



## MH3511 Data Analysis with Computer Group Project

### Numbers in Football: Ratings & Wages

| Name                | Matriculation Number |
|---------------------|----------------------|
| Chloe Soh Sze Yee   | U2040534J            |
| David Tay Ang Peng  | U1910603L            |
| Donovan Lim Wei Bin | U2040151L            |
| Fabian Koh Ye Jun   | U2040541A            |
| Kaneko Yoshiki      | U2040883K            |

#### **Abstract:**

*With millions of fans globally, the football leagues are indisputably widely celebrated and successful professional sports leagues. Success in this industry has resulted in exorbitant player wages. This elevated level of income is traditionally justified by their skill and performance on the pitch. However, modern day football has evolved to the point where other potential factors start to come into play. Hence, we would like to examine the relationship between players' wages and various non-performance measures in the English Premier League through basic data analysis techniques.*

# Content Page

|   |           |
|---|-----------|
| <b>1. Introduction</b>  | <b>3</b>  |
| <b>2. Data description</b>  | <b>4</b>  |
| <b>3. Description and cleaning of dataset</b>   | <b>5</b>  |
| 3.1 Summary statistics for the main variable of interest - Wage   | 5         |
| 3.2 Summary statistics for other variables  | 7         |
| 3.2.1 Age   | 7         |
| 3.2.3 Reaction  | 8         |
| 3.2.4 Best Overall Rating   | 9         |
| 3.2.5 Height  | 9         |
| 3.2.6 Weight  | 10        |
| 3.3 Final Dataset for Analysis  | 10        |
| <b>4. Statistical Analysis</b>  | <b>11</b> |
| 4.1 Correlations between $\ln(\text{Wage})$ and other Continuous Variables                              | 11        |
| 4.2 Statistical Tests   | 12        |
| 4.2.1 Relationship between Wage and Best Position   | 12        |
| 4.2.2 The single most important measure that is affecting $\ln(\text{wage})$                            | 15        |
| 4.2.2.1 Univariate Linear Regression  | 15        |
| 4.2.2.2 Multiple Linear Regression  | 17        |
| 4.2.3 Relationship between players' wages and their clubs   | 19        |
| 4.2.4 Relationship between Wages categorised into quartiles and the International Reputation of players | 22        |
| 4.2.5 Relationship between Wages and Nationality  | 23        |
| <b>5. Conclusion And Discussion</b>   | <b>27</b> |
| <b>6. Appendix</b>  | <b>28</b> |
| 6.1 Kaggle Dataset  | 28        |
| 6.2 Code For Project  | 28        |
| <b>7. References</b>  | <b>35</b> |

# 1. Introduction

With millions of fans globally, the football league is now one of the most popular and successful professional sports leagues. Huge profits have led to the superstars in the game being paid lavishly - Cristiano Ronaldo being paid 1.1 million euros per week and Lionel Messi being paid 800 thousand euros per week (Marca, 2021).

In our project, a dataset contains the player ratings of footballers in FIFA21 as well as their wages and we want to answer the question of whether players are paid just based on their performance, or based on non-performance factors as well.

Based on this dataset, we seek to answer the following questions with regards to players in the English Premier League, which is widely regarded as the most successful league in the world (Bonte-Friedheim, 2018):

- 1) Is the wage of the player dependent on the position he plays?
- 2) Is the wage of the player dependent on the best overall rating that he has?
- 3) Is the wage of the player dependent on the club he plays for?
- 4) Is the wage of the player dependent on his international reputation?
- 5) Is the wage of the player dependent on their nationality?
- 6) What is the single most important variable that affects wage?

This report will cover the data descriptions, data visualisation and analysis using R language. For each of our research objectives, we will perform statistical analysis and draw conclusions in the most appropriate approach, together with explanations and elaborations.

## 2. Data description

The dataset titled 'FIFA21\_Dataset' is obtained from the online data library Kaggle. The original data set consists of a csv data frame containing 17108 observations of 65 variables.

Before proceeding to data analysis, we first performed a preliminary data cleaning to ensure that:

1. Irrelevant columns are eliminated, e.g. "jersey number" and "body type";
2. Converted all columns related to money to numeric datatype
3. Redundant information is cut out, e.g. the characters "lbs" under the "weight" column as we only need the number for analysis;
4. Filtered the dataset to only the English Premier League
5. Added a variable "lwage"

After all the preparation, 803 observations (players) of 12 variables are retained for analysis:

1. Age: Player's age
2. Nationality: Player's nation of origin
3. Potential: Player's potential rating (ceiling rating)
4. Club: Player's current club
5. Wage: Player's wage
6. International Reputation: Player's reputation around the world, rated upon 5, with 5 being the highest.
7. Height: Player's height
8. Weight: Player's weight
9. Reactions: Player's reaction speed
10. Best Position: Player's most suitable position
11. Best Overall Rating: Player's best overall rating
12. lwage: Taking the logarithm of Player's wage

### 3. Description and cleaning of dataset

In this section, we shall look into the data in more detail. We first investigate variables in groups to weigh their correlation to the variable 'Wage'. This will allow us to do a preliminary analysis of the dataset and narrow down which variables affect players' wages the most.

#### 3.1 Summary statistics for the main variable of interest - Wage

| Minimum | Median | Mean  | Max    | Skewness | Kurtosis |
|---------|--------|-------|--------|----------|----------|
| 1000    | 40000  | 46365 | 370000 | 2.083522 | 10.53048 |

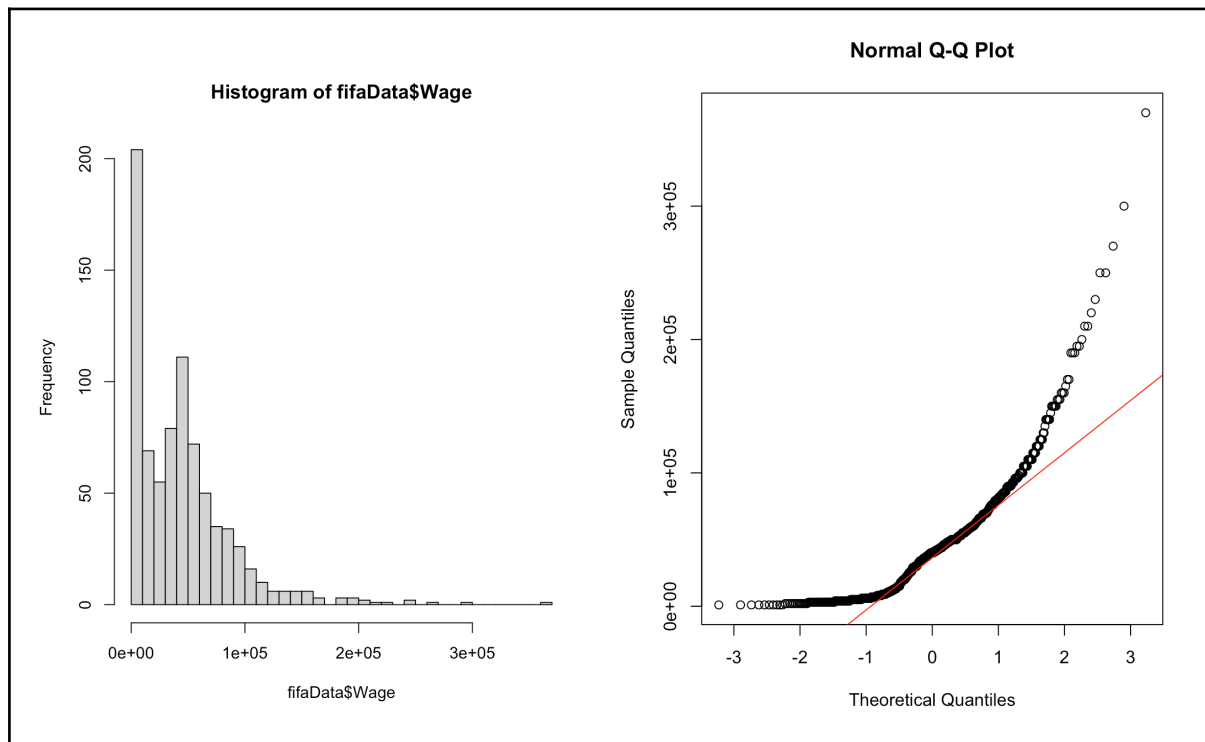


Figure 1. Distribution of Players' Wage

As shown in Figure 1, there is a strong right tail in the distribution of Players' Wage which suggests a right-skew, which is coherent with our skewness value of 2.08 (Positive and Large) and kurtosis value of 10.53 (more than 3). This is further confirmed by the QQ-plot, which shows a long right tail and short left tail. This prompted the team to take the log-distribution instead. The summary statistics for  $\ln(\text{wage})$  is tabulated in the table below.

| Minimum | Median | Mean   | Max    | Skewness   | Kurtosis |
|---------|--------|--------|--------|------------|----------|
| 6.908   | 10.597 | 10.200 | 12.821 | -0.6511001 | 2.556963 |

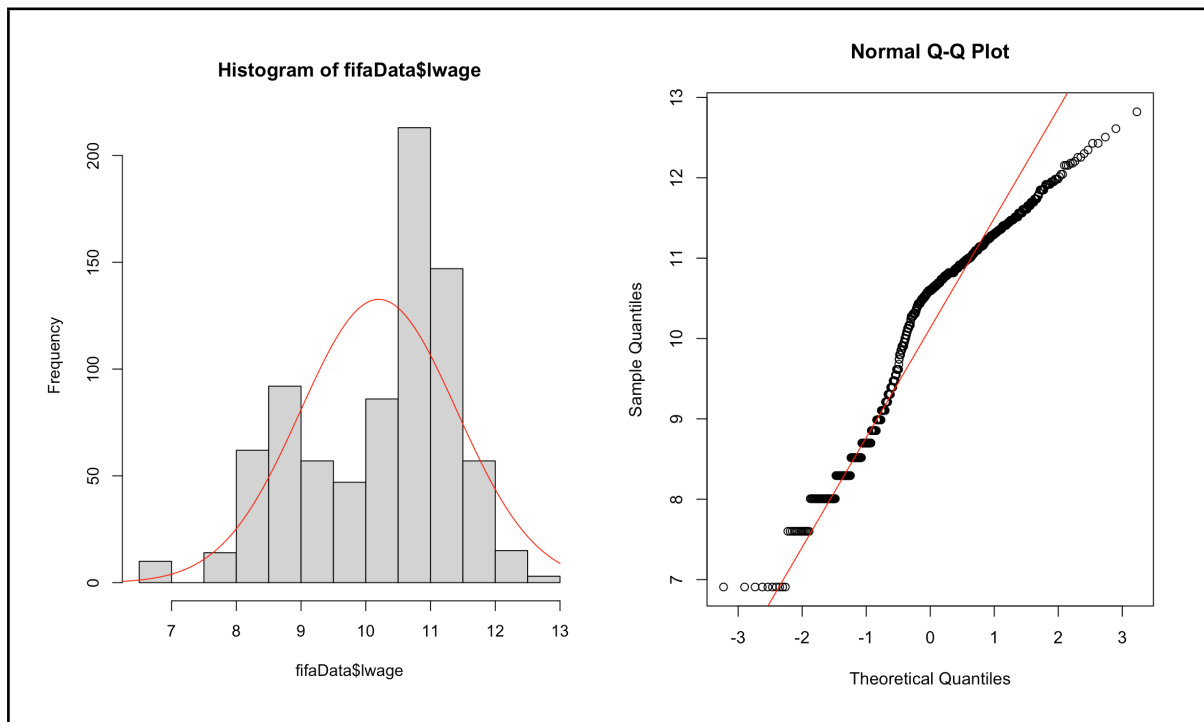


Figure 2. Distribution of Players' lwage

As we take log-transformation of wage, we note a change in skewness from a positive value to a negative value. There is a smaller magnitude in the absolute value of skewness. In addition, the kurtosis value decreases from 10.53 to 2.56. As our skewness and kurtosis values approach 0 and 3 respectively after taking log-transformation, it resembles more of a normal distribution (Kallner, 2017). Therefore, we will be working with the  $\ln(\text{wage})$  for our following analysis.

