**Statistical Analysis:**

**UEFA 2019 Quarter Final Teams**

****

**ANLY 502: Analytical Methods I**

**Under Guidance of**: **Prof. Dr. Scott Layden**

**Submitted by:** Prabath Wanaguru, Gaurav Gade, Yaqi Yang & Ketaki Kale

**Abstract**

This paper presents a statistical analysis of UEFA 2019 player’s data spread across various teams. This paper covers some statistical analysis like Exploratory Data Analysis, Hypothesis Testing, ANOVA, t-Test, Multi-Linear Model and The Best Fit Model using backward elimination. Using these techniques, we wanted to analyze correlation between Overall Rating of the player with their skills, wages and age. These factors help to understand the correlation amongst them and help to decide the strategy, which player to invest in for the next football championship. The objective of this paper is to derive relation between parameters and select younger players with high ratings and low wages; reducing our cost of investment.

1. **Introduction:**

After the recent UEFA Champions League held in June 2019, we decided to invest on best players to form a new team for FIFA 2020. Before we come up with our ideal team, we had many questions to be answered, for instance, efficient wages of the player, performance of the player, Age & physical fitness of the player, and so on. We followed the following steps for arriving at our sample data.

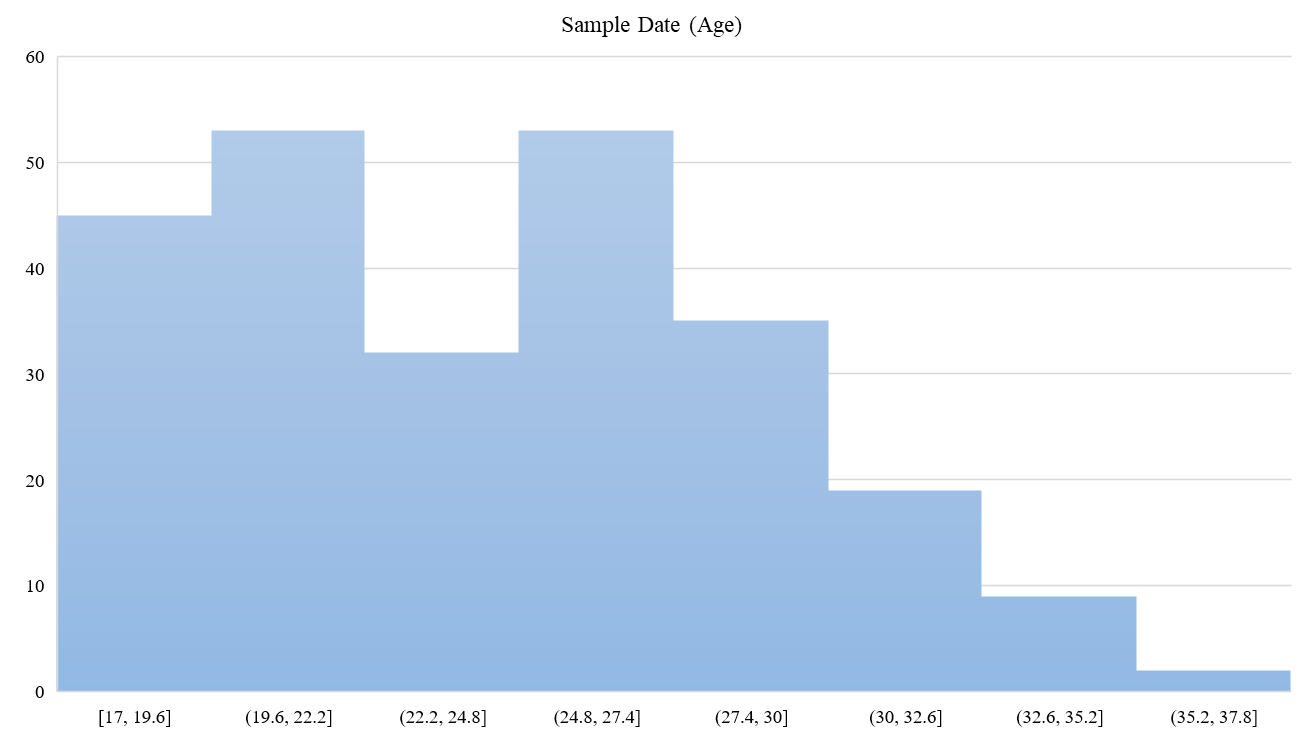
1. **Sourcing our data:** We found the required data set on [www.kaggle.com](http://www.kaggle.com) – ‘FIFA 2019 complete player dataset’. We considered this data set as this captured the most recent available information of the players, the clubs they represent, wages, physical fitness, and so on.
2. **Carving sample data:** The total data we obtained was for 18,206 players analyzed across various parameters for instance, age, nationality, leg strength, stamina, jump, skills, wage, club, investment, and so on. We decided to select the quarter finalist teams to arrive at a defined scope for analysis. There were 248 number of players selected across these eight teams, namely FC Porto, Ajax, Tottenham Hotspur, Liverpool, Juventus, Manchester United, Manchester City and FC Barcelona.
3. **Data Screening and Data Analysis:**

