Analysis of FIFA 22 Player's Dataset

Atharva Jadhav
Data Science
Illinois Institute of Technology
ajadhav16@hawk.iit.edu

Harshali Gaikwad
Data Science
Illinois Institute of Technology
hgaikwad@hawk.iit.edu

Sriujan Harihar
Data Science
Illinois Institute of Technology
sharihar1@hawk.iit.edu

Illinois Institute of Technology
CSP571-Data Preparation and Analysis
Professor: Jawahar Panchal

Table of Contents

1. ABSTRACT

This study presents a comprehensive analysis of player performance in FIFA 22, a popular football simulation video game. Leveraging a dataset encompassing diverse player attributes and in-game statistics, we employ advanced analytical techniques to uncover patterns, trends, and correlations within the virtual football landscape. Our investigation delves into key aspects such as player ratings, skill moves, preferred positions, and overall effectiveness on the virtual pitch. By applying statistical methods and machine learning algorithms, we aim to provide valuable insights into the dynamics of player performance within the FIFA 22 gaming environment. The findings of this analysis contribute to a deeper understanding of virtual football strategies and comparing the common statistics between the male and female football players following multiple model predictions.

Keywords. Football, statistical models, predictions, Exploratory Data Analysis (EDA)

2. INTRODUCTION

In the realm of virtual football, FIFA 22 stands as a prominent simulation video game that captivates millions of players worldwide. This study embarks on a comprehensive exploration of player performance within this dynamic gaming environment. Drawing upon an extensive dataset that encapsulates a myriad of player attributes and in-game statistics, we employ sophisticated analytical techniques to unravel intricate patterns, discern trends, and identify correlations. Our investigation delves into crucial facets of the virtual football landscape, including player ratings, skill moves, preferred positions, and overall effectiveness on the digital pitch. Through the application of statistical methods and machine learning algorithms, our primary objective is to furnish valuable insights into the nuanced dynamics of player performance within the FIFA 22 gaming realm.

As an extension of our inquiry, we extend our analysis to the realm of gender, comparing common statistics between male and female football players through multiple model predictions, thus enriching our understanding of virtual football strategies across diverse player profiles.

3. PROPOSED METHODOLOGY

The steps for analysis and prediction involve understanding the data, exploring the data, investigating the patterns and creating plots and graphs. The initial analysis involves cleaning the data. This step includes exploring the datatypes of the data features, exploring the dependent and independent variables. We studied the data and found that there are NULL data values which clearly increased the complexity for the models to interpret the data and pose difficulty for the analysis as well. The huge volume of data packed a lot of missing values and empty cells which increased the complexity of analysis. We cleaned the data by some methods like replacing the missing data with the mean values, removing unwanted columns from the data. After cleaning and wrangling the data, we implemented different

types of models and calculated the accuracy of each model. We implemented the following models, since the data is quantitative we implemented Linear Regression Model, followed by Linear Discriminant Analysis, Random Forest and then K-Nearest Neighbor Model. We used formulas to calculate the accuracy of each model. We have used Confusion Matrix, Mean Standard Error, R-squared value to find the efficiency and accuracy of the model.

4. DATA

4.1 Data Properties

The data that we used in this project are collected from Kaggle, https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset. This data set consists of data over years from 2017 to 2022. Each of the dataset consists of over 15,000 records for the male players and over 200 records for the female players. The dataset that we are going to analyze and evaluate consists of over 19,000 records for male and over 300 records for females. Moreover the number of features/predictors in each of this dataset is 110, i.e. p=110.

Let's have a look at the dataset for the male and female players. Following is the male players dataset.

```
str(fifaMale)
```

```
## 'data.frame':    19239 obs. of  110 variables:
##  $ sofifa_id                 : int  158023 188545 20801 190871 192985 200389 231747 167495 192448 202126 ...
##  $ player_url                : chr  "https://sofifa.com/player/158023/lionel-messi/220002" "https://sofifa.co
m/player/188545/robert-lewandowski/220002" "https://sofifa.com/player/20801/c-ronaldo-dos-santos-aveiro/220002" "h
ttps://sofifa.com/player/190871/neymar-da-silva-santos-jr/220002" ...
##  $ short_name                : chr  "L. Messi" "R. Lewandowski" "Cristiano Ronaldo" "Neymar Jr" ...
##  $ long_name                 : chr  "Lionel Andrés Messi Cuccittini" "Robert Lewandowski" "Cristiano Ronaldo d
os Santos Aveiro" "Neymar da Silva Santos Júnior" ...
##  $ player_positions          : chr  "RW, ST, CF" "ST" "ST, LW" "LW, CAM" ...
##  $ overall                   : int  93 92 91 91 91 91 90 90 90 ...
##  $ potential                 : int  93 92 91 91 91 93 95 90 92 90 ...
##  $ value_eur                 : num  7.80e+07 1.20e+08 4.50e+07 1.29e+08 1.26e+08 ...
##  $ wage_eur                  : num  320000 270000 270000 270000 350000 130000 230000 86000 250000 240000 ...
##  $ age                       : int  34 32 36 29 30 28 22 35 29 27 ...
##  $ dob                       : chr  "1987-06-24" "1988-08-21" "1985-02-05" "1992-02-05" ...
##  $ height_cm                 : int  170 185 187 175 181 188 182 193 187 188 ...
##  $ weight_kg                 : int  72 81 83 68 70 87 73 93 85 89 ...
##  $ club_team_id              : num  73 21 11 73 10 240 73 21 241 18 ...
##  $ club_name                 : chr  "Paris Saint-Germain" "FC Bayern München" "Manchester United" "Paris Saint
-Germain" ...
##  $ league_name               : chr  "French Ligue 1" "German 1. Bundesliga" "English Premier League" "French L
igue 1" ...
##  $ league_level              : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ club_position             : chr  "RW" "ST" "ST" "LW" ...
##  $ club_jersey_number        : int  30 9 7 10 17 13 7 1 1 10 ...
##  $ club_loaned_from          : chr  "" "" "" "" ...
##  $ club_joined               : chr  "2021-08-10" "2014-07-01" "2021-08-27" "2017-08-03" ...
##  $ club_contract_valid_until : int  2023 2023 2023 2025 2025 2023 2022 2023 2025 2024 ...
##  $ nationality_id            : int  52 37 38 54 7 44 18 21 21 14 ...
##  $ nationality_name          : chr  "Argentina" "Poland" "Portugal" "Brazil" ...
##  $ nation_team_id            : num  1369 1353 1354 NA 1325 ...
##  $ nation_position           : chr  "RW" "RS" "ST" "" ...
##  $ nation_jersey_number      : int  10 9 7 NA 7 NA 10 1 NA 9 ...
##  $ preferred_foot            : chr  "Left" "Right" "Right" "Right" ...
##  $ weak_foot                 : int  4 4 4 5 5 3 4 4 4 5 ...
##  $ skill_moves               : int  4 4 5 5 4 1 5 1 1 3 ...
```