### 3.2 Summary statistics for other variables

The histogram, the boxplot, the transformation applied and the outliers removed from the variables are tabulated in the following subsections.

For outliers, as the team are unsure of the reasons that caused these outliers, the team will be removing them as it is the best option with our limited information (Frost, 2021):

- There are sufficient observations in the dataset (800 after removal)
- Only a small proportion of the data has to be removed (3 out of 803)

#### 3.2.1 Age

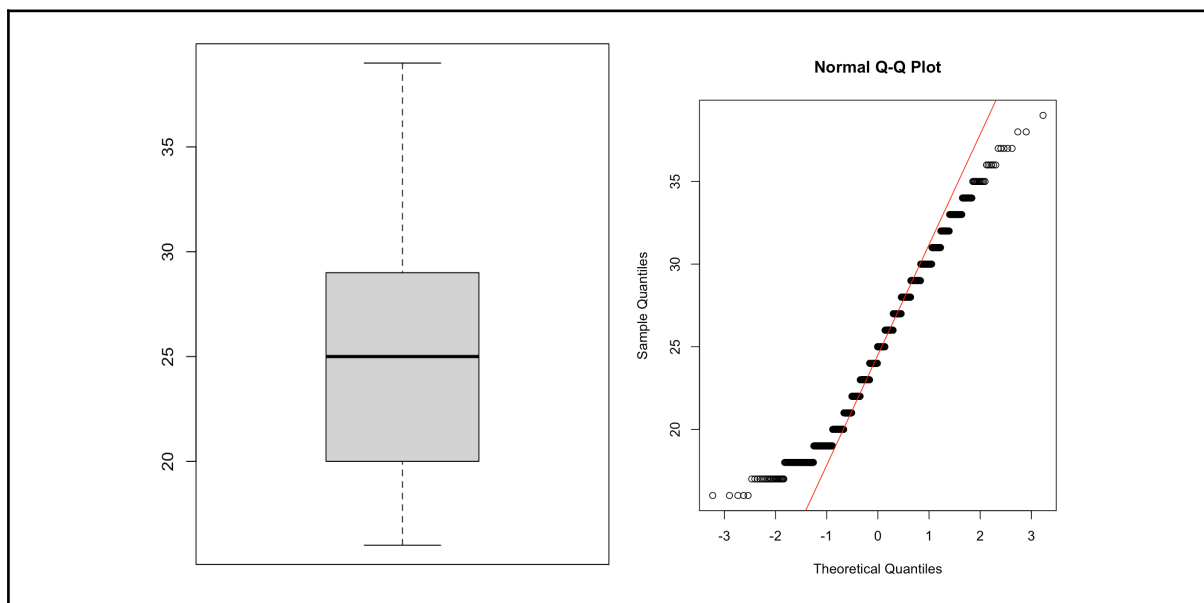


Figure 3. Distribution of Players' Age

We observe that there are no outliers for the players' age. We also observe that the players' age is approximately normal as the tails are relatively close to the red line.

### 3.2.2 Potential

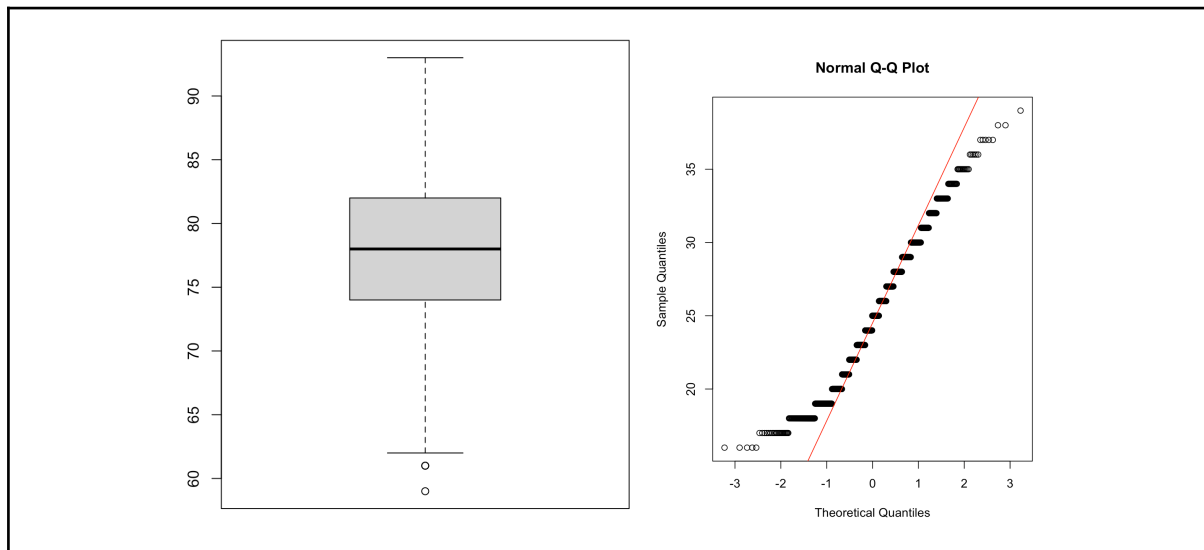


Figure 4. Distribution of players' Potential

The team chose to remove the two outliers identified. It can also be observed that the distribution of potential has a short tail on both ends.

### 3.2.3 Reaction

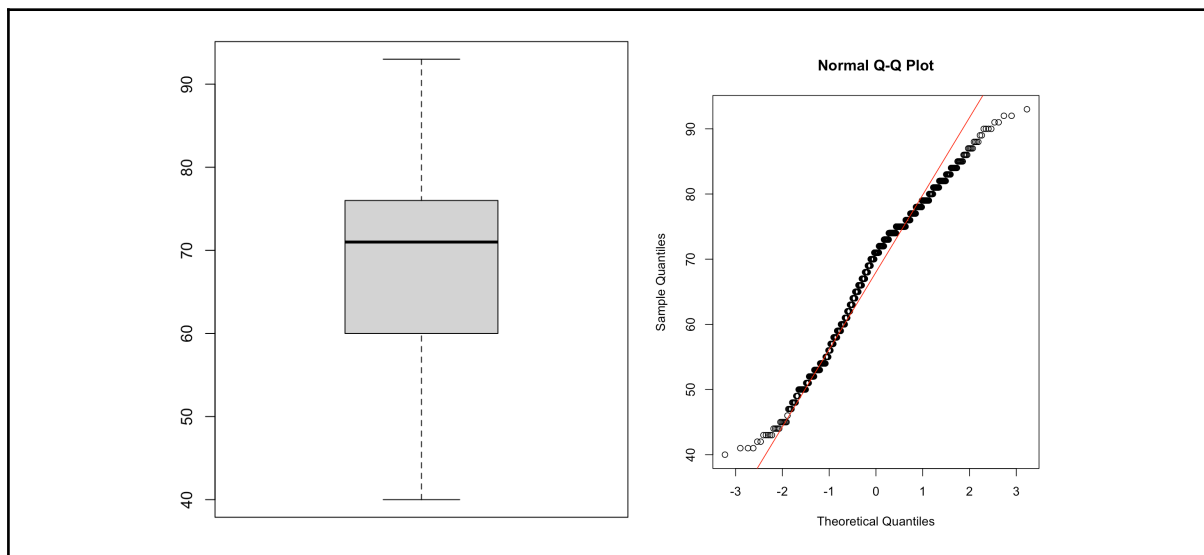


Figure 5. Distribution of players' Reaction

We observe that there are no outliers for the players' reaction. We also observe that the players' reaction is approximately normal as the tails are close to the red line.



### 3.2.4 Best Overall Rating

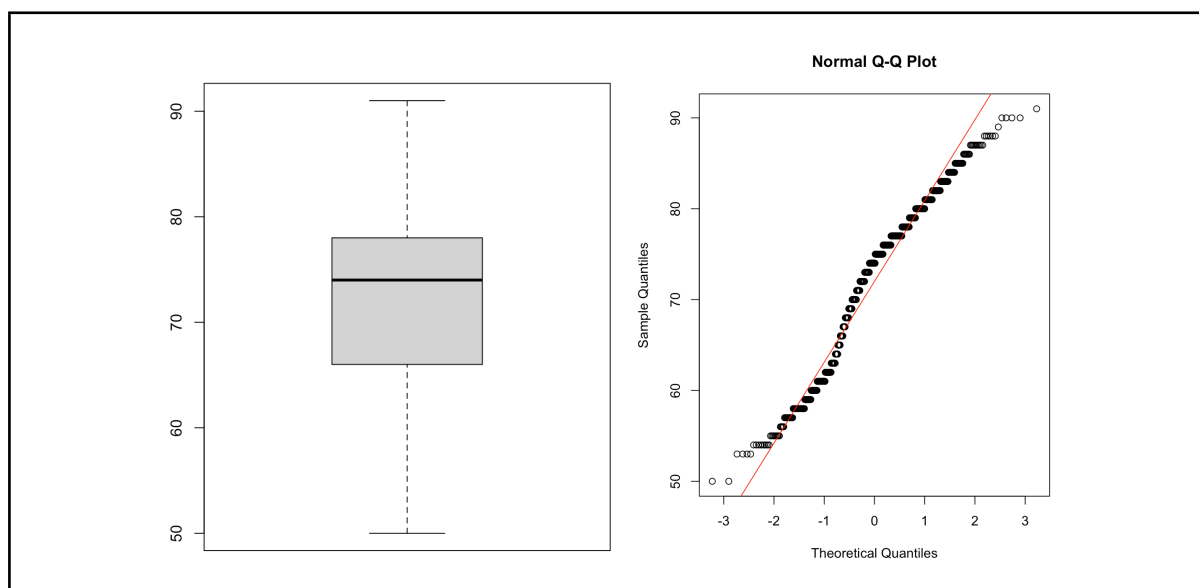


Figure 6. Distribution of players' Best Overall Rating

We observe that there are no outliers for the players' best overall rating. We also observe that the players' best overall rating is approximately normal as the tails are close to the red line.

