



## **MH3511 Data Analysis with Computer**

### **Group Project**

Football Players: What makes their transfer value high?

Name	Matriculation Number
Jameerul Kader Faizan	U2023863D
Nguyen Tung Bach	U2120390E
Pham Minh Quan	U2140810C

# Table of Contents

<b>1. Introduction</b>	<b>3</b>
<b>2. Data description</b>	<b>4</b>
<b>3. Dictionary, Cleaning and Preparation of Dataset</b>	<b>6</b>
3.1 Summary Statistics for the main variable of interest: Value	6
3.2 Summary statistics for other variables	7
3.2.1 Finishing	7
3.2.2 Passing	8
3.2.3 Indirect shot contribution	9
3.2.4 Defensive actions, ball recovery	9
3.2.5 Touches	11
3.2.6 Ball carrying	12
3.2.7 Ball received	12
3.2.8 Fouls	12
3.2.9 Aerial Ability	12
3.3 Final Dataset for Analysis	13
<b>4. Statistical Analysis</b>	<b>14</b>
4.1 Correlation between numerical variables and log(price)	14
4.2. Statistical Tests	15
4.2.1 Relation between Start and Value	15
4.2.2 Relation between Min and Value	16
4.2.3 Relation between MP and Value	16
4.2.4 Relation between Goals and Value	17
4.2.6 Relation between PasTotCmp. and Value	17
4.2.7 Relationship between PasTotCmp and Value	18
4.2.8 Relationship between ToSuc. and Value	19
4.2.9 Relationship between Carries and Value	20
4.2.5 Relationship between GCA and Value	20
4.3. Multiple Linear Regression	22
4.4. Skills Analysis by Position	23
4.5. Comparison between regular and irregular players	23
<b>5 Conclusion and Limitations</b>	<b>24</b>
<b>6 Appendix</b>	<b>26</b>

# 1. Introduction

Football as a game has always attracted tens of millions of people to watch on average for a single game. With a multitude of leagues and competitions, there's always something happening in the football world. So there is no doubt that there is a demand for skillful football players in the market today. This in turn leads to the said skillful players being bought and sold by the clubs for millions of dollars. For the 2022-2023 Premier League season alone, 10 clubs spent a combined 2.4 billion dollars in buying players for their clubs.

In our project, we chose a dataset containing 2689 players with 126 columns. The columns talk about individual aspects of football players with the main categories mainly being Pass, Shot, Touch, Dribble, Passing, Recovery and Defensive Actions. This dataset not only focuses on players from the Premier League but talks about all major leagues. Here, we are trying to determine what aspect of a football player's skills primarily influences their market price.

Based on this dataset, we aim to answer the following questions:

1. What characteristics of the players affect their transfer value?
2. Are there some characteristics that can dominate the effect on the value of players?
3. Do the important skills vary when it comes to different positions?

## 2. Data description

Our dataset is about football players' stats and their respective details. We retrieved it from <https://www.kaggle.com/datasets/vivovinco/20222023-football-player-stats>. It contains players of all ages and from leagues such as the Premier League, Ligue 1, Bundesliga, Serie A and La Liga. As for the value of each player, we had to source it from <https://data.world/dcereijo/player-scores> and merge it with our dataset.

Before proceeding with data analysis, we performed data cleaning to ensure that the data is fair and balanced to use. The dataset contains 2689 rows and 126 columns. Though most columns have a vast amount of zeros in them, removing them won't make sense as most footballers are not all-rounders and they might lack in several departments hence the zeros in some columns. Hence we will be converging the 125 columns (not including the value column) into several categories and try to find a correlation between each column in the respective category with the value of the player. After which we will remove columns that have minuscule correlation values and focus on the other columns to find which column/feature of the player affects their price the most.

For the first part, we are focusing on the player stats alone and trying to find which stat influences the price. In our second part, we will be viewing it from a different angle: By position to obtain different results as to what might influence the values then.

### 3. Dictionary, Cleaning and Preparation of Dataset

#### 3.1 Summary Statistics for the main variable of interest: Value

The following plots show the distribution of the variable: value.

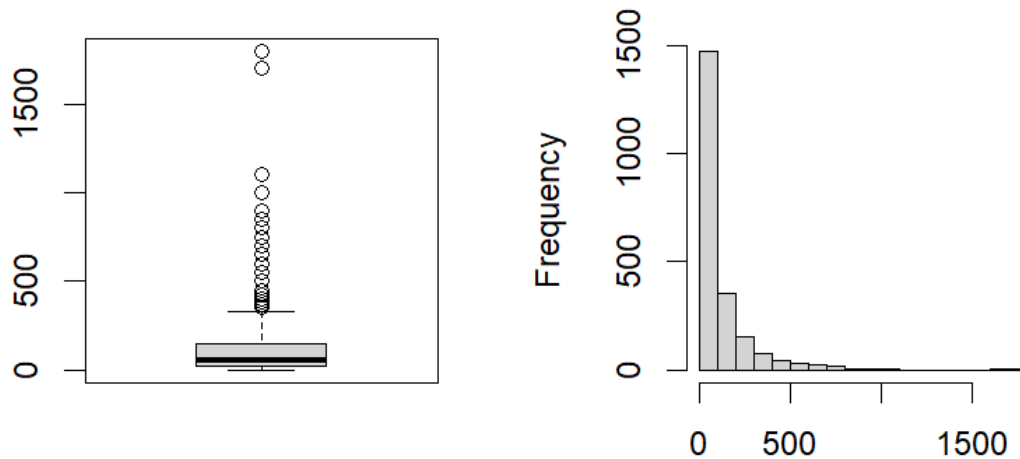
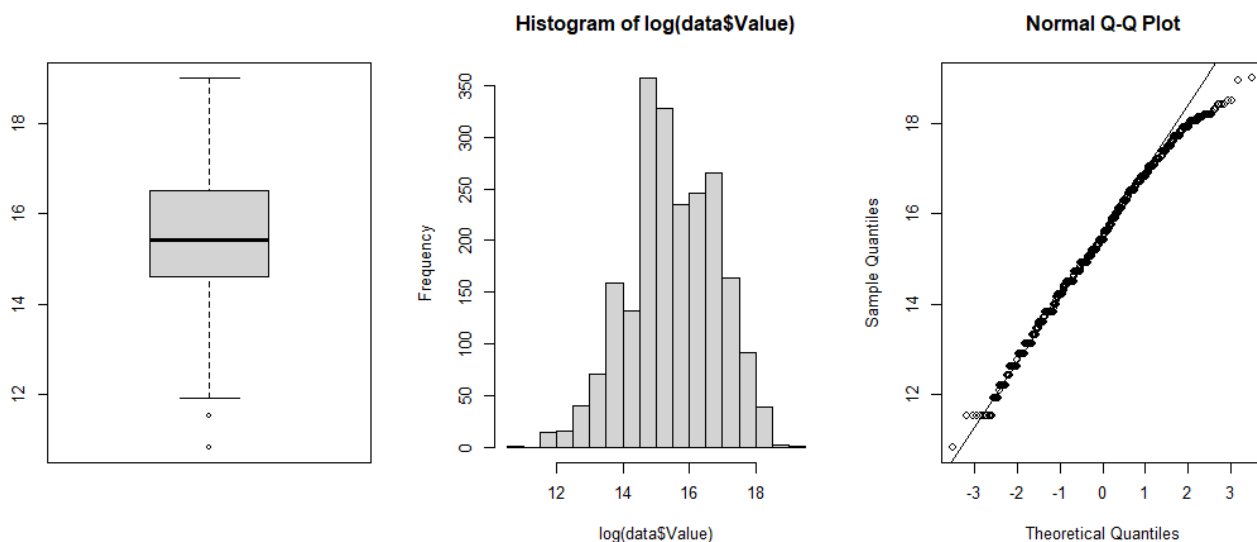


Figure 3.1.1 *Value (units of 100,000)*

We are choosing not to remove any of the outliers for this variable as they might provide some insights as to why some players are valued exceptionally higher than others. However, a log of base e transformation is performed onto the value column to normalize the table and better assist in the later analysis parts. This resulted in a neater-looking value histogram.



**Figure 3.1.2 Log(Value)**

From the QQ plot, it can be seen that log(price) is not completely normally distributed, with the right tail deviating more. Moving forward, this variable has been used. The statistical summary of *log(Value)* is shown in Table 3.1.4 below.

Minimum	1 <sup>st</sup> Quartile	Median	Mean	3 <sup>rd</sup> Quartile	Max
50000	2200000	5000000	11521952	15000000	180000000

**Table 3.1.3 Summary Statistics for Value**

Minimum	1 <sup>st</sup> Quartile	Median	Mean	3 <sup>rd</sup> Quartile	Max
10.82	14.60	15.42	15.49	16.52	19.01

**Table 3.1.4 Summary Statistics for log(Value)**

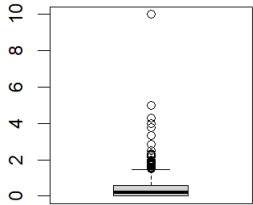
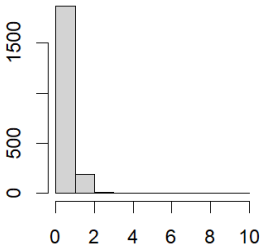
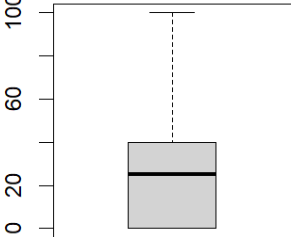
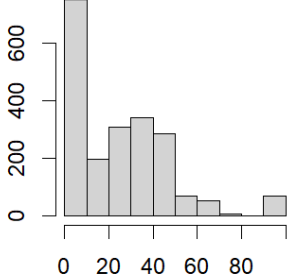
### 3.2 Summary statistics for other variables

In the next subsections, we dropped from 125 variables to 34 variables. We removed variables that were deemed as not important (not relevant much to player ability/profile) or overlapped with some other similar columns. From these 34 columns, we will be streamlining them further by taking a look at the importance and relevance of each of the columns to see which can/will affect the value of the price. We will be omitting all other columns not mentioned below. Histogram and boxplot for each variable that we will be using under each category are shown. We also applied normalization for every variable.

