# 

# Report of Group 3

Contents

[List of Figures 2](#_Toc60941391)

[List of Tables 3](#_Toc60941392)

[Introduction 4](#_Toc60941393)

[Visualizations 5](#_Toc60941394)

[Analysis 9](#_Toc60941395)

[Computation 10](#_Toc60941396)

[Anova Test 11](#_Toc60941397)

[Conclusion 12](#_Toc60941398)

[References 13](#_Toc60941399)

## List of Figures

[Figure 1: Box plot between stamina and position 3](#_Toc60938650)

[Figure 2: Histogram of stamina with curve 4](#_Toc60938651)

[Figure 3: Distribution of stamina of Defenders 5](#_Toc60938652)

[Figure 4: Distribution of Stamina of Midfielders 6](#_Toc60938653)

[Figure 5: Distribution of Stamina of Forward players 7](#_Toc60938654)

## List of Tables

[Table 1: Summary of Different Groups 7](#_Toc60938658)

[Table 2: Wilcox Test for position (Defender & Forward) and Stamina 8](#_Toc60938659)

[Table 3: Wilcox Test for position (Defender & Midfielder) and Stamina 8](#_Toc60938660)

[Table 4: Wilcox Test for position (Forward & Midfielder) and Stamina 9](#_Toc60938661)

[Table 5: Anova test for all 3 Groups 9](#_Toc60938662)

## 

## Introduction

FIFA 20 is a football video game dataset published as a part of FIFA series by Electronic Arts. Football, because of the thrill and excitement it offers, is the favorite game of millions of people across the world. FIFA 20 is a one of the best-selling sports video game in the world. The improved features in every series of FIFA have attracted not only gamers, but also football lovers. FIFA holds information about many famous football players across the world by obtaining FIFPro (International Federation of Professional Footballers) License.

Every year, ahead of the latest FIFA video game release, EA Sports reveals the FUT (FIFA Ultimate Team) card player ratings sending social media into meltdown. While fans debate the ratings with excitement and fervor, players tag EA Sports in their posts demanding an explanation for their underrated skills. Many people play this game to improve the ratings of their favorite team or players. The choice of position of players is a new added feature in the FIFA series.

The Analysis is performed to find the difference between stamina and the position they play on the field in FIFA 20. The performance of a player in the field mainly depends on their stamina, age and skills. The stamina of a player must be known to help gamers to decide position of that player. Hence, the research Question is:

“Is there a difference between the stamina of a player and position they play on the field in FIFA 20? “

The dataset includes 16,042 observations and 30 variables. Only outfield football players i.e. forward, defender and midfielder are included in this dataset.The dependent variable is stamina scaled in percentage and most of values are in the range 29% - 97%. The independent variable is the position variable with 3 values, Midfielder, Forward and Defender as we consider only outfield football players. The data comprises 78% Midfielders, 17% Defenders and 6% Forward Position Players.

The Null Hypothesis and Alternative Hypothesis are

H0: There is no difference between the stamina of a player and position they play on the field in FIFA 20.

H1: There is a difference between the stamina of a player and position they play on the field in FIFA 20.

## Visualizations

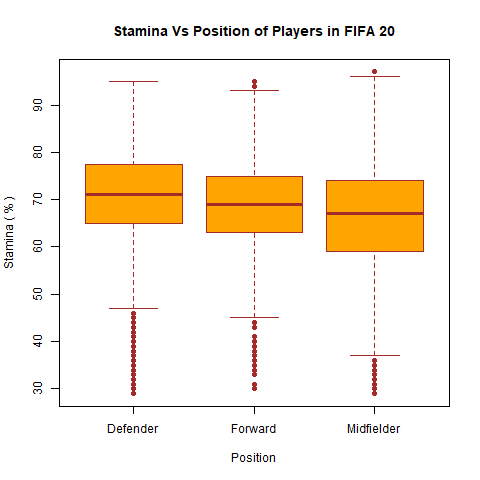


Figure 1: Box plot between stamina and position

The analysis is performed to find if there is any significant difference between the stamina and position of the players in the field. This is a comparison between nominal and interval data. Since we are using nominal independent variable and interval dependent variable we are using box plot for visualizing the data. Figure 1 shows that the central tendency the median of each category and box represents the middle 50% of data. Whiskers show the maximum and minimum value within 1.5 times. All the categories are evenly distributed along the median.