### 3.2.5 Height

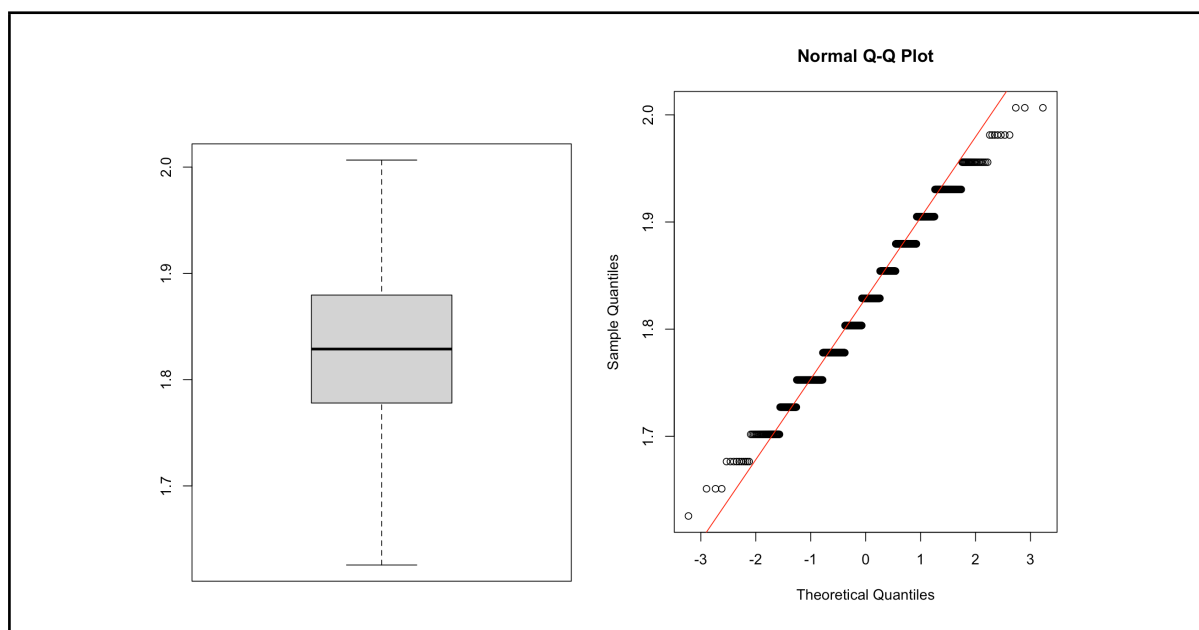


Figure 7. Distribution of players' Height

We observe that there are no outliers for the players' height. We also observe that the players' height is approximately normal as the tails are close to the red line.

### 3.2.6 Weight

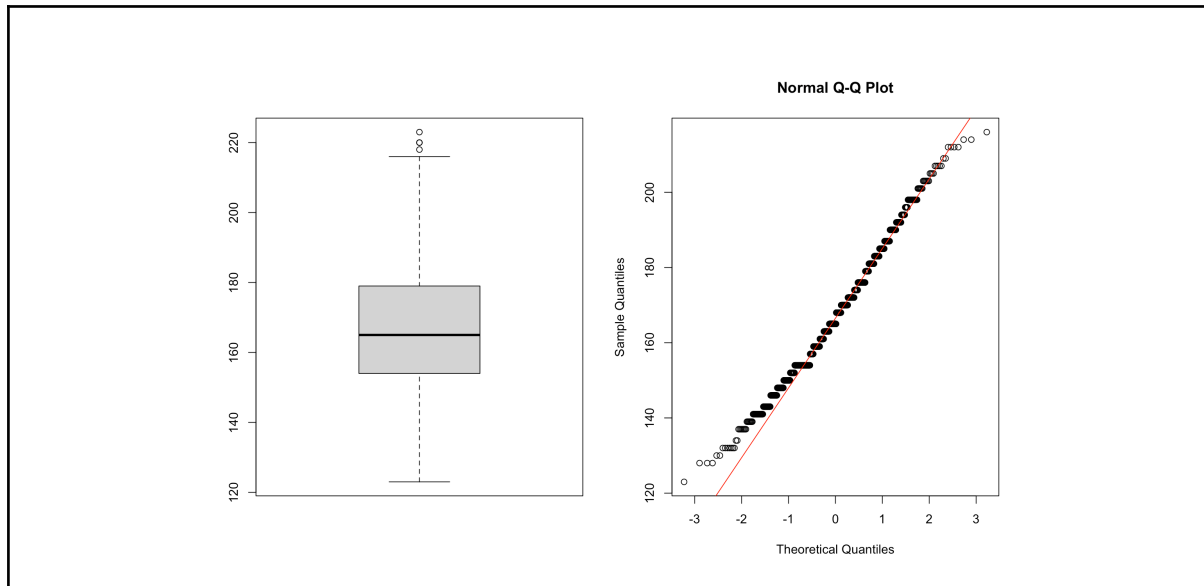


Figure 8. Distribution of players' Weight

The team chose to remove the four outliers identified. It can also be observed that the distribution of weight has a short left tail.

### 3.3 Final Dataset for Analysis

Based on the above analysis, the team has removed the outliers. The log-transformation (base e) was applied to variables. The updated dataset remains as 796 observations.

```
> str(fifaData)
Classes 'data.table' and 'data.frame': 796 obs. of 12 variables:
 $ Age          : int  29 25 29 27 30 21 26 30 28 28 ...
 $ Nationality  : chr   "Belgium" "Portugal" "Netherlands" "France" ...
 $ Potential    : int   91 90 85 87 86 92 85 83 90 87 ...
 $ Club         : chr   "Manchester City" "Manchester United" "Liverpool" "Manchester United" ...
 $ Wage         : num  370000 195000 150000 190000 140000 110000 110000 130000 250000 190000 ...
 $ International Reputation: num  4 2 3 4 2 2 2 4 3 3 ...
 $ Height       : num   1.8 1.78 1.75 1.91 1.83 ...
 $ Weight       : num  154 152 152 185 176 152 163 181 157 168 ...
 $ Reactions    : num   91 86 86 81 86 83 83 79 92 90 ...
 $ Best Position: chr    "CAM" "CAM" "CM" "CM" ...
 $ Best Overall Rating: num   91 88 85 86 86 87 84 83 90 87 ...
 $ lwage        : num   12.8 12.2 11.9 12.2 11.8 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

Figure 9. Internal structure of fifaData

## 4. Statistical Analysis

### 4.1 Correlations between $\ln(\text{Wage})$ and other Continuous Variables

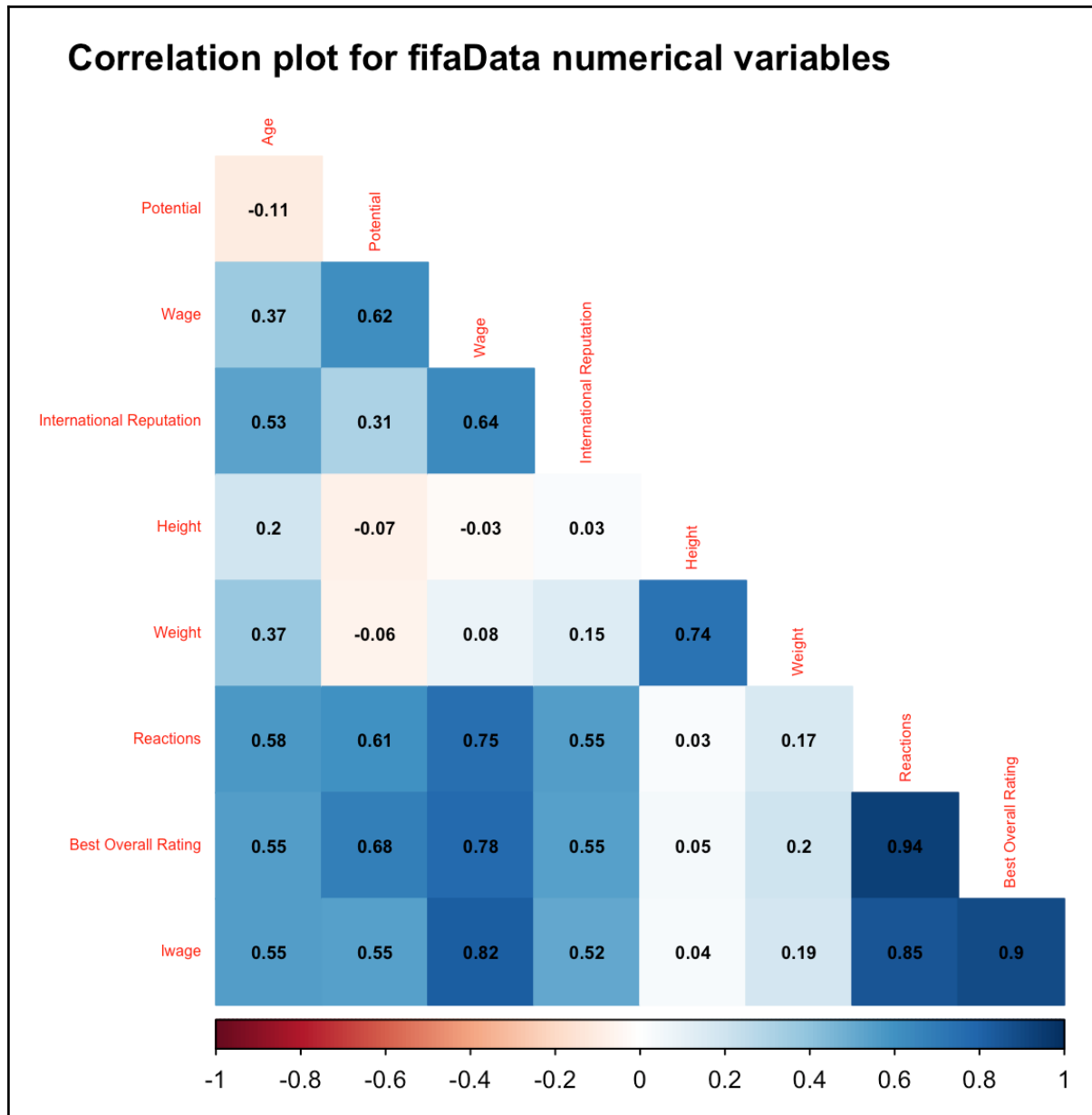


Figure 10. Corplot between  $\ln(\text{Wage})$  and other continuous variables

Correlation coefficients are useful in identifying whether two numerical variables have a linear relationship to each other. The higher the magnitude of the correlation coefficient, the stronger the linear association between the two variables.

From the plots, it appears that  $\ln(\text{Wage})$  is highly correlated to four variables, less Wage which it was derived from:

1. Best Overall Rating ( $r = 0.9$ ) has a very strong positive linear relationship
2. Reactions ( $r = 0.85$ ) has a very strong positive linear relationship
3. Age ( $r = 0.55$ ) has a strong positive linear relationship
4. Potential ( $r = 0.55$ ) has a strong positive linear relationship

From the correlation plot, we also observe that height and weight have small correlation coefficients with  $\ln(\text{wage})$  with correlation coefficients of 0.04 and 0.19 respectively, suggesting weak relationships to  $\ln(\text{wage})$ , thus they will be omitted for subsequent analysis.

## 4.2 Statistical Tests

### 4.2.1 Relationship between Wage and Best Position

In this section we try to answer the question “Is the wage of a soccer player dependent on his best position in the game?”

Since "Best Position" is a categorical variable, we will determine whether  $\ln(\text{wage})$  varies with the best positions of players through the use of an analysis of variance (ANOVA) test. The distributions of  $\ln(\text{wage})$  across different best positions of players are illustrated by the following boxplot.

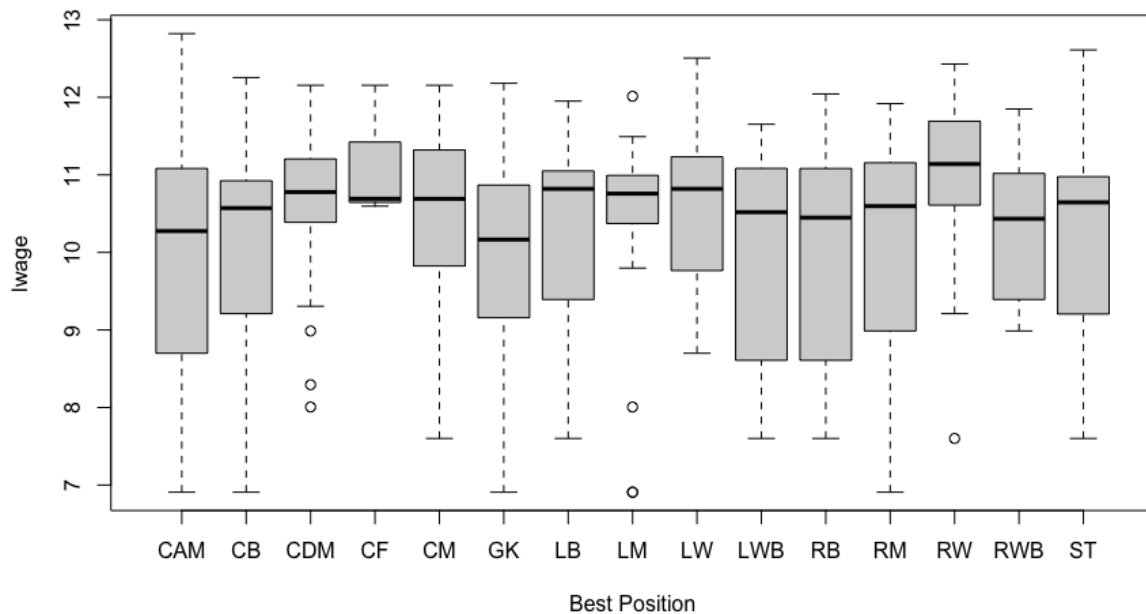


Figure 11. Boxplot of  $\ln(\text{wage})$  vs Best Position

Looking at the above boxplot, we can see that the distributions of  $\ln(\text{wage})$  vary to a large extent when analysed based on the different best positions of players. Hence, the ANOVA test is appropriate in testing the equality of the means ( $\mu_i$ ) of  $\ln(\text{wage})$  for each best position.

We test the following:

$$H_0: \mu_{\text{CAM}} = \mu_{\text{CB}} = \dots = \mu_{\text{ST}} \text{ against } H_1: \text{not all } \mu \text{ are equal}$$