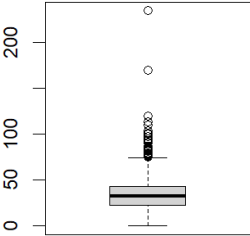
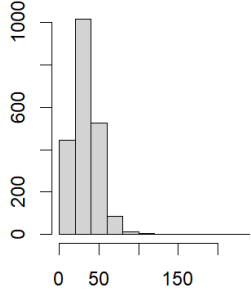
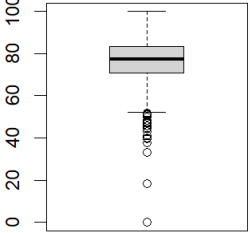
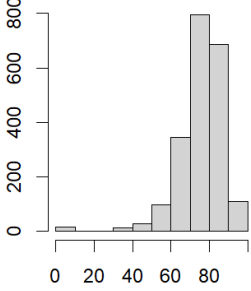
Below are the variables that we did not use for our analysis as they only provided information about each of our players.

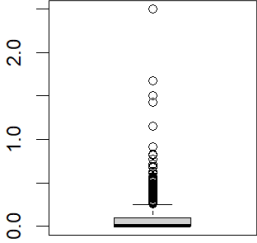
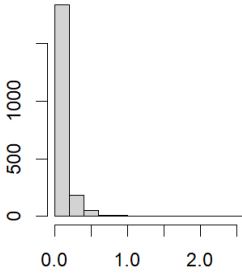
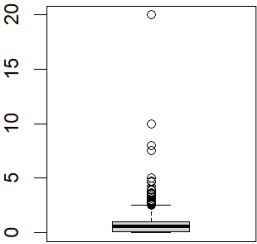
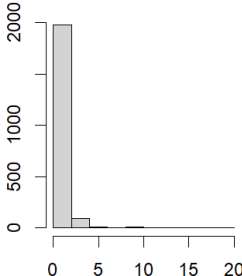
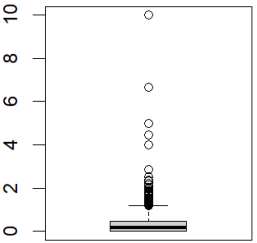
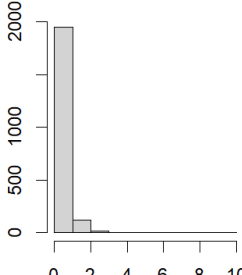
- Rk: Rank
- Player: Player's name
- Nation: Player's nation
- Pos: Position
- Squad: Squad's name
- Comp: League that squad occupies
- Age: Player's age

### 3.2.1 Finishing

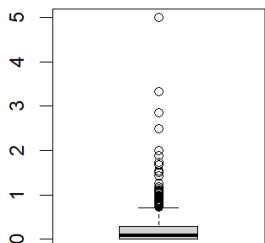
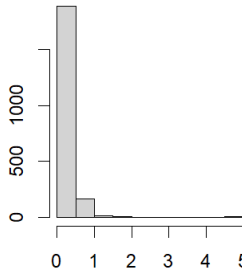
<b>Shots:</b> Total shots	 A box plot showing the distribution of total shots. The y-axis ranges from 0 to 10. The median is approximately 0.5, with the interquartile range (IQR) between 0 and 1. There are several outliers between 2 and 10.	 A histogram showing the frequency of total shots. The x-axis ranges from 0 to 10, and the y-axis ranges from 0 to 1500. The distribution is highly right-skewed, with a peak frequency of over 1500 for 1 shot.
<b>SoT.:</b> Shot on target percentage	 A box plot showing the distribution of shot on target percentage. The y-axis ranges from 0 to 100. The median is approximately 25%, with the IQR between 0% and 35%. There are several outliers between 40% and 100%.	 A histogram showing the frequency of shot on target percentage. The x-axis ranges from 0 to 80, and the y-axis ranges from 0 to 600. The distribution is right-skewed, with a peak frequency of approximately 700 for percentages between 0 and 10.

### 3.2.2 Passing

<b>PasTotCmp:</b> Total passes completed	 A box plot showing the distribution of total passes completed. The y-axis ranges from 0 to 200. The median is approximately 30, with the IQR between 20 and 45. There are several outliers between 50 and 200.	 A histogram showing the frequency of total passes completed. The x-axis ranges from 0 to 150, and the y-axis ranges from 0 to 1000. The distribution is right-skewed, with a peak frequency of approximately 1000 for between 0 and 25 passes.
<b>PasTotCmp.:</b> Pass completion percentage	 A box plot showing the distribution of pass completion percentage. The y-axis ranges from 0 to 100. The median is approximately 75%, with the IQR between 70% and 85%. There are several outliers between 0% and 50%.	 A histogram showing the frequency of pass completion percentage. The x-axis ranges from 0 to 80, and the y-axis ranges from 0 to 800. The distribution is right-skewed, with a peak frequency of approximately 800 for percentages between 60 and 70.

<p><b>Assists:</b> Number of assists per 90 minutes</p>		
<p><b>PPA:</b> Completed pass to 18-yard box</p>		
<p><b>Sw:</b> Side switch pass</p>		