Figure 4a: Male players Dataset

The dataset of female players also has a similar structure as that of the male players dataset. Following are

some of the records.

```
head(fifaFemale,5)
```

```
##   sofifa_id                                          player_url short_name
## 1    227246     https://sofifa.com/player/227246/lucy-bronze/220002  L. Bronze
## 2    227316    https://sofifa.com/player/227316/wendie-renard/220002  W. Renard
## 3    233746 https://sofifa.com/player/233746/vivianne-miedema/220002 V. Miedema
## 4    227125    https://sofifa.com/player/227125/samantha-kerr/220002    S. Kerr
## 5    226301     https://sofifa.com/player/226301/alex-morgan/220002  A. Morgan
##                   long_name player_positions overall potential value_eur
## 1 Lucia Roberta Tough Bronze           RB, CM      92        92        NA
## 2  Wéndèleine Thérèse Renard               CB      92        92        NA
## 3           Vivianne Miedema               ST      92        93        NA
## 4          Samantha May Kerr           ST, LW      91        91        NA
## 5  Alexandra Morgan Carrasco               ST      90        90        NA
##   wage_eur age        dob height_cm weight_kg club_team_id club_name
## 1       NA  29 1991-10-28       171        67           NA        NA
## 2       NA  30 1990-07-20       187        70           NA        NA
## 3       NA  24 1996-07-15       178        65           NA        NA
## 4       NA  27 1993-09-10       167        55           NA        NA
## 5       NA  31 1989-07-02       170        62           NA        NA
##   league_name league_level club_position club_jersey_number club_loaned_from
## 1          NA           NA            NA                 NA               NA
## 2          NA           NA            NA                 NA               NA
## 3          NA           NA            NA                 NA               NA
## 4          NA           NA            NA                 NA               NA
## 5          NA           NA            NA                 NA               NA
##   club_joined club_contract_valid_until nationality_id nationality_name
## 1          NA                       NA             14          England
## 2          NA                       NA             18           France
## 3          NA                       NA             34      Netherlands
## 4          NA                       NA            195        Australia
## 5          NA                       NA             95    United States
##   nation_team_id nation_position nation_jersey_number preferred_foot weak_foot
## 1         113002             RB                    2          Right         3
## 2         113003            LCB                    3          Right         2
## 3         113011             ST                    9          Right         4
## 4         112998             ST                   20          Right         4
## 5         113009             ST                   13           Left         4
```

Figure 4b: sample data from the female data set

4.2 Data Pre-processing, Cleaning and Wrangling

After the initial analysis of the dataset we found that there are a total of 110 variables/features. However, the data is not perfect, there are a lot of missing values. Both the male and female data sets have a lot of missing values. Moreover, some of the columns consist of hefty missing values making them irrelevant for the analysis. Parameters like goalkeeping_speed have a lot of missing values and are irrelevant for the analysis, so we remove this column from the data set. The next step is cleaning the data by removing the unwanted columns, replacing the missing values and splitting the data set into 3 small datasets which further easies the analysis of the data.

```
names(fifaFemale)
```

```
##   [1] "sofifa_id"                "player_url"
##   [3] "short_name"               "long_name"
##   [5] "player_positions"         "overall"
##   [7] "potential"                "value_eur"
##   [9] "wage_eur"                 "age"
##  [11] "dob"                      "height_cm"
##  [13] "weight_kg"                "club_team_id"
##  [15] "club_name"                "league_name"
##  [17] "league_level"             "club_position"
##  [19] "club_jersey_number"       "club_loaned_from"
##  [21] "club_joined"              "club_contract_valid_until"
##  [23] "nationality_id"           "nationality_name"
##  [25] "nation_team_id"           "nation_position"
##  [27] "nation_jersey_number"     "preferred_foot"
##  [29] "weak_foot"                "skill_moves"
##  [31] "international_reputation" "work_rate"
##  [33] "body_type"                "real_face"
##  [35] "release_clause_eur"       "player_tags"
##  [37] "player_traits"            "pace"
##  [39] "shooting"                 "passing"
##  [41] "dribbling"                "defending"
##  [43] "physic"                   "attacking_crossing"
##  [45] "attacking_finishing"      "attacking_heading_accuracy"
##  [47] "attacking_short_passing"  "attacking_volleys"
##  [49] "skill_dribbling"          "skill_curve"
##  [51] "skill_fk_accuracy"        "skill_long_passing"
##  [53] "skill_ball_control"       "movement_acceleration"
##  [55] "movement_sprint_speed"    "movement_agility"
##  [57] "movement_reactions"       "movement_balance"
##  [59] "power_shot_power"         "power_jumping"
##  [61] "power_stamina"            "power_strength"
##  [63] "power_long_shots"         "mentality_aggression"
##  [65] "mentality_interceptions"  "mentality_positioning"
##  [67] "mentality_vision"         "mentality_penalties"
##  [69] "mentality_composure"      "defending_marking_awareness"
##  [71] "defending_standing_tackle" "defending_sliding_tackle"
##  [73] "goalkeeping_diving"       "goalkeeping_handling"
##  [75] "goalkeeping_kicking"      "goalkeeping_positioning"
##  [77] "goalkeeping_reflexes"     "goalkeeping_speed"
##  [79] "ls"                       "st"
##  [81] "rs"                       "lw"
##  [83] "lf"                       "cf"
##  [85] "rf"                       "rw"
##  [87] "lam"                      "cam"
##  [89] "ram"                      "lm"
##  [91] "lcm"                      "cm"
##  [93] "rcm"                      "rm"
##  [95] "lwb"                      "ldm"
##  [97] "cdm"                      "rdm"
##  [99] "rwb"                      "lb"
## [101] "lcb"                      "cb"
## [103] "rcb"                      "rb"
## [105] "gk"                       "player_face_url"
## [107] "club_logo_url"            "club_flag_url"
## [109] "nation_logo_url"          "nation_flag_url"
```

Figure 4c: All of the features/predictors in the dataset

The following is the data after the cleaning and wrangling of the data set.

```
fifaFemale <- data.frame(fifaFemale)
fifaFemaleS1 <- subset(fifaFemale, select = c(sofifa_id, short_name, long_name, player_positions, overall, potenti
al, age, dob, height_cm, weight_kg, nationality_name, nation_jersey_number, body_type))

fifaFemaleS2 <- subset(fifaFemale, select = c(sofifa_id, overall, wage_eur, value_eur, preferred_foot, weak_foot,
skill_moves, pace, shooting, passing, dribbling, defending, physic, attacking_crossing, attacking_finishing, attac
king_heading_accuracy, attacking_short_passing, attacking_volleys, skill_dribbling, skill_curve, skill_fk_accurac
y, skill_long_passing, skill_ball_control, movement_acceleration, movement_sprint_speed, movement_agility, movemen
t_reactions, movement_balance, power_shot_power, power_jumping, power_stamina, power_strength, power_long_shots, m
entality_aggression, mentality_interceptions, mentality_positioning, mentality_vision, mentality_penalties, mental
ity_composure, defending_marking_awareness, defending_standing_tackle, defending_sliding_tackle, goalkeeping_divin
g, goalkeeping_handling, goalkeeping_kicking, goalkeeping_positioning, goalkeeping_reflexes, goalkeeping_speed))

fifaFemaleS3 <- subset(fifaFemale, select = c(sofifa_id, ls, st, rs, lw, lf, cf, rf, rw, lam, cam, ram, lm, lcm, c
m, rcm, rm, lwb, ldm, cdm, rdm, rwb, lb, lcb, cb, rcb, rb, gk))

fifaFemaleS2$preferred_foot_new <- ifelse(fifaFemaleS2$preferred_foot == "Right", 1, ifelse(fifaFemaleS2$preferred
_foot == "Left", 2, 0))
fifaFemaleS2$preferred_foot_new <- factor(fifaFemaleS2$preferred_foot_new, levels = c(1, 2))

head(fifaMale,5)
```

Figure 4d: Cleaning, splitting and preparing the dataset

```
##   sofifa_id                                                player_url
## 1    158023              https://sofifa.com/player/158023/lionel-messi/220002
## 2    188545         https://sofifa.com/player/188545/robert-lewandowski/220002
## 3     20801 https://sofifa.com/player/20801/c-ronaldo-dos-santos-aveiro/220002
## 4    190871  https://sofifa.com/player/190871/neymar-da-silva-santos-jr/220002
## 5    192985            https://sofifa.com/player/192985/kevin-de-bruyne/220002
##          short_name                          long_name player_positions
## 1          L. Messi      Lionel Andrés Messi Cuccittini       RW, ST, CF
## 2    R. Lewandowski                 Robert Lewandowski               ST
## 3 Cristiano Ronaldo Cristiano Ronaldo dos Santos Aveiro           ST, LW
## 4         Neymar Jr       Neymar da Silva Santos Júnior          LW, CAM
## 5      K. De Bruyne                    Kevin De Bruyne          CM, CAM
##   overall potential value_eur wage_eur age        dob height_cm weight_kg
## 1      93        93  78000000   320000  34 1987-06-24       170        72
## 2      92        92 119500000   270000  32 1988-08-21       185        81
## 3      91        91  45000000   270000  36 1985-02-05       187        83
## 4      91        91 129000000   270000  29 1992-02-05       175        68
## 5      91        91 125500000   350000  30 1991-06-28       181        70
##   club_team_id           club_name            league_name league_level
## 1           73 Paris Saint-Germain         French Ligue 1            1
## 2           21   FC Bayern München   German 1. Bundesliga            1
## 3           11   Manchester United English Premier League            1
## 4           73 Paris Saint-Germain         French Ligue 1            1
## 5           10     Manchester City English Premier League            1
##   club_position club_jersey_number club_loaned_from club_joined
## 1            RW                30                   2021-08-10
## 2            ST                 9                   2014-07-01
## 3            ST                 7                   2021-08-27
## 4            LW                10                   2017-08-03
## 5           RCM                17                   2015-08-30
```

Figure 4e: The data set after the data cleaning and wrangling

4.3 Problem Statement

What this project seeks to address

- What is the diverse range of ages of the players?
- What is the distribution of the player's potential?
- What is the distribution of the player's overall attribute?
- Which country produces the best players?
- What is the distribution of the player wages?
- Make a prediction on what will be the next overall attribute.