* **Data Dimensions**: Our sample data includes 248 rows across 12 variables.
* **Sample Data**: After understanding the available data set; we excluded missing information and outliers using Mahanopolis test in our statistical analysis.

1. **Statistical conclusions:** Lastly, our analysis will be concluded by our findings.
2. **Exploratory Analysis for Entire Data vs Sample Data:**

To understand data better and for easy decision making, we analyzed some of the parameters from both Overall Data and Quarter Final Sample Data, as below:

**Age - Sample Data:**

****

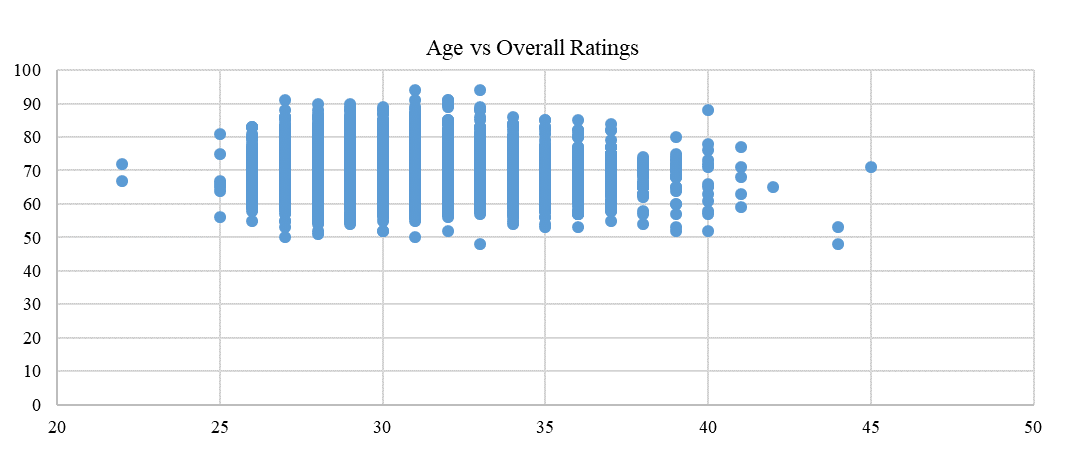
The above histogram represents bi-modal average of age across selected players, in case of the entire data it peaks at approx. 21 to 23 and then at 24-25, similarly, in case of the selected sample data it peaks at approx. 19-22 and then at 24-27.

**Nationality:**

The table besides captures Nationality of all the players from both the data sets. Of the total number of players from the Entire Data set, we observe that maximum number of players are from England which are maximum in the winning teams, followed by Spain, Brazil and Netherlands. However, players from Germany, Colombia and Japan do not contribute to winning teams.

