

**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

MH3511 Data Analysis with Computer

Group Project

FIFA23 Stats: Examining the Numbers Behind the World's Most Popular Video Game

Name	Contribution	Matriculation Number
Tio Guo Yong	Shapiro tests, Pairwise variance tests, ANOVA, Pairwise T-tests, References	
Ivan Lim Khai Ze	Paired T-tests, Two-sample T-tests, Kruskal-Wallis test, Explanations for findings	
Max Tan Han Xian	Summary Statistics, Data Visualization, Chi-square test of Independence, Linear Regression	
Yeoh Ming Wei	Abstract, Introduction, One-sample T-test, Proportion test, Conclusion	

Abstract:

FIFA (International Federation of Association Football) has millions of fans worldwide and the quadrennial World Cup always has everyone excited, and fans all over the world stay up all night to watch their favourite football player play. In our FIFA dataset, there are a lot of exciting variables to see. Through our study and analysis, we aim to explore as much of the dataset as possible and produce insights into some of the burning questions that fans may have about their favourite sport and league. Our study addresses questions such as 'Does the jersey number of the players mean anything?' 'Does the prestige level of the player's team affect his value?' and so on. We will examine the relationships between variables and what meaningful insights they can provide using basic data analysis techniques.

Table of Contents

1. Introduction	3
2. Data Description	3
3. Data Analysis	4
3.1 Description of Dataset	4
3.1.1 Data Cleaning	4
3.1.2 Summary Statistics and Normality Check	5
3.1.3 Data Transformation	9
3.1.3.1 Log Transformation	9
3.1.3.2 BMI of FIFA players	9
3.2 Statistical Analysis	10
3.2.1 Numerical Data Analysis	10
3.2.1.1 One-Sample T-Test of Average BMI of Players	10
3.2.1.2 Proportion Test of Overall Ratings of Players	11
3.2.1.3 Paired T-Test of LongShots vs Finishing	12
3.2.2 Categorical Data Analysis	13
3.2.2.1 Chi-Square Test of JerseyNumber vs BestPosition	13
3.2.3 Mixed Data Analysis	14
3.2.3.1 ANOVA Test of Crossing vs BestPosition	14
3.2.3.2 Two-Sample T-Test of IntPrestige vs Value (Unequal Variance)	17
3.2.3.3 Kruskal-Wallis Test of Overall vs Skill Moves	18
3.2.3.4 Two-Sample T-Test of Preferred Foot vs Shot Power (Unequal Variance)	20
3.3 Correlation and Regression	21
3.3.1 Correlation	21
3.3.2 Simple Linear Regression	21
3.3.3 Residual Analysis	22
4. Conclusion and Discussion	23
5. Appendix	24
6. References	27

1. Introduction

Comprising over 200 national associations in the International Federation of Association Football (FIFA) and millions of fans willing to wake up at 4 am to watch matches of their favourite team, FIFA is a remarkably successful professional sports league. Lots of legends appear and fade, and teams win and lose. Players such as Lionel Messi, Cristiano Ronaldo and Mbappe have become famous favourites of many fans.

Our project uses a dataset containing the statistics of most of the players in FIFA up to the year 2023. Some of the statistics are the salaries, demographic information, and performance statistics of the player. Based on this dataset, we are going to explore the dataset and aim to answer the following questions:

1. Does all the FIFA players' BMI fall between Normal (18.5 to 24.9)?
2. What proportion of players have an overall value that is 80 and above?
3. Does a player's ability to score from long range significantly differ from their ability to score from close range?
4. Does the jersey number have any significant relationship with the player's position?
5. Does crossing value differ by best position?
6. Does the prestige level of the player's team affect his value?
7. Does a player's skill move have any effect in his overall rating?
8. Is there a significant difference in shot power between left-footed and right-footed players in FIFA?
9. Which numerical variable can be used to predict the player's market value?

2. Data Description

To analyze the performance of FIFA 23 players, two datasets were collected from Kaggle, a popular online Data Science community. The original data was obtained by scraping EA Sports' Official Game, FIFA 23 and consisted of two data frames: 'FIFA 23 Players Data.csv' and 'teams_fifa23.csv'. The first dataset contained player data for 18539 players, while the latter had team data for 672 teams in FIFA.

The first dataset served as the primary data source for player information, and only a subset of the second dataset was utilized. By merging these datasets with the club's name, we obtained 18534 observations. After preparation, we selected a subset of 13 variables for analysis. Additionally, we will rename certain variables to make the dataset more organized and understandable and can also make it easier to reference and manipulate variables during analysis.

Original Variable Name	Renamed Variable Name	Description
Overall	Overall	Average key attributes of the player
Value.in.Euro.	Value	Market Value of the players in Euros
Crossing	Crossing	How accurate the player crosses the ball
Height.in.cm	Height	Height of the players in cm
Weight.in.kg	Weight	Weight of the players in kg
Shot.Power	ShotPower	The shooting power rating of the player
Long.Shots	LongShots	Player's ability to shoot accurately from a distance
Finishing	Finishing	Player's ability to score from close range
Best.Position	BestPosition	Best position of the player
Club.Jersey.Number	JerseyNumber	Number wore by the player in his club
IntPrestige	IntPrestige	Prestige level of the player's team at international level
Skill.Moves	SkillMoves	The skill moves rating of the player
Preferred.Foot	PreferredFoot	The foot, which the player strong at

3. Data Analysis

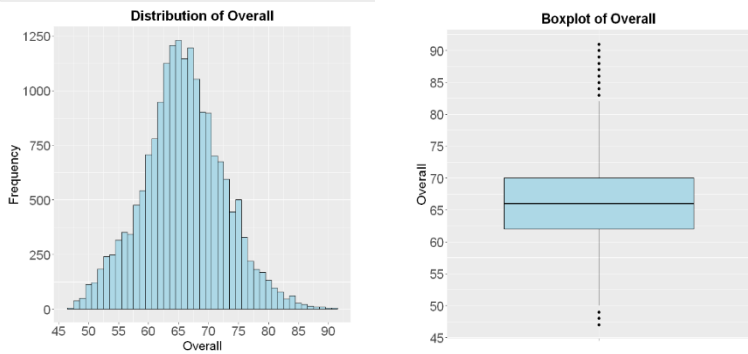
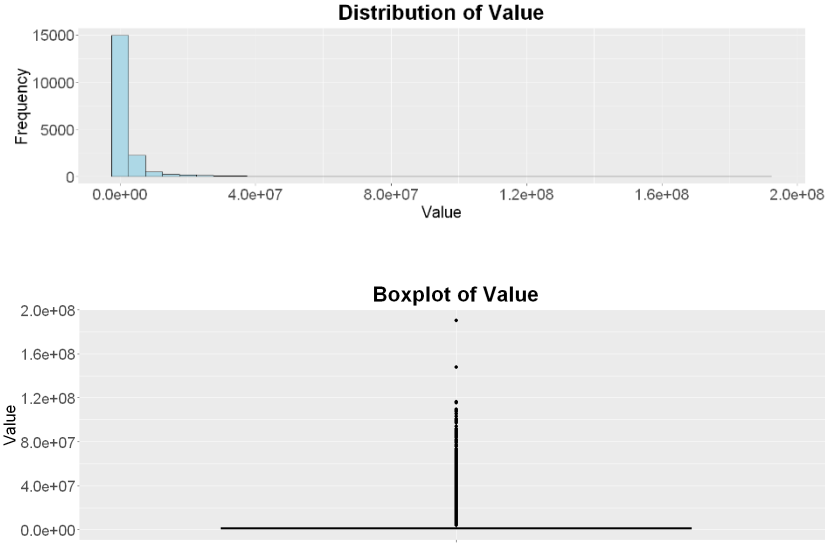
3.1 Description of Dataset