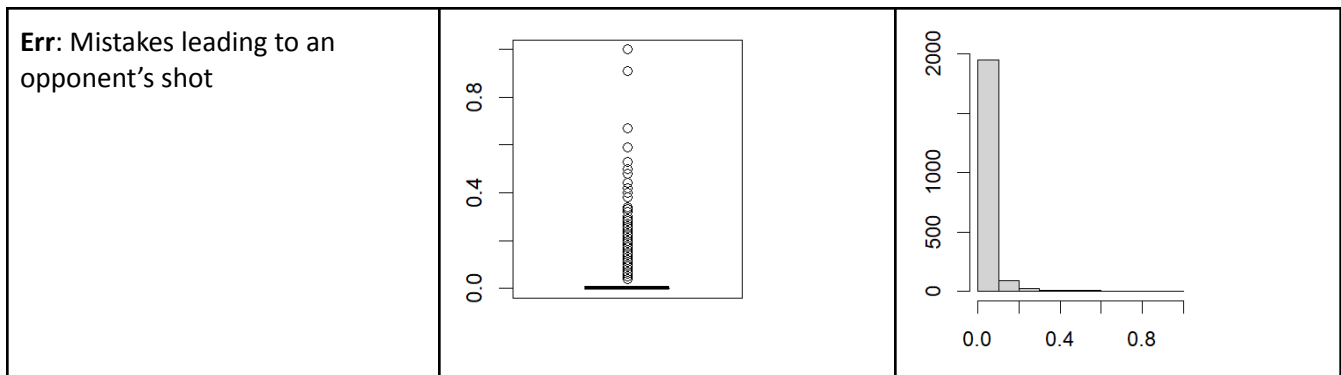
### 3.2.3 Indirect shot contribution

<p><b>GCA:</b> Goal-creating actions</p>		
--	---	---

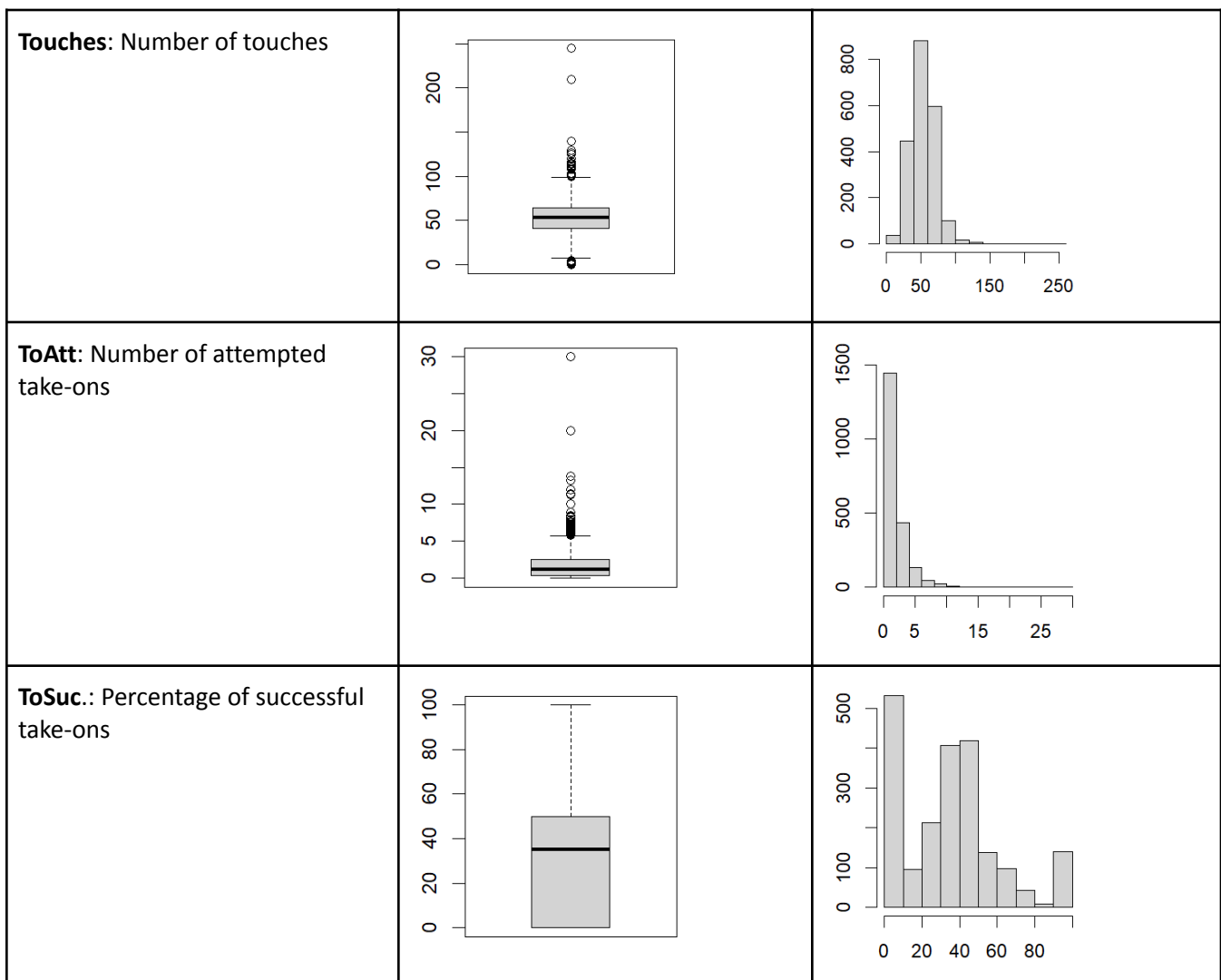


3.2.4 Defensive actions, ball recovery

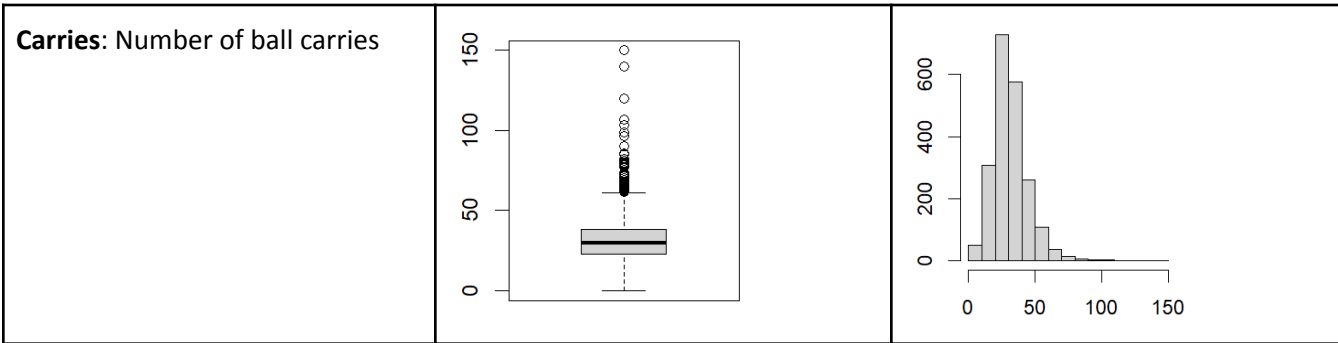
<b>TklWon:</b> Tackles that win back the possession		
<b>Blocks:</b> Number of times blocking the ball		
<b>Int:</b> Interceptions		
<b>Clr:</b> Clearances		



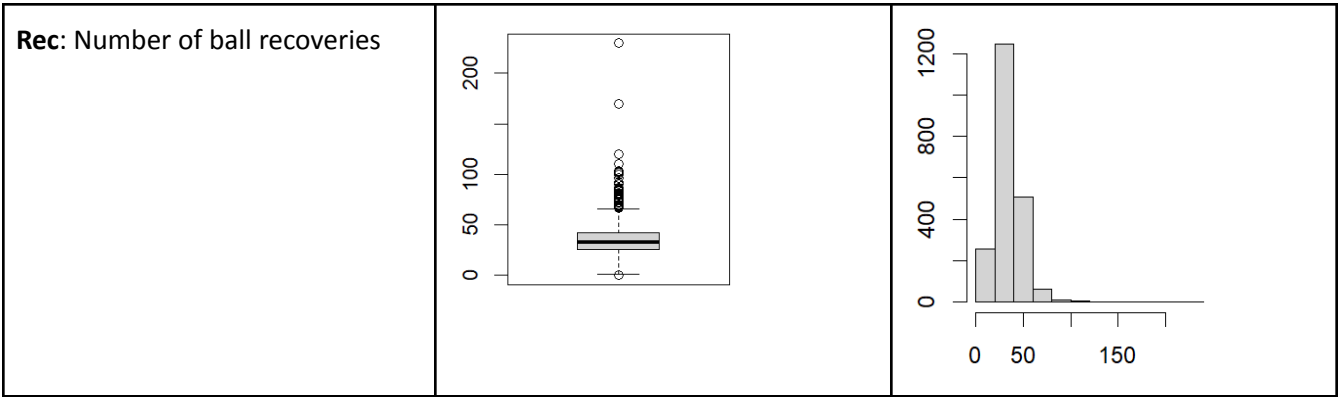
### 3.2.5 Touches



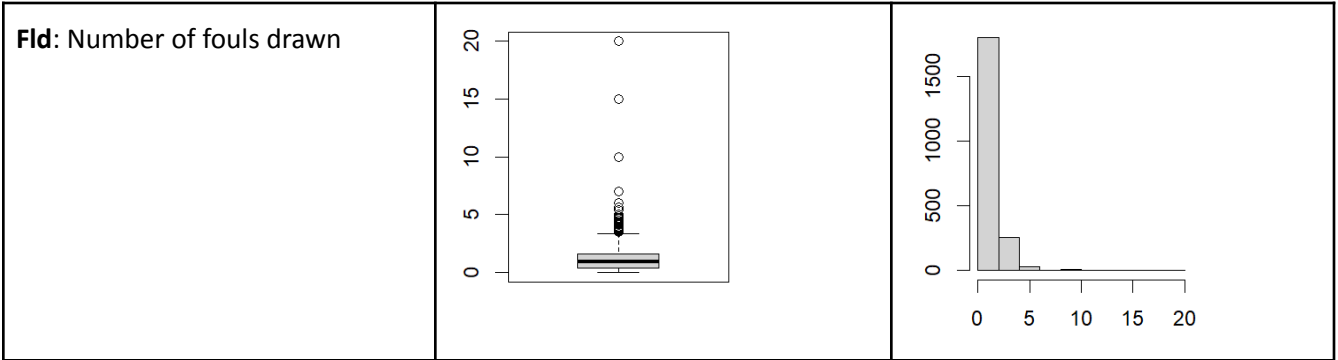
3.2.6 Ball carrying



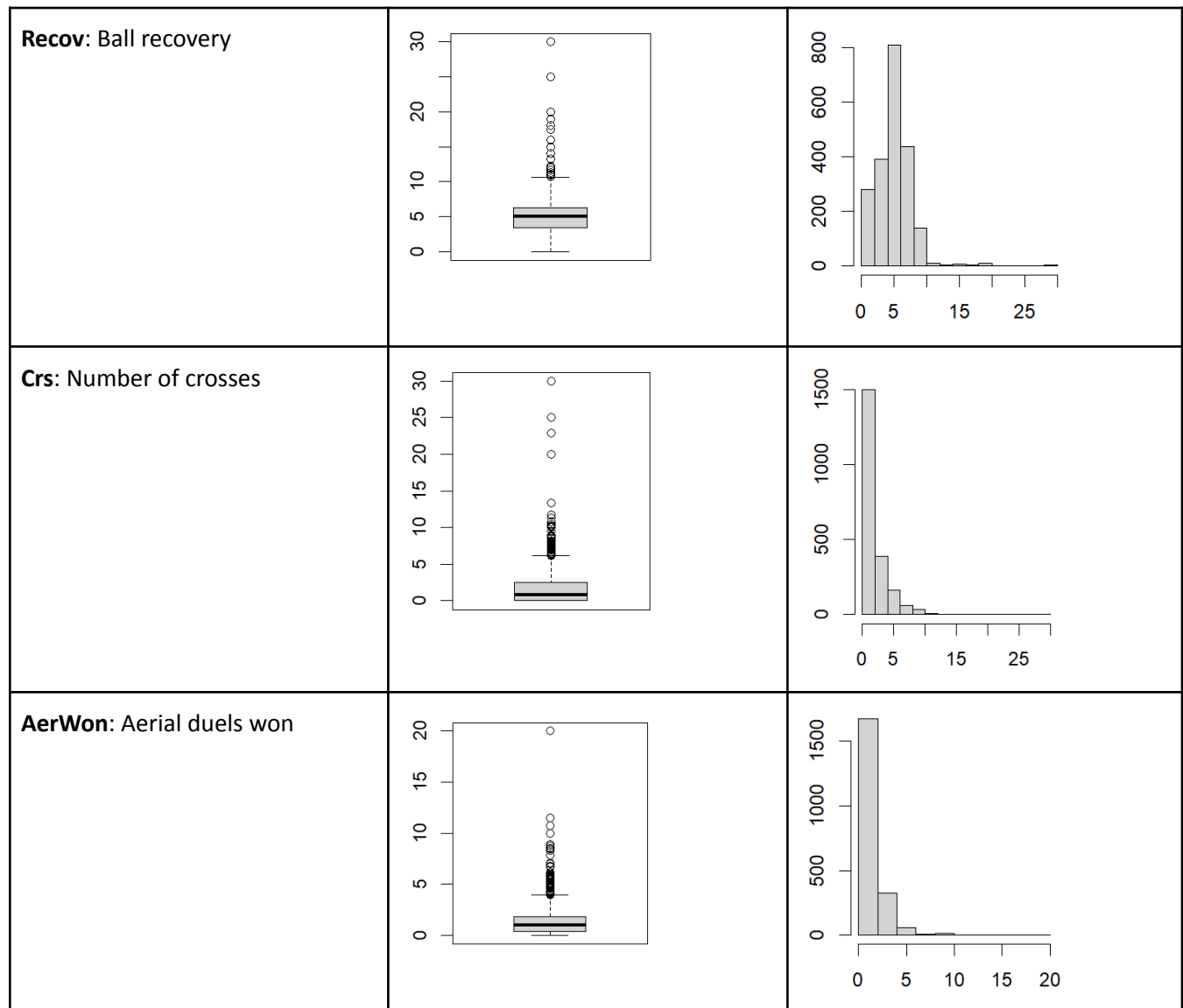
3.2.7 Ball received



3.2.8 Fouls



### 3.2.9 Aerial Ability



### 3.3. Cleaning of Position column

Inside the dataset, there are several players who can play two positions among forwards, midfielders, and defenders. To simplify the categories later, we decide to only take the first two letters of the position column as their official roles. For example, a player with the value on the position column as **FWMF** will be mapped to **FW** only.