5. EXPLORATORY DATA ANALYSIS

We have used the ggplot library to plot the graphs based on the cleaned dataset to generate graphs and analyze the data.

Boxplot of Player Attributes

Figure 5(a) shows us the plot for the boxplot of player attributes. Notice that some of the attributes have a lot of outliers which indicates that these attributes can be dropped and not used for the model predictions to increase the model effectiveness.

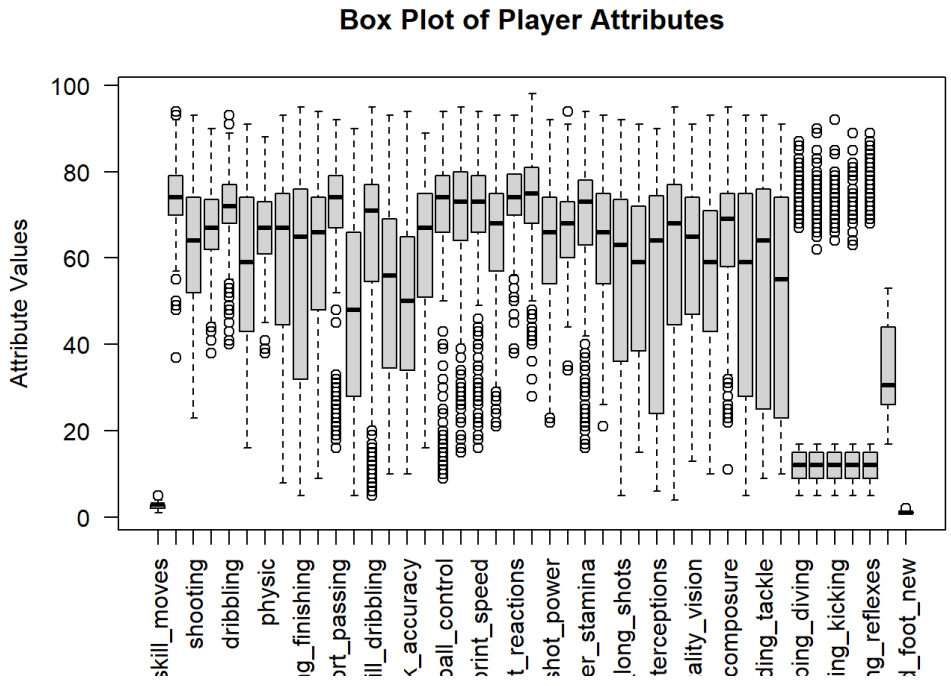

Figure 5a: Box Plot of Player Attributes

Distribution of the ages of the players

Figure 5(b) and 5(c) shows the distribution of players according to the ages. From the graphs it is clear that the male players play up to a higher age limit than that of the the female players

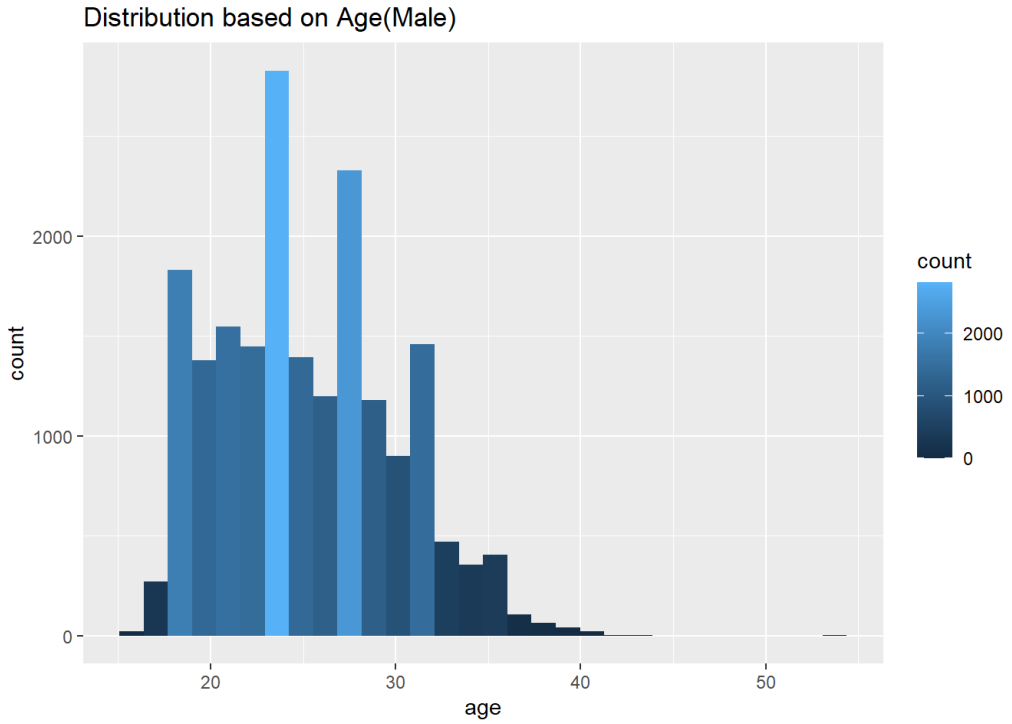

Figure 5b: Distribution based on age(male)

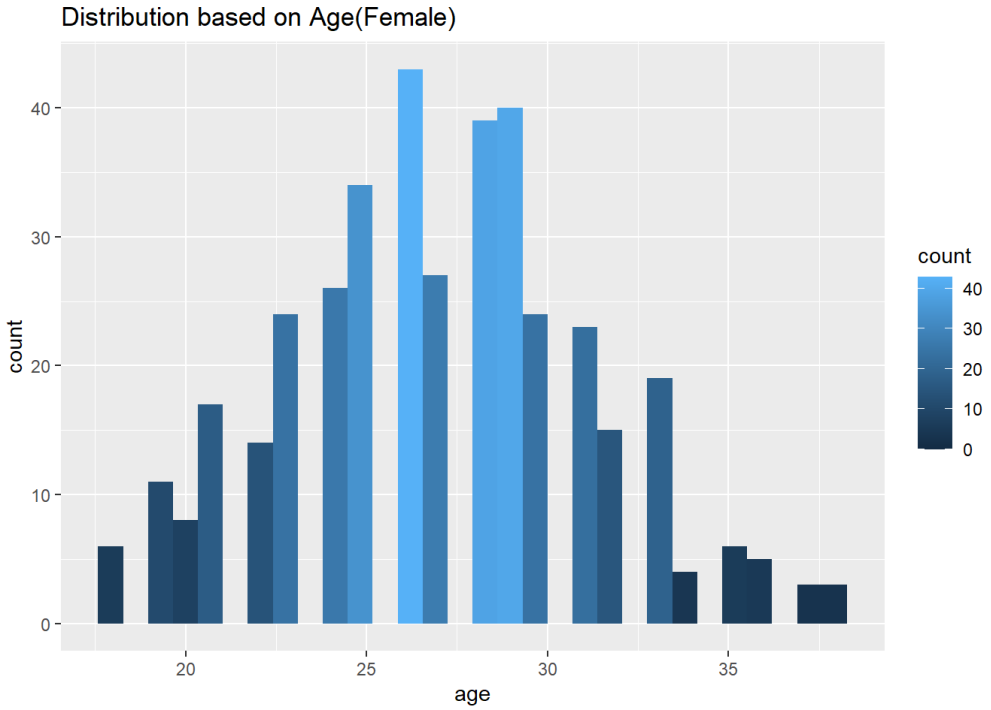

Figure 5c: Distribution based on age(female)

Distribution of the players based on the Overall Rating

Figure 5(d) and 5(e) show us the distribution of players based on the overall rating. The graph is the player overall rating vs the count of the players.