```

                                Df Sum Sq Mean Sq F value Pr(>F)
factor(fifaData$`Best Position`) 14   47.8    3.412    2.409 0.00264 **
Residuals                        781 1106.2    1.416
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 12. Summary of ANOVA test

The ANOVA test returns a p-value of 0.00264, which shows that the means are different at a significance level of 0.05. Therefore, we conclude that the salary of a FIFA player does in fact vary with the best position he plays in the game.

```

Pairwise comparisons using t tests with pooled SD

data:  fifaData$lwage and fifaData$`Best Position`

   CAM   CB   CDM   CF   CM   GK   LB   LM   LW   LWB   RB   RM   RW   RWB
CB 0.0681 -     -     -     -     -     -     -     -     -     -     -     -
CDM 9.7e-05 0.0061 -     -     -     -     -     -     -     -     -     -
CF 0.0671 0.1434 0.4702 -     -     -     -     -     -     -     -     -
CM 0.0015 0.0542 0.4429 0.3351 -     -     -     -     -     -     -
GK 0.2438 0.6919 0.0060 0.1238 0.0443 -     -     -     -     -     -
LB 0.0395 0.3674 0.1941 0.2446 0.5287 0.2732 -     -     -     -     -
LM 0.0379 0.2745 0.3984 0.3033 0.8048 0.2085 0.7789 -     -     -     -
LW 0.0501 0.1984 0.9409 0.4871 0.7200 0.1586 0.4732 0.6267 -     -
LWB 0.2910 0.6688 0.4394 0.2896 0.7018 0.5714 0.9783 0.8362 0.5727 -
RB 0.6489 0.4100 0.0055 0.0970 0.0322 0.6415 0.1745 0.1362 0.1105 0.4362 -
RM 0.3817 0.7199 0.0180 0.1265 0.0844 0.9607 0.3215 0.2444 0.1735 0.5718 0.7204 -
RW 0.0014 0.0149 0.4550 0.7314 0.2064 0.0117 0.1035 0.1932 0.5456 0.2395 0.0081 0.0177 -
RWB 0.1061 0.3901 0.5353 0.3348 0.8816 0.3134 0.7900 0.9689 0.6892 0.8287 0.2190 0.3309 0.2771 -
ST 0.0169 0.3725 0.0668 0.2078 0.2977 0.2676 0.8187 0.6062 0.3686 0.9250 0.1671 0.3455 0.0532 0.6617

P value adjustment method: none

```

Figure 13. Pairwise T-test comparing lwage and Best Position

From the pairwise T-test, it is very difficult to get a reasonable conclusion as there are too many pairs of positions and no clear pattern. Therefore, the team decided to group up the positions by their category - "Goalkeeper", "Defender", "Midfielder" and "Forward" to test the following:

$$H_0: \mu_{GK} = \mu_{DF} = \mu_{MF} = \mu_{FW} \text{ against } H_1: \text{not all } \mu \text{ are equal}$$

|                       | Df  | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------------------|-----|--------|---------|---------|--------|
| factor(fifaData\$cat) | 3   | 7      | 2.331   | 1.61    | 0.186  |
| Residuals             | 792 | 1147   | 1.448   |         |        |

Figure 14. Summary of ANOVA test

The ANOVA test returns a p-value of  $0.186 > 0.05$ , which shows that the means cannot be conclusively said to be different at a significance level of 0.05. Therefore, we reject the null hypothesis and conclude that the salary of a FIFA player does not vary with the category of position that he plays in the game.

| Pairwise comparisons using t tests with pooled SD |          |         |            |
|---|----------|---------|------------|
| data: fifaData\$lwage and fifaData\$cat           |          |         |            |
|   | Defender | Forward | Goalkeeper |
| Forward   | 0.058    | -       | -          |
| Goalkeeper  | 0.578    | 0.053   | -          |
| Midfielder  | 0.560    | 0.145   | 0.341      |
| P value adjustment method: none                   |          |         |            |

Figure 15. Pairwise comparisons using t-tests with pooled standard deviation

Then, we repeated the same analysis and found that at 0.05 level of significance, we do not have sufficient evidence to reject the null hypothesis, thus we conclude that the category of the players do affect their  $\ln(\text{wage})$ .

## 4.2.2 The single most important measure that is affecting $\ln(\text{wage})$

### 4.2.2.1 Univariate Linear Regression

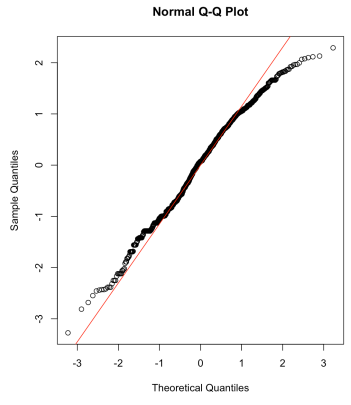
In section 4.1, we have identified that certain non-performance measures are quite strongly correlated to  $\ln(\text{wage})$ . We now perform a simple linear regression analysis to determine which of the four variables could be used to model  $\ln(\text{wage})$  in a linear fashion using the following equation:

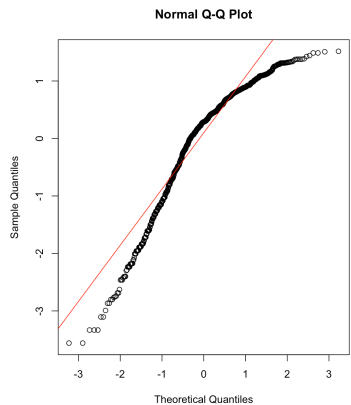
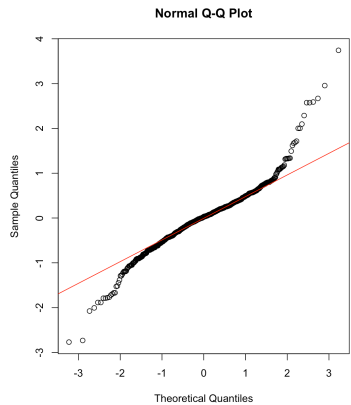
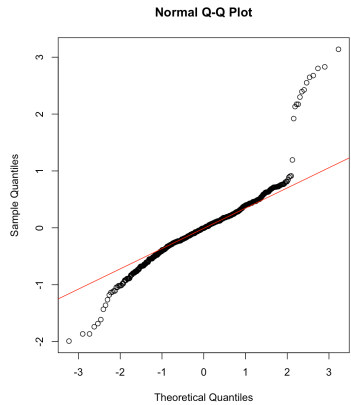
$\ln(\text{wage}) = \beta_0 + \beta_1 * X + \varepsilon$  where X could be any one of the following:

- Age
- Potential
- Reactions
- Best Overall Rating

The summary of the analysis is listed in the table below.