### **3.4. Final dataset for analysis**

Based on the further reduction from above, we are left with 34 columns of 2165 rows for our data analysis. We did not remove outliers in our dataset as they may assist in showing any strong correlations or relations between the supposed column and the Value of the player. In the world of football, certain characteristics of a player might contribute to their insane price tag and we wished to delve into that. However, we did remove many columns that we found to be overlapping/similar in nature or did not provide much of an insight or correlation coefficient to warrant their existence in future analysis.

## 4. Statistical Analysis

### 4.1 Correlation between numerical variables and log(price)

We will be taking the top 10 highest correlated values out of the 34 columns from our dataset as we perceived them to be more important than the bottom 24 columns/variables. Then we can conduct further analysis on those factors below.

	Starts	Min	MP	Goals	Rec	PasTotCmp.	PasTotCmp	ToSuc.	Carries	GCA	Value
Starts	1.000										
Min	0.991	1									
MP	0.875	0.899	1								
Goals	0.365	0.355	0.385	1							
Rec	0.014	0.014	0.022	-0.024	1						
PasTotCmp.	0.167	0.168	0.127	-0.066	0.463	1					
PasTotCmp	0.113	0.116	0.049	-0.173	0.892	0.605	1				
ToSuc.	0.203	0.209	0.232	0.084	0.166	0.174	0.148	1			
Carries	0.045	0.047	0.03	-0.07	0.927	0.473	0.876	0.148	1		
GCA	0.023	0.019	0.09	0.232	0.116	-0.017	-0.009	0.088	0.083	1	
Value	0.384	0.383	0.37	0.337	0.247	0.199	0.193	0.183	0.174	0.172	1

**Table 4.1.1 Correlation coefficients (r) between the top 10 factors**

We're unable to use Scatterplots to efficiently showcase the dynamic between each of the variables against the log(value) of the dataset as it does not efficiently show the relation or the regression lines owing to the many variables that our dataset has.

However, from our correlation table, we have selected the top ten variables that affect the log(value) the most. Studying each of the 34 variables does not seem feasible hence why we selected the top ten variables to streamline our process. In this, we see that Starts (Matches started ) has the highest correlation with the log(value) which makes plenty of sense as the number of matches started by a player has the ability to influence their salary and value more than any other variable. However, variables such as Min (Minutes Played) and MP (Matches played) come a close second and third which also seems viable as all 3 factors go hand in hand when deciding the value of a player. You look at the matches played and the minutes played before determining a good value for the said player.

Among the other characteristics of the car, there is an interesting observation that was made:

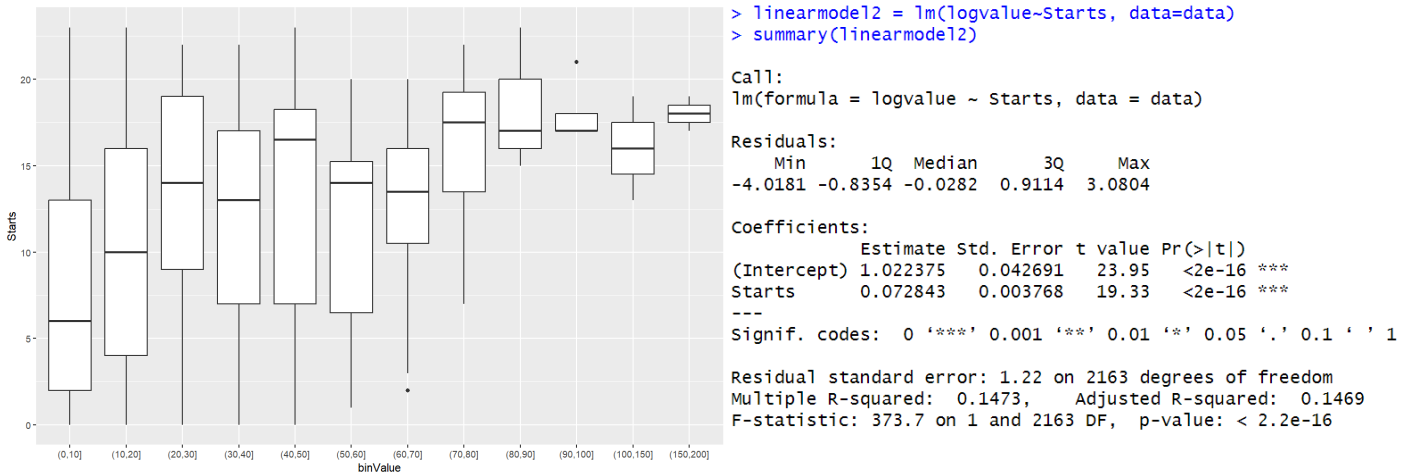
- As mentioned above the variable MP has a high correlation with Starts and Min which shows how closely related all 3 variables are
- The Variable PasTotCmp (Passes completed) has a high correlation with Rec (Number of times a player successfully received a pass). In football, this is explained by the general phenomenon where players are highly likely to pass the ball after receiving it from another player which is proven by this high correlation score.

In the following section, we perform statistical tests to confirm some of these observations.

## 4.2. Statistical Tests

### 4.2.1 Relation between *Start* and *Value*

The boxplot for *Start* (Matches started) against *Value* is shown below.

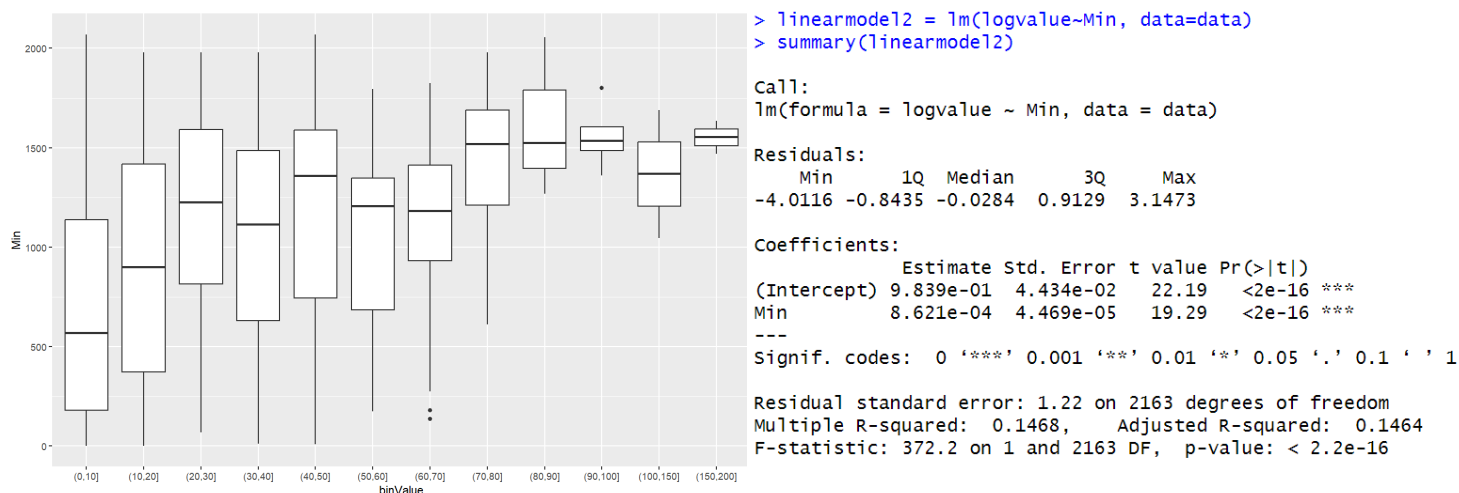


The boxplot shows us that the price range is higher for a player whose median is higher for the number of matches started. And inversely, the value of a player whose number of matches started is low is also low in the 0 to 100 (0 to 10 million) ranges. And the median subsequently increases as the number of matches starts to increase as well. Though when the value reaches above 90 million dollars, we see the median stagnating and even dipping a bit in the 100 to 150 million range.

With the p-value of  $2.2 \times 10^{-16} \ll \alpha = 0.05$  which is a very minuscule amount, we can say that the variable *Starts* has a very high impact on the value of the player. However, the R-squared value ( $R^2=0.1473$ ) tells us that this high impact is seen only in ~14%. Hence we can conclude that while the number of matches started has a high effect on the value of the player it's seen only in 14% of the players.