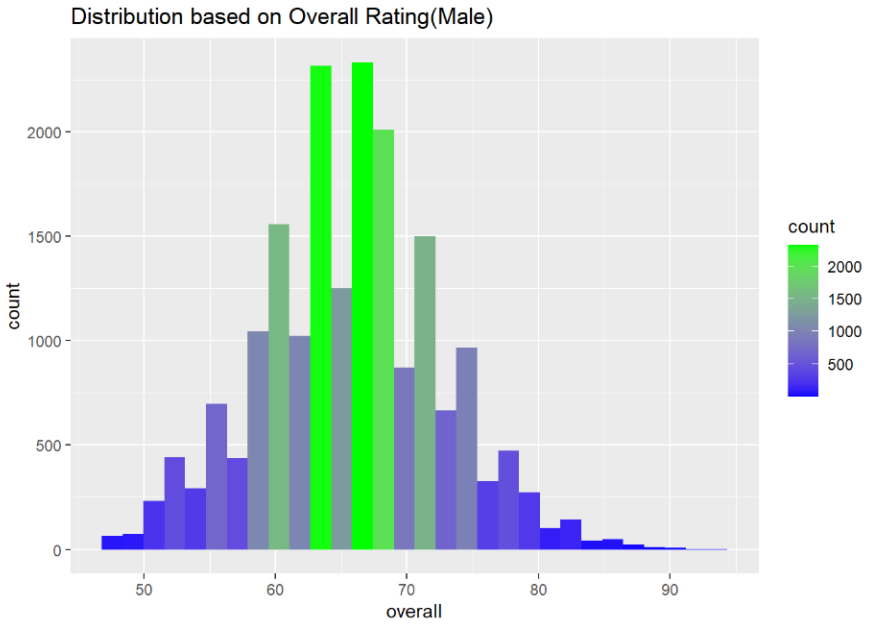

Figure 5d: Distribution of players based on overall ratings male

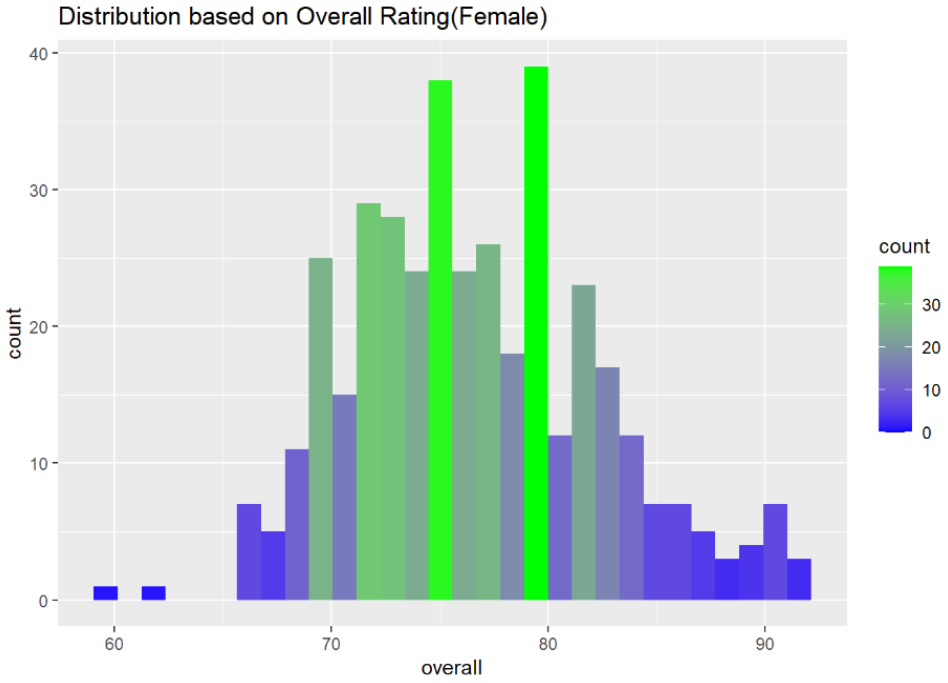

Figure 5e: Distribution of players based on overall ratings female

Distribution of the players based on their Potential Rating

Figure 5(f) and 5(g) show us the distribution of players based on the potential rating. The graph is the player potential rating vs the count of the players.