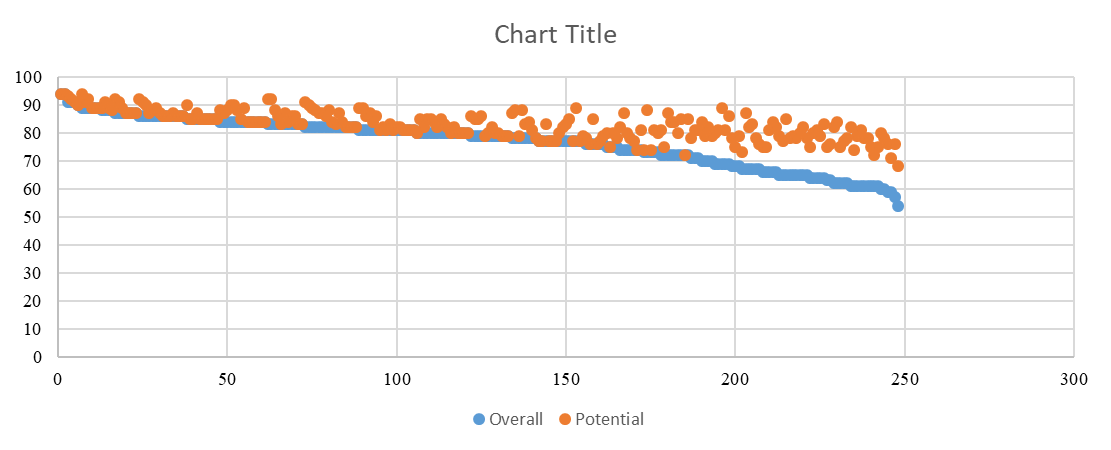
**Ratings:**

Of the total 18,207 players, 6160 players have zero difference in the Expectation gap in this case, Potential Rating – Overall Ratings. The top 2 performers are L. Messi followed by Cristiano Ronaldo with highest rating of 94 each.

****

This scatter plot maps Overall Ratings for players between age group 22 to 45. As seen in the graph above, we observe that there are couple of outliers.

**Overall & Potential Ratings – Sample Date:**

****

The graphs above represent the potential rating of a player as against his actual overall rating. In both the graphs we can observe a significant similarity. However, in the Sample Data, the expectation gap, in this case; difference between Potential Rating – Overall Rating, increases for people less than Overall Ratings of 70.

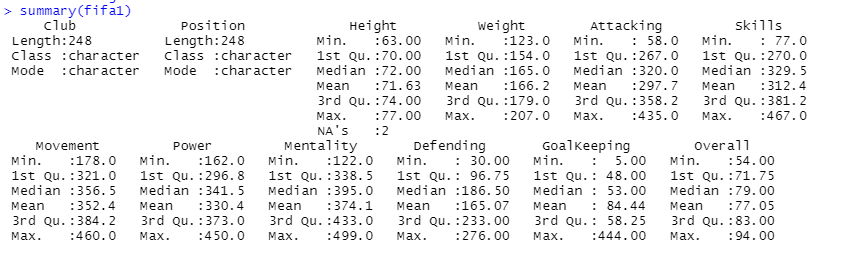
**Overall Ratings, Age and Wage:**

|  |  |  |
| --- | --- | --- |
| **Club** | **Wage of top player**  **(overall >80)** | **Mean Wage offered to a**  **top player (K/week)** |
| FC Porto | 138,000 | 15,333 |
| Ajax | 140,000 | 20,000 |
| Tottenham Hotspur | 1,494,000 | 87,882 |
| Liverpool | 1,506,000 | 88,588 |
| Juventus | 2,161,000 | 113,737 |
| Manchester United | 2,032,000 | 127,000 |
| Manchester City | 2,830,000 | 157,222 |
| FC Barcelona | 3,520,000 | 195,556 |

When we compared wages, we observed a direct correlation between wage, age and overall ratings. A player with low age has lower wage as compared with the player with higher age. However low age with higher overall rating shows a significant increase in the wage for the younger player. Of all the quarter final teams, 4 of the teams are low wage teams, namely, Ajax, FC Porto, Liverpool and Tottenham Hotspur.