### 4.2.2 Relation between *Min* and *Value*

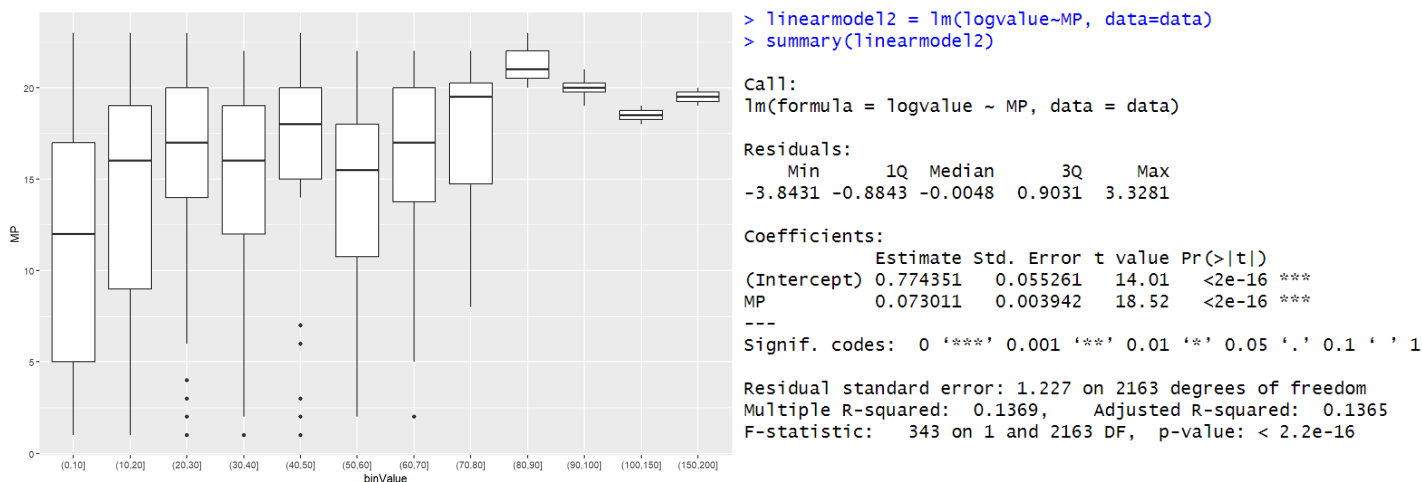
The boxplot follows a similar path to that of the previous boxplot (*Starts* vs *Value*). As we had deduced earlier that *Min*, *Starts* and *MP* shared a strong correlation amongst themselves, so this comes off as no surprise that they affect the value of the players strongly and similarly.



Similar to the comparison above, the relation between Min against Value has a p-value of  $2.2 \times 10^{-16}$   $\ll \alpha = 0.05$  which goes on to show that this variable has a strong effect on the value of the player. Also, the R-squared value ( $R^2=0.1468$ ) tells us that this strong effect plays on 14% of the total cohort which is again similar to the pattern observed above.

### 4.2.3 Relation between MP and Value

The boxplot for MP (Matches played) against Value is below



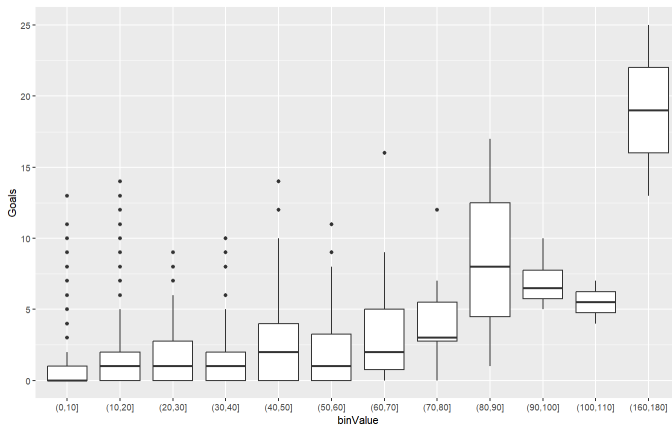
Similar to the above 2 comparisons, the value increases as the median number of matches played increases across the board. We also see that the number of players who managed to play a lot of matches and secure a higher value ( $>90$  million) is comparatively low compared to the range of 60 to 90 million. This can explain why the number of matches played might not be a strong contender to consider when deciding the value of the player. A higher number of matches played might not equate to the strength, versatility or overall performance of a player.



However, our linear regression model paints a similar picture to that of the 2 comparisons above. With the p-value =  $2.2 \times 10^{-16} \ll \alpha = 0.05$ , it shows a very strong effect on the value of a player. Also, the R squared value ( $R^2=0.1369$ ) shows that this effect plays on 13% of the total cohort.

#### 4.2.4 Relation between Goals and Value

The boxplot for Goals (Goals scored or allowed) against Value is below:



```
> linearmodel2 = lm(logvalue~Goals, data=data)
> summary(linearmodel2)
```

Call:  
lm(formula = logvalue ~ Goals, data = data)

Residuals:

	Min	1Q	Median	3Q	Max
	-4.4390	-0.7698	-0.0570	0.9091	2.8742

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.44332	0.03009	47.97	<2e-16 ***
Goals	0.21254	0.01275	16.67	<2e-16 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

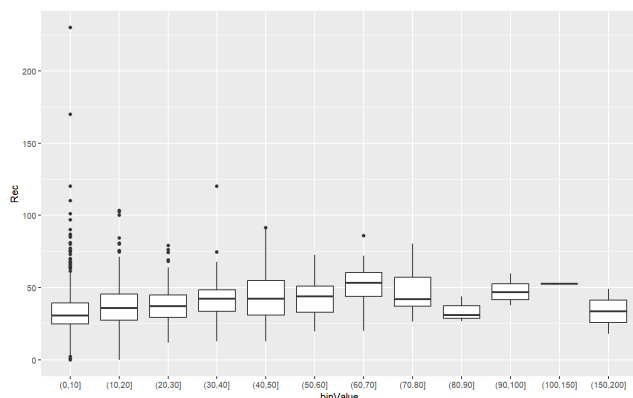
Residual standard error: 1.244 on 2163 degrees of freedom  
Multiple R-squared: 0.1138, Adjusted R-squared: 0.1134  
F-statistic: 277.8 on 1 and 2163 DF, p-value: < 2.2e-16

This boxplot tells us a very convincing story on how the value of the player has increased exponentially when the number of goals scored is higher than the median of the other ranges. With a median score of around 20 goals in this category, they are worth around 160 to 180 million which puts them in a league of their own away from all other ranges. Also, the range of players in the top category is significantly higher than all the other ranges except for 80 to 90 million.

As for the linear regression model, it shows a p-value of  $2.2 \times 10^{-16} \ll \alpha = 0.05$  which shows a very strong effect on the value of the player. The R squared value ( $R^2 = 0.1138$ ) also tells us that this strong effect plays out on 11% of the total cohort.

#### 4.2.5 Relation between Rec and Value

The boxplot shows Rec (Number of times a player successfully received a pass) against Value below:



```
> linearmodel2 = lm(logvalue~Rec, data=data)
> summary(linearmodel2)
```

Call:  
lm(formula = logvalue ~ Rec, data = data)

Residuals:

	Min	1Q	Median	3Q	Max
	-5.2094	-0.8293	-0.0110	0.9817	3.8188

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.929598	0.068548	13.56	<2e-16 ***
Rec	0.021401	0.001806	11.85	<2e-16 ***

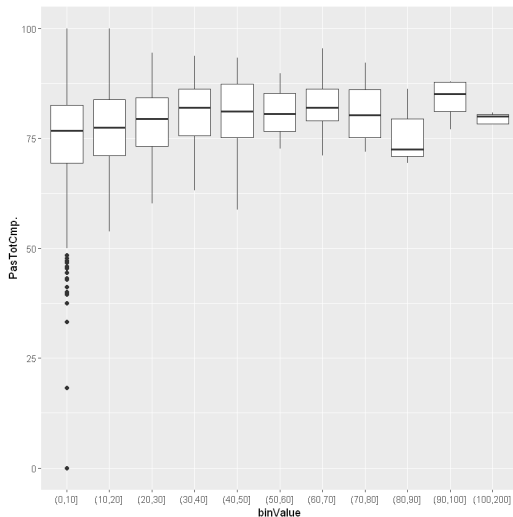
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.28 on 2163 degrees of freedom  
Multiple R-squared: 0.06098, Adjusted R-squared: 0.06055  
F-statistic: 140.5 on 1 and 2163 DF, p-value: < 2.2e-16

The median for the number of times a player successfully received a pass seems quite uniform throughout the ranges of the values.

The linear regression model shows a p-value of  $2.2 \times 10^{-16} \ll \alpha = 0.05$  which shows a strong effect of the Rec against the value of the player. However, the R squared value ( $R^2 = 0.060$ ) shows that this strong effect only impacts 6% of the total cohort which is significantly lower than the first four variables that we were seeing.

#### 4.2.6 Relation between *PasTotCmp.* and *Value*



```
model <- lm(log(Value) ~ PasTotCmp., data = data)
summary(model)
```

```
Call:
lm(formula = log(Value) ~ PasTotCmp., data = data)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-4.7819 -0.7900  0.0037  0.9861  2.6919
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.441965   0.168171   14.521  <2e-16 ***
PasTotCmp.   0.019060   0.002195    8.683  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.226 on 2088 degrees of freedom
Multiple R-squared:  0.03485,    Adjusted R-squared:  0.03439
F-statistic: 75.39 on 1 and 2088 DF,  p-value: < 2.2e-16
```

The boxplot of *PasTotCmp.* and different bins of *Value* is shown below. Players with lower worth generally have a lower successful passing percentage, which is quite visible for values in the range of 0 to 100, with a large number of outliers in the lower range. There are players in the lower value range with quite a high passing percentage, however, this is most likely due to variance stemming from them having a very low playing time, which results in a lower number of pass attempts. We can see this trend being repeated for other percentage-based statistics.

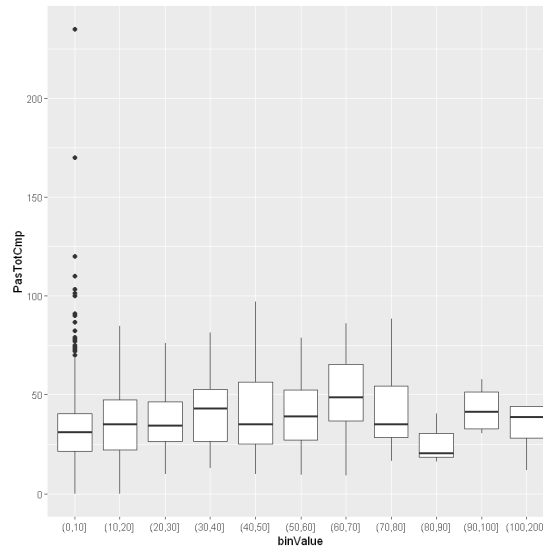
We perform a simple linear regression to test for any significant relationship between *PasTotCmp.* and  $\log(\text{Value})$ .