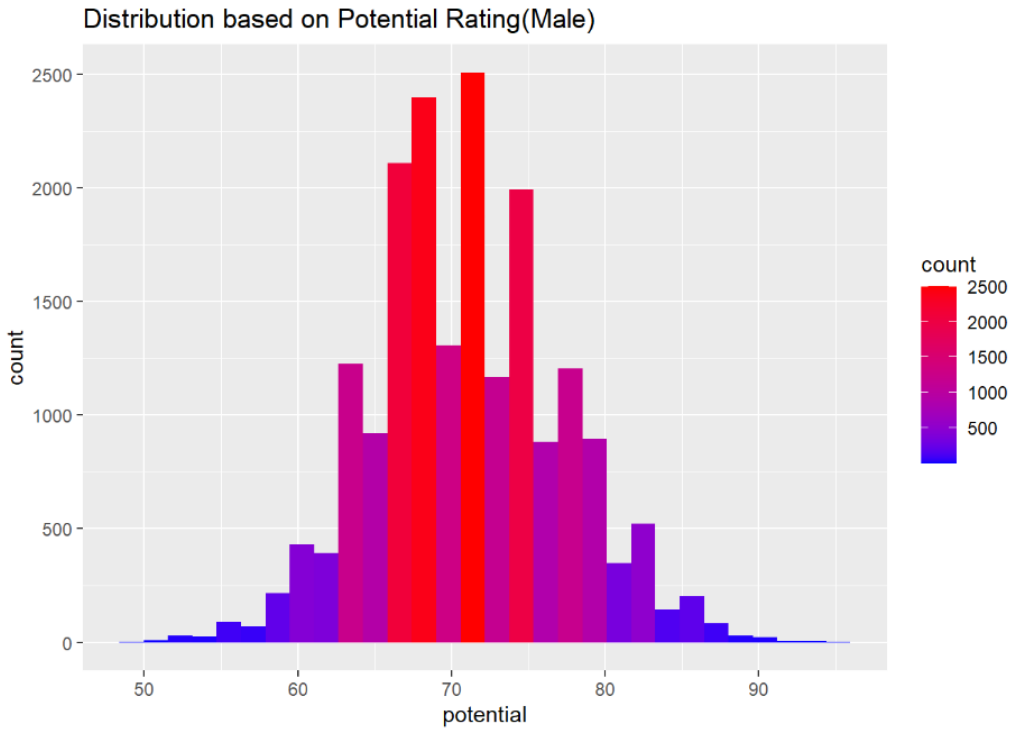

Figure 5f: Distribution of players based on potential ratings male

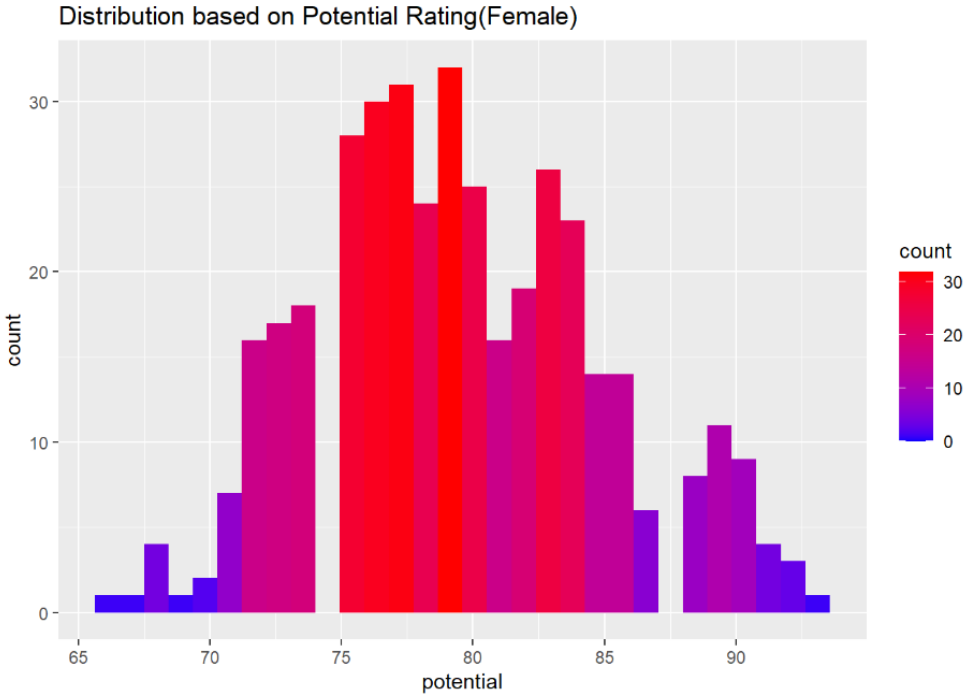

Figure 5g: Distribution of players based on potential ratings female

Top Countries producing best talent

Figure 5(h) and 5(i) shows us the top players and their nationality. For the male players the best country producing talent is England while that of female players is China according to the data.

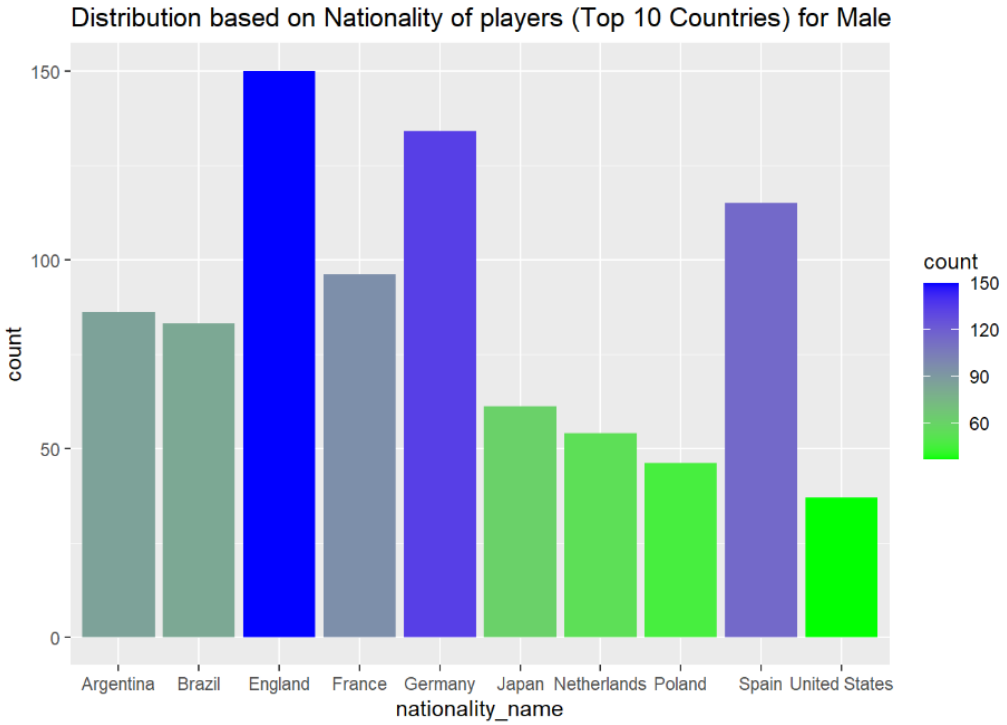

Figure 5h: Distribution based on nationality for males

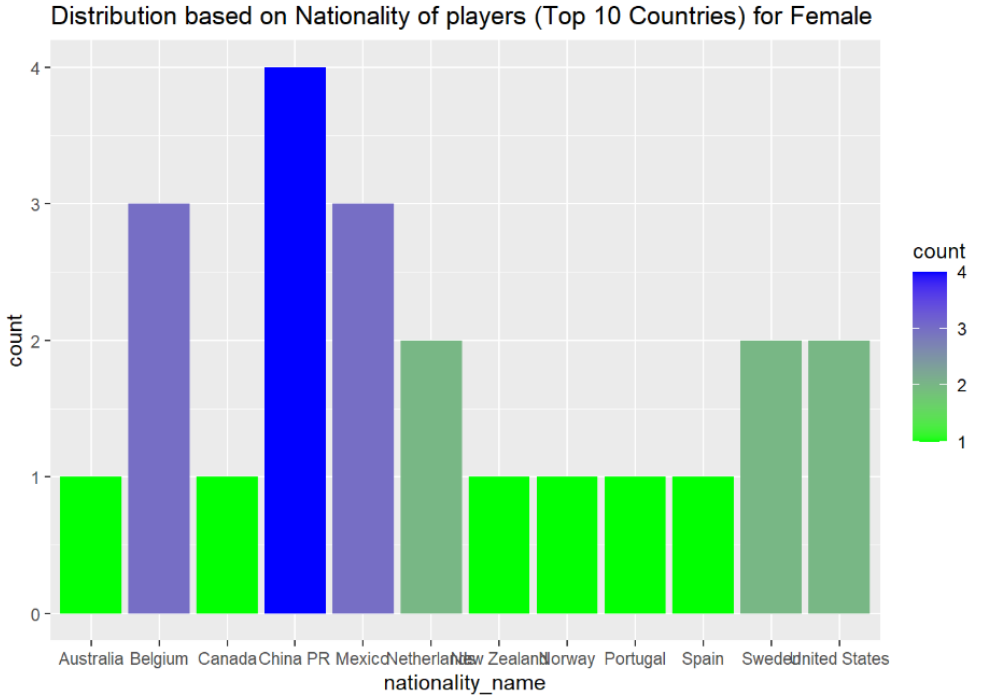

Figure 5i: Distribution based on nationality for females

Distribution of the top wages

The following figure shows us the distribution of the top wages of the elite players.

```
top_1_percent_wage_male  <- quantile(fifaMaleS2$wage_eur, probs=0.99)
filtered_wage_male <- filter(fifaMaleS2, wage_eur > top_1_percent_wage_male)

g_value_male <- ggplot(filtered_wage_male, aes(wage_eur))
g_value_male + geom_histogram(aes(fill=..count..)) + ggtitle("Distribution of top 1% wage(Male)")
```