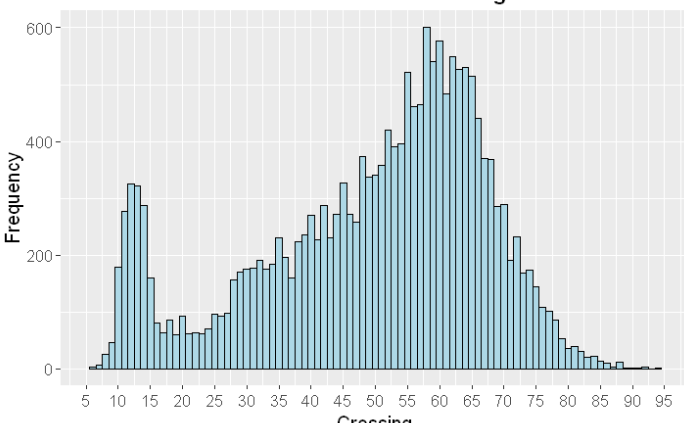
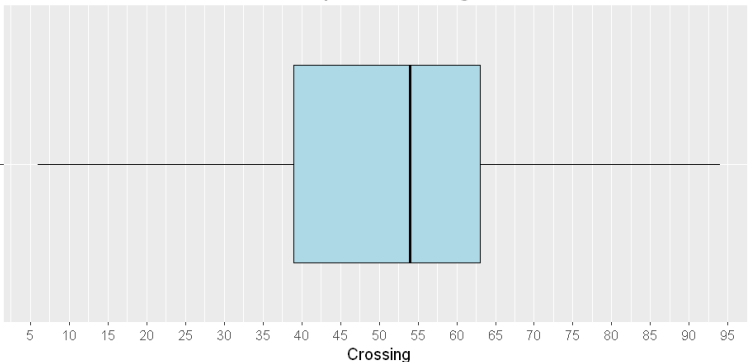
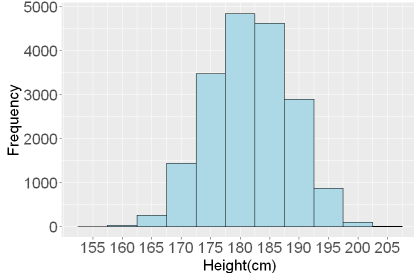
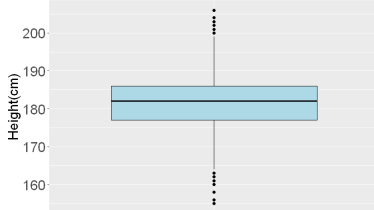
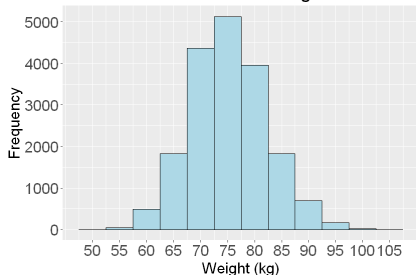
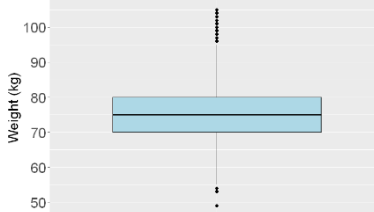
In this section, we will examine the overall distribution of each variable and compute basic summary statistics to gain insights into the data, detect any null values and assess the degree of skewness present in the data.

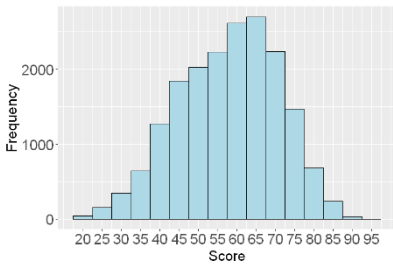
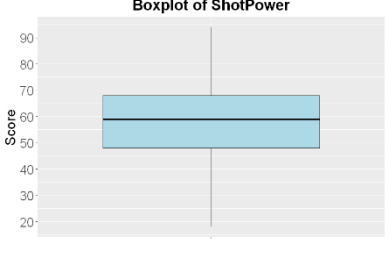
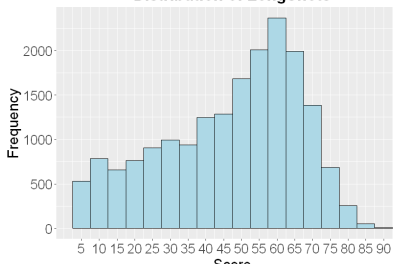
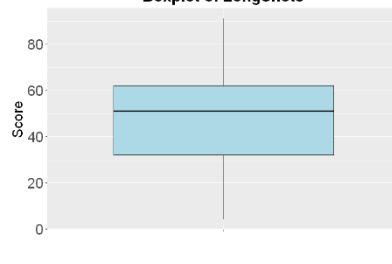
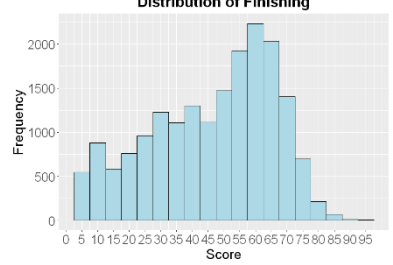
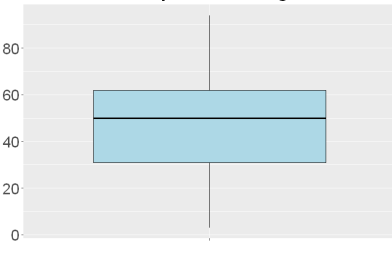
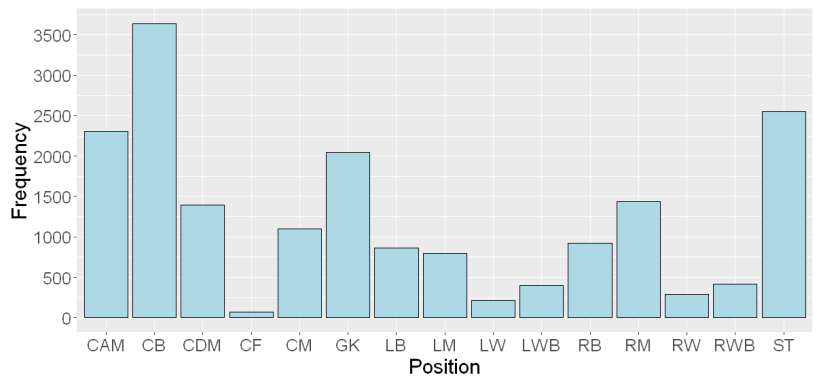
3.1.1 Data Cleaning

Data cleaning is an essential process in data analysis that involves identifying and handling errors, inconsistencies, and missing values in datasets. In this dataset, which is relatively complete, a few variables have a limited number of null values, including 'JerseyNumber' and 'IntPrestige', which has 92 and 28 null values respectively, and 'Value', which has 104 zero values. To ensure data quality, these values will be removed. After removing the null and zero values, the dataset remains with 18407 data points, which should provide a more reliable foundation for data analysis. (Refer to Appendix)

3.1.2 Summary Statistics and Normality Check

Variable Name	Data Visualization	Description
Overall	 <p>The figure displays two plots for the 'Overall' variable. The left plot, titled 'Distribution of Overall', is a histogram showing a symmetrical, bell-shaped distribution of data points ranging from approximately 45 to 90, with a peak frequency of about 1250. The right plot, titled 'Boxplot of Overall', shows a boxplot where the median is around 65, the first quartile is near 62, and the third quartile is near 70, with whiskers extending from 45 to 90 and no significant outliers.</p>	<p>The histogram displays a symmetrical bell-shaped curve, while the boxplot shows that the median, first quartile, and third quartile are closely clustered together, with significantly fewer significant outliers. Thus, Overall is quite normally distributed.</p>
Value	 <p>The figure displays two plots for the 'Value' variable. The left plot, titled 'Distribution of Value', is a histogram showing a heavily right-skewed distribution where most data points are concentrated near 0.0e+00, with a few outliers extending up to 2.0e+08. The right plot, titled 'Boxplot of Value', shows a boxplot where the median is near 0.0e+00, the first quartile is near 0.0e+00, and the third quartile is near 0.0e+00, with numerous outliers extending up to 2.0e+08.</p>	<p>From the histogram and boxplot, it is evident that the data is heavily right-skewed. The histogram reveals that most of the data is concentrated on the left-hand side, while the boxplot indicates the presence of numerous outliers. Given the significant deviation from normality suggested by these observations, data transformation may be necessary to facilitate further analysis.</p>