The regression model provides a p-value of  $2.2 \times 10^{-16} < \alpha = 0.05$ , indicating a statistically significant relationship between *PasTotCmp.* and  $\log(\text{Value})$ . However, the R-squared value for this model is around 0.035, or 3.5%, meaning that the pass completion percentage can only account for 3.5% of the variation in  $\log(\text{Value})$ , suggesting that it may not be too significant.

#### 4.2.7 Relationship between *PasTotCmp* and *Value*

Similarly, we now examine the number of passes completed by a player instead. A boxplot between different bins of *Value* and *PasTotCmp* is shown below, suggesting that the number of passes a player made will increase based on price up to 60-70m, afterward taking a slight drop, which could be due to a

higher number of forwards having higher prices. A large number of outliers can be observed for the lower end of *Value*.



We perform Pearson's product-moment correlation test to examine if there is a statistically significant relationship between the number of passes made and a player's value.

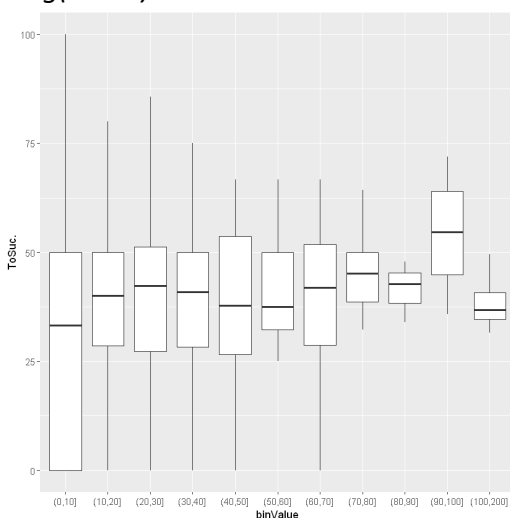
The test provided a p-value of 2.2e-16, smaller than our significance level of 0.05, thus, we can reject the null hypothesis, and there might be a statistically significant relationship between the number of passes made and a player's value.

#### 4.2.8 Relationship between *ToSuc.* and *Value*

From the box plot between *ToSuc.* and *Value*, we observe that the successful take-on percentage of a player generally increases as their value goes up. Some players with lower value have 100% successful take-on percentage, likely to be a result of having few games played or making few take-ons. Aside from that, there seem to be more anomalies this time, with the player with the highest percentage seeming to be in the middle range of value.

Players in the price range of 90-100m have the highest percentage of successful take-ons.

We perform a simple linear regression to test for any significant relationship between *ToSuc.* and  $\log(\text{Value})$ .



```
model <- lm(log(Value) ~ ToSuc., data = data)
summary(model)
```

```
Call:
lm(formula = log(Value) ~ ToSuc., data = data)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-4.3580 -0.8873 -0.0118  0.9690  3.5526
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.362312   0.045457  29.969  <2e-16 ***
ToSuc.       0.008825   0.001017   8.679  <2e-16 ***
```

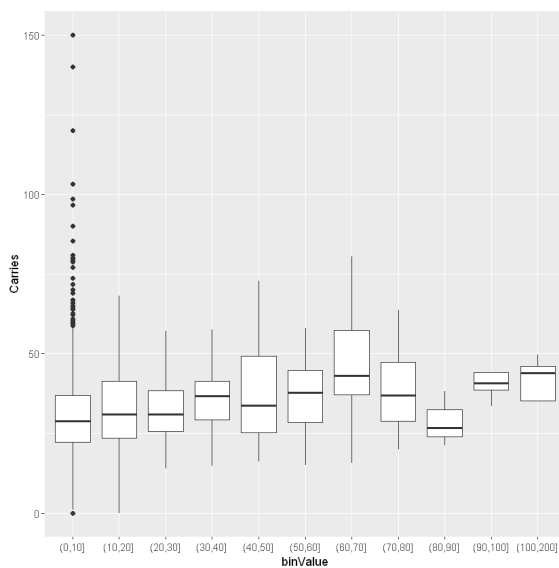
```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.299 on 2163 degrees of freedom
Multiple R-squared:  0.03365,    Adjusted R-squared:  0.0332
F-statistic: 75.32 on 1 and 2163 DF,  p-value: < 2.2e-16
```

The p-value of  $2.2e-16 < \alpha = 0.05$  implies there might be a significant relationship between a player's successful take-on percentage and their price. However, the R-squared value is only 0.0332, meaning that this statistic can only explain slightly over 3% of the variation in *Value*, implying a small level of significance.

#### 4.2.9 Relationship between *Carries* and *Value*

From the boxplot between *Carries* and *Value*, we observe that as a player's value goes up, the number of carries that they complete also goes up. There is quite a large number of outliers in the 0-10 m bin, which could be a result of a large number of defenders and goalkeepers, or the low number of games played. Likewise, there is a sudden drop in the 80-90 m bin, made up of primarily forwards.



```
summary(aov(data$Carries ~ factor(data$binValue)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(data\$binValue)	10	13654	1365.4	7.183	3.12e-11 ***
Residuals	2154	409438	190.1		

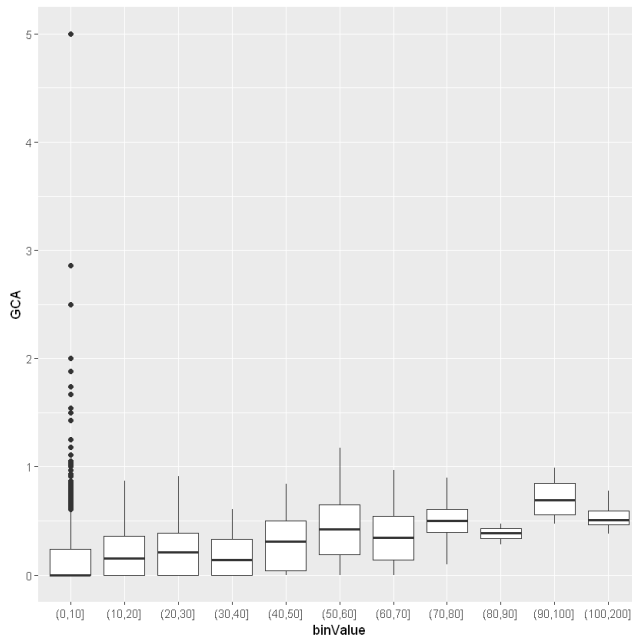
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We examine the relationship between different values and carries using ANOVA between *Carries* and *binValue*, our categorical data.

$\text{Pr}(>F)$  is  $3.12e-11 < 0.05$  our significance level, thus we may reject the null hypothesis, and conclude that there exists some form of correlation between the number of carries and the value of a player.

#### 4.2.10 Relationship between *GCA* and *Value*

As observed from the boxplot between *GCA* and *Value*, as players' valuation increases, their goal-creating actions also increase. Once again there is a large number of outliers in the 0-10m bins, likely caused by variance due to the low number of matches played.



```
model <- lm(log(Value) ~ GCA, data = data)
summary(model)
```

```
Call:
lm(formula = log(Value) ~ GCA, data = data)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-4.5432 -0.8543 -0.0434  0.9743  3.3524
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.54745    0.03201  48.349  < 2e-16 ***
GCA          0.62088    0.07653   8.113 8.17e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.301 on 2163 degrees of freedom
Multiple R-squared:  0.02953, Adjusted R-squared:  0.02909
F-statistic: 65.83 on 1 and 2163 DF, p-value: 8.166e-16
```

We perform a simple linear regression to test the relationship between these two variables.

The model returns a p-value of  $8.166e-16 < 0.05$  significance level, meaning that there is some statistical significance in the relationship between the number of goal-creating actions, and the value of a player. However, R-squared is quite low at slightly less than 0.03, meaning that it can only account for 3% of the variation in value, thus, it is not that significant of a predictor.

### 4.3. Multiple Linear Regression

We can make a new dataset with name **data2** from the original **data** by removing the first 6 columns which are only related to the profile but not skills. Then by using Linear Regression to estimate Transfer Value from the 28 skill factors with the backward elimination method, we get the outcome below.

```
# LINEAR REGRESSION WITH BACKWARD ELIMINATION
```

```
model5 =lm(log(data2$Value)~., data=data2)
```

```
step(model5, direction="backward")
```

```
Call:
```

```
lm(formula = log(data2$Value) ~ Age + Min + Goals + PasTotCmp. +
    Assists + PPA + GCA + Blocks + Int + ToSuc. + Carries + Rec +
    Fld + Crs, data = data2)
```

