**MIDDLE EAST TECHNICAL UNIVERSITY**

**DEPARTMENT OF STATISTICS**

**Analysis of Premier League Players’ Statistics**

**PROJECT REPORT SUBMITTED**

**IN FULFILMENT OF THE REQUIREMENTS FOR THE COURSE**

**STAT 250 – APPLIED STATISTICS**

**BY**

**İsmail Çağan Yılmaz 2502490**

**Batuhan Üstün 2290997**

**Mert Nuri Gülay 2429074**

**June 2023**

**ABSTRACT**

Player statistics constitute the most important part of the observation reports in today's sports world. Our study, it is aimed to present a detailed report to the observers by analyzing the statistics of the players of the English Football League, such as goals, assists, and number of matches. As a result of the report, the analyzes obtained to support the talent scouts and club officials' search for talented players. In short, this study assists the talent search in the football world based on knowledge and data.

1. **Introduction**

Detailed interpretation of player statistics is important so that new talents can be discovered by scouts. In the old football world, talent research is based on subjective data rather than analysis and information. Today, objective results can be achieved through statistical analysis. Our research includes extensive player statistics. Using these statistics, we aim to conduct detailed data analysis where scouts can provide player comparison and team analysis. In the next sections, we'll examine comprehensive data analysis and results for scouting.

* 1. **Data Description**

The data we provided shows the English Premier League players' statistics from 2020 to 2021. There are 532 distinct players and 18 variables with 4 categorial, 5 continuous, and 9 discrete types about each player in the data set. In categorical variables; "Names" is the name of each player. "Club" is the name of the club that each player is playing in. "Nationality" is the nationality or country of each player. "Position" is the position played by each player. Examples include forward midfielder, defender, or goalkeeper. In discrete variables; "Age" is the age of the player. "Matches" is the number of matches the player has played. "Starts" is the number of times the player was named in the starting 11 by the manager. "Goals" is the number of Goals scored by the player. "Assists" is the number of times the player has assisted another player in scoring the goal. "Penalty\_Goals" is the number of times the player has assisted another player in scoring the goal. "Penalty\_Attempted" is the number of penalty attempts made by the player. "Yellow\_Cards" is the number of yellow cards received by the player. "Red\_Cards" is the number of red cards received by the player. In continuous variables; "Mins" is the total minutes played by the player. "Passes\_Attempted" is the number of passes attempted by the player. "Perc\_Passes\_Completed" is the number of passes that the player accurately passed to his teammate. "xG (Expected Goals)" is the expected number of goals from the player in a match. "xA (Expected Number of Assists)" is the expected number of assists from the player in a match.

* 1. **Research Questions**

1. Is the goal average of Manchester City, the champion of the premier league in the 2020-2021 season, above the league average?
2. Are defensive players (DF) more prone to yellow cards than midfield players (MF)?
3. Is the probability of Harry Kane, the top scorer in the 2020-2021 season, to score more than 50% per match?
4. Is the percentage of successful passes of the starting 11 players aged 28 and over higher than the starting 11 players under 28 years old? So, is the reason why a player is preferred in the starting 11 because he has a higher percentage of successful passes?
5. Can we explain the number of goals (Goals) with goal expectation (xG)?
6. Can we explain the number of assists (Assist) in terms of assist expectation (xA), percentage of successful passes (Perc\_Passes\_Complated), and total pass attempts (Passes\_Attempted)?
7. Is the number of yellow cards (Yellow\_Cards) a factor in starting the starting 11 (Starts)? Does age (Age) have an effect on starting the first 11(Starts)?
   1. **Aim of the Study**

This study aims to obtain a more scientific and rational data-based report against traditional talent hunting in today's football world. In short, this report provides convenience to football clubs by interpreting player recruitment and team analysis.

1. **Methodology/Analysis**
2. The null hypothesis claims that the goal average of Manchester City, the champion of the premier league in the 2020-2021 season, below the league average.

The alternative hypothesis claims that the goal average of Manchester City, the champion of the premier league in the 2020-2021 season, above the league average.

The average goal statistic in the data is 1,853. According to the statistics of the 24 players in the Manchester City team, Manchester City's goal average is 3,416 and the standard deviation is 3,999. As a result of the t-test we performed with these statistics, we found the t score as 1.915 and the p-value as 0.034.