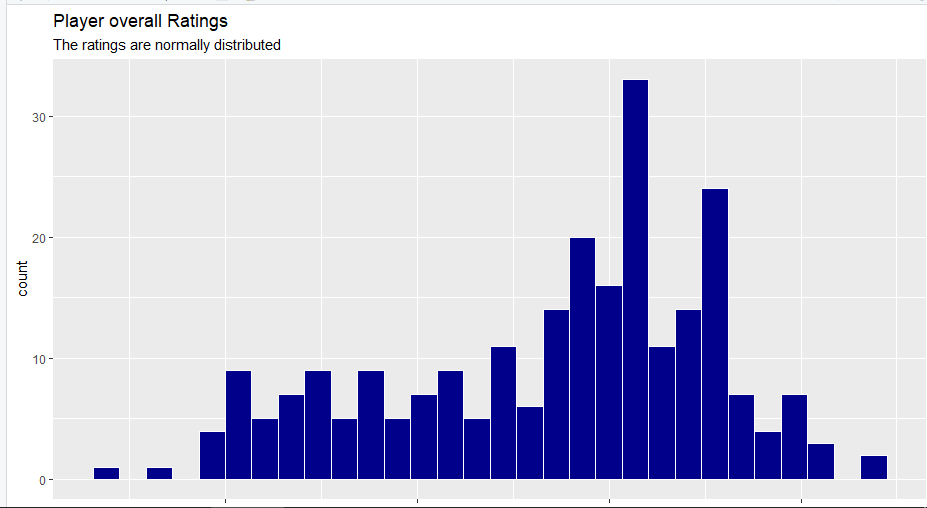
1. **Statistical Result using R Studio:**

**Summary of UEFA Dataset:**



This dataset captures various skillsets of players across clubs. There is a significant variation in the Attacking, Movement, Power, Mentality, Defending and GK skills which is expected, as a forward will score high in the attacking areas but will have less points in defense and goalkeeping. Similarly, weight and height in comparison have a much even spread, where the median height was 71.63 inches and weight was 165 pounds across the 8 quarter final teams.

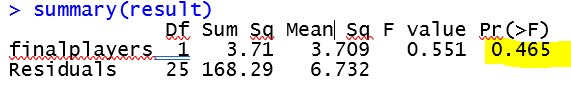
**Distribution of Overall Ratings:**



The histogram shows a bimodal average peaking at approx. 82 and approx. 86

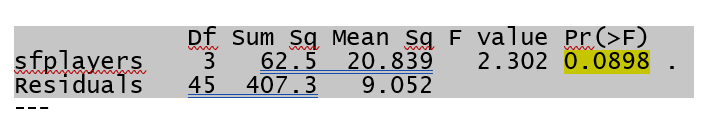
**Hypothesis Testing:**

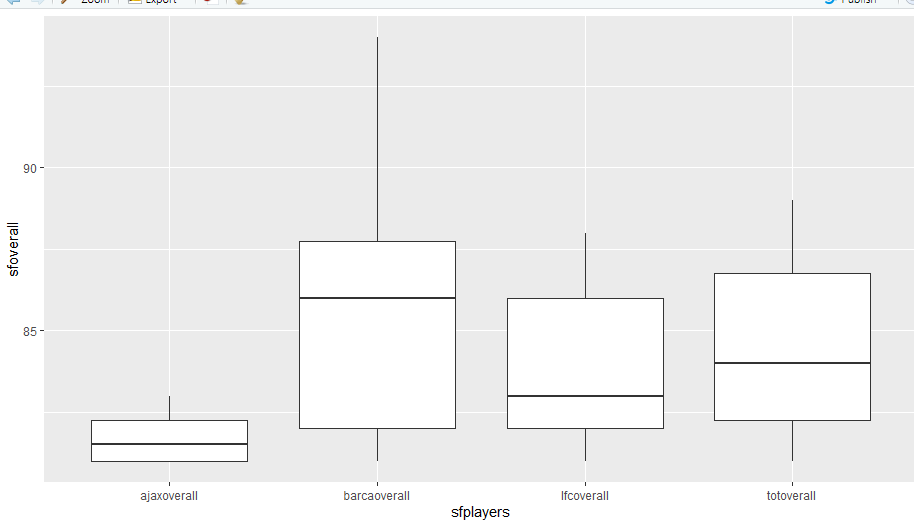
* **H0: Tottenham and Liverpool have equal overall ratings**



Using ANOVA comparison, the box plot above show that even if the area covered by the box plot is almost similar there still is a slight difference in the mean and a small p value. Hence, we will not reject null hypothesis, as it shows that tactics and strategy are relevant and can make a significant impact on the winning or losing team.