Figure 1 shows that the median of the defender is high compared to the other two positions of players. The stamina of the players of each position differs slightly. The defenders have more stamina and the midfielders have comparatively low stamina. There are so many outliers in the Defenders. From the size of the box, it is evident that the data contain more midfielder values than other two positions of players.

The boxes in the plot represent the lower and upper quartiles of the respective observations and the bar within the box denotes the median value of each distribution.

Histograms are plotted to visualize the distribution of stamina of all players and also for each type of players.

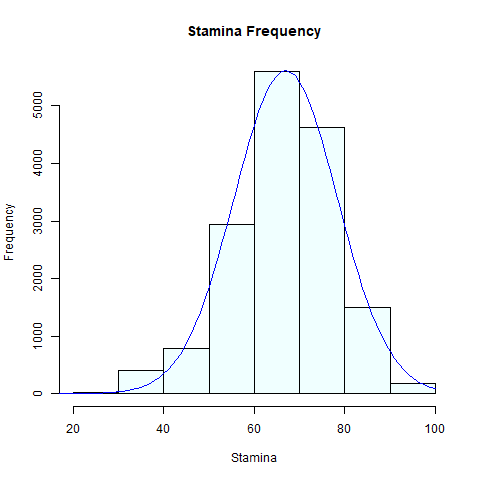


Figure : Histogram of stamina with curve

Figure 2 shows the distribution of the stamina of the players irrespective of their positions. This is close to normal distribution but skewed slightly towards right side. This is plotted between stamina and the count of players who was in the particular range of stamina.

From the Figure 2, it can be seen that stamina data are symmetrical.  The x-axis denotes the stamina of the players and y-axis denotes its frequency. The total count of players in data set is 16,042.The stamina of players starts from around 30% to 100%. More than 5000 players have stamina in the range 60% to 70%. About 75% of players’ stamina ranges from 50-80 %.

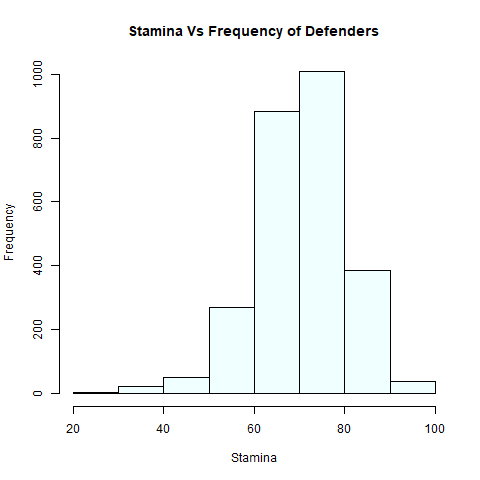


Figure : Distribution of stamina of Defenders

Figure 3 shows the data distribution of stamina of defenders in the dataset. There are 2659 Defenders. About 75% of the players in defending position have stamina in the range of 60% to 80%.