1. The null hypothesis claims that defensive players (DF) less prone to yellow cards than midfield players (MF).

The alternative hypothesis claims that defensive players (DF) more prone to yellow cards than midfield players (MF).

For players who can play defense, we found the following stats:

Mean: 2.468085 | Standard deviation: 2.253992 | Sample size: 47.

For players who can play midfield, we found the following stats:

Mean: 2.410256 | Standard deviation: 2.917443 | Sample size: 39.

We used f-test to equate variances. We found the ratio of variances 0.5968981. Then, we

use z-test.

1. The null hypothesis claims that the probability of Harry Kane, the top scorer in the 2020-2021 season, to score less than 50% per match.

The alternative hypothesis claims that the probability of Harry Kane, the top scorer in the 2020-2021 season, to score more than 50% per match.

We found that:

Percentage of goals per game: 0.657 | Standard deviation: 0.0845 | Z value: 1.859.

1. The null hypothesis claims that there is no significant difference between the percentage of successful passes for players aged 28 and over and the percentage of successful passes for players under 28.

The alternative hypothesis claims that the percentage of successful pass of the players aged 28 and over is lower than the players under the age of 28.

We found that statistics for players aged 28 and over:

Proportion of successful passes: 0.7867238

Standard deviation: 0.09465945 | Sample size: 181

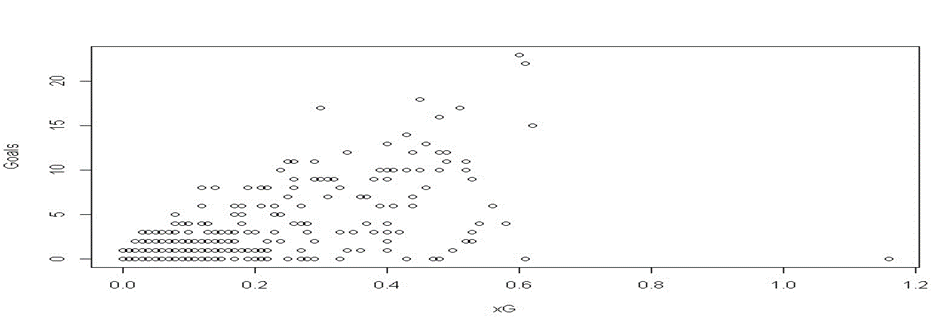
We found that statistics for players aged 28 and over:

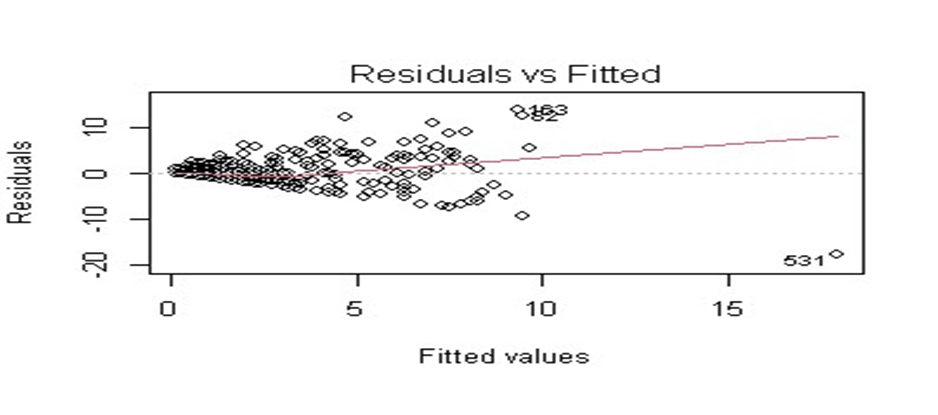
Proportion of successful passes: 0.7738632 | Standard deviation: 0.1449861

Sample size: 351

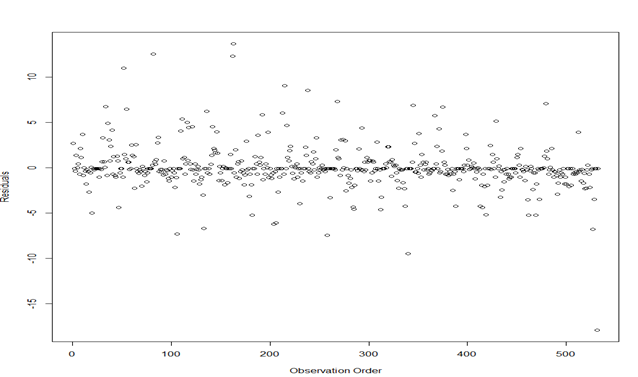
We also found the z value as 0.3406132.