By comparing the R-squared and the residual plot, 'Best Overall Rating' is determined to be the single most important performance measure to model  $\ln(\text{wage})$  using a simple linear model.

| Variable (X) | Fitted Model (Y = $\ln(\text{Wage})$ ) | p-value   | R-squared | qq-plot of residuals  |
|--------------|--|-----------|-----------|---|
| Age          | $Y = 0.133X + 6.89$                    | $< 2e-16$ | 0.309     |  <p>Figure 16.</p> |

|                     |                      |           |       |  |
|---------------------|----------------------|-----------|-------|--|
| Potential           | $Y = 0.113X + 1.38$  | $< 2e-16$ | 0.293 |  <p>Normal Q-Q Plot</p> <p>The plot shows Sample Quantiles on the y-axis (ranging from -3 to 1) and Theoretical Quantiles on the x-axis (ranging from -3 to 3). The data points closely follow the diagonal red line, indicating a normal distribution.</p> <p>Figure 17.</p> |
| Reactions           | $Y = 0.0954X + 3.67$ | $< 2e-16$ | 0.725 |  <p>Normal Q-Q Plot</p> <p>The plot shows Sample Quantiles on the y-axis (ranging from -3 to 4) and Theoretical Quantiles on the x-axis (ranging from -3 to 3). The data points follow the diagonal red line, indicating a normal distribution.</p> <p>Figure 18.</p>        |
| Best Overall Rating | $Y = 0.126X + 1.06$  | $< 2e-16$ | 0.808 |  <p>Normal Q-Q Plot</p> <p>The plot shows Sample Quantiles on the y-axis (ranging from -2 to 3) and Theoretical Quantiles on the x-axis (ranging from -3 to 3). The data points follow the diagonal red line, indicating a normal distribution.</p> <p>Figure 19.</p>       |



#### 4.2.2.2 Multiple Linear Regression

In this section, we attempt to build a multiple linear model for  $\ln(\text{wage})$  based on the 4 given indicators, namely Age, Potential, Reactions and Best Overall Rating. We use a backward elimination method to select the most appropriate model. The result is shown in the R output below.

```
Call:
lm(formula = lwage ~ Potential + `Best Overall Rating`, data = fifaData)

Residuals:
    Min       1Q   Median       3Q      Max
-1.9067 -0.2694  0.0047  0.2267  3.2592

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.269777   0.249531   9.096  < 2e-16 ***
Potential     -0.026950   0.004349  -6.197 9.27e-10 ***
`Best Overall Rating` 0.138594   0.002914  47.563  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5169 on 793 degrees of freedom
Multiple R-squared:  0.8164,    Adjusted R-squared:  0.8159
F-statistic: 1763 on 2 and 793 DF,  p-value: < 2.2e-16
```

Figure 20. Summary of Multiple Linear Regression Without Restrictions

We obtained a model where  $Y = 2.2698 - 0.02695 * \text{Potential} + 0.1386 * \text{'Best Overall Rating'}$ , with a R-Squared value of 0.8159, which indicates a relatively good fit (Wikipedia, 2022).

We obtain a negative coefficient for potential which logically does not make sense. As potential is a subjective quantity, it may not be a suitable feature to be used in our MLR model. Hence, we run the code to generate a MLR model which excludes potential.

```

Call:
lm(formula = lwage ~ Age + `Best Overall Rating`, data = fifaData)

Residuals:
    Min       1Q   Median       3Q      Max
-1.8812 -0.2648  0.0062  0.2252  3.1796

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.038820   0.157826   6.582 8.43e-11 ***
Age             0.019474   0.004441   4.385 1.32e-05 ***
`Best Overall Rating` 0.119976   0.002605  46.049 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.523 on 793 degrees of freedom
Multiple R-squared:  0.812,    Adjusted R-squared:  0.8116
F-statistic: 1713 on 2 and 793 DF,  p-value: < 2.2e-16

```

Figure 21. Summary of Multiple Linear Regression Without Potential

We obtained a model where  $Y = 1.03882 + 0.01947 * \text{Age} + 0.11998 * \text{'Best Overall Rating'}$ , with a R-Squared value of 0.8116, which indicates a relatively good fit (Wikipedia, 2022) as well, but less so than the model above with potential.

However, there are certain limitations to using linear regression to model the wage:

- We are unable to model non-linear relationships.
- We are unable to account for non-numerical factors such as the club the player plays for.

To overcome these limitations, deep learning models such as neural networks and other forms of machine learning models such as random forest can be used.

### 4.2.3 Relationship between players' wages and their clubs

In this section, we try to answer the question “Does the player’s wage depend on the club that he is playing for?”

At the end of a Premier League season, clubs receive prize money according to their final league position (Ambille, 2022). Hence we hypothesise that the top few clubs are then able to contribute a greater amount to their players’ wages. To simplify our analysis, we separate players into two groups – players in one of the top 6 clubs, and players that are not in one of the top 6 clubs. To determine whether a club has the status of being a top 6 club in the English Premier League, we have two criteria that we look at.

Firstly, the ‘Big 6’ of the English Premier League refers to Arsenal, Chelsea, Liverpool, Manchester City, Manchester United and Tottenham Hotspur, in no particular order, owing to their consistent successes as well as top 6 placements in the league at the end of every season (Kelly, R. (2021, April 21)).

The second criteria in which we determine the top 6 of the league through their mean wages which is consistent with our first criteria as seen in the boxplot below, with the clubs sorted alphabetically.

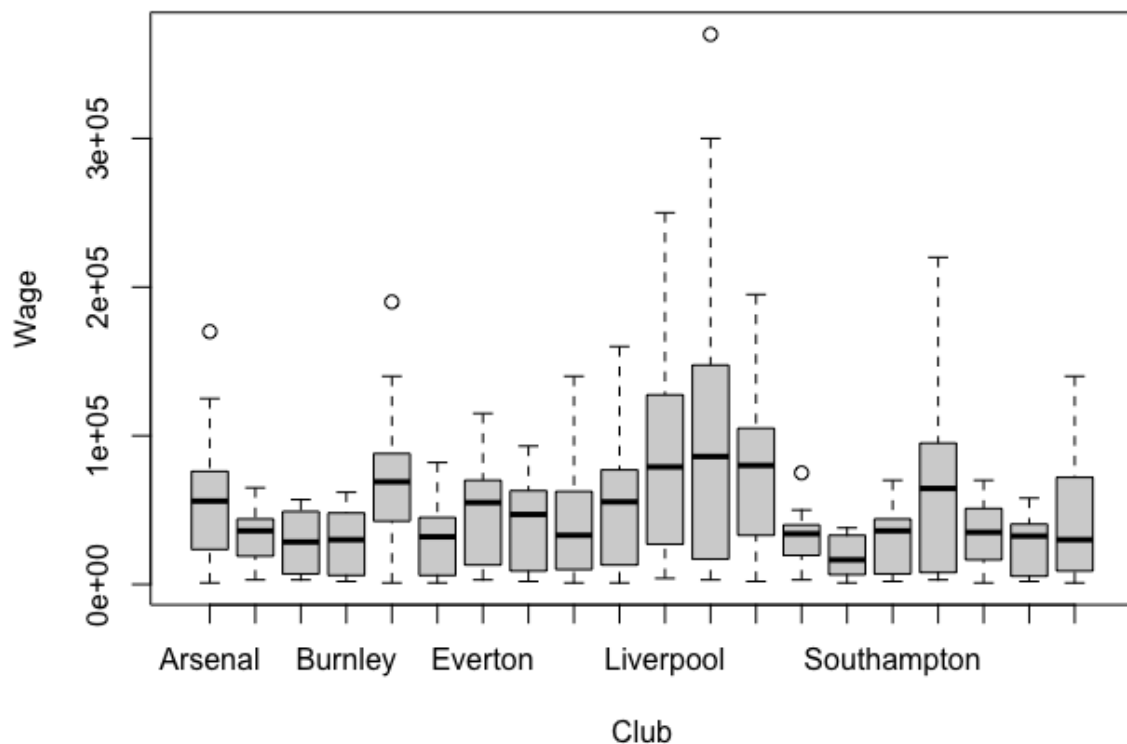


Figure 22. Boxplot of the wages of the clubs

After subsetting the clubs into top 6 clubs and non-top 6 clubs, we first run an F test to determine if the variances of the wages are equal.

$H_0$ : The variance of the wages between 'top6players' and 'nontop6players' are equal.

$H_1$ : The alternative hypothesis is that the variance of the wages between 'top6players' and 'nontop6players' are not equal.

We set the level of significance to 0.05.

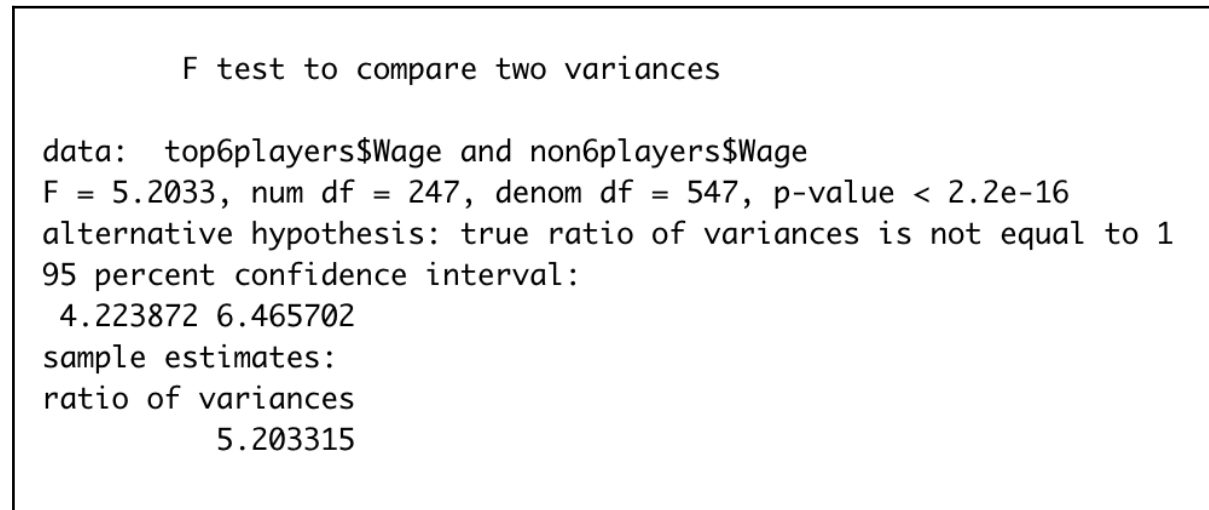


Figure 22. F test comparing wages between players in the top 6 clubs and player that are not in the top 6 clubs

From the test, we can understand that the p-value is extremely small, and we can reject the null hypothesis, realising that the variances of the two subsets are not equal. In fact, the players in the top 6 clubs have greater variance in their wages as compared to the non top 6 clubs.

Next, we run the Welch Two Sample t-test to determine if the mean of the wages of the two subsets are equal.

$H_0$ : The mean of the wages between 'top6players' and 'nontop6players' are equal.

$H_1$ : The mean of the wages between 'top6players' and 'nontop6players' are not equal.

We set the level of significance at 0.05.

### Welch Two Sample t-test

```
data: top6players$Wage and non6players$Wage
t = 9.7951, df = 290.84, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 31271.04 46997.74
sample estimates:
mean of x mean of y
 73435.48 34301.09
```

Figure 23. Welch two sample t-test comparing wages between players in the top 6 clubs and player that are not in the top 6 clubs

From the results, we can see that there is indeed a disparity between the mean of the players in the top 6 clubs as compared to those in non-top 6 clubs. In fact, the players from the top 6 clubs are better paid as compared to those that are not from the top 6 clubs. This leads us to conclude that it is likely that the players' wages are inflated if they were to be from a top 6 club as compared to from a non-top 6 club.