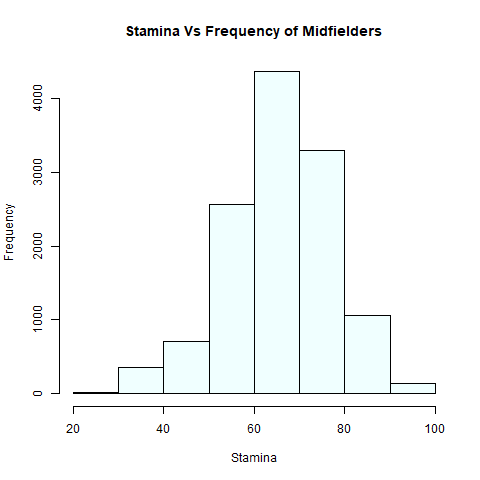


Figure : Distribution of Stamina of Midfielders

Figure 4 shows the distribution with a high frequency. The data is somewhat normally distributed. 80% of the player’s stamina is the range of 50% to 80%. The frequency scale range is higher than other distribution which shows that that the dataset contains more midfielders’ data.

The Distribution of data of forward players is shown in Figure 5. The frequency scale of the histogram plot has low range compared to other plots. Hence the dataset contains less data about the forward players. Majority of the players have stamina range from 60% to 70%. Most of the forward players have good stamina.

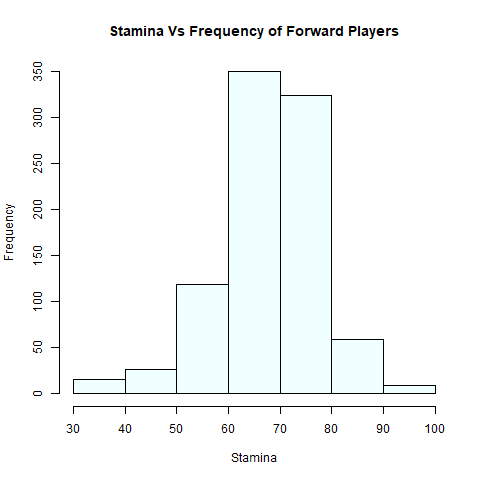


Figure : Distribution of Stamina of Forward players

## Analysis

The analysis is based on the comparison of dependent variable interval and independent variable nominal. Hence considering the medians of these variables the non-parametric test (**Wilcox test**) is performed to prove our hypothesis.

Mann–Whitney U test (or) Wilcoxon test is a non-parametric test of the null hypothesis that, if X and Y from two populations, the probability of X being greater than Y is equal to the probability of Y being greater than X and alternative hypothesis is that the probability of X being greater than Y is not equal to the probability of Y being greater than X.

### 

### Computation

We are calculating the Mean, Median and Interquartile range for all the positions. The Median differs for the groups as 71, 69 and 67.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| position | Count | Mean | median | sd | InterQRange |
| <chr> | <int> | <dbl> | <dbl> | <dbl> | <dbl> |
| Defender | 2659 | 70.8 | 71 | 10.1 | 12.5 |
| Forward | 902 | 68.3 | 69 | 9.85 | 12 |
| Midfielder | 12481 | 66 | 67 | 11.6 | 15 |

Table 1: Summary of Different Groups

Step 1: On performing Wilcox test with 2 groups (forward & Defender) and stamina the output of p-value is = 4.819e-11

|  |
| --- |
| result <- filter(dffifa, position == "Defender" | position =="Forward" )  u<-result$position  v<-result$stamina  res <- wilcox.test(round(as.numeric(v))~u)  res  Wilcoxon rank sum test with continuity correction  data: round(as.numeric(v)) by u  W = 1374584, p-value = 4.819e-11  alternative hypothesis: true location shift is not equal to 0 |

Table 2: Wilcox Test for position (Defender & Forward) and Stamina

Step 2: On performing Wilcox test with 2 groups (Defender & Midfielder) and stamina the output of p-value is < 2.2e-16

|  |
| --- |
| result <- filter(dffifa, position == "Defender" | position =="Midfielder" )  u<-result$position  v<-result$stamina  res <- wilcox.test(round(as.numeric(v))~u)  res  Wilcoxon rank sum test with continuity correction  data: round(as.numeric(v)) by u  W = 20772473, p-value < 2.2e-16  alternative hypothesis: true location shift is not equal to 0 |