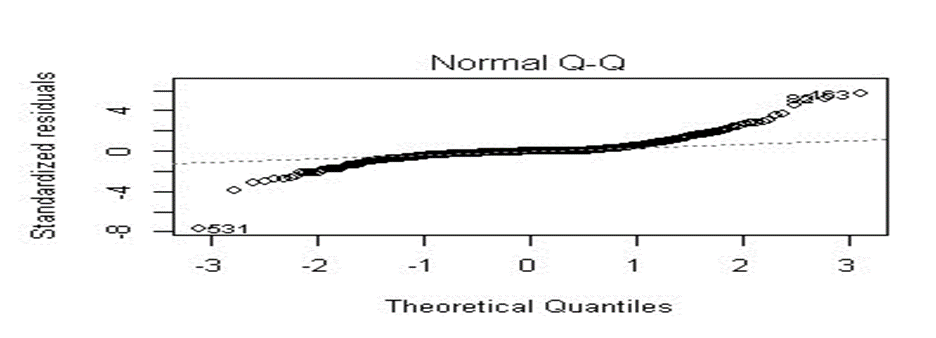
1. In order to examine the linear relationship between the variables, first of all, the correlation coefficient is examined. This coefficient specifies the linear power and direction between the two variables. If the correlation is positive, there is a positive linear relationship between the two variables. In our analysis, we found a correlation coefficient of 0.682. This number indicates a positive correlation. Now let's check the linearity.



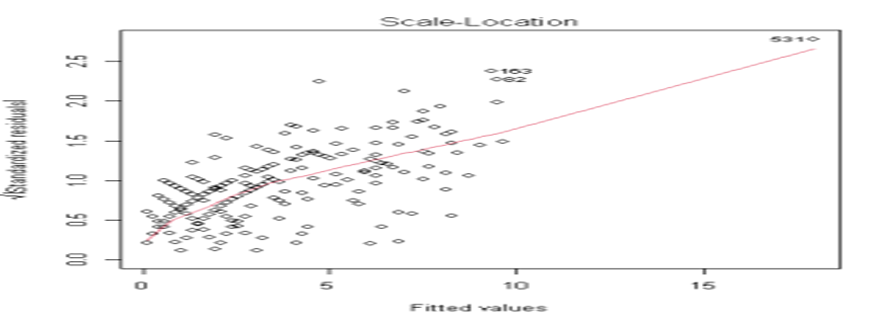
This graph shows the linearity between the two variables. When we connect the dots, we see an increasing line direction. According to these data, there is an important linear relationship between the variables. Also, we can check the Residuals vs Fitted plot to make interpretations about linearity:

In this graph, we examine whether the residues exhibit a random pattern within the range of predicted or appropriate values. Linearity is satisfied as the residuals are close to the zero line and randomly distributed. Also, to examine the independence assumption, we need to examine the independence of residuals.



So, does the order of observations help visualize the pattern and slope? Certainly. For the independence assumption to be satisfied, it is essential that the graph show a random distribution of points without any discernible patterns or trends around the zero line. In addition, since the third assumption is normality, it means to control if the variables follow the normal distribution or not.

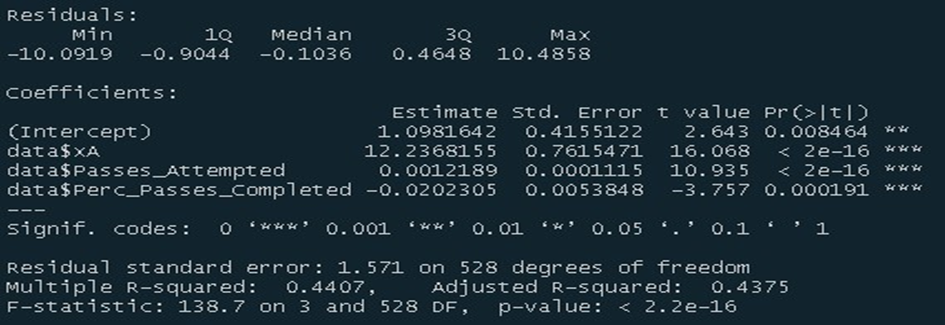
In our analysis, we examined the normality of our data using the Shapiro-Wilk test and assessed the relationship between quantiles and standardized residuals through a Quantiles vs Residuals plot.