<p>Crossing</p>	<div data-bbox="483 205 1170 663"> <p>Distribution of Crossing</p>  </div> <div data-bbox="453 705 1203 1094"> <p>Boxplot of Crossing</p>  </div>	<p>The histogram exhibits a bimodal distribution, with a small peak on the left side and a larger peak on the right. The left side of the histogram displays a fat tail, indicating that a significant amount of data falls in the lower range of values.</p> <p>The boxplot displays a distribution skewed towards the lower values, as indicated by the left side of the plot being larger than the right side. However, no outliers are present in the data, suggesting a relatively consistent spread of the values within the interquartile range.</p>
<p>Height</p>	<div data-bbox="391 1144 808 1444"> <p>Distribution of Height</p>  </div> <div data-bbox="829 1157 1208 1394"> <p>Boxplot of Height</p>  </div>	<p>Based on the histogram and boxplot, the data appear normally distributed, as evidenced by the bell-shaped curve on the histogram and the symmetric boxplot with only a few outliers.</p>
<p>Weight</p>	<div data-bbox="396 1501 813 1801"> <p>Distribution of Weight</p>  </div> <div data-bbox="829 1514 1208 1751"> <p>Boxplot of Weight</p>  </div>	<p>Similar to Height, the data seems to follow a normal distribution as suggested by the histogram's bell-shaped curve and the symmetric boxplot with limited outliers.</p>

ShotPower	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>Distribution of ShotPower</p>  </div> <div style="text-align: center;"> <p>Boxplot of ShotPower</p>  </div> </div>	<p>Based on the histogram, the data distribution appears to be normal, with a bell-shaped curve and a symmetrical distribution. Moreover, the boxplot shows that there are no outliers, suggesting that there are no extreme values.</p>
LongShots	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>Distribution of LongShots</p>  </div> <div style="text-align: center;"> <p>Boxplot of LongShots</p>  </div> </div>	<p>Based on the histogram, the distribution of the data is skewed to the left, with most of the data concentrated on the left side and very little data on the right side.</p> <p>However, there are no outliers shown in the boxplot.</p>
Finishing	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>Distribution of Finishing</p>  </div> <div style="text-align: center;"> <p>Boxplot of Finishing</p>  </div> </div>	<p>Similar to LongShots, the distribution is skewed to the left, with most of the data concentrated on the left side and only a small amount on the right side.</p> <p>No outlier is shown based on the boxplot.</p>
BestPosition	<p style="text-align: center;">Distribution of BestPosition</p> 	<p>The variable comprises 15 distinct categories, and the distribution among them is uneven, with 'CB' having the highest frequency and 'CF' having the lowest frequency.</p>

JerseyNumber	<p>Distribution of JerseyNumber</p>	<p>The dataset consists of 99 categories, with most of the data concentrated in the first half of the category range. However, it is notable that categories '77' and '99' stand out the most in the latter half of the range.</p>
IntPrestige	<p>Distribution of IntPrestige</p>	<p>The variable consists of 10 ordinal categories, each labelled from 1 to 10. The number of observations decreases as the rating increases, with 1 containing the largest data points and 10 having the fewest.</p>
SkillMoves	<p>Distribution of SkillMoves</p>	<p>The rating consists of 5 ordinal categories. The higher the rating, the more skilled the player is. Based on the chart, most players have a skill level of 2, while a minority have a maximum skill level of 5.</p>
PreferredFoot	<p>Pie Chart of Preferred Foot</p>	<p>Based on the pie chart, it can be inferred that approximately one-quarter of the players favour using their left foot, while the remaining prefer using their right foot.</p>

3.1.3 Data Transformation

3.1.3.1 Log Transformation

The 'Value' variable is highly skewed with a long tail on the right side. As a result, we applied log transformation to this variable to reduce the skewness and make the distribution more symmetrical.

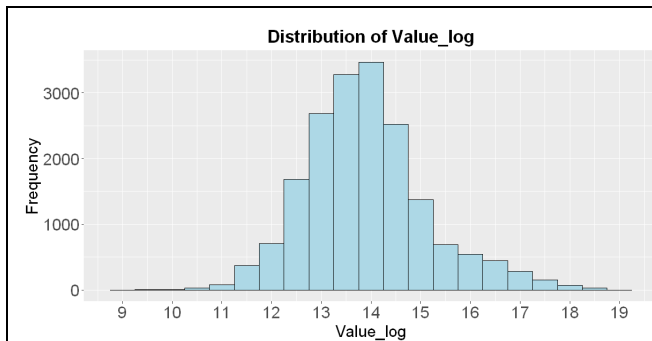


Figure 1: Histogram of Value_log

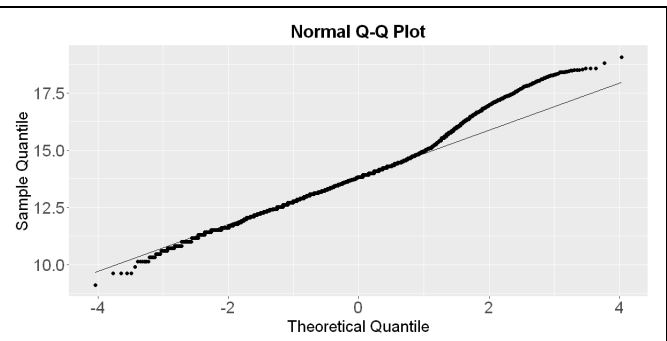


Figure 2: Normal Q-Q plot of Value_log

After the transformation, we can observe that the data points are more evenly distributed, and the extreme values no longer dominate the distribution. This makes analysing and visualising the relationship between 'Value_log' and other variables easier.

3.1.3.2 BMI of FIFA players

In the following test, we will need to use the BMI of FIFA players. Thus, we extract the height and weight of the players from the dataset to calculate the BMI using the subset function.

Now, we will create a new column called BMI using the formula $\text{Weight} / (\text{Height}/100)^2$.

```
1 stats <- stats %>%  
2   mutate (BMI = Weight / (Height/100)^2)  
3 BMI <- stats$BMI
```

Figure 3: Formula to create a column containing BMI

This BMI column will be used in the one-sample t-test below.

3.2 Statistical Analysis

3.2.1 Numerical Data Analysis

3.2.1.1 One-Sample T-Test of Average BMI of Players

Does all FIFA players' BMI fall within the range of Normal (18.5 - 24.9)?

It has always been a provocative question whether all football athletes must have a BMI within a certain range, typically within the 'Normal' range. Therefore, we will perform a one-sample t-test to see if the average BMI of the FIFA player falls on the value of 21.7, which is the average value of 18.5 to 24.9. Then, we will construct a confidence interval with $\alpha = 0.05$ and see if the BMI falls in the stipulated range.

We will carry out our one-sample t-test using the BMI obtained from the data transformation above.

One-sample T-test

$$H_0: \mu_{BMI} = 21.7$$

$$H_1: \mu_{BMI} \neq 21.7$$

```
1 t.test(BMI, mu = 21.7)

One Sample t-test

data: BMI
t = 105.79, df = 18538, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 21.7
95 percent confidence interval:
 22.76332 22.80346
sample estimates:
mean of x
22.78339
```

Figure 4: Result of one-sample t-test of BMI

From the p-value obtained, the null hypothesis is rejected, and it can be concluded that the average BMI of the players is not 21.7.

Conclusion

A fascinating fact that we can observe here from the one-sample t-test result is that the 95 per cent confidence interval of the mean of the BMI is significantly tight, with a width of only 0.04014. The mean BMI is also shown to be 22.78339. This means the players' BMI is on the upper end of the spectrum of the range 18.5 to 24.9, giving us the insight that FIFA players are mostly quite heavy for their height. This is likely due to the muscle mass they have trained for a long time.

3.2.1.2 Proportion Test of Overall Ratings of Players

What proportion of the players have an Overall rating of more than 80?

Overall are the ratings of the average key attributes of the players, and players with a high Overall rating are seen as players with exceptional skills and talent. Therefore, to produce more valuable insight, we will use the proportion test to see what proportion of players have a rating over 80.