We also test the relationship between clubs and international reputation. We want to see if players with a higher international reputation (from a scale of 1 to 4) have a higher probability to be in one of the top six clubs. We create a matrix for top6 against International Reputation, and compute the expected values.

$H_0$ : There is a strong relationship between "top6" and "International Reputation".

$H_1$ : There is a weak relationship between "top6" and "International Reputation".

We then perform the chi-square test to understand the relationship between the two categorical variables as shown in the image below, with the level of significance at 0.05:

### Pearson's Chi-squared test

```
data: IR_Top6
X-squared = 77.394, df = 3, p-value < 2.2e-16
```

Figure 24. Chi-square test

From the test, we can see that the p-value is insignificant, thus we do not reject the null hypothesis. The strong relationship between the international reputation of players and whether they are in the top six clubs is also another possible reason why players in the top six clubs are paid more.

#### 4.2.4 Relationship between Wages categorised into quartiles and the International Reputation of players

In this section, we want to test if the international reputation of all players is closely related to their wages.

Firstly, since wage is a continuous variable, we will create a new categorical variable called WageQuartile so that it can be compared with the International Reputation variable.

$H_0$ : Proportion of “WageQuartile” is independent of “International Reputation”

$H_1$ : Proportion of “WageQuartile” is dependent on “International Reputation”

We set the level of significance at 0.05.

We also create a 2-way contingency table that will compare WageQuartile and International Reputation. The resulting table is as follows:

|         | int rep 1 | int rep 2 | int rep 3 | int rep 4 |
|---------|-----------|-----------|-----------|-----------|
| 1st Qu. | 195       | 6         | 0         | 0         |
| 2nd Qu. | 157       | 38        | 5         | 0         |
| 3rd Qu. | 131       | 51        | 13        | 1         |
| 4th Qu. | 61        | 70        | 58        | 10        |

Figure 25. Table comparing WageQuartile and International Reputation

From the table above, we can see that generally, the percentage of players being paid at the fourth quartile range of wages increases as their international reputation increases. We then find the expected values for each category and perform the chi-square test to understand the relationship between the two categorical variables as shown in the image below:

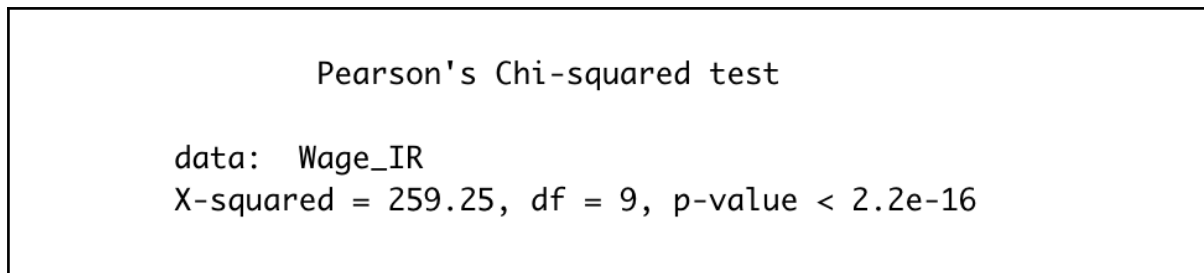


Figure 26. Chi-square test

From the test, we can see that the p-value is very small and insignificant, thus we do not reject the null hypothesis. This makes sense and shows that players who are more reputable will also receive higher wages because of their reputation.

#### 4.2.5 Relationship between Wages and Nationality

In this section, we want to test if the nationality of players is closely related to their wages. As players with differing skill levels are expected to be paid differently, we wish to eliminate the effects of skill on wages by obtaining the 'lwage per overall', whereby we divide a player's wage's logarithm by their best overall rating. After doing so, we would compare the mean lwage per overall for all the nationalities to check whether a player's wage is determined by their nationality. We would do so using a boxplot, which is shown in figure 16:

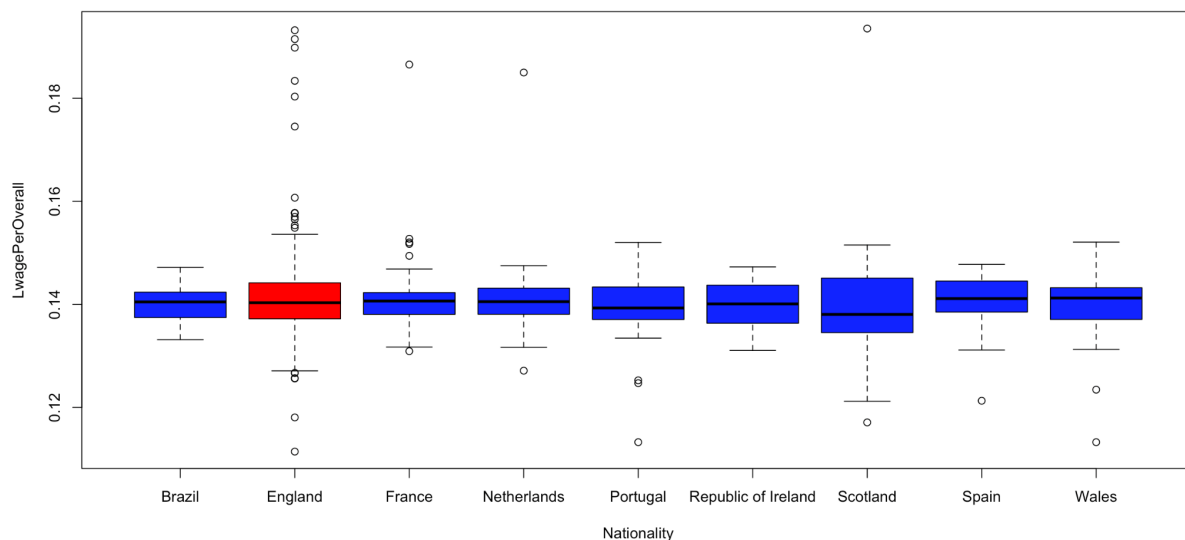


Figure 27. Boxplot of lwage per overall against nationality

From figure 16, we can observe that the distribution of LwagePerOverall for the English players are significantly different from those of other nations. To confirm this, we performed an ANOVA test, which yielded a p-value of less than 0.05, which

means that there is conclusive evidence to conclude that the wages for players of the different nationalities have different distributions.

$H_0: \mu_{\text{Brazil}} = \mu_{\text{England}} = \dots = \mu_{\text{Wales}}$  against  $H_1$ : not all  $\mu$  are equal

```
> summary(aov(Wage ~ Nationality, data = requiredData))
              Df      Sum Sq   Mean Sq F value Pr(>F)
Nationality    8 1.771e+11 2.214e+10   15.27 <2e-16 ***
Residuals   569 8.249e+11 1.450e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 28. Summary of ANOVA test of Wage against Nationality

To find out which nations have different distributions from the others, we did a pairwise t-test and it was observed that the distribution of player wages for players from England, Republic of Ireland, Wales and Scotland are significantly different from the other nations with at least 15 players in the English Premier League due to their low p-values (highlighted in yellow) while having high p-values with each other, signifying that they have similar distributions (highlighted in blue):

```
> pairwise.t.test(requiredData$Wage, requiredData$Nationality, p.adjust.method = "none")

Pairwise comparisons using t tests with pooled SD

data: requiredData$Wage and requiredData$Nationality
```

|                     | Brazil  | England | France  | Netherlands | Portugal | Republic of Ireland | Scotland | Spain  |
|---------------------|---------|---------|---------|-------------|----------|---------------------|----------|--------|
| England             | 2.2e-12 | -       | -       | -           | -        | -                   | -        | -      |
| France              | 0.1323  | 2.9e-09 | -       | -           | -        | -                   | -        | -      |
| Netherlands         | 0.0043  | 0.0036  | 0.1075  | -           | -        | -                   | -        | -      |
| Portugal            | 0.0991  | 1.0e-06 | 0.7934  | 0.2018      | -        | -                   | -        | -      |
| Republic of Ireland | 3.8e-09 | 0.4594  | 1.4e-06 | 0.0053      | 2.2e-05  | -                   | -        | -      |
| Scotland            | 2.2e-05 | 0.1928  | 0.0022  | 0.2035      | 0.0086   | 0.1283              | -        | -      |
| Spain               | 0.0904  | 8.5e-08 | 0.8086  | 0.1738      | 0.9751   | 7.3e-06             | 0.0055   | -      |
| Wales               | 1.3e-05 | 0.6263  | 0.0010  | 0.0998      | 0.0039   | 0.3836              | 0.6224   | 0.0025 |

P value adjustment method: none

Figure 29. Pairwise t-test between wage and nationality

Therefore, the team decided to group the players from the United Kingdom and the Republic of Ireland together as they have similar distributions. We conducted the whole analysis again using this new data and observed that players from the group have a significantly different distribution of lwage per overall compared to players of other nationalities, as seen in figure X:



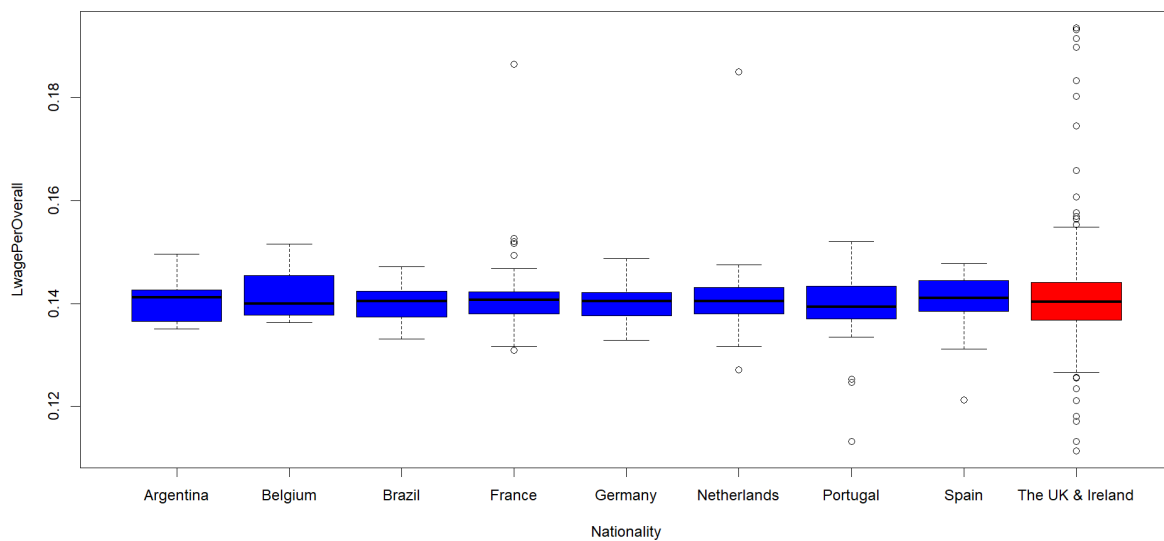


Figure 30. Boxplot of wage per overall against nationality using new data

We conducted the ANOVA test and pairwise t-test, this time using alternative = “less”, and found the following:

- There is conclusive evidence to show that the United Kingdom and Republic of Ireland players’ lwage per overall distribution is greater than those of other nations, with p-values of less than 0.05 (highlighted in yellow)
- The lwage per overall distribution between the other nations cannot be conclusively determined to be different as their p-values are all more than 0.05

$H_0: \mu_{\text{Belgium}} = \mu_{\text{Brazil}} = \dots = \mu_{\text{TheUK\&Ireland}}$  against  $H_1$ : not all  $\mu$  are equal