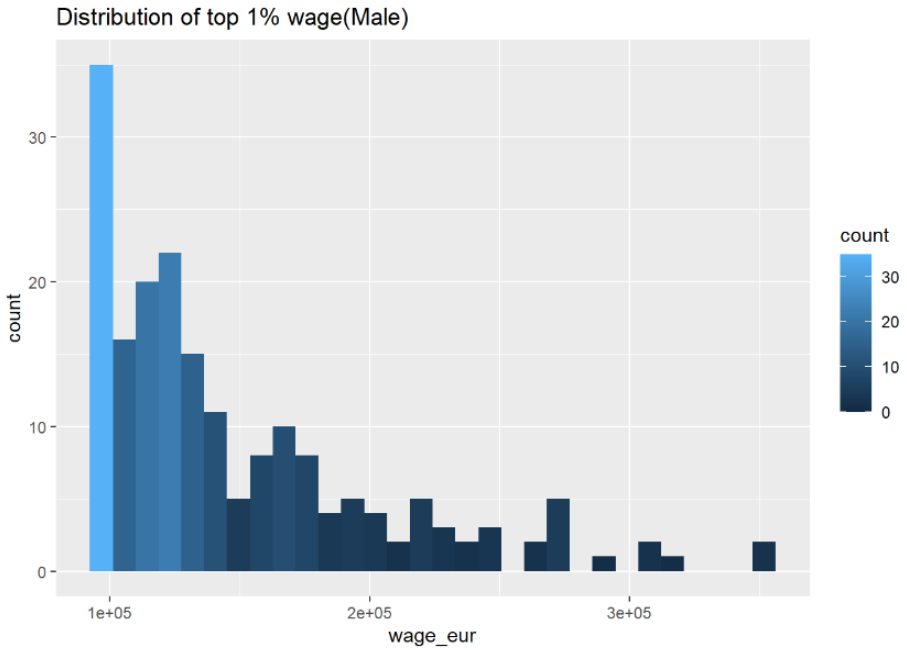

Figure 5j: Distribution of the top wages

Top 10 most valuable players

Figure 5(k) shows us the top 10 most valuable players in the data set. We can see that the average price of each player in this graph is more than 100 million euros. Kylian Mbappe stands out with the highest market value followed by Erling Haaland. It is very intriguing to know that both of them play as Forwards.

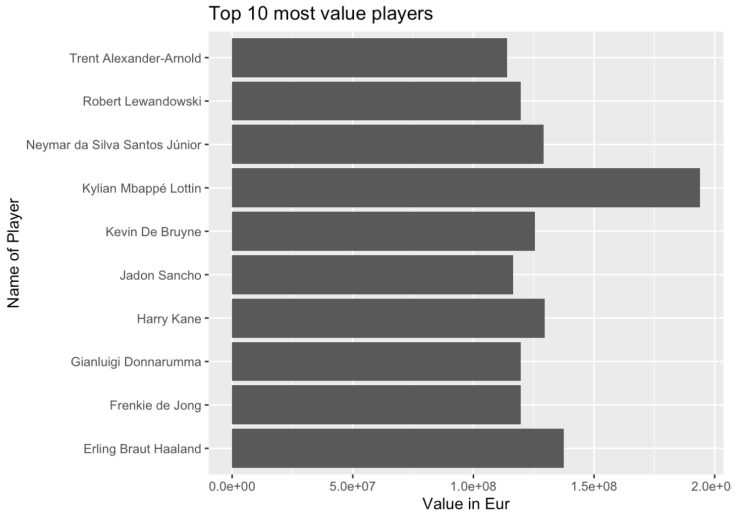

Figure 5(k): Top 10 most valuable players

Age vs Overall rating

Figure 5(l) is a very important graph as it shows us the dependent variable Overall against the age of the players. From the scatterplot it is clear that most of the top players lie between the age of 23 to 33. Moreover, as the age increases above 35 we gradually see a decrease in the quality of the players.

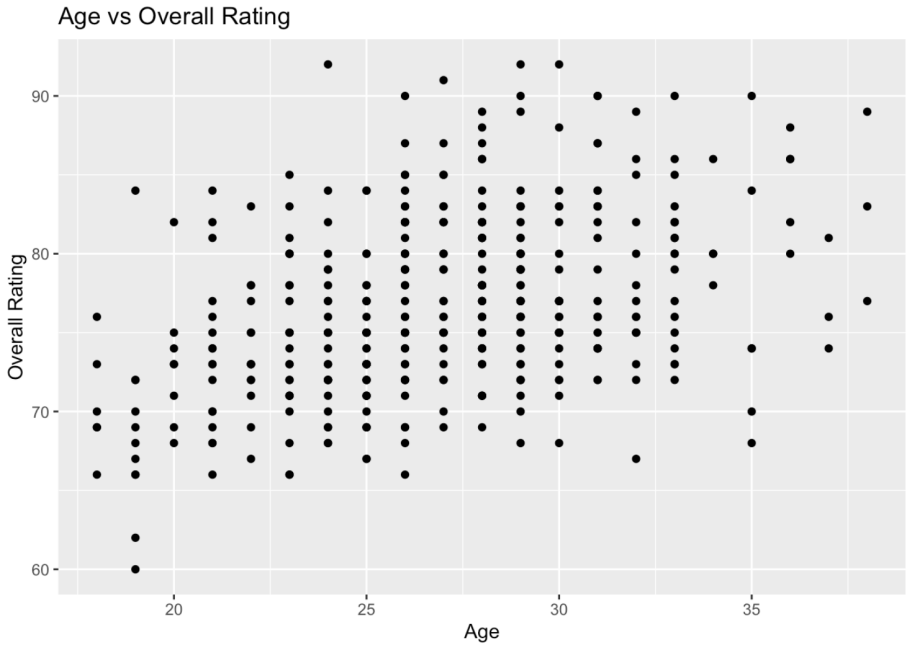

Figure 5l: Age vs Overall Rating

Top 10 FIFA players

This is a summary table showing the best players from the dataset. We notice that Lionel Messi comes out to be on the top with an overall rating of '93' followed by Robert Lewandowski at '92' and Cristiano Ronaldo at '91'. It is shocking to know that these top 3 players are aged 32 and above yet outshine young and more valuable players.

Top 10 FIFA Players

short_name	age	overall
L. Messi	34	93
R. Lewandowski	32	92
Cristiano Ronaldo	36	91
Neymar Jr	29	91
K. De Bruyne	30	91
J. Oblak	28	91
K. Mbappé	22	91
M. Neuer	35	90
M. ter Stegen	29	90
H. Kane	27	90

Figure 5m: Table of top 10 FIFA players

6. MODELING AND ANALYSIS

6.1 Correlations

Following is the correlation plot of all the features with the dependent variable **"Overall"**. Since there are more than 100 features considered for the plot the graph's compactness is justified. From the

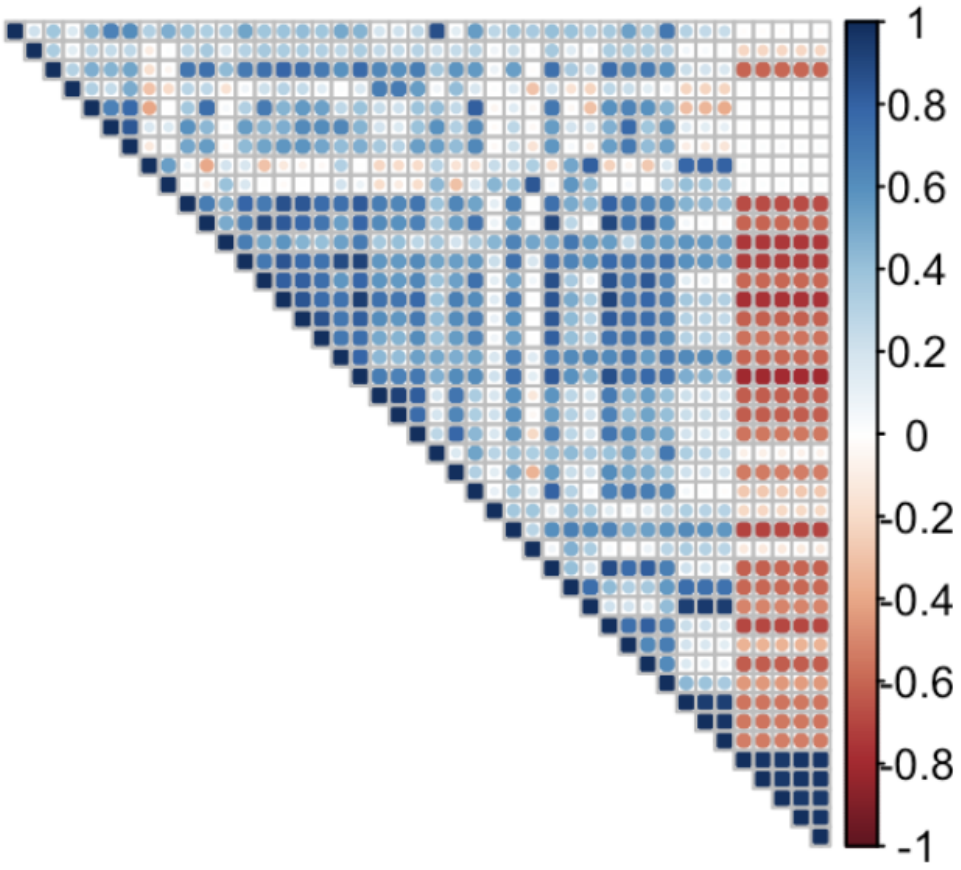

Figure 6a: Correlation plot

6.2 Train Test Split

For applying the different models we first need to split our data into training and testing sets. The split we have used is 80% and 20%. The training data set is trainMale and the testing data set is testMale. Following is the code for splitting data using the caret package.

```
index <- createDataPartition(fifaMaleS2$overall, times = 1,p=0.8,list = FALSE)
trainMale <- fifaMaleS2[index,]
testMale <- fifaMaleS2[-index,]
```

```
## 'data.frame':    15392 obs. of  49 variables:
## 'data.frame':    3847 obs. of  49 variables:
```