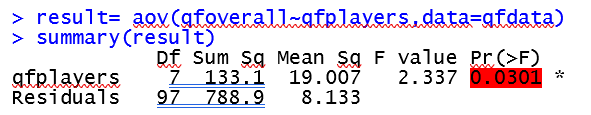
* **H0: All four teams in semi-final are evenly matched**

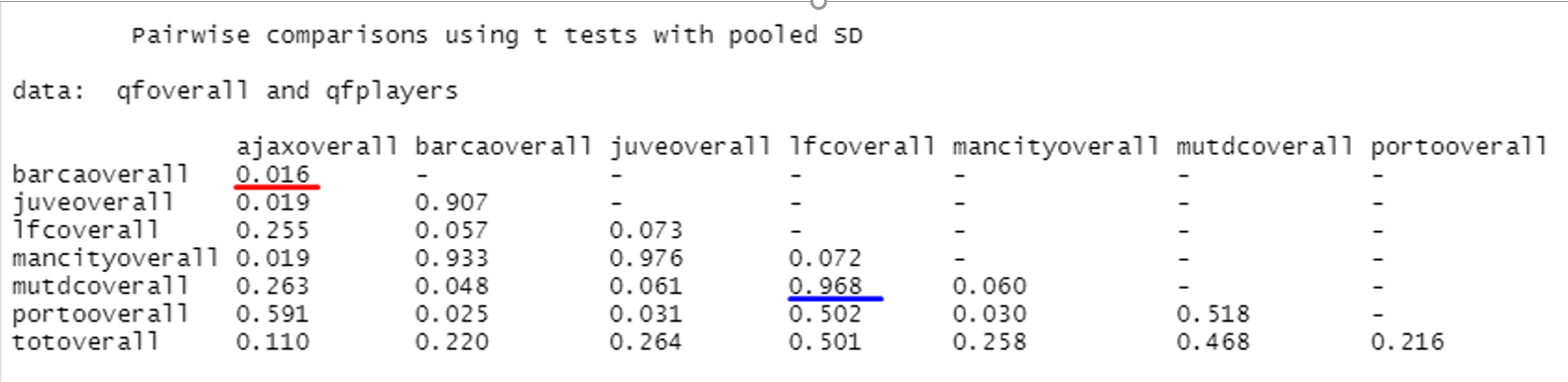
****

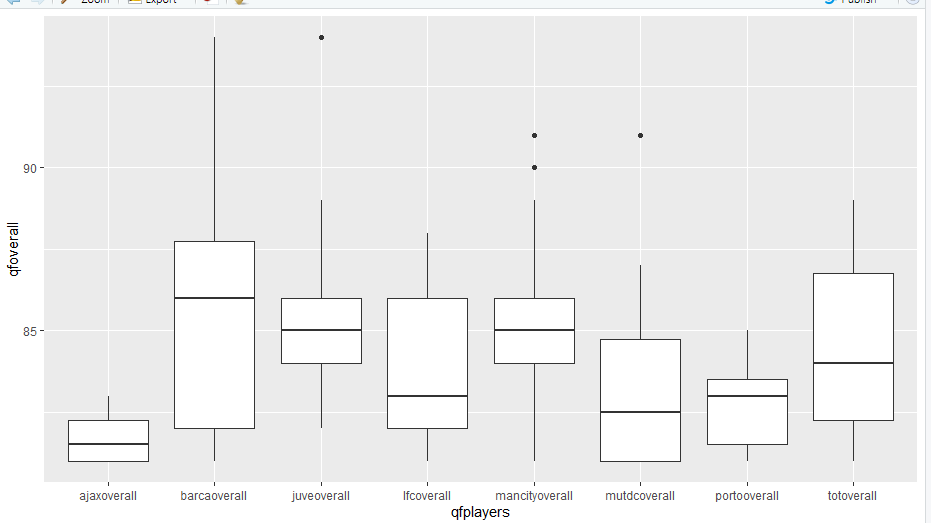


Using ANOVA comparison, the box plot above show that the area differs by each of the box plot and so there mean, also the p value is very small at 0.08. Hence, we will not reject null hypothesis. This shows that Liverpool and Tottenham were the most even teams.