```
> pairwise.t.test(requiredData$Wage, requiredData$Nationality, p.adjust.method = "none", alternative = "less")
```

Pairwise comparisons using t tests with pooled SD

data: requiredData\$Wage and requiredData\$Nationality

|                  | Argentina | Belgium | Brazil  | France  | Germany | Netherlands | Portugal | Spain   |
|------------------|-----------|---------|---------|---------|---------|-------------|----------|---------|
| Belgium          | 0.6718    | -       | -       | -       | -       | -           | -        | -       |
| Brazil           | 0.7542    | 0.5681  | -       | -       | -       | -           | -        | -       |
| France           | 0.3488    | 0.1767  | 0.0799  | -       | -       | -           | -        | -       |
| Germany          | 0.6508    | 0.4805  | 0.4115  | 0.8006  | -       | -           | -        | -       |
| Netherlands      | 0.0647    | 0.0218  | 0.0038  | 0.0664  | 0.0277  | -           | -        | -       |
| Portugal         | 0.2890    | 0.1418  | 0.0617  | 0.4034  | 0.1613  | 0.8834      | -        | -       |
| Spain            | 0.2914    | 0.1400  | 0.0568  | 0.4105  | 0.1601  | 0.8980      | 0.5116   | -       |
| The UK & Ireland | 5.1e-05   | 3.6e-06 | 2.4e-11 | 1.4e-08 | 9.8e-06 | 0.0036      | 2.7e-06  | 2.9e-07 |

P value adjustment method: none

Figure 31. Pairwise comparisons using t-test with pooled standard deviation

Therefore, through this test, we can conclusively conclude that unless a player is from England, Scotland, Wales or the Republic of Ireland, their wage is not affected by nationality. If a player is from the above-stated countries, their wage would likely be higher. Upon some investigation, it was found that the English Premier League has a Homegrown Player Rule where at least 8 players at each club must have trained in the United Kingdom, which lead to an “English Premium” where players in the region are in higher demand and thus are paid better (Wikipedia, 2020), which is consistent with our findings.

## 5. Conclusion And Discussion

The football industry has a reputation for being one of the most lucrative of the major professional sports leagues, earning money from a combination of television rights, merchandising, ticket sales, and more. In order to sustain its business model, teams will have to attract the most talented players based on their performance, and reward them with attractive salary. In this report, we attempted to answer some of the basic questions related to player wages based on 2021 season data with a variety of variables.

We conclude that:

- Wage of a FIFA player varies according to the best position they play in the game
- “Best Overall Rating” is determined to be the single most important performance measure to model  $\ln(\text{wage})$
- Players in the top 6 clubs in the Premier League are generally paid higher wages than their counterparts not in the top 6 clubs
- Players that are more well-known internationally are generally better paid than their counterparts
- Players are likely to be better paid if their nationality is among English, Welsh, Scottish or Irish in the English Premier League

Although the results of this report are thought-provoking, it must be noted that this report is only based on one single season of data published on the Internet. Furthermore, our data studies are only limited to the English Premier League and that the results may differ among different leagues, such as the Spanish League and more.

A further study into the relationship between other factors such as the height and weight of players could also give a deeper insight into how players' wages are determined, but this project is limited in the understanding of how these variables are viewed by those determining the wages that players receive. There are also factors involved in the negotiations between agents and clubs that will also slightly affect the final wages determined which is not recorded in the dataset.

Conclusively, we believe that a deeper and wider analysis of the data with more advanced analytical techniques would be needed to make a stronger statement about the relationship between the variables and the players' wages.

## 6. Appendix

### 6.1 Kaggle Dataset

The dataset titled 'FIFA21\_Dataset' is obtained from the online data library Kaggle. The original data set consists of a csv data frame containing 17108 observations of 65 variables.

The dataset can be obtained from:

<https://www.kaggle.com/datasets/umeshkumar017/fifa-21-player-and-formation-analysis>

### 6.2 Code For Project

```
#### Import Required Libraries
```

```
library(data.table)
```

```
library(corrplot)
```

```
library(dplyr)
```

```
library(moments)
```

```
library(rpart)
```

```
#### Import Data
```

```
fifaData <- fread("FIFA21_Dataset.csv", header = TRUE, encoding = "UTF-8")
```

```
#### Data Description - Preliminary data cleaning ####
```

```
## Eliminate irrelevant columns
```

```
fifaData <- fifaData[, c('Age','Nationality','Potential','Club','Wage','International  
Reputation','Height','Weight','Reactions','Best Position','Best Overall Rating')]
```

```
## Convert all clumns related to money to numerical data type
```

```
# Remove Wage = 0
```

```
fifaData <- fifaData[fifaData$Wage != "€0"]
```

```
# Function for cleaning the columns related to money
```

```
cleanMoney <- function(x) {
```

```
  x <- gsub("\\€", "", x) # Remove the euro symbol
```

```
  suffix <- substr(x, nchar(x), nchar(x)) # Get the suffix of "K" / "M"{
```

```
  if (suffix == "K") {
```

```
    x <- as.numeric(substr(x, 1, nchar(x) - 1)) * 10^3
```

```
  } else if (suffix == "M") {
```

```
    x <- as.numeric(substr(x, 1, nchar(x) - 1)) * 10^6
```

```
  } else {
```

```
    x <- as.numeric(x)
```

```
  }
```

```
  return(x)
```

```
}
```

```

## Clean all columns related to money
fifaData$Wage <- cleanMoney(fifaData$Wage)

## Remove redundant information
# Converting Weight - In Pounds
fifaData$Weight <- gsub("[a-z]", "", fifaData$Weight)
fifaData$Weight <- as.numeric(fifaData$Weight)

# Converting Height - From Feet To Cm
fifaData$Height <- gsub("[']", "", fifaData$Height)
fifaData$Height <- as.numeric(fifaData$Height)
n <- nchar(fifaData$Height)
fifaData$Height <- as.numeric(substr(fifaData$Height, 1, 1)) * 0.3048 +
as.numeric(substr(fifaData$Height, 2, n)) * 0.0254 # Converting feet & inches to meters

## Filtering for clubs in English Premier League
english <- c("Arsenal",
            "Aston Villa",
            "Brighton & Hove Albion",
            "Burnley",
            "Chelsea",
            "Crystal Palace",
            "Everton",
            "Fulham",
            "Leeds United",
            "Leicester City",
            "Liverpool",
            "Manchester City",
            "Manchester United",
            "Newcastle United",
            "Sheffield United",
            "Southampton",
            "Tottenham Hotspur",
            "West Bromwich Albion",
            "West Ham United",
            "Wolverhampton Wanderers")
fifaData <- fifaData[fifaData$Club %in% english, ]

## Creating new variable lwage (Attempt to normalize wage)
fifaData[, 'lwage'] = log(fifaData$Wage)

summary(fifaData)
str(fifaData)

#### Description and Cleaning of data set

# Exploratory Data Analysis For Dependent Variable - Wage

# 3.1 Summary statistics for wage
summary(fifaData$Wage)
skewness(fifaData$Wage) # 2.08 > 0 Thus Right-Skewed

```

kurtosis(fifaData\$Wage) # 10.53 > 3 Thus Leptokurtic Based On  
<https://www.geeksforgeeks.org/skewness-and-kurtosis-in-r-programming/>

```
hist(fifaData$Wage, breaks = 50) # Right-Skewed
qqnorm(fifaData$Wage)
qqline(fifaData$Wage,col='red')
```

```
# 3.1 Summary statistics for lwage
summary(fifaData$lwage)
skewness(fifaData$lwage)
kurtosis(fifaData$lwage)
```

```
hist(fifaData$lwage)
xpt = seq(from=6,to=13,by=0.01)
ypt = dnorm(xpt,mean=mean(fifaData$lwage),sd=sd(fifaData$lwage))
ypt = ypt * length(fifaData$lwage) * 0.5
lines(xpt,ypt,col='red')
qqnorm(fifaData$lwage)
qqline(fifaData$lwage,col='red')
```

```
# 3.2 - Summary statistics for other variables
```

```
# 3.2.1 - Age
```

```
boxplot(fifaData$Age)
qqnorm(fifaData$Age)
qqline(fifaData$Age,col='red')
```

```
# 3.2.2 - Potential
```

```
boxplot(fifaData$Potential)
qqnorm(fifaData$Potential)
qqline(fifaData$Potential,col='red')
```

```
# 3.2.3 - Reaction
```

```
boxplot(fifaData$Reactions)
qqnorm(fifaData$Reactions)
qqline(fifaData$Reactions,col='red')
```

```
# 3.2.4 - Best Overall Rating
```

```
boxplot(fifaData$`Best Overall Rating`)
qqnorm(fifaData$`Best Overall Rating`)
qqline(fifaData$`Best Overall Rating`,col='red')
```

```
# 3.2.5 - Height
```

```
boxplot(fifaData$Height)
qqnorm(fifaData$Height)
qqline(fifaData$Height,col='red')
```

```
# 3.2.6 - Weight
```

```
boxplot(fifaData$Weight)
qqnorm(fifaData$Weight)
qqline(fifaData$Weight,col='red')
```

```
# remove three outlier as suggested from the boxplot on potential
fifaData = fifaData[!(fifaData$Potential < quantile(fifaData$Potential,0.25) - 1.5 *
IQR(fifaData$Potential))]
# remove four outlier as suggested from the boxplot on weight
fifaData = fifaData[!(fifaData$Weight > quantile(fifaData$Weight,0.75) + 1.5 * IQR(fifaData$Weight))]
```

```
# 3.3 - Should have 796 observations of 12 variables by now
str(fifaData)
```

### ### Statistical Analysis ###

```
# 4.1 - Correlation between ln(Wage) and other continuous variables
fifaData_numeric <- fifaData %>% dplyr::select(where(is.numeric))
fifaData_clean <- cor(fifaData_numeric, use = "pairwise.complete.obs")
corrplot(fifaData_clean, type = "lower", method = "color", addCoef.col = "black", number.cex = 0.6,
tl.cex = 0.5,diag = F,title='Correlation plot for fifaData numerical variables',mar=c(0,0,2,0))
```

### # 4.2 Statistical Tests