After observing the distribution of Overall, we estimate that the threshold for our proportion test will be 80 and the predefined proportion value of players with ratings over 80 is 0.05, since there should be very few players with such a high rating.

Proportion Test

$H_0: p_{\text{Overall} > 80} = 0.05$

$H_1: p_{\text{Overall} > 80} \neq 0.05$

Then, we will create a column consisting of players with a rating that is more than the threshold.

```
1 overall <- players[c('Overall')]
2 threshold <- 80
3 predefined_proportion <- 0.05
4
5 players_above_threshold <- overall %>%
6   filter(Overall == threshold)
```

Figure 5: R code for data preparation

Then, we find the number of players above the threshold as well as the total number of players.

```
1 nrow(players_above_threshold)
2 nrow(overall)

131
18539
```

Figure 6: Number of players above threshold and total number of players

Now, we carry out the proportion test with continuity correction.

```
1 prop.test(131, 18539, p = predefined_proportion, alternative = 'less')

1-sample proportions test with continuity correction
data: 131 out of 18539, null probability predefined_proportion
X-squared = 718.53, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is less than 0.05
95 percent confidence interval:
 0.000000000 0.008181366
sample estimates:
      p
0.007066185
```

Figure 7: Result of one-tailed proportion test of Overall > 80

Since the p value is less than 0.05, the null hypothesis is rejected, and it can be concluded that the proportion of players with an 'Overall' score that is above 80 is less than 0.05.

Conclusion

From the findings, the prop.test gives us valuable insight that there is less than 5% of the players, which is less than around 926 players who have a rating of more than 80. Realistically, the number should be significantly lower. Therefore, this shows that exceptional FIFA players are exceedingly rare.

3.2.1.3 Paired T-Test of LongShots vs Finishing

Does a player's ability to score from long range significantly differ from their ability to score from close range?

LongShots is a measure of a player's ability to shoot accurately from a distance, while Finishing is a measure of a player's ability to score from close range or in one-on-one situations with the goalkeeper. By comparing these two attributes, we can gain insight into a player's overall shooting ability and whether they are better suited to scoring from distance or from close range.

Since both attributes are numerical, we will perform a paired t-test to compare their mean.

<p>F test to compare two variances</p> <p>data: longShots and finishing F = 0.97374, num df = 18406, denom df = 18406, p-value = 0.07105 alternative hypothesis: true ratio of variances is not equal to 1 95 percent confidence interval: 0.9460056 1.0022844 sample estimates: ratio of variances 0.9737385</p> <p>Figure 8: Result of variance test of LongShots vs Finishing</p>	<p>Paired t-test</p> <p>data: longShots and finishing t = 9.368, df = 18406, p-value < 2.2e-16 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: 0.4494925 0.6873577 sample estimates: mean of the differences 0.5684251</p> <p>Figure 9: Result of paired t-test of LongShots vs Finishing</p>
--	---

Variance test

$$H_0: \sigma_{LongShots}^2 = \sigma_{Finishing}^2$$

$$H_1: \sigma_{LongShots}^2 \neq \sigma_{Finishing}^2$$

At $\alpha = 0.05$, since $0.07105 > 0.05$, we do not reject H_0 . Their variances are equal.

Paired T-test

$$H_0: \mu_{LongShots} = \mu_{Finishing}$$

$$H_1: \mu_{LongShots} \neq \mu_{Finishing}$$

Using a paired t-test with equal variance, the p-value is less than $2.2e^{-16}$. At $\alpha = 0.05$, since $p\text{-value} < 0.05$, we reject H_0 and conclude that there is significant difference in the average Long Shots rating and Finishing rating of players.

Conclusion

From the findings, both stats are different from each other. This implies that players who are good at taking long-range shots may need to improve at finishing scoring opportunities in the box and vice versa. Coaches and teams should consider both stats when assessing a player's overall scoring ability and playing style.

3.2.2 Categorical Data Analysis

3.2.2.1 Chi-Square Test of JerseyNumber vs BestPosition

Does each jersey number belong to a specific best position?

Both variables under consideration are categorical. Therefore the appropriate statistical test to use is the chi-square test of independence. However, due to the substantial number of categories in the data, it is necessary to modify and extract a subset of the data for analysis.

To achieve this, we will extract JerseyNumber values between 1 and 11, which typically correspond to the positions of the starting players in a soccer match (Kipsta, 2021). We will also group the categories in the following format to reduce the number of categories and simplify the analysis. For more information on the meaning of each category, please refer to the appendix.

For BestPosition, we will group the categories as follows:

'Gk' (Goalkeeper): 'GK'

'Def' (Defender): 'CB', 'LB', 'RB', 'LWB', 'RWB'

'Mid' (Midfielder): 'CDM', 'CAM', 'LM', 'RM', 'CM'

'Atk' (Attacker): 'LW', 'RW', 'ST', 'CF'

For JerseyNumber, we will group the categories based on the squad numbers as below (Goal, 2021):

'Group 1': 2, 3, 4, 5

'Group 2': 6, 8, 10

'Group 3': 7, 9, 11

'Group 4': 1

Subsequently, the data is converted into a contingency table. Notably, the categories 'Gk' and number '1' always co-occur; hence, we can drop both categories from the analysis. Afterwards, the expected frequencies of each cell are computed.

Observed Frequencies (Before)					Observed Frequencies (After)					Expected Frequencies				
	Atk	Def	Mid	Gk		Atk	Def	Mid		Atk	Def	Mid		
Group 1	20	1861	315	1	Group 1	20	1861	315	Group 1	432.60	879.78	883.62		
Group 2	199	349	1215	0	Group 2	199	349	1215	Group 2	347.30	706.31	709.39		
Group 3	908	82	772	0	Group 3	908	82	772	Group 3	347.10	705.91	708.99		
Group 4	0	0	0	577										

Now, we verify the conditions for the chi-square test of independence.

- I. Both the variables are categorical variables with more than two categories.
- II. Each observation is considered independent of one another since they correspond to distinct players.
- III. All the cells have more than 5 expected frequencies.

Since all the conditions are met, we can conduct the chi-square test of independence.

Chi-square test of Independence

H_0 : There is no difference in the distribution of grouped jersey number and the position.

H_1 : There is a significant difference in the distribution of grouped jersey number and the position.

```
Pearson's Chi-squared test
data: data_table_2
X-squared = 3921.6, df = 4, p-value < 2.2e-16
```

Figure 10: Result of chi-square test of Independence of grouped jersey number vs position

At a significance level of $\alpha = 0.05$, the obtained p-value $< \alpha$, provides compelling evidence to reject H_0 , indicating that there is a significant association between the grouped jersey number and the position.

Conclusion

The findings suggest that football players who wear jersey numbers between 1 to 11 tend to perform optimally in specific positions on the field. The underlying reasons for this phenomenon may include factors such as tradition and coaching strategies. The study's implications may benefit coaches and team managers in selecting a solid lineup for their team.

3.2.3 Mixed Data Analysis

3.2.3.1 ANOVA Test of Crossing vs BestPosition

Does the crossing value differ by best position?

Since Crossing is a numerical variable and BestPosition is a categorical variable, an analysis of variance (ANOVA) test is adequate to answer the statement. We will narrow down to left and right positions as these positions play a significant role in crossing to other positions. From the boxplots below, we observe that the distribution of crossings is similar among left and right positions.

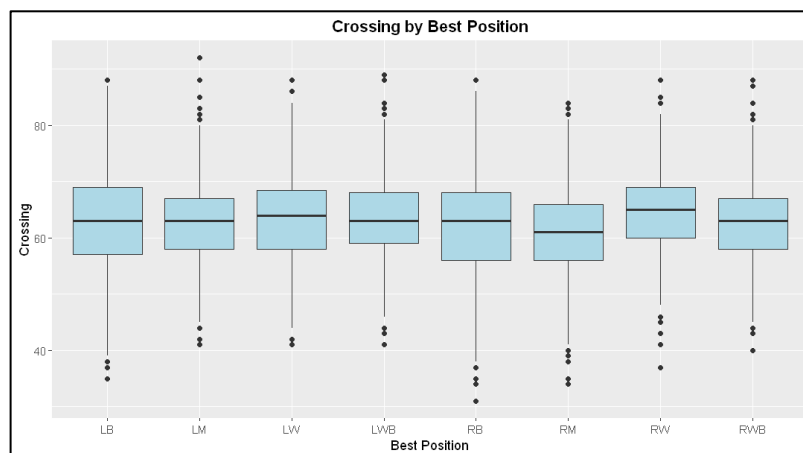


Figure 11: Boxplots of crossing vs best position

Before performing the ANOVA test, the data must fulfill three conditions.