6.3 MODELING

6.3.1 Linear Regression

We first used the traditional linear regression model to predict the dependent variable. The linear regression model evaluation is done calculating the mean squared error and the r-squared value.

```
linearReg_overall <- lm(overall ~ weak_foot + skill_moves + pace + shooting + passing + dribbling + defending + ph
ysic + attacking_crossing + attacking_finishing + attacking_heading_accuracy + attacking_short_passing + attacking
_volleys + skill_dribbling +skill_curve + skill_fk_accuracy + skill_long_passing + skill_ball_control + movement_a
cceleration + movement_sprint_speed + movement_agility + movement_reactions + movement_balance + power_shot_power
+ power_jumping + power_stamina + power_strength + power_long_shots + mentality_aggression + mentality_interceptio
ns + mentality_positioning + mentality_vision + mentality_penalties + mentality_composure + defending_marking_awar
eness + defending_standing_tackle + defending_sliding_tackle + goalkeeping_diving + goalkeeping_handling + goalkee
ping_kicking + goalkeeping_positioning + goalkeeping_reflexes + preferred_foot_new, data = trainMale)


predictLinearRegression <- predict(linearReg_overall, testMale)

actual_values <- testMale$overall

mse <- mean((actual_values - predictLinearRegression)^2)

mean_actual <- mean(actual_values)
ss_total <- sum((actual_values - mean_actual)^2)
ss_residual <- sum((actual_values - predictLinearRegression)^2)
r_squared <- 1 - (ss_residual / ss_total)

# Print the results
cat("Mean Squared Error (MSE):", mse, "\n")
```

```
## Mean Squared Error (MSE): 4.891431
```

```
cat("R-squared (R2):", r_squared, "\n")
```

```
## R-squared (R2): 0.8940189
```