```
# 4.2.1 - Relation between Wage and Best Position
boxplot(lwage ~ `Best Position`, data = fifaData, main="Boxplot of log(wage) vs Best Position")
# 4.2.1 - ANOVA Model
aov(fifaData$lwage~factor(fifaData$`Best Position`))
summary(aov(fifaData$lwage~factor(fifaData$`Best Position`))) #pvalue=0.00356, reject null hyp -->
not all means are equal
# 4.2.1 - Pairwise t-test by position
pairwise.t.test(fifaData$lwage, fifaData$`Best Position`, p.adjust.method = "none")
```

```
# Group positions into categories
fifaData$cat = 0
fwd = c("LW", "RW", "RF", "LF", "ST", "CF")
mf = c("CAM", "CM", "CDM", "RM", "LM")
def = c("RWB", "LWB", "RB", "LB", "CB")
gk = "GK"
fifaData$cat[fifaData$`Best Position` %in% fwd] <- "Forward"
fifaData$cat[fifaData$`Best Position` %in% mf] <- "Midfielder"
fifaData$cat[fifaData$`Best Position` %in% def] <- "Defender"
fifaData$cat[fifaData$`Best Position` %in% gk] <- "Goalkeeper"
```

```
#Pairwise t test
# 4.2.1 - ANOVA model
aov(fifaData$lwage~factor(fifaData$cat))
summary(aov(fifaData$lwage~factor(fifaData$cat)))
# 4.2.1 - Pairwise t-test by position category
pairwise.t.test(fifaData$lwage, fifaData$cat, p.adjust.method = "none")
```

### # 4.2.2 - The single most important measure that is affecting ln(Wage)

#### # 4.2.2.1 - Univariate linear regression

```

# against Age
model1 = lm(lwage~Age,data = fifaData)
summary(model1)
qqnorm(model1$residuals)
qqline(model1$residuals,col='red')

# against Potential
model2 = lm(lwage~Potential,data = fifaData)
summary(model2)
qqnorm(model2$residuals)
qqline(model2$residuals,col='red')

# against Reaction
model3 = lm(lwage~Reactions,data = fifaData)
summary(model3)
qqnorm(model3$residuals)
qqline(model3$residuals,col='red')

# against Best overall rating
model4 = lm(lwage~`Best Overall Rating`,data = fifaData)
summary(model4)
qqnorm(model4$residuals)
qqline(model4$residuals,col='red')

# 4.2.2.2 - Multiple Linear Regression
model5 = lm(lwage~Age + Potential + Reactions + `Best Overall Rating`,data=fifaData)
model5_step <- step(model5,direction='backward')
summary(model5_step)

# Multivariate linear regression without potential
model6 = lm(lwage~Age + Reactions + `Best Overall Rating`,data=fifaData)
model6_step <- step(model6,direction='backward')
summary(model6_step)

# 4.2.3 - Relationship between players' wages and their clubs

boxplot(Wage ~ Club, data = fifaData)

# Top 6 clubs = Arsenal, Chelsea, Liverpool, Manchester City, Manchester United, Tottenham Hotspur
top6Clubs <- c('Arsenal', 'Chelsea', 'Liverpool', 'Manchester City', 'Manchester United', 'Tottenham
Hotspur')
fifaData$top6 <- ifelse(fifaData$Club %in% top6Clubs, 1, 0)

# Top 6 clubs players
top6players <- fifaData[fifaData$top6 == 1]
non6players <- fifaData[fifaData$top6 == 0]

var.test(top6players$Wage, non6players$Wage)
# P-value < 2.2e-16, reject null hyp --> the variances are not equal

t.test(top6players$Wage, non6players$Wage, var.equal = FALSE)
# P-value < 2.2e-16, reject null hyp --> the means are not equal

```



# 4.2.3 - Create a table comparing International Reputation and Top6

```
IR_Top6 <- table(fifaData$top6, fifaData$`International Reputation`)
rownames(IR_Top6)=c("non6", "top6")
colnames(IR_Top6)=c("int rep 1", "int rep 2", "int rep 3", "int rep 4")
```

# 4.2.3 - Expected values for International Reputation and Top6 (Optional)

```
colsum = matrix(colSums(IR_Top6), ncol=4)
rowsum = matrix(rowSums(IR_Top6), ncol=1)
exIR_Top6 = rowsum %*% colsum / sum(colsum)
```

# 4.2.3 - Chi-square test

```
chisq.test(IR_Top6)
```

# 4.2.4 - Wage quartiles vs International Reputation

# 4.2.4 - Adding new column WageQuartile

summary(fifaData\$Wage) # To find wage quartiles

```
fifaData$WageQuartile <- ifelse(fifaData$Wage <= 40000, (ifelse(fifaData$Wage <= 10000, 1, 2)),
(ifelse(fifaData$Wage <= 63250, 3, 4)))
```

# 4.2.4 - Create a table comparing International Reputation and Wage quartiles

```
Wage_IR <- table(fifaData$WageQuartile, fifaData$`International Reputation`)
rownames(Wage_IR)=c("1st Qu.", "2nd Qu.", "3rd Qu.", "4th Qu.")
colnames(Wage_IR)=c("int rep 1", "int rep 2", "int rep 3", "int rep 4")
```

# 4.2.4 - Finding expected values (Optional)

```
colsum = matrix(colSums(Wage_IR), ncol=4)
rowsum = matrix(rowSums(Wage_IR), ncol=1)
exWage_IR = rowsum %*% colsum / sum(colsum)
```

# 4.2.4 - Chi-square test

```
chisq.test(Wage_IR)
```

# 4.2.5 - Log wage to best overall rating has correlation coefficient of 0.9

```
fifaData$LwagePerOverall <- fifaData$lwage / fifaData$`Best Overall Rating`
```

# 4.2.5 - All nationalities

```
nationalities <- unique(sort(fifaData$Nationality))
col <- rep("Blue", length(nationalities))
col[nationalities == "England"] <- "Red"
boxplot(LwagePerOverall ~ Nationality, data = fifaData, col = col) # England players have inflated wage per overall
```

# 4.2.5 - Countries with more than 15 players

```
newData <- fifaData %>% count(Nationality, sort = TRUE) %>% filter(n > 15)
nationalities <- sort(newData$Nationality)
col <- rep("Blue", length(nationalities))
col[nationalities == "England"] <- "Red"
requiredData <- fifaData[fifaData$Nationality %in% nationalities, ]
boxplot(LwagePerOverall ~ Nationality, data = requiredData, col = col) # England players have inflated wage per overall
```

```

# ANOVA For Wage Vs Nationality
summary(aov(Wage ~ Nationality, data = requiredData))
pairwise.t.test(requiredData$Wage, requiredData$Nationality, p.adjust.method = "none")

# We see that Republic of Ireland, Scotland and Wales have high p-value with england, meaning they
# have similar distributions, thus we shall group all the Great Britain players together
great_britain <- c("England", "Wales", "Scotland", "Republic of Ireland")
fifaDataGB <- copy(fifaData)
fifaDataGB[fifaDataGB$Nationality %in% great_britain, "Nationality"] <- "The UK & Ireland"

# All nationalities - Great Britain
nationalities <- unique(sort(fifaDataGB$Nationality))
col <- rep("Blue", length(nationalities))
col[nationalities == "The UK & Ireland"] <- "Red"
boxplot(LwagePerOverall ~ Nationality, data = fifaDataGB, col = col) # England players have inflated
wage per overall

# Countries with more than 10 players
newData <- fifaDataGB %>% count(Nationality, sort = TRUE) %>% filter(n > 10)
nationalities <- sort(newData$Nationality)
col <- rep("Blue", length(nationalities))
col[nationalities == "The UK & Ireland"] <- "Red"
requiredData <- fifaDataGB[fifaDataGB$Nationality %in% nationalities, ]
boxplot(LwagePerOverall ~ Nationality, data = requiredData, col = col) # England players have
inflated wage per overall

# ANOVA For Wage Vs Nationality
summary(aov(Wage ~ Nationality, data = requiredData))
pairwise.t.test(requiredData$Wage, requiredData$Nationality, p.adjust.method = "none")
pairwise.t.test(requiredData$Wage, requiredData$Nationality, p.adjust.method = "none", alternative =
"less")

```

## 7. References

- Ambille, I. (2022, January 20). *Premier league prize money 2021/2022: EPL Teams Prize by position, table*. Interesting Football. Retrieved April 5, 2022, from <https://interestingfootball.com/premier-league-prize-money-epl-teams-prize-by-position-table/>
- Bonte-Friedheim, J. (2018, June 18). Premier League or La Liga: Which is the Best Soccer League? theperspective.com/. Retrieved April 1, 2022, from <https://www.theperspective.com/debates/sports/premier-league-or-la-liga-which-is-the-best-soccer-league/#:~:text=The%20English%20Premier%20League%20is%20the%20Best%20at%20Soccer&text=The%20English%20Premier%20League%20is%20the%20world's%20best%20league%20because,quick%2Dpressing%20soccer%20is%20valued>
- Marca. (2021, August 31). *Top 10 highest-paid soccer players in the world in 2022*. MARCA. Retrieved April 1, 2022, from <https://www.marca.com/en/football/international-football/2021/08/31/612e3f93e2704ee36d8b463e.html>
- Wikipedia. (2021, September 20). *Homegrown player rule (england)*. Wikipedia. Retrieved April 1, 2022, from [https://en.wikipedia.org/wiki/Homegrown\\_Player\\_Rule\\_\(England\)#:~:text=It%20forms%20part%20of%20the,at%20least%20eight%20homegrown%20players](https://en.wikipedia.org/wiki/Homegrown_Player_Rule_(England)#:~:text=It%20forms%20part%20of%20the,at%20least%20eight%20homegrown%20players)
- Frost, J. (2021, April 5). Guidelines for removing and handling outliers in data. Statistics By Jim. Retrieved April 5, 2022, from <https://statisticsbyjim.com/basics/remove-outliers/>
- Kelly, R. (2021, April 21). *Who are the Premier League 'big six'? top english clubs & nickname explained*. Who are the Premier League 'big six'? Top English clubs & nickname explained | Goal.com. Retrieved April 5, 2022, from <https://www.goal.com/en/news/who-are-premier-league-big-six-top-english-clubs-nickname/130iokmi8t8dt1k3kudou73s1k>
- Kallner, A. (2017, November 3). Formulas. Laboratory Statistics (Second Edition). Retrieved April 5, 2022, from <https://www.sciencedirect.com/science/article/pii/B9780128143483000010>
- Wikipedia. (2022, March 29). *Coefficient of determination*. Wikipedia. Retrieved April 5, 2022, from [https://en.wikipedia.org/wiki/Coefficient\\_of\\_determination#:~:text=R2%20is%20a%20measure,predictions%20perfectly%20fit%20the%20data](https://en.wikipedia.org/wiki/Coefficient_of_determination#:~:text=R2%20is%20a%20measure,predictions%20perfectly%20fit%20the%20data)