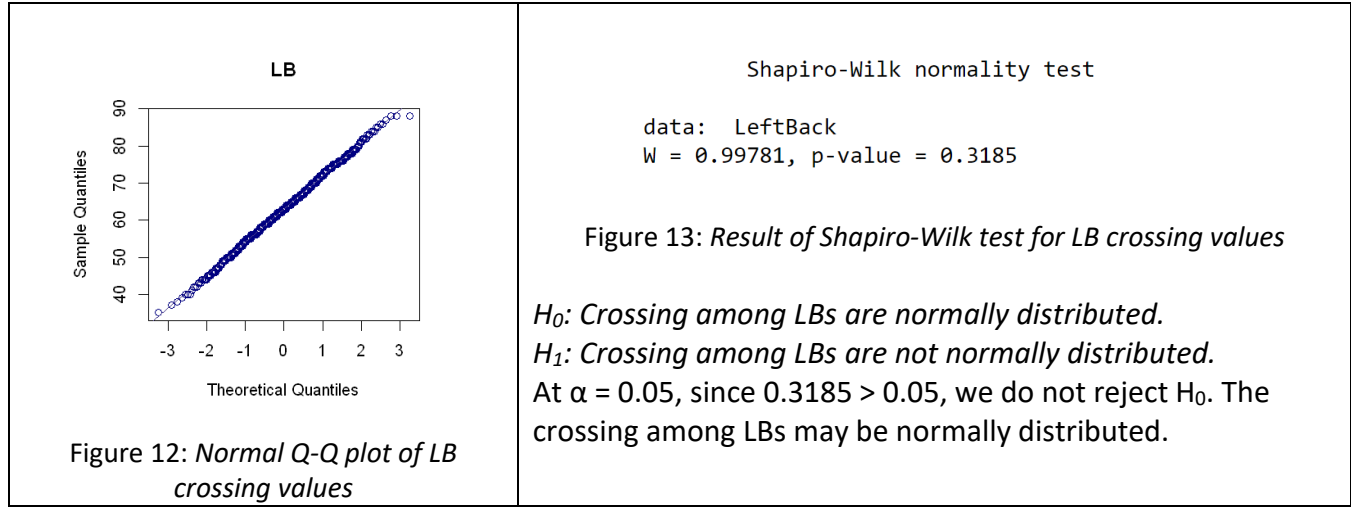
(1) Independence between various positions.

It is satisfied as each player can only have one best position in the data.

(2) Normality.

We use both Q-Q plot and Shapiro-Wilk test. One such example of QQ-plot and Shapiro test for the left back (LB) position is shown below:

Normality test



Even though there are certain positions that may not have normally distributed data by Shapiro test (see Appendix 2), the sample size for each position exceeds 30. Hence, we can use CLT to assume the normality of the samples and condition (2) is also satisfied.

(3) Homogeneity of variance.

We use F-test to check whether two positions have the same variance in crossing. One such example is for the left back (LB) versus right wing (RW) position as shown below:

Variance test

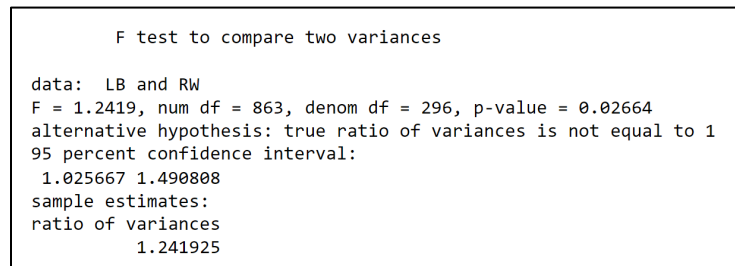


Figure 14: Result of variance test of LB vs RW crossing values

$$H_0: \sigma_{LB}^2 = \sigma_{RW}^2$$

$$H_1: \sigma_{LB}^2 \neq \sigma_{RW}^2$$

At $\alpha = 0.05$, since $0.02664 < 0.05$, we reject H_0 . Their variances are not equal.

This test is performed on every combination of left and right positions (see Appendix 3). In the end, we are only left with 4 positions: left wing back (LWB), right wing back (RWB), left wing (LW) and right wing (RW) which satisfy all conditions.

Then, we proceed with ANOVA to test for the equality of the crossing means.

ANOVA test

```

              Df Sum Sq Mean Sq F value Pr(>F)
BestPosition    3    393   130.85    2.129  0.0947 .
Residuals   1334   81978    61.45
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 15: Result of ANOVA test of crossing vs left and right positions

$H_0: \mu_{LWB} = \mu_{RWB} = \mu_{LW} = \mu_{RW}$

H_1 : At least one mean is not equal.

At $\alpha = 0.05$, since $0.0947 > 0.05$, we do not reject H_0 . Hence, all means are equal. Since the p-value is close to 0.05, let us do pairwise t-tests to check whether there is any pair that has unequal means.

Pairwise t-test