Figure 6b: Linear regression model

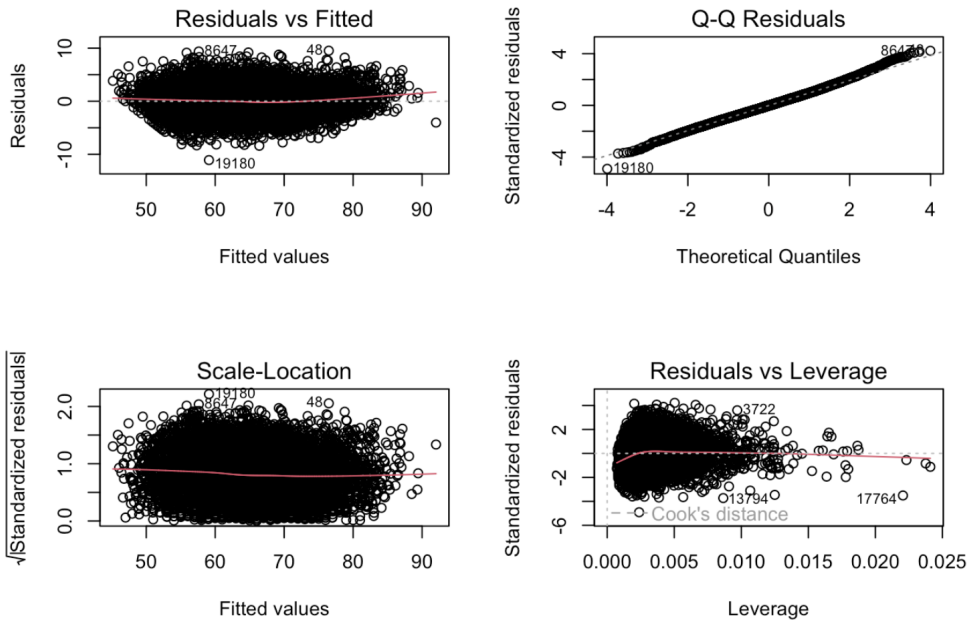

Figure 6c: Plot for Linear Regression Model

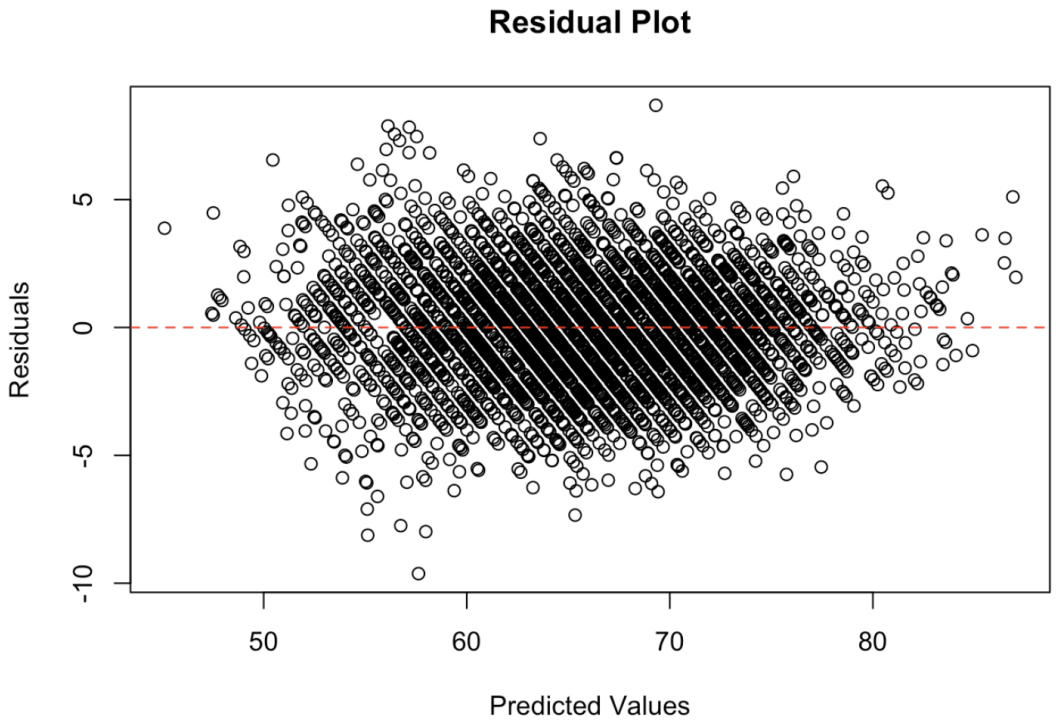

Figure 6d: Residual Plot for the Linear Regression Model

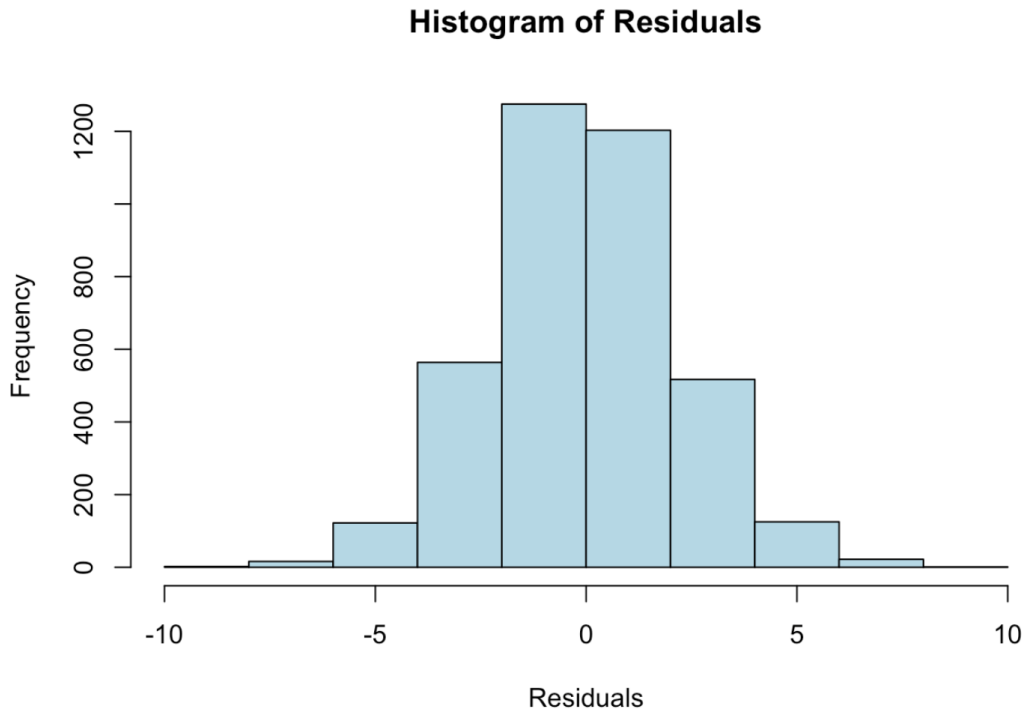

Figure 6e: Histogram of Residuals for Linear Regression Models

6.3.2 Linear Discriminant Analysis

In order to improve the results we decided to give different models an approach and tried the linear discriminant analysis method and predict the dependent variable. The accuracy obtained using the linear discriminant analysis is 56%.

```
modelLDA <- lda(overall ~ weak_foot + skill_moves + pace + shooting + passing
+ dribbling + defending + physic + attacking_crossing + attacking_finishing +
attacking_heading_accuracy + attacking_short_passing + attacking_volleys +
skill_dribbling +skill_curve + skill_fk_accuracy + skill_long_passing +
skill_ball_control + movement_acceleration + movement_sprint_speed +
movement_agility + movement_reactions + movement_balance + power_shot_power +
power_jumping + power_stamina + power_strength + power_long_shots +
```

```
mentality_aggression + mentality_interceptions + mentality_positioning +
mentality_vision + mentality_penalties + mentality_composure +
defending_marking_awareness + defending_standing_tackle +
defending_sliding_tackle + goalkeeping_diving + goalkeeping_handling +
goalkeeping_kicking + goalkeeping_positioning + goalkeeping_reflexes +
preferred_foot_new, data = trainMale)

predictions <- predict(modelLDA, newdata = testMale)
conf_matrix <- table(predictions$class, testMale$overall)
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
accuracy

## [1] 0.563764
```

Figure 6f: Linear Discriminant Analysis

6.3.3 K-Nearest Neighbors

The k-nearest neighbor model is the next technique we tried. With the value of k set to 5 we trained the model on the training dataset and evaluated it on the test data set. The model accuracy came out to be better than that of the linear discriminant analysis. The model accuracy is 67%.

```
k_neighbors <- 5

knn_model <- knn(train = trainMale[, predictors], test = testMale[,
predictors], cl = trainMale$overall, k = k_neighbors)

conf_matrix_knn <- table(knn_model, testMale$overall)
```

```
accuracy <- sum(knn_model == testMale$overall) / length(testMale$overall)

print(paste("Accuracy:", round(accuracy, 4)))

## [1] "Accuracy: 0.6791"
```

Figure 6g: K-Nearest Neighbor Model

6.3.4 Random Forest Model

The next model is the random forest model. This model produced the best result than that of the other models except that of the linear regression. Since the data set is quite huge and has high dimensionality, the random forest model performs much better than the other models. The high-dimensionality causes the KNN model to fail due to the curse of dimensionality which is not the case in case of random forest. The model accuracy comes out to be 74.65%.

```
rf_model <- randomForest(factor(overall) ~ ., data = trainMale[, c("overall",
predictors)], ntree = 500)

rf_predictions <- predict(rf_model, newdata = testMale[, predictors])

conf_matrix_rf <- table(rf_predictions, testMale$overall)
```

```
accuracy_rf <- sum(diag(conf_matrix_rf)) / sum(conf_matrix_rf)

print(paste("Random Forest Accuracy:", round(accuracy_rf, 4)))

## [1] "Random Forest Accuracy: 0.7465"
```

Figure 6h: Random Forest Model

6.4 Results Analysis:

We have experimented with 4 models and we have also captured the metric of model accuracy, as we can see, Random Forest has an accuracy of 74.65%. Then comes another model KNN (k-Nearest Neighbors) with approximately 68% accuracy. We also have Linear Discriminant Analysis with 56.3% accuracy and Linear Regression Model with R-squared value of 0.89. Out of all the 4 models, Linear Regression has the most variance explained and indicates that there is a strong relationship between the dependent and independent variables. A linear regression with high R-Squared value(0.89) and a low Mean Squared Error (MSE - 4.89) fits the data accurately. We have the least accuracy in the linear discriminant analysis since, the data we have is highly dimensional in nature and similar to KNN, LDA's performance also degrades due to the "Curse of Dimensionality", since we have a high number of parameters for each dataset.

7. CONCLUSION AND FUTURE WORK

This project investigated the 2022 FIFA Player dataset, extracting useful insights and evaluating the efficacy of four statistical models for predicting football players overall rating. Linear regression was the most accurate model, attaining 0.89 R-Squared value and 4.89 as the Mean Squared Error (MSE). But, we can see that Random Forest found skill moves and dribbling as the most significant predictors, rather than other models that relied on physic and movement reactions. The trials also demonstrated that KNN suffered from the "curse of dimensionality" issue, exposing its limits in this setting. Overall, the investigation shows that Random Forest is also successful at forecasting football player overall rating, with useful insights derived by evaluating Skill moves and dribbling. The findings also serve as a warning about the possible hazards of KNN in high-dimensional data processing.

In terms of future work, we can use other predictors like player performance in a particular position to predict what position the player should play in for maximum efficient results. Additionally, if we could use other techniques like Web Scraping to obtain useful information like Total Passes, Penalties statistics, maximum speed and so on, we might generate a far more efficient prediction model. Including goals and team data in the dataset would also allow us to generate better predictions based on top performers or players. Though loading large datasets will cost us, this can be explored in future development.

REFERENCES

[1] FIFA 23 Exploratory Data Analysis

https://medium.com/@seyiogar/fifa-23-exploratory-data-analysis-ed56ea424f48

[2] Football Insights from FIFA Datasets: Player Valuation

https://medium.com/@ofirmagdaci/football-insights-from-fifa-data-player-valuation-5 5b1b748e05d

[3] M. Burch, G. Wallner, S. L. Angelescu and P. Lakatos, "Visual Analysis of FIFA World Cup Data," 2020 24th International Conference Information Visualisation (IV), Melbourne, Australia, 2020, pp. 114-119, doi: 10.1109/IV51561.2020.00028.

[4] Clemente FM, Silva F, Martins FML, Kalamaras D, Mendes RS. Performance Analysis Tool for network analysis on team sports: A case study of FIFA Soccer World Cup

2014. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology. 2016;230(3):158-170. doi:10.1177/1754337115597335

[5] Cea, S., Durán, G., Guajardo, M. *et al.* An analytics approach to the FIFA ranking procedure and the World Cup final draw. *Ann Oper Res* 286, 119–146 (2020). https://doi.org/10.1007/s10479-019-03261-8.