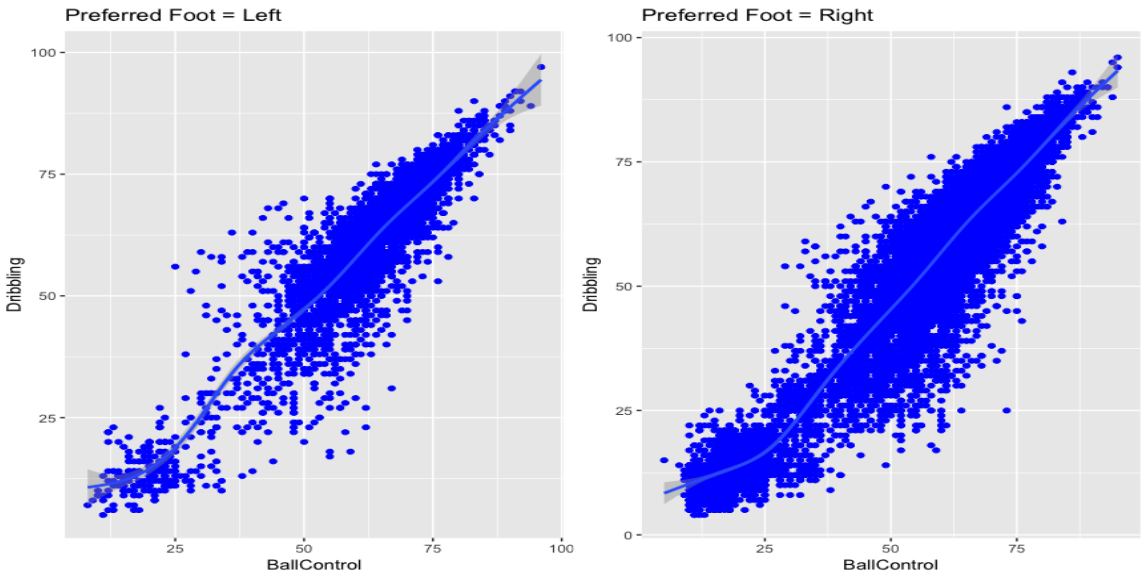
```
> cat("Top left-footed players:")
```

```
Top left-footed players:
```

```
> df_left <- df_fifa19[which(df_fifa19["Preferred.Foot"] == "Left"), ]
> head(df_left[,c("Name","Overall")])
```

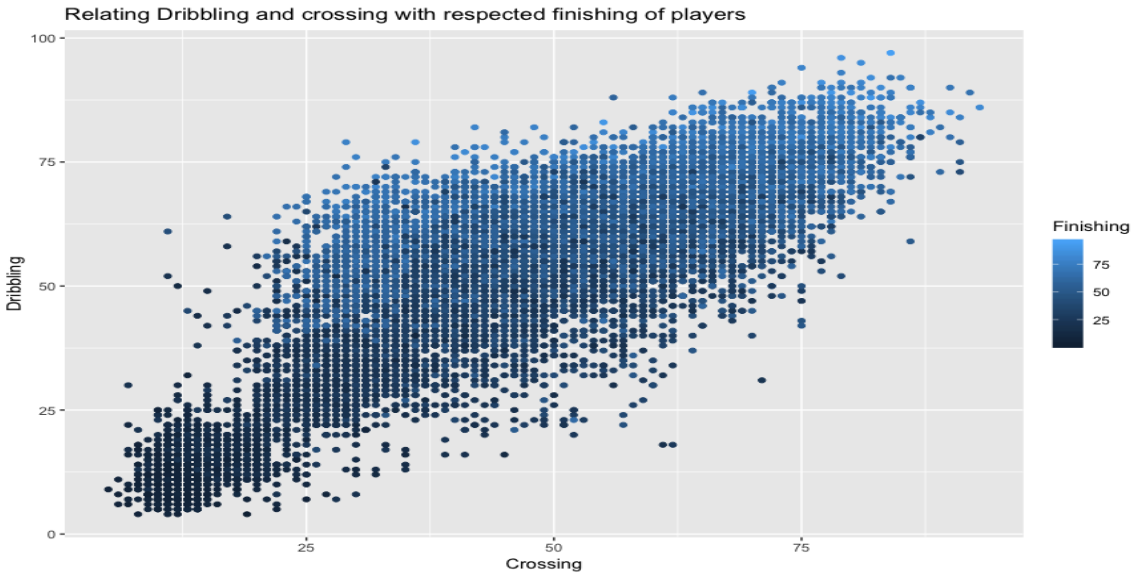
```
           Name Overall
1      L. Messi      94
14  David Silva      90
16    P. Dybala      89
18 A. Griezmann      89
20  T. Courtois      89
25 G. Chiellini      89
```

```
> left <- ggplot(data=df_left, aes(x=df_left$BallControl,y=df_left$Dribbling)) +
+   geom_point(color="blue") + xlab("BallControl") + ylab("Dribbling") +
+   geom_smooth() + ggtitle("Preferred Foot = Left")
> right <- ggplot(data=df_right, aes(x=df_right$BallControl,y=df_right$Dribbling)) +
+   geom_point(color="blue") + xlab("BallControl") + ylab("Dribbling") +
+   geom_smooth() + ggtitle("Preferred Foot = Right")
> ggarrange(left,right,nrow=1,ncol=2)
> drib_cross<- ggplot(data=df_fifa19, aes(x=df_fifa19$Crossing,y=df_fifa19$Dribbling,color=P
+   geom_point() + xlab("Crossing") + ylab("Dribbling") +
+   ggtitle("Relating Dribbling and crossing with respected finishing of players")
>
```




```
> cat("Dribbling vs Crossing with respect to finishing:")
```

```
Dribbling vs Crossing with respect to finishing:
```



h. Analysis-8

Snapshots of Shiny app made for displaying multiple plots like Age Stamina, Age Potential, Age Agility and Age SprintSpeed