```

Pairwise comparisons using t tests with pooled SD

data: Crossing and BestPosition

      LW      LWB      RW
LWB 0.980 -      -
RW  0.341 0.276 -
RWB 0.221 0.133 0.013

P value adjustment method: none

```

Figure 16: Result of pairwise t-tests of crossing values vs LW, RW, RWB, LWB positions

Take RWB vs RW as an example:

$H_0: \mu_{RWB} = \mu_{RW}$

$H_1: \mu_{RWB} \neq \mu_{RW}$

At $\alpha = 0.05$, since $0.013 < 0.05$, we reject H_0 . Hence, their means are not equal.

We can see that all combinations of LW, RW, LWB, and RWB have equal means while RW vs RWB does not have equal mean.

Conclusion

The findings suggest that right backs (RWB) have significantly lower attacking attributes than right wings (RW). This implies that while both positions are responsible for attacking down the right flank, the role of the RWB may require more defensive responsibilities than the RW. This could mean that RWBs have slightly different skill sets and physical attributes that impact their ability to perform in this position, resulting in a difference in their mean ratings compared to RWs.

3.2.3.2 Two-Sample T-Test of IntPrestige vs Value (Unequal Variance)

Does the prestige level of the player's team affect his value?

Since prestige level is a categorical variable, and players' value is a numerical variable, a two-sample t-test can compare the average player values of two groups and determine if they have a significant difference. In this analysis, we will compare the mean player values of the two groups based on their team prestige level. Specifically, we will focus on teams with international prestige levels of 10 (Tier 1 Club) and 9 (Tier 2 Club) to determine if there is a significant difference in player values between these two groups.

<pre>F test to compare two variances data: tier2Value and tier1Value F = 0.66245, num df = 179, denom df = 115, p-value = 0.0134 alternative hypothesis: true ratio of variances is not equal to 1 95 percent confidence interval: 0.4721045 0.9182839 sample estimates: ratio of variances 0.662453</pre> <p>Figure 17: Result of variance test of tier 1 value vs tier 2 value</p>	<pre>Welch Two Sample t-test data: tier2Value and tier1Value t = -3.9103, df = 209.61, p-value = 6.225e-05 alternative hypothesis: true difference in means is less than 0 95 percent confidence interval: -Inf -8754281 sample estimates: mean of x mean of y 21562917 36722241</pre> <p>Figure 18: Result of two-sample t-test of tier 1 value vs tier 2 value</p>
---	--

Variance test

$$H_0: \sigma_{Tier1Clubs}^2 = \sigma_{Tier2Clubs}^2$$

$$H_1: \sigma_{Tier1Clubs}^2 \neq \sigma_{Tier2Clubs}^2$$

At $\alpha = 0.05$, since $0.0134 < 0.05$, we reject H_0 . Their variances are not equal.

Two-sample t-test

$$H_0: \mu_{Tier1Clubs} = \mu_{Tier2Clubs}$$

$$H_1: \mu_{Tier1Clubs} \neq \mu_{Tier2Clubs}$$

Using a two-sample t-test with unequal variance, the p-value is $6.225e^{-05}$. At $\alpha = 0.05$, since p-value < 0.05 , we reject H_0 and conclude that the players in Tier 1 Club have significantly higher values than those in Tier 2 Clubs.

Conclusion

A team's prestige is based on its history, achievements, and reputation. From the findings, players in higher prestige teams would have a higher market value than players in lower prestige teams. Understanding the relationship between a team's prestige level and its players' value could help players, agents, and teams make more informed decisions about player transfers, contracts, and negotiations.

3.2.3.3 Kruskal-Wallis Test of Overall vs Skill Moves

Does a player's skill move have any effect in his overall rating?

Since Overall is a numerical variable and Skill Moves is a categorical variable, an analysis of variance (ANOVA) test is adequate to answer the statement. The Skill Moves column has a categorical value of 1 to 5, with 1 being the lowest skill rating and 5 being the highest. From the boxplots below, we observe that the distribution of Overall is based on players with different skill ratings.

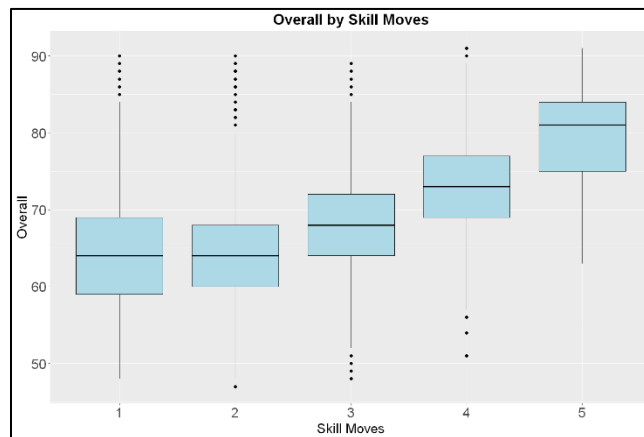


Figure 19: Boxplots of overall vs skill moves

Before performing ANOVA test, the data must fulfill three conditions.

(1) Independence between different skill moves category.

It is satisfied as each player can only belong to skill moves category.

(2) Normality.

Since the sample size for each skill moves rating exceeds 30 (Refer to Appendix). Hence, we can use CLT to assume the normality of the samples and thus condition (2) is satisfied.

(3) Homogeneity of variance.

We use F-test to check whether two different skill moves rating have the same variance in Overall. One such example is for the players with skill moves rated 3 versus players with skill moves rated 4 as shown below:

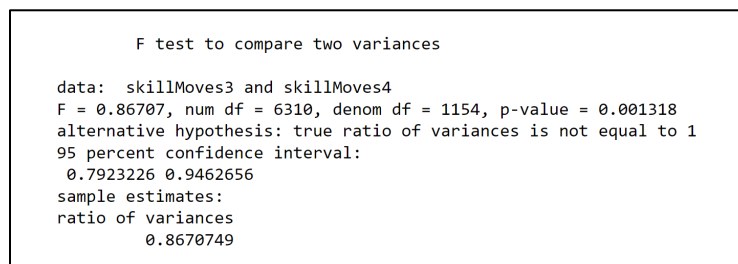


Figure 20: Result of variance test of skillMoves3 and skillMoves4

Variance test

$$H_0: \sigma_{skillMoves3}^2 = \sigma_{skillMoves4}^2$$

$$H_1: \sigma_{skillMoves3}^2 \neq \sigma_{skillMoves4}^2$$

At $\alpha = 0.05$, since $0.001318 < 0.05$, we reject H_0 . Their variances are not equal.

This test is performed on every combination of skill moves rating.

	1	2	3	4	5	6	7	8	9	10
SkillRatingPair	1 2	1 3	1 4	1 5	2 3	2 4	2 5	3 4	3 5	4 5
PValue	0.000	0.000	0.000	0.119	0.000	0.041	0.942	0.001	0.198	0.599

Figure 21: Result of pairwise-variance test of Overall vs Skill Moves

Since most combination have different variance so the assumption for the homogeneity of variance has failed. The assumptions for parametric test have failed, we move on to conduct non-parametric tests, specifically the Kruskal-Wallis Test.

Kruskal-Wallis rank sum test

data: overall and skillMoves

Kruskal-Wallis chi-squared = 3012.1, df = 4, p-value < 2.2e-16

Figure 22: Result of Kruskal-Wallis rank sum test of left foot shot power vs right foot shot power

Pairwise comparisons using Wilcoxon rank sum test

data: overall and skillMoves

1	2	3	4
2	0.00028	-	-
3	< 2e-16	< 2e-16	-
4	< 2e-16	< 2e-16	< 2e-16
5	< 2e-16	< 2e-16	< 2e-16

P value adjustment method: none

Figure 23: Result of pairwise-Wilcoxon rank sum test of left foot shot power vs right foot shot power

Kruskal-Wallis Test

H_0 : Overall Rating is independent of Skill Moves

H_1 : Overall Rating is not independent of Skill Moves

Using a Kruskal-Wallis Test, the p-value is less than $2.2e^{-16}$. At $\alpha = 0.05$, since p-value < 0.05, we reject H_0 and conclude that there is an impact of Skill Moves on Overall Rating of the player.

Conclusion

Skill Moves have a significant impact on a player's Overall Rating, from the boxplot and statistical testing, there is a positive relationship between the two variables. This supports the idea that skilled players are perceived to be more valuable on the field as they can create more opportunities for their team, leading to better performance and more wins. This finding can be useful for teams, coaches, fans, and media in evaluating and assessing player performance and value.

3.2.3.4 Two-Sample T-Test of Preferred Foot vs Shot Power (Unequal Variance)

Is there a significant difference in shot power between left-footed and right-footed players in FIFA?

Shot power is a crucial skill in soccer that determines whether a player can score a goal. Since a player's preferred foot can influence their shot power, it is essential to investigate whether there is a significant difference between left-footed and right-footed players in FIFA. Therefore, we will perform a two-sample t-test on the ShotPower by matching left and right-footed players.

F test to compare two variances

```
data: leftFooted and rightFooted
F = 1.188, num df = 4465, denom df = 13940, p-value = 6.417e-13
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.133050 1.246389
sample estimates:
ratio of variances
 1.188017
```

Figure 24: Result of variance test of left foot shot power vs right foot shot power

Welch Two Sample t-test