* **H0: All quarter final teams had equal overall ratings**

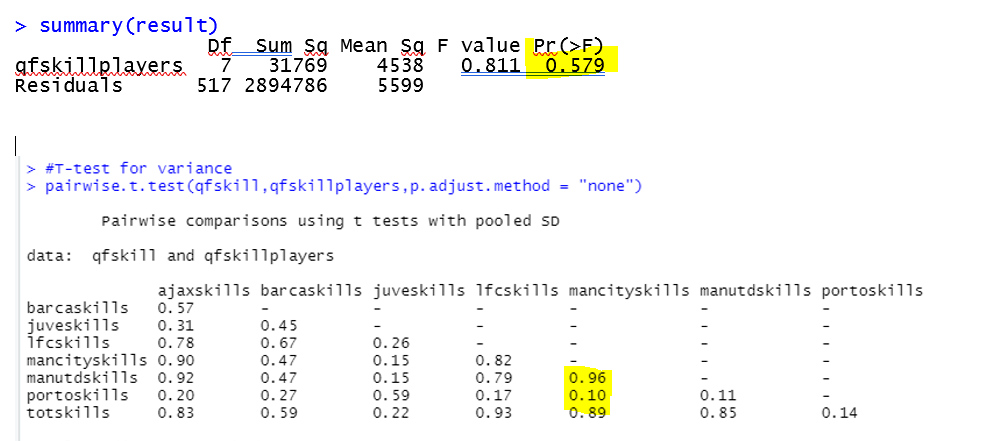
****

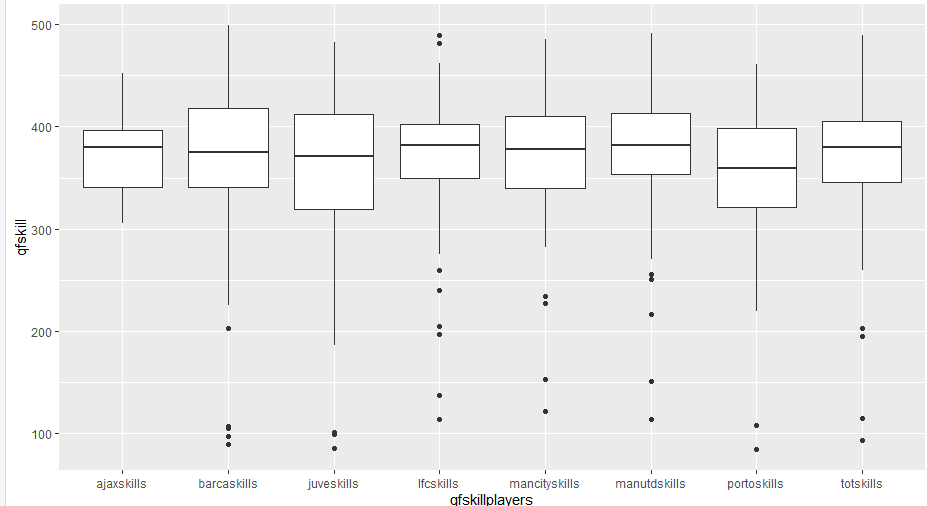




Using t-test, the above table shows that Barcelona as against Ajax would have been the most lopsided contestant. However, Liverpool as against Manchester United would have been the most even. This highlights the significance of having sound strategy and a good team work under a successful manager.

* **H0: All quarter final teams had equal skill levels**

****



Using Pairwise T-test, we observed that the 8 teams were equal in terms of skills, which means that even though a team like Ajax has less expensive players they were smart and skilled players. Also, Ajax players were given less overall ratings as compared to their skill. Porto FC and Manchester City had the widest disparity in skill. Manchester United and Manchester City were both in the high spending clubs have equally skilled players but Manchester City has more achievements which signifies the importance of good team management, tactics and football philosophy.