The p-value (< 2.2e-16) of the Shapiro-Wilk test allowed us to reject the null hypothesis that our data followed a normal distribution. We also observed a strong, near-linear pattern on the Quantiles vs. Residuals plot, showing that our data aligns well with the expected quantities of the normal distribution. Thus, we found strong evidence that our data met the normality assumption. As a result, our linear regression model can measure the relationship between variables. To assess the homogeneity of variance, we need to examine the variance across the plot.

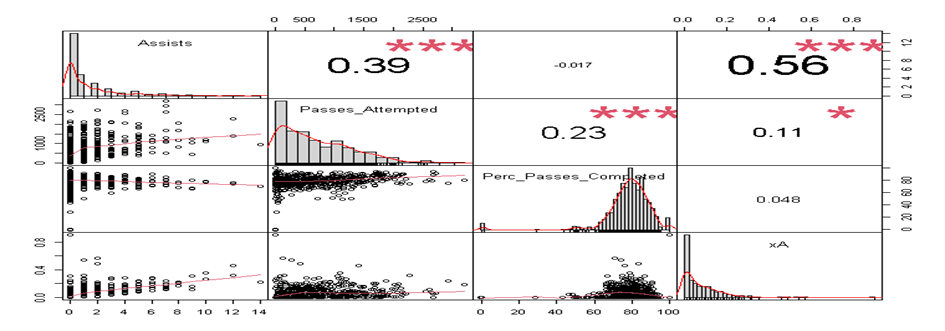
Scale-Location plots are used to make comments on residual points increasing with the value of the appropriate result variable, which now shows non-constant variances in errors. A horizontal line with equally spread points is a good indication of homoscedasticity. The model is suitable for this situation. Based on all these observations, we can construct a linear equation.

y = 15.363\*xG + 0.113

1. For this analysis, we expect to predict the number of assists in terms of assist expectancy, pass attempts, and percentage of completed passes for each player. To decide whether or not there is a multi-linear relationship between these variables, the significance of the model, i.e., the fit of the model, should be tested.



The F-statistic of our multi-regression model is 138.7 and the p-value is less than 2.2e-16 to make statements about the significance of the model. To analyze the significance of the predictor variables, t-statistical p-values are helpful. T-statistics for assist expectancy, pass attempt, and percentage of passes completed are 16.068, 10.935, and -3.757, respectively, with corresponding p-values less than 2e-16(xA and pass attempts) and 0.0000191. Apart from this, in multilinear regression, correlation analysis between variables is also important to determine the fit of the model.



If you look at the correlation diagram, it is clear that there is a fairly strong and significant correlation between the individual variables. Furthermore, it can be seen that there is a strong, linear relationship between the variables, which also supports multilinearity. Since we have enough results to assert the significance and fit of the model, we can construct the multilinear regression equation as follows;

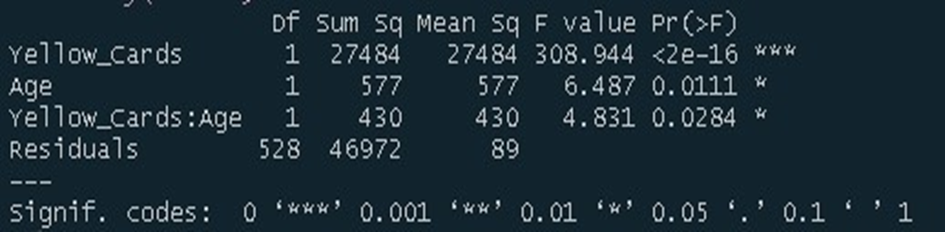
Assists: xA\*12.23+Passes\_Attempted\*0.001+Perc\_Passes\_Complated\*-0.0202+1.0981

1. H01: Yellow Cards does not have an effect on Starts 11.

H11: Yellow Cards has an effect on Starts 11.

H02: Age does not have an effect on Starts 11.

H12: Age has an effect on Starts 11.