```
data: leftFooted and rightFooted
t = 1.5553, df = 7033.8, p-value = 0.1199
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09416966  0.81737184
sample estimates:
mean of x mean of y
58.08486  57.72326
```

Figure 25: Result of two-sample t-test of left foot shot power vs right foot shot power

Variance test

$$H_0: \sigma_{\text{LeftFooted}}^2 = \sigma_{\text{RightFooted}}^2$$

$$H_1: \sigma_{\text{LeftFooted}}^2 \neq \sigma_{\text{RightFooted}}^2$$

At $\alpha = 0.05$, since $6.417e^{-13} < 0.05$, we reject H_0 . Their variances are not equal.

Two-sample t-test

$$H_0: \mu_{\text{LeftFooted}} = \mu_{\text{RightFooted}}$$

$$H_1: \mu_{\text{LeftFooted}} \neq \mu_{\text{RightFooted}}$$

Using a two-sample t-test with unequal variance, the p-value is 0.1199. At $\alpha = 0.05$, since p-value > 0.05 , we accept H_0 and conclude that the shot power of left-footed players and right-footed players have equal mean.

Conclusion

Regarding shooting power, being left-footed or right-footed does not appear to significantly impact a player's ability to score goals in FIFA. Coaches and scouts may want to focus on factors such as a player's preferred position, playing style, physical attributes, and skill set when evaluating their suitability for a particular team or role.

3.3 Correlation and Regression

3.3.1 Correlation

To determine which numerical variable is appropriate for predicting Value_log via linear regression, we generate a correlation heatmap to depict the correlation coefficient between each variable.

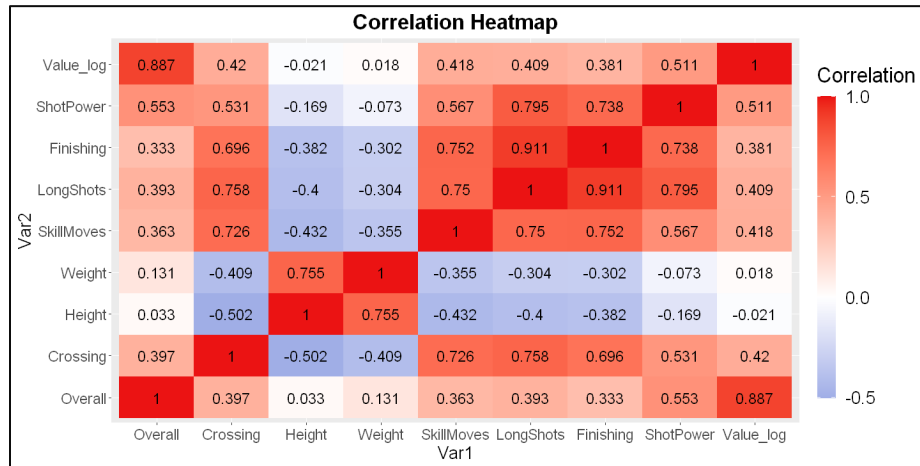


Figure 26: Correlation Heatmap of all numerical data

Based on the correlation heatmap, it is evident that Overall exhibits a strong positive correlation with Value_log, indicated by the correlation coefficients of 0.887. This suggests that Overall can be considered for linear regression analysis to predict Value_log.

3.3.2 Simple Linear Regression

Prior to fitting the variables into the linear regression model, it is imperative to ensure the linearity of the data. A scatterplot can be utilized for this purpose.

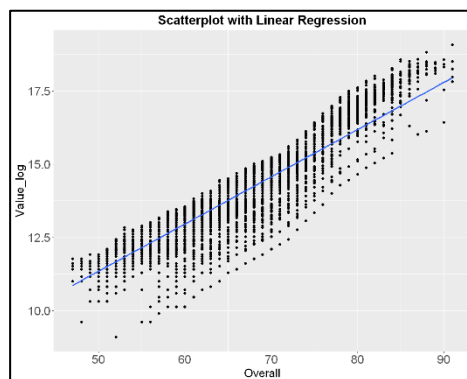


Figure 27: Scatterplot of Value_log vs Overall

Upon examination of the scatterplot, a clear linear pattern is observed, which indicates the suitability of performing a simple linear regression with Overall as the predictor and Value_log as the response variable.

Suggested Model:

$$Value_log = \beta_0 + \beta_1 * Overall + \varepsilon$$

β_0 = Intercept of the linear model

β_1 = Coefficient of the predictor variable Overall

ε = the residual term of the linear model

We can proceed with fitting the data into the linear regression model and assessing its reliability.

```
Call:
lm(formula = Value_log ~ Overall, data = num_data)

Residuals:
    Min       1Q   Median       3Q      Max
-2.82839 -0.29768  0.06221  0.42362  1.58992

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.2970230   0.0409452   80.52  <2e-16 ***
Overall      0.1609666   0.0006187  260.17  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5701 on 18433 degrees of freedom
Multiple R-squared:  0.786,    Adjusted R-squared:  0.786
F-statistic: 6.769e+04 on 1 and 18433 DF,  p-value: < 2.2e-16
```

Figure 28: Summary of Linear Regression Model

The linear regression model shows that Overall is a significant predictor of Value_log (Estimate = 0.161, t-value = 260.17, p-value < 2.2e⁻¹⁶). The model has a high coefficient of determination (R-squared = 0.786), indicating that the model explains a significant portion of the variance in the data. The intercept has an estimate of 3.297 and is also significant (t-value = 80.52, p-value < 2e⁻¹⁶). The residuals have a relatively small standard error of 0.5701, indicating that the model has a good fit to the data.

Fitted Model:

$$Value_log = 3.297 + 0.161 * Overall$$

3.3.3 Residual Analysis

To assess the reliability of the model, we need to verify the constancy of residuals' variance and their normality.

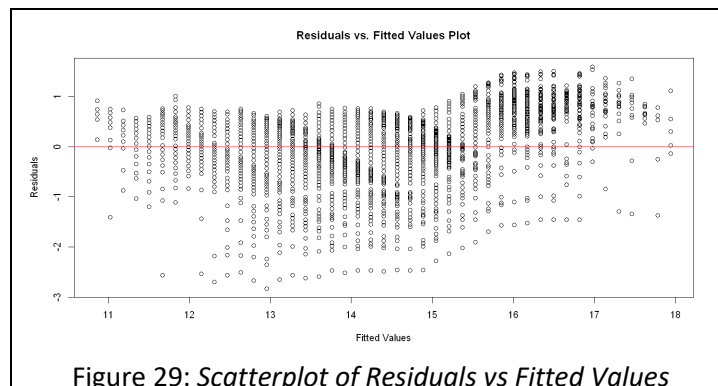


Figure 29: Scatterplot of Residuals vs Fitted Values

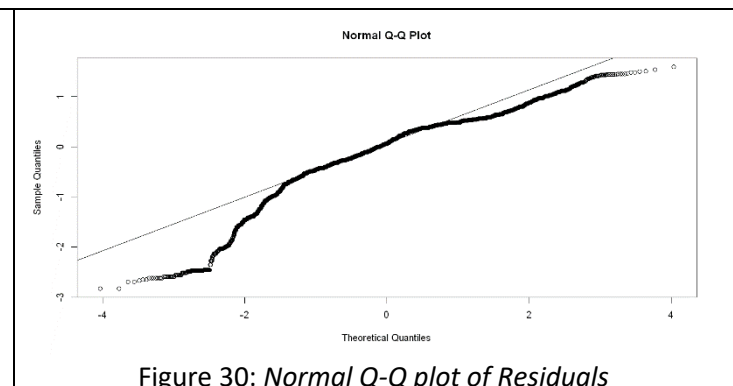


Figure 30: Normal Q-Q plot of Residuals

Based on the observation from the residuals' scatterplot, the variance appears not constant across the range of the predictor variable. In addition, the Q-Q plot suggests that the residuals do not follow a normal distribution. These findings indicate that the linear regression model may not be appropriate for the data, as violations of the assumptions of constant variance and normality can lead to biased and inefficient estimates of the regression coefficients. However, given the large sample size, we can still apply the Central Limit Theorem. Thus, slight deviations from normality may not be a major concern.

4. Conclusion and Discussion

The International Federation of Association Football or Federation Internationale de Football Association (FIFA) is one of the most prestigious sports organizations in the world. The FIFA video game is also listed in the Guinness World Records as the world's best-selling sports video game franchise, resulting in a massive fanbase for the sport. In this report, we attempt to answer some of the questions that the fans of the famous sport may be curious about based on the FIFA23 dataset, which contains statistics of every football player from the game FIFA23.

We conclude that:

- Most of the players' BMI fall within the range of Normal (18.5 - 24.9) with most of them fall in the upper end of the range.
- The proportion of players who have an 'Overall' rating of 80 and above is less than 5%.
- There is a significant difference between a player's ability to score from long range and their ability to score from close range.
- There is a significant association between the players' jersey number and their positions.
- Typically, left and right positions have similar average crossing values.
- Players in higher prestige teams would have a higher market value than players in lower prestige teams.
- There is an impact of a player's Skill Moves on the Overall Rating of that player.
- The shot power of left-footed players and right-footed players have no significant differences hence being left-footed or right-footed do not have significant impact on players' ability to score goals.

Additionally, from the correlation of players' statistics with Value_log of the players, we found that the 'Overall' variable of the players can be used to model the players' market value by using linear regression. On the other hand, the players' height and weight do not seem to have any effect on the Value of the player.

Although the results we have obtained in this report are interesting, the results are based on statistics obtained from the video game franchise FIFA, which has all the football players modelled to match the statistics of football players in FIFA. Furthermore, FIFA has been able to generate more sophisticated and detailed statistics about the players in FIFA. However, the questions answered here may not be suitable for making prominent decisions. A more in-depth analysis of the FIFA data and advanced

analytical techniques are needed to provide more accurate and specific insights to infer and make decisions.

5. Appendix

Appendix 1. Abbreviations Meaning of BestPosition

Abbreviation	Position
<chr>	<chr>
CAM	Centre_Defensive_Midfielder
CB	Centre_Back
CDM	Right_Defensive_Midfielder
CF	Central_Forward
CM	Central_Midfielder
GK	Goal_Keeper
LB	Left_Back
LM	Left_Midfielder
LW	Left_Winger
LWB	Left_Winger_Back
RB	Right_Back
RM	Right_Midfielder
RW	Right_Winger
RWB	Right_Winger_Back
ST	Striker

Appendix 2. Complete p-values of Shapiro-Wilk test for crossing by left and right positions

BestPosition	LM	RM	LB	RB	LWB	RWB	LW	RW
Size	797	1443	864	925	404	422	215	297
PValue	0.001	0.000	0.318	0.000	0.001	0.030	0.375	0.135

Appendix 3. Complete p-values of variance test for crossing by each pair combination of left and right positions

	2	3	8	9	15	16	18	19	20	22									
BestPositionsPair	LM LB	LM RB	RM LB	RM RB	LB LWB	LB RWB	LB RW	RB LWB	RB RWB	RB RW									
PValue	0.000	0.000	0.001	0.000	0.000	0.000	0.027	0.000	0.000	0.012									
	1	4	5	6	7	10	11	12	13	14	17	21	23	24	25	26	27	28	
BestPositionsPair	LM RM	LM LWB	LM RWB	LM LW	LM RW	RM LWB	RM RWB	RM LW	RM RW	LB RB	LB LW	RB LW	LWB RWB	LWB LW	LWB RW	RWB LW	RWB RW	LW RW	
PValue	0.162	0.548	0.553	0.165	0.444	0.083	0.081	0.549	0.873	0.673	0.205	0.126	0.988	0.089	0.245	0.089	0.246	0.549	

Appendix 4. Removing NA in JerseyNumber

<pre>summary(playersTeamCleaned\$JerseyNumber)</pre>	<pre># clean jerseyNumber playersTeamCleaned\$JerseyNumber <- as.numeric(as.character(playersTeamCleaned\$JerseyNumber)) playersTeamCleaned <- playersTeamCleaned[!is.na(playersTeamCleaned\$JerseyNumber),] summary(playersTeamCleaned\$JerseyNumber)</pre>
<pre>- 92 1 577 10 579 11 593 12 352 13 381 14 521 15 451 16 473 17 543 18 536 19 527 2 509 20 531 21 530 22 512 23 526 24 436</pre>	<pre>Min. 1st Qu. Median Mean 3rd Qu. Max. 1.00 9.00 18.00 21.11 27.00 99.00</pre>

Appendix 5. Removing zero values in Value

<pre>summary(playersTeamCleaned\$Value)</pre>
<pre>Min. 1st Qu. Median Mean 3rd Qu. Max. 0 500000 1000000 2889802 2000000 190500000</pre>
<pre># remove players with no value playersTeamCleaned <- playersTeamCleaned[playersTeamCleaned\$Value != 0,] summary(playersTeamCleaned\$Value)</pre>
<pre>Min. 1st Qu. Median Mean 3rd Qu. Max. 9000 500000 1000000 2891683 2000000 190500000</pre>

Appendix 6. Removing NA in IntPrestige

<pre>colSums(is.na(playersTeamCleaned))</pre>	<pre>playersTeamCleaned <- na.omit(playersTeamCleaned) colSums(is.na(playersTeamCleaned))</pre>
<pre>Overall 0 Value 0 BestPosition 0 Height 0 Weight 0 JerseyNumber 0 Crossing 0 ShotPower 0 IntPrestige 28 PreferredFoot 0 SkillMoves 0 LongShots 0 Finishing 0</pre>	<pre>Overall 0 Value 0 BestPosition 0 Height 0 Weight 0 JerseyNumber 0 Crossing 0 ShotPower 0 IntPrestige 0 PreferredFoot 0 SkillMoves 0 LongShots 0 Finishing 0</pre>

Appendix 7. Summary of each variable