**Multi-Linear Regression Model:**

Below is the presentation of the multiple linear regression model used to understand the Sample Data and decision making. We have considered 7 physical performance based variables like attacking, movement, skills, power, mentality stability, defending, goal keeping and other physical attributes like height and weight, which are all independent variables. The dependent variable is overall rating of the player.

1. Linear Regression Model for Overall Rating and Skills of the Player:

A close up of a map

Description automatically generated

1. Linear Regression Model for Overall Rating and Mentality of the Player:

A close up of a map

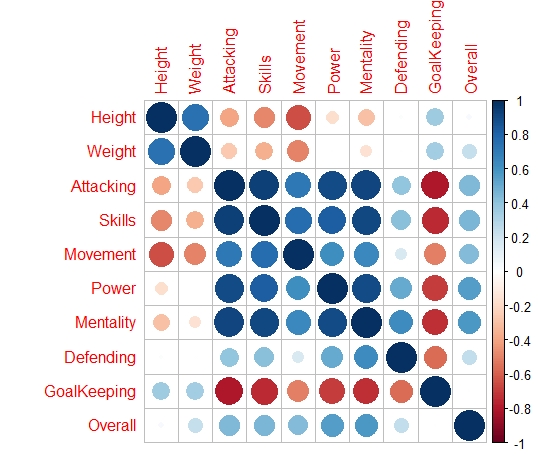
Description automatically generated

1. Linear Regression Model for Overall Rating and Goal Keeping of the Player:

A close up of a map

Description automatically generated

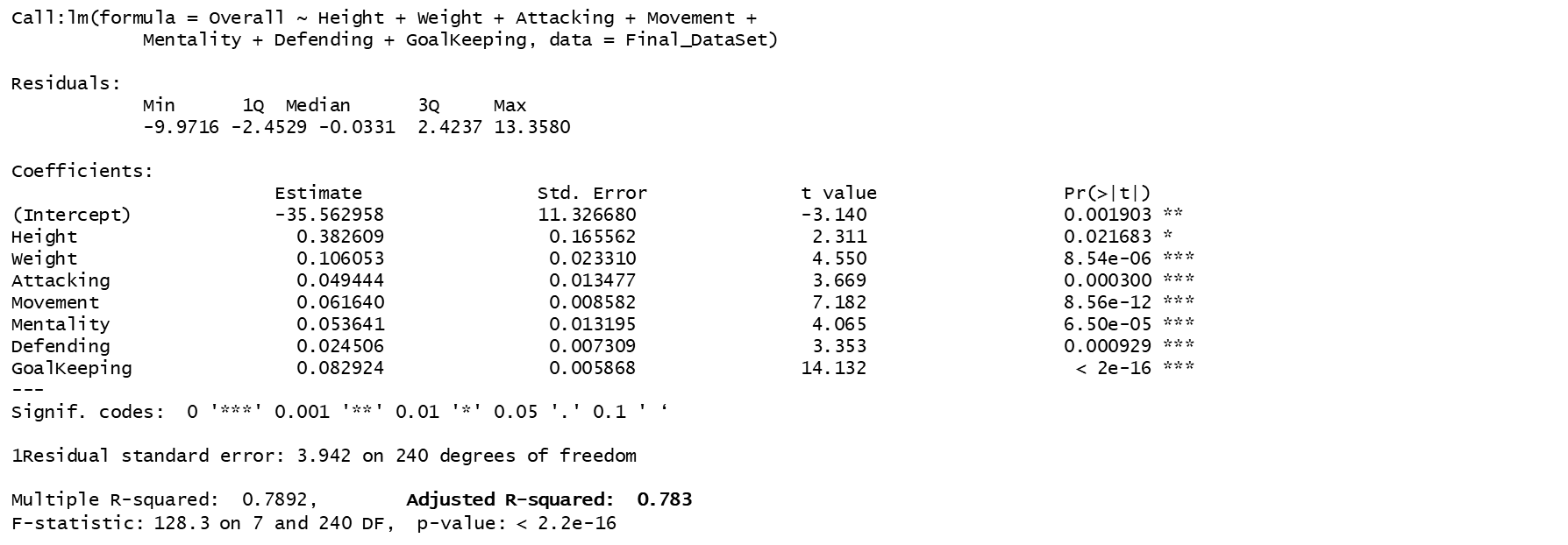
To understand the above graphs, we need to understand the dataset capturing skills, mentality and goalkeeping. These scores have been derived from the various other parameters factored. For instance, skills capture each player’s scores on dribbling, curve, FK accuracy, long passing and ball control. Similarly, movement captures acceleration, sprint speed, agility, reactions and balance. And goal keeping captures diving, handling, kicking, positioning and reflexes. Other such graphs have been capture under Annexure. Further with our analysis, we used ‘Reversed Elimination Technique’ and excluded skills and power.