```
Coefficients:
```

(Intercept)	Age	Min	Goals	PasTotCmp.
2.64595	-0.10300	0.10605	0.16641	0.06505
Assists	PPA	GCA	Blocks	Int
0.06531	0.06096	0.05251	-0.06034	-0.03742
ToSuc.	Carries	Rec	Fld	Crs
0.01536	-0.40540	0.78137	0.04599	-0.05327

We can see that the fitted model only uses 14 out of the 28 columns given in the formula below.

$$\begin{aligned} \text{Log(Value)} = & 2.64 - 0.103 * \text{Age} + 0.106 * \text{Min} + 0.166 * \text{Goal} + 0.065 * \text{PasTotCmp} \\ & + 0.065 * \text{Assist} + 0.06 * \text{PPA} + 0.052 * \text{GCA} - 0.06 * \text{Blocks} - 0.037 * \text{Int} \\ & + 0.015 * \text{ToSuc.} - 0.405 * \text{Carries} + 0.781 * \text{Rec} + 0.045 * \text{Fld} - 0.053 * \text{Crs} \end{aligned}$$

#### 4.4. Skills Analysis by Position

In football, the nature of each position is quite different. Therefore, the players will train themselves on some particular skills that can increase their efficiency in the match. Besides using correlation and regression for the whole dataset as above, we can conduct analysis based on positions.

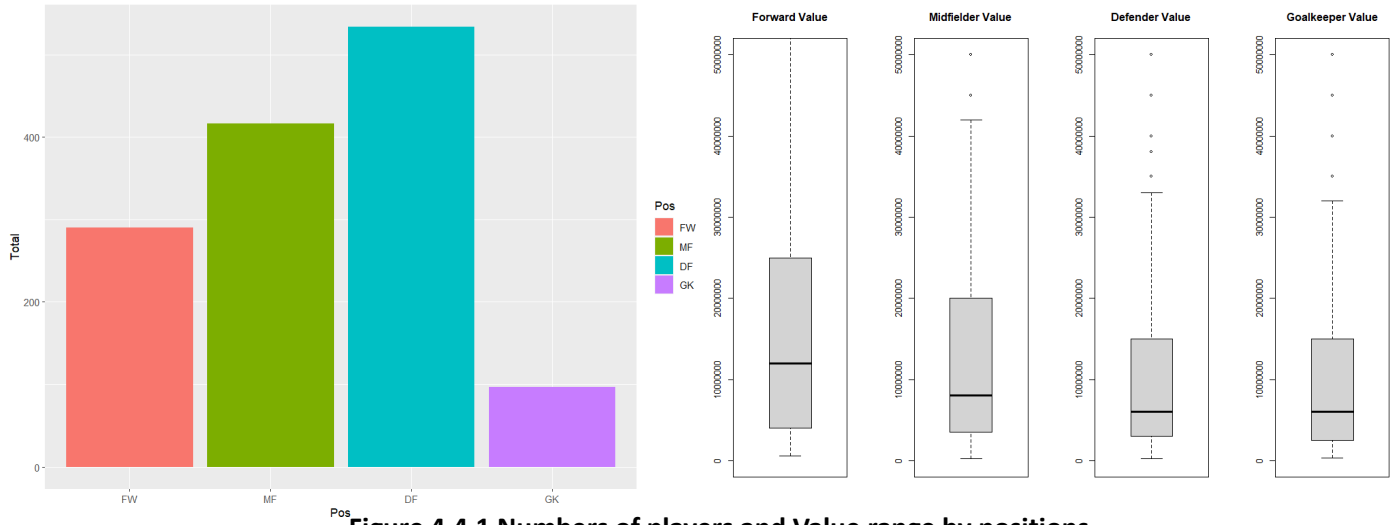


Figure 4.4.1 Numbers of players and Value range by positions

We will conduct Linear Regression again on positions Forward, Midfielder and Defender.

Residuals:  
Min 1Q Median 3Q Max  
-0.20916 -0.05283 -0.00457 0.04043 0.45112

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.050674	0.077162	0.657	0.512101
Age	-0.218046	0.035565	-6.131	0.00000000445 ***
MP	-0.106296	0.048592	-2.188	0.029842 *
Starts	-0.050875	0.123735	-0.411	0.681383
Min	0.163410	0.138589	1.179	0.239733
Goals	0.263321	0.076254	3.453	0.000674 ***
Shots	0.250306	0.061374	4.078	0.00006500346 ***
SoT.	-0.010836	0.046817	-0.231	0.817186
PasTotCmp	0.888652	0.251118	3.539	0.000498 ***
PasTotCmp.	-0.056254	0.093254	-0.603	0.547021
Assists	-0.004425	0.050773	-0.087	0.930640
PPA	0.168084	0.065589	2.563	0.011107 *
Sw	-0.062617	0.054661	-1.146	0.253318
GCA	0.020980	0.057339	0.366	0.714822
Tklwon	0.056196	0.057825	0.972	0.332291
Blocks	0.017074	0.043914	0.389	0.697819
Int	0.042635	0.069427	0.614	0.539833
Clr	0.074682	0.042565	1.755	0.080842 .
Err	-0.064217	0.069382	-0.926	0.355765
Touches	-0.967970	0.344531	-2.810	0.005444 **
ToAtt	0.248119	0.066095	3.754	0.000227 ***
ToSuc.	-0.096588	0.056149	-1.720	0.086913 .
Carries	-0.323302	0.126029	-2.565	0.011026 *
Rec	0.452038	0.245688	1.840	0.067238 .
Fld	-0.044805	0.047414	-0.945	0.345791
Crs	0.021484	0.064039	0.335	0.737607
Recov	-0.095565	0.070157	-1.362	0.174653
AerWon	0.005165	0.080186	0.064	0.948706

(Forward)

Residuals:  
Min 1Q Median 3Q Max  
-0.26789 -0.07065 -0.01678 0.04094 0.46746

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.03802	0.11143	0.341	0.73320
Age	-0.21747	0.03654	-5.952	0.00000000730484 ***
MP	-0.09831	0.05548	-1.772	0.07740 .
Starts	-0.25069	0.13178	-1.902	0.05807 .
Min	0.42774	0.13010	3.288	0.00113 **
Goals	0.16361	0.06222	2.629	0.00899 **
Shots	0.06353	0.06768	0.939	0.34862
SoT.	-0.01688	0.04442	-0.380	0.70415
PasTotCmp	0.66312	0.39616	1.674	0.09519 .
PasTotCmp.	-0.08255	0.12604	-0.655	0.51297
Assists	0.10587	0.05162	2.051	0.04114 *
PPA	0.30586	0.08276	3.696	0.00026 ***
Sw	-0.03463	0.04385	-0.790	0.43030
GCA	0.07155	0.06307	1.134	0.25749
Tklwon	0.07694	0.04962	1.551	0.12204
Blocks	0.11502	0.05504	2.090	0.03748 *
Int	0.02226	0.04487	0.496	0.62014
Clr	0.09681	0.06306	1.535	0.12576
Err	0.03755	0.03671	1.023	0.30717
Touches	-1.07801	0.48044	-2.244	0.02557 *
ToAtt	0.06765	0.07357	0.920	0.35852
ToSuc.	0.04745	0.04150	1.143	0.25380
Carries	-1.04305	0.14345	-7.271	0.0000000000303 ***
Rec	1.57637	0.36671	4.299	0.00002316520514 ***
Fld	0.02489	0.04191	0.594	0.55307
Crs	-0.05196	0.06598	-0.788	0.43154
Recov	0.02238	0.08482	0.264	0.79203
AerWon	0.04808	0.05317	0.904	0.36654

(Midfielder)

Residuals:  
Min 1Q Median 3Q Max  
-0.40381 -0.08091 -0.01771 0.04497 0.59039

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.125071	0.073097	-1.711	0.08785 .
Age	-0.249681	0.039403	-6.337	0.0000000000635 ***
MP	0.165188	0.077398	2.134	0.03343 *
Starts	0.207673	0.172396	1.205	0.22906
Min	-0.195627	0.186054	-1.051	0.29368
Goals	0.059582	0.049433	1.205	0.22880
Shots	0.046839	0.065404	0.716	0.47432
SoT.	-0.004682	0.035184	-0.133	0.89420
PasTotCmp	1.452010	0.591112	2.456	0.01446 *
PasTotCmp.	-0.150071	0.153030	-0.981	0.32735
Assists	-0.017005	0.069298	-0.245	0.80628
PPA	0.149732	0.075711	1.978	0.04865 *
Sw	0.036249	0.050504	0.718	0.47334
GCA	0.034412	0.084500	0.407	0.68405
Tklwon	0.073277	0.054863	1.336	0.18243
Blocks	-0.004018	0.055853	-0.072	0.94269
Int	0.037395	0.049137	0.761	0.44708
Clr	0.194313	0.102796	1.890	0.05944 .
Err	0.078580	0.042117	1.866	0.06281 .
Touches	-1.053276	0.456155	-2.309	0.02145 *
ToAtt	0.179608	0.068656	2.616	0.00923 **
ToSuc.	-0.020192	0.028598	-0.706	0.48057
Carries	-0.611586	0.125333	-4.880	0.00001538708 ***
Rec	0.656839	0.262032	2.507	0.01258 *
Fld	-0.062229	0.041734	-1.491	0.13673
Crs	0.005138	0.080750	0.064	0.94930
Recov	0.070012	0.056068	1.249	0.21251
AerWon	0.035144	0.052384	0.671	0.50268

(Defender)

Using a significance level of 0.05, It can be seen that Age is all important despite different positions. For Forwards, the most significant predictors are Goals, Shots and ToAtt. For Midfielders, PPA, Carries and Rec are the most noticeable. Meanwhile, for defenders, Carries, Touches and Rec are the most outstanding factors. The R-squared value of both is around 0.5, which is better than regression on the original dataset without grouping of around 0.35.

We can also conduct tests on means and variance by groups such as t-test and var-test. With null hypotheses respectively the means and variances of two samples are equal, only the transfer value of defenders and goalkeepers with a high p-value larger than 0.05 will favor that hypothesis. On the other hand, forward and midfielder do not really indicate any similarity in these statistics.

```
> t.test(data_df$Value, data_gk$Value)
```

Welch Two Sample t-test

```
data: data_df$Value and data_gk$Value
t = 0.32406, df = 142.72, p-value = 0.7464
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2314587  3222286
sample estimates:
mean of x mean of y
 11542509 11088660
```

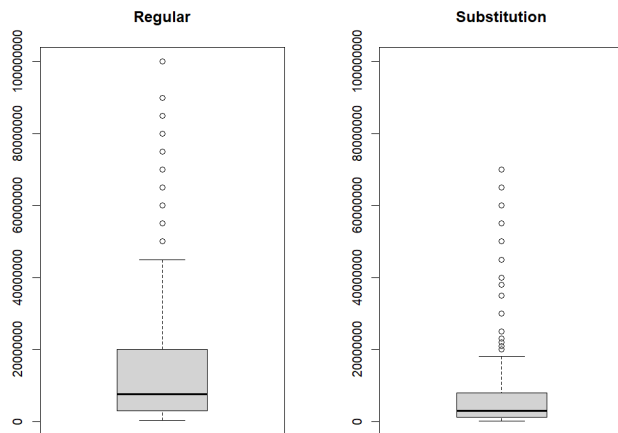
```
> var.test(data_df$Value, data_gk$Value)
```

F test to compare two variances

```
data: data_df$Value and data_gk$Value
F = 1.2376, num df = 533, denom df = 96, p-value = 0.1976
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.893773 1.657400
sample estimates:
ratio of variances
 1.237634
```

## 4.5. Comparison between regular and irregular players

There is often a lower threshold for minutes played so that player statistics can be analyzed properly. We can separate original **data** into **data\_reg** and **data\_sub** to compare if these two categories have an impact on the transfer value.