```
summary(playersTeamCleaned)
dim(playersTeamCleaned)
```

Overall	Value	BestPosition	Height	
Min. :47.00	Min. : 9000	CB :3632	Min. :155.0	
1st Qu.:62.00	1st Qu.: 500000	ST :2548	1st Qu.:177.0	
Median :66.00	Median : 1000000	CAM :2300	Median :182.0	
Mean :65.83	Mean : 2894661	GK :2039	Mean :181.5	
3rd Qu.:70.00	3rd Qu.: 2000000	RM :1434	3rd Qu.:186.0	
Max. :91.00	Max. :190500000	CDM :1395	Max. :206.0	
		(Other):5059		
Weight	JerseyNumber	Crossing	ShotPower	
Min. : 49.00	Min. : 1.0	Min. : 6.00	Min. :18.00	
1st Qu.: 70.00	1st Qu.: 9.0	1st Qu.:39.00	1st Qu.:48.00	
Median : 75.00	Median :18.0	Median :54.00	Median :59.00	
Mean : 75.17	Mean :21.1	Mean :49.48	Mean :57.81	
3rd Qu.: 80.00	3rd Qu.:28.0	3rd Qu.:63.00	3rd Qu.:68.00	
Max. :105.00	Max. :99.0	Max. :94.00	Max. :94.00	
IntPrestige	PreferredFoot	SkillMoves	LongShots	Finishing
Min. : 1.000	Left : 4466	Min. :1.000	Min. : 4.00	Min. : 3.00
1st Qu.: 1.000	Right:13941	1st Qu.:2.000	1st Qu.:32.00	1st Qu.:31.00
Median : 1.000		Median :2.000	Median :51.00	Median :50.00
Mean : 2.483		Mean :2.366	Mean :46.82	Mean :46.25
3rd Qu.: 4.000		3rd Qu.:3.000	3rd Qu.:62.00	3rd Qu.:62.00
Max. :10.000		Max. :5.000	Max. :91.00	Max. :94.00

18407 13

6. References

- Alex. (2022, Sept 29). *FIFA 23 Complete Player Dataset [UPD:29/09/22]*. Kaggle. https://www.kaggle.com/datasets/cashncarry/fifa-23-complete-player-dataset?select=teams_fifa23.csv
- Cox M. (2018, Sept 1). *Attacking left-backs like Benjamin Mendy, Marcos Alonso revolutionising the Premier League*. ESPN. <https://www.espn.com/soccer/english-premier-league/23/blog/post/3618643/attacking-left-backs-like-benjamin-mendy-and-marcos-alonso-revolutionising-the-premier-league>
- FIFAUTeam. (2021). *FIFA 21 PLAYER ATTRIBUTES*. <https://fifauteam.com/fifa-21-attributes-guide/>
- Goal. (2021, March 4). *Football squad numbers explained: How positions are traditionally numbered & player roles*. Goal. <https://www.goal.com/en-sg/news/football-squad-numbers-explained-how-positions-are-traditionally-numbered--player-roles/1rwwpkupqgbczd542hkjiwl2>
- Guinness World Records. (2018, Jan 29). *Best-selling sports videogame series*. Guinness World Records: Best-selling sports videogame series. <https://www.guinnessworldrecords.com/world-records/92071-best-selling-sports-videogame-series>
- Kipsta. (2021, September 17). *All the positions on a football club's team sheet*. Kipsta. <https://www.kipsta.com/all-the-positions-on-a-football-clubs-team-sheet>
- Naik S.S. (2022). *Fifa 23 Players Dataset*. Kaggle. <https://www.kaggle.com/datasets/sanjeetsinghnaik/fifa-23-players-dataset?select=Fifa+23+Players+Data.csv>
- Selini D. (2021, Apr 18). *The Untapped Potential of Wrong-Footed Full-Backs*. Running The Show. <https://runningtheshowblog.wordpress.com/2021/04/18/tactical-analysis-inverted-full-backs-wrong-footed-tactics/>