By using “aov” function we got a two-way anova model for our question. Then, we obtained a summary of it. From here, degrees of freedom of Yellow Cards, Age and an Intercept are 1,1,1. Sum of squares are equal to 27484,577 and 430 respectively. Mean of squares are equal to the sum of squares because the degrees of freedom are equal to 1. F- values are shown as 308.944, 6.487,4.831 accordingly.

1. **Results and Findings**
2. As a result of the evaluation, the null hypothesis is rejected because the p value is 0.034<0.05. According to the results of the data we have, the goal average of Manchester City, the champion of the first league in the 2020-2021 season, is above the league average. In other words, being a champion in the league has a positive correlation with the number of goals scored.
3. According to the results, we found the z-value of 0.1012 and the p-value of 0.4597. Z0 < zTable so we do not have enough evidence to reject The Null Hypothesis. So, we have proven that the tendency to get yellow cards is not dependent on positions.
4. Harry Kane, the top scorer in the 2020-2021 season, scores 0.657 goals per game. So, the probability of Harry Kane scoring a goal in a match is higher than the probability of not scoring.
5. According to the results, we found the z-value of 0.3406132. Z0 < zTable so, we cannot reject the Null Hypothesis. This means that there is no particular correlation between the percentage of successful passes and age.
6. Through further analysis and findings, it was found that goal expectancy is a good explanatory factor for the number of goals scored by a player.
7. By evaluating our results to test the significance and effectiveness of the multilinear model, our multilinear model between the variables of assists, assists expectation, pass attempt, and percentage of passes is significant because the t-values between the variables are closer (except percentage of passes) and the p-values are small. Multilinearity is established by looking at the correlation graph and is also significant. The equation shows that a 1.0981 increase in assists values can be explained by a 0 increase in the variables xA, Passes\_Attempted, and Perc\_Passes\_Completed. In addition, a 1 unit increase in xA can be explained by a 12.23-point increase in assists, and a 1000 unit increase in Passes\_Attempted can be explained by a 1 point increase in assists, and a 1000 unit increase in Perc\_Passes\_Completed can be explained by a 202 point increase in assists.
8. As a result, we found p-values for yellow cards, age, and their intercept of 2e-16, 0.111, and 0.0284, respectively. So, we reject all the null hypothesis because our alpha is 0.05 and all the p values lay in rejection area. In other words, we can claim that there is an effect of Yellow Cards, Age, and their interception on Starts 11.
9. **Discussion/Conclusion**

In summary, the analysis of the results presented highlights several important findings. First, the goal average of Manchester City, the champions of the 2020-2021 season, exceeds the league average, indicating a positive correlation between the championship title and the number of goals scored. Second, the tendency to receive yellow cards does not depend on position, as shown by the lack of statistical significance. Third, Harry Kane, the top scorer, has a high scoring rate, which increases the likelihood that he will score a goal in a match. Fourth, there is no particular correlation between the percentage of successful passes and age. In addition, goal expectancy is identified as a reliable predictor of the number of goals scored by a player. In addition, the multilinear model shows the importance of variables such as assists, assist expectancy, pass attempt, and percentage of passes in explaining variance. Finally, the p-values of yellow cards, age, and their interaction support the rejection of the null hypothesis, indicating an effect on starting lineup inclusion. These results contribute to a comprehensive understanding of the dynamics and performance factors in the English Premier League and provide valuable insights for decision makers in the areas of player recruitment, team strategy, and player evaluation.

**References**

Arnholt A, Evans B (2021). \_BSDA: Basic

Statistics and Data Analysis\_. R package

version 1.2.1,

<https://CRAN.R-project.org/package=BSDA>.

Peterson BG, Carl P (2020).

\_PerformanceAnalytics: Econometric Tools for

Performance and Risk Analysis\_. R package

version 2.0.4,

<https://CRAN.R-project.org/package=PerformanceAnalytics>.

Rajat Chauhan (2021). English Premier League 2020-2021. Kaggle dataset. Retrieved from <https://www.kaggle.com/datasets/rajatrc1705/english-premier-league202021>.

R Core Team (2023). \_R: A Language and

Environment for Statistical Computing\_. R

Foundation for Statistical Computing, Vienna,

Austria. <https://www.R-project.org/>.

Taiyun Wei and Viliam Simko (2021). R package

'corrplot': Visualization of a Correlation

Matrix (Version 0.92). Available from

<https://github.com/taiyun/corrplot>