Table 3: Wilcox Test for position (Defender & Midfielder) and Stamina

Step 3: On performing Wilcox test with 2 groups (Forward & Midfielder) and stamina the output of p-value is = 2.166e-10

|  |
| --- |
| result <- filter(dffifa, position == "Forward" | position =="Midfielder" )  u<-result$position  v<-result$stamina  res <- wilcox.test(round(as.numeric(v))~u)  res  Wilcoxon rank sum test with continuity correction  data: round(as.numeric(v)) by u  W = 6340099, p-value = 2.166e-10  alternative hypothesis: true location shift is not equal to 0 |

Table 4: Wilcox Test for position (Forward & Midfielder) and Stamina

Instead of iterating through two groups, we can use anova test to compare more than two groups. Same result is produced when the same data is tested using anova test and Wilcox Test.

### Anova Test

|  |
| --- |
| atest<- aov(y~x,data=dffifa)  summary(atest)  Df Sum Sq Mean Sq F value Pr(>F)  x 2 51400 25700 202.5 <2e-16 \*\*\*  Residuals 16039 2035811 127  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

Table 5: Anova test for all 3 Groups

In all the cases p-value is < 0.05 (Significance value). When a **P value is less than or equal to the significance level, we reject the null hypothesis**. This means that there is a difference between stamina and position of the players in the field.

If we consider a way that null hypothesis is that the two populations have the same distribution with the same median. Here we are rejecting the null hypothesis that means we have evidence that the medians of the two populations differ. The R Statistical environment, which we use to implement Wilcoxon rank sum test above, refers to this a ‘location shift’.

Also, we notice that p-value is < 0.05. Based on this result we conclude that medians of the two distributions differ. The Alternative Hypothesis states that “True location shift is not equal to 0”. This also means that distribution of one population is shifted to left or right of the other which concludes different medians.

## Conclusion

FIFA 20 is a favorite video game of many football fans across the world. The game is developed using data from the original players and other features related to the game. Every year the data is updated and new series of FIFA is introduced to the gaming world. The dataset of this video is used for the analysis. The players are categorized into different types based on the position they play in the field.

The winning probability of the team can be improved if the best players are chosen and categorized based on their skills. The stamina of the player is a measure of their performance in previous matches which may help to rate their performance. How the players are assigned to different positions? On what measures are their skills evaluated?. These are some interesting questions asked by the football lover.

The research was carried out to know the difference between the stamina and the position of the players in the field. The dataset is visualized using boxplot and the data distribution using histogram. The dataset contains more midfielder players data and few number of forward players data. The stamina of all the players playing in different position seems to vary in the range of 65% to 75 %. To prove the hypothesis, Wilcox test is performed to estimate the p-value based on various positions. As there are 3 groups (levels) namely Defender, Forward and Midfielder, to consider all 3 groups- iterate the Wilcox test on 3 iterations between 2 groups.

From the analysis, we conclude that there is a significant difference between the stamina and position of the players. Hence we accept the alternate hypothesis. The players are placed in various positions based on their stamina. Even a small variation in stamina makes a huge difference in identifying their positions in the field. The data is close to normal distribution but not normal as it is skewed at the lower end. Hence chosen Wilcox test as it would be appropriate for comparing the medians.

## References

1. *(‘Durango, Ana & Refugio, Craig. (2018). An Empirical Study on Wilcoxon Signed Rank Test. 10.13140/RG.2.2.13996.51840.’, no date)*
2. *(Wilcoxon Test in R: The Ultimate Guide - Datanovia, no date)*
3. (*The Wilcoxon Rank Sum Test | University of Virginia Library Research Data Services + Sciences*, no date)
4. (*One-Way ANOVA Test in R - Easy Guides - Wiki - STHDA*, no date)
5. (*Mann–Whitney U test - Wikipedia*, no date)
6. (*FIFA 20 Review - IGN*, no date)
7. (*FIFA 20 - Soccer Video Game - EA SPORTS Official Site.*, no date)