```
> t.test(data_sub$Value, data_reg$Value, alt='greater')
```

Welch Two Sample t-test

```
data: data_sub$Value and data_reg$Value
t = -12.412, df = 2144.2, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -8593659      Inf
sample estimates:
mean of x mean of y
 6836141 14423859
```

We take the regular players as playing more than 500 minutes, while the remaining will be irregular or substitution players. From the boxplot and t-test, we can clearly indicate the amount of time played is different, even though some statistics of regular players may be smaller than the substitution group. A possible reason can be that it is very hard to maintain excellent skill statistics through many matches.

## 5. Conclusion and Limitation

With the above analysis we are able to come to several conclusions. The top ten factors that we took from the correlation table played a significant part in impacting the value of the player in descending order from Starts to GCA.

- We see that the number of matches started, minutes played and matches played all have an equal R-squared value and p-value. Hence we will be considering them to be a heavy indicator when it comes to deciding the values of the players given the dataset that we have
- Though other variables all have an impact whatsoever on the value of the player, we consider the top 3 variables of *Start*, *Min* and *MP* to be the best indicator when it comes to determining the value of the player

However when it comes to deciding the value of the player by the position that they play in, each position seems to favor a different subset of attributes, which is quite expected as different roles in football are very distinct, occupy different parts of the field, and serve widely different purposes during a game, though, all three roles favor younger players. While the goals or attacking actions can dominate the value for the forward position, it is not that clear for midfielder, defender and goalkeeper. Those positions often favor passing skills, chances created and touches.

Our dataset does come with certain limitations. In the world of football, not one variable determines the price tag of a player and there are several factors beyond the players' talents that come into determining the price of a player.

- Data accuracy: The quality of the dataset is crucial. Our dataset derives from the football seasons of 2022-2023. One year of data might not be enough to build a solid consensus on the type of variable that truly affects a value of a player. Not all data available on the internet is accurate, complete, or up-to-date. Therefore, using a dataset that is not reliable or not taken from a longer time period could lead to inaccurate valuations.
- Data bias: Another issue is the bias that can exist in the data. For instance, the data may only include certain leagues, countries, or seasons, which may skew the valuation. In this case, the dataset derives its data from the Big 5 European Leagues: Bundesliga (Germany), La Liga (Spain), Ligue 1 (France), Premier League (England), and Serie A (Italy). With the dataset omitting other international football leagues such as FIFA World Cup, Liga MX (Mexico), Major League Soccer (MLS) (USA and Canada) and more, conclusions derived from this dataset might not be the norm and could change when other factors and leagues are taken into consideration. Additionally, the data may not account for factors such as injuries, team performance, or off-field issues that can affect a player's value.



- Lack of context: Data alone cannot provide a complete picture of a player's value. Other factors such as the player's age, experience, potential, and marketability, should be considered. For instance, a young player with high potential may have a higher value than an older player with similar stats. Similarly, a young player with a higher fan following could incur a higher value for themselves for the brand that they possess which football clubs might take into consideration.
- Subjectivity: Player valuation is not an exact science, and different people may have different opinions on the value of a player. Football player valuation is also influenced by market perception. For example, a player who is highly sought after by multiple teams may be valued higher than a similar player who has less interest from teams. Therefore, it is essential to consider multiple sources of data and opinions to arrive at a more accurate valuation.
- External factors: Finally, the market itself is subject to external factors such as economic conditions, supply and demand, and geopolitical events that can impact a player's value. Therefore, it is essential to consider these factors when evaluating a player's worth.
- The nature of the sport: Football is more complex than sports such as tennis or basketball, due to its continuity nature with 22 players constantly on the move for 90 minutes, as well as its low-scoring nature, leading to a lot of variances in a single match.

Also, we do not observe a high value of R-squared for multiple linear regression. This is a point of future improvement that we can apply log, square root, or some other mappings on each column variable to test the efficiency change of the fitted model.

With such limitations, coming to a concrete conclusion on what determines a player's value is challenging. However, with the dataset that we chose, we tried to answer the question and hence provide the relevant analysis to support our answer.

We believe the dataset could be improved by incorporating more modern and advanced football statistics, such as expected goals (xG) for forwards, expected goals against goalkeepers, or metrics that could measure hard-to-quantify impacts such as positioning, defensive impacts, and tactics of a particular club.

## 6. Appendix

```
#data preparation
data = read.csv('complete.csv')
data = data[complete.cases(data),]
data$Pos = substr(data$Pos,1,2)
col = c('Rk', 'Player', 'Nation', 'Pos', 'Squad', 'Comp',
        'Age', 'MP', 'Starts', 'Min', 'Goals', 'Shots', 'SoT.',
        'PasTotCmp', 'PasTotCmp.', 'Assists', 'PPA',
        'Sw', 'GCA', 'TklWon', 'Blocks', 'Int', 'Clr',
        'Err', 'Touches', 'ToAtt', 'ToSuc.', 'Carries', 'Rec',
        'Fld', 'Crs', 'Recov', 'AerWon', 'Value')
data = data[, col]
#summary of value
par(mfrow = c(1,3))
boxplot(log(data$Value))
hist(log(data$Value))
qqnorm(log(data$Value)); qqline(log(data$Value))
#create bins for Value
bins <- c(0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200)
data$binValue <- cut(data$Value, bins)
#4.2.1
ggplot(data, aes(x=binValue, y=Starts)) + geom_boxplot(fill='white')
linearmodel2 = lm(Value~Starts, data=data)
summary(linearmodel2)
#4.2.2
ggplot(data, aes(x=binValue, y=Min)) + geom_boxplot(fill='white')
linearmodel2 = lm(logvalue~Min, data=data)
summary(linearmodel2)
#4.2.3
ggplot(data, aes(x=binValue, y=MP)) + geom_boxplot(fill='white')
linearmodel2 = lm(logvalue~MP, data=data)
summary(linearmodel2)
#4.2.4
ggplot(data, aes(x=binValue, y=Goals)) + geom_boxplot(fill='white')
linearmodel2 = lm(logvalue~Goals, data=data)
summary(linearmodel2)
#4.2.5
ggplot(data, aes(x=binValue, y=Rec)) + geom_boxplot(fill='white')
```

```

linearmodel2 = lm(logvalue~Rec, data=data)
summary(linearmodel2)
#4.2.6
ggplot(data, aes(x=binValue, y=PasTotCmp.)) + geom_boxplot(fill='white', group='clarity')
model <- lm(log(Value) ~ PasTotCmp., data = data)
summary(model)
#4.2.7
ggplot(data, aes(x=binValue, y=PasTotCmp)) + geom_boxplot(fill='white', group='clarity')
cor.test(data$PasTotCmp, log(data$Value))
#4.2.8
ggplot(data, aes(x=binValue, y=ToSuc.)) + geom_boxplot(fill='white', group='clarity')
model <- lm(log(Value) ~ ToSuc., data = data)
summary(model)
#4.2.9
ggplot(data, aes(x=binValue, y=Carries)) + geom_boxplot(fill='white', group='clarity')
summary(aov(data$Carries ~ factor(data$binValue)))
#4.2.10
ggplot(data, aes(x=binValue, y=GCA)) + geom_boxplot(fill='white', group='clarity')
model <- lm(log(Value) ~ GCA, data = data)
summary(model)
#4.3
model5 =lm(log(data2$Value) ~ ., data=data2)
step(model5, direction="backward")
#4.4
data_fw = subset(data, Pos=='FW')
data_mf = subset(data, Pos=='MF')
data_df = subset(data, Pos=='DF')
data_gk = subset(data, Pos=='GK')
par(mfrow = c(1,4))
boxplot(data_fw$Value, main = 'Forward Value', ylim=c(0,50000000))
summary(data_fw$Value)
boxplot(data_mf$Value, main = 'Midfielder Value', ylim=c(0,50000000))
summary(data_mf$Value)
boxplot(data_df$Value, main = 'Defender Value', ylim=c(0,50000000))
summary(data_df$Value)
boxplot(data_gk$Value, main = 'Goalkeeper Value', ylim=c(0,50000000))
summary(data_gk$Value)
ggplot(count, aes(x = Pos, y = Max_Value, fill = Pos)) + geom_bar(stat = "identity")

linreg = lm(Value~., data=data_fw); summary(linreg)
linreg = lm(Value~., data=data_mf); summary(linreg)
linreg = lm(Value~., data=data_df); summary(linreg)
var.test(data_df$Value, data_gk$Value)
var.test(data_mf$Value, data_fw$Value)

```

```
t.test(data_df$Value, data_gk$Value)
t.test(data_mf$Value, data_fw$Value)
```

#4.5

```
data_reg = subset(data, Min>=500)
data_sub = subset(data, Min<500)
t.test(data_sub$Value, data_reg$Value, alt='greater')
```

#Contribution

Faizan contributed on introduction and data description, plotting of all 34 columns in final dataset and statistical analyzing 5 highest correlation factors

Bach contributed on ideation, merging dataset with transfer value and analyzing 5 other highest correlation factors

Quan contributed on cleaning the columns, skills analyzing by position, comparing value by minutes played and multiple linear regression.

The remaining tasks such as conclusion/limitation, exploratory data analysis and editing report are divided equally among us.