**Covariance:**

Some of the previous results prompt us to calculate covariance among all numerical variables. It is fascinating as only Height and Weight have positive correlation for goalkeeping. All other variables have negative correlation with the goal keeping.

**The “Best Fit” Model:**

****

We performed the “backward elimination method,” where we started with the all independent variables and found that our best model is;

Best Model= Height + Weight + Attacking + Movement + Mentality + Defending + GoalKeeping 🡪 with this model 78% of overall rating can be explained.

A close up of a map

Description automatically generated

In order to our model to hold, we need to have the following conditions to be hold:

* The residuals of the model are nearly normal,
* The variability of the residuals is nearly constant,
* The residuals are independent, and
* Each variable is linearly related to the outcome.

1. **Conclusion:**

In our analysis above, we observed that skills for each player would differ depending on the role he performs within his team, for instance, the physical fitness requirement for a goal keeper will differ with that of a forward player. Various skills such as age, height, weight, jumping, strength, running, stamina, and so on. Hence, considering all these various aspects we have concluded that skills alone cannot determine a winning team/player. Strategy is also equally important.

The data we obtained from our source was in depth and vast, however information on team strategies followed was missing. The other challenges that we faced were in data cleaning activity, where we had to exclude missing information and irrelevant information.

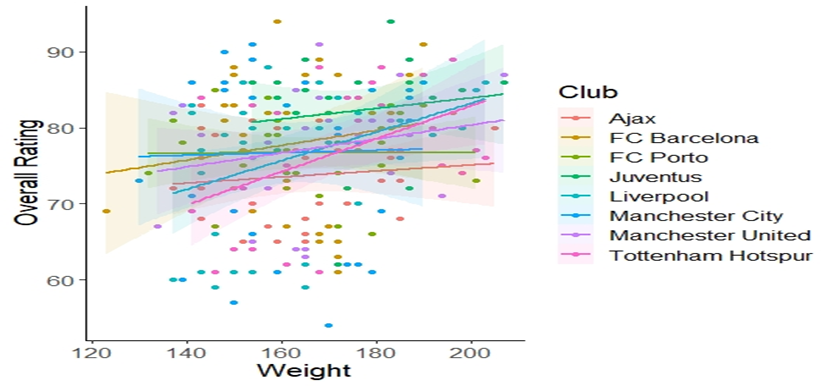
**Annexure:**

Linear Regression Model for Overall Rating and Height of the Player:

A close up of a map

Description automatically generated

Linear Regression Model for Overall Rating and Weight of the Player:



Linear Regression Model for Overall Rating and Attacking of the Player:

A close up of a map

Description automatically generated

Linear Regression Model for Overall Rating and Movement of the Player:

A close up of a map

Description automatically generated

Linear Regression Model for Overall Rating and Power of the Player:

A close up of a map

Description automatically generated

Linear Regression Model for Overall Rating and Defence of the Player:

A close up of a map

Description automatically generated

**Bibliography:**

Website for sourcing dataset from Kaggle, retrieved latest available data from[*https://www.kaggle.com/karangadiya/fifa19/version/4#*](https://www.kaggle.com/karangadiya/fifa19/version/4)