A Brief Analysis of Football Statistics

DEPARTMENT OF STATISTICS OF MIDDLE EAST TECHNICAL UNIVERSITY

BY

Arda Palıt

Batuhan Kalaycı

Emre Taşkın

Gürdal Safel

JUNE 2024

1. Introduction

In today’s world, most fields use statistics and datasets to develop their fields, especially in sports; data analysis is vital to improving athlete performance. In this project, we have focused on analyzing football statistics to make inferences. We have analyzed players’ statistics to investigate how their features affect their performance and scores by separating footballers playing in the top 3 leagues of Europe. We have collected a dataset about football players from the football statistics database of the Sofascore website to conduct a brief analysis. The dataset contains data about some player features and performance statistics. While working on the project, we have faced with some problems. Initially, we could not find relevant and sufficient raw datasets in CSV form on reliable sources to analyze. We have solved this problem by preparing our raw dataset as a CSV file using the data from Sofascore, which is not in raw form. Another problem was obtaining all the data on football players. We have solved this by using sampling techniques by taking samples from the top 3 leagues of Europe. Even we prepared the dataset there were some differences in structure of our dataset and we realized that during the analysis, then we have solved that by applying some data cleaning procedures. In some parts we have faced with not satisfying assumptions of hypothesis testing. Since it is quitely challenging to find already satisfied assumptions data in real life we have made assumptions to apply analysis. Our main goal in this project is to conclude how factors change players’ performances and how the demographic structure changes by leagues. To investigate this aim, we have prepared some research questions. We have used different analyzing methods in RStudio, such as proportion tests, t-tests, ANOVA (Analysis of Variance), and simple linear regression, to look for answers to these questions. After analyzing the methods, we made inferences about the relationship between players’ performances and features, such as their leagues, preferred feet, heights, and positions. We have also concluded domestic and youth player ratios of leagues.

1.1. Data Description

The dataset was created in 2024 by Sofascore, it has 250 observations with 16 variables which are:

- Name.Surname: Name and Surname of the players
- Nation: Nations of players
- Age: Age of players

- Height: Height of players
- Position: Position of players
- Leg: Left or Right
- League: Premier League, LaLiga, SerieA
- xG: Goal expectations
- xA: Assist expectations
- Touches: Number of touches
- Accurate.pass.per.game: Number of accurate passes per game of player
- Tackles: Number of tackles
- Aerial.duels.won: Number of aerial duels won
- Possession.lost: Number of possessions lost
- Fouls.per.game: Number of fouls per game
- Yellow.Card: Total number of yellow cards

There are 9 numerical and 7 categorical variables in this dataset.

Research Questions

We have 6 research questions to investigate in our project.

- Is there a statistically significant difference between the mean goal expectations of players by leagues and mean goal expectations of players?
- Is there a significant difference in average of accurate passes per game between right- and left-footed players?
- Is the proportion mean of nativeness by leagues are significantly higher than 0.4?
- Is there a significant difference according to young player proportions between English and Spanish League?
- Is there a significant difference in mean goal expectations according to positions of players?
- Is there a significant simple linear relationship between heights of players and aerial duels won?

1.2. Aim of The Study

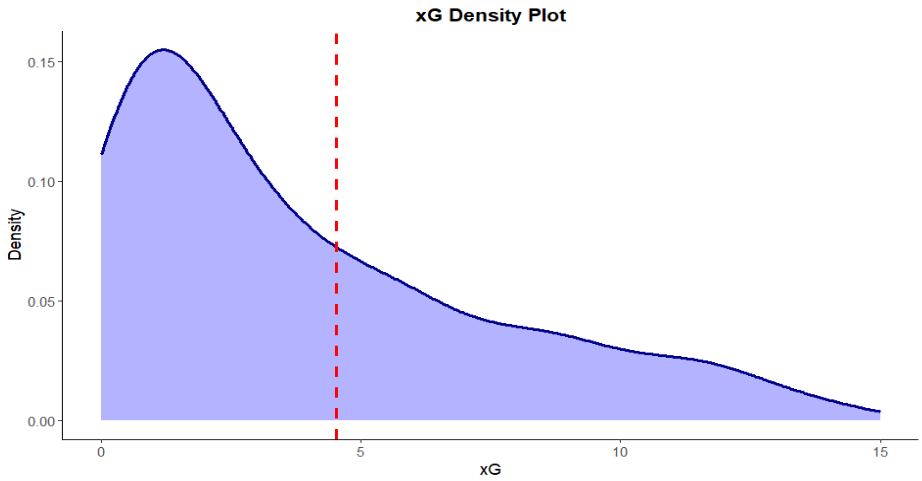
The study aims to see factors interesting goal expectations, to see how preferred foot affects the pass accuracy, to see whether leagues have high domestic player rates, to see if there is a difference in youth player rates change by leagues, and to see how aerial duels won changes by heights. Conclusively, the aim is to gain an extensive perceptive of the factors that affect to football players' performance and demographic structures of the leagues.

2. Methodology/Analysis

In order to make conclusion about first research question, we have decided to apply one sample t-test with the t.test() function in RStudio. Before we apply the t-test we have checked the assumptions for one sample t-test with shapiro.test() function in R, even the result of Shapiro-Wilk normality test did not resulted normally, we used Central Limit Theorem to assume population normally distributed. We aimed to find if there is difference between the mean goal expectations of players by leagues and mean goal expectations of players. For the second question, our aim was to see if there is difference in mean pass accuracy according to preferred foot. Firstly, we have controlled the assumptions in order to apply two sample t-test. We have proved that populations are normally distributed by Shapiro-Wilk test, populations are independent and assumed variances of populations are not equal and unknown. Then, we conducted two sample t-test by t.test() function in R in line with our goal. In the analysis of third research question, we purposed to find whether mean nativeness proportions of leagues is higher than 0.40. Thus, we have used prop.test() function in R in order to apply hypothesis testing about population proportion. Before applying the hypothesis test, we have checked whether the sample proportion $\hat{p}$ is approximately normally distributed or not, then we proved it is distributed normally by CLT (np & nq >5). To evaluate if there is significant difference in mean goal expectations according to positions of players, we have applied hypothesis about two population proportions for independent samples with the prop.test() function in R since $n1\hat{p}$, $n1(1 - \hat{p})$, $n2\hat{p}$, and $n2(1 - \hat{p})$ are greater than 5. In order to make inferences about fifth research question, we have conducted ANOVA (Analysis of Variance) to compare whether all the mean goal expectations of positions are equal. In order to apply ANOVA test, we have checked the assumptions of ANOVA which are normally distributed by Shapiro-Wilk test, samples are independent, and variances are assumed equal with the bartlett.test() function in R. After that, we applied ANOVA test by the aov() function in R. In the last research question, we have fitted a simple linear regression model to consider how heights of players affects the aerial duels won with lm() function in RStudio. We have checked the assumptions by using residuals, we have proved linearity, normality, homoscedascity and independence by applying some tests with the qqnorm(), qqplot(). In order to test significance of our simple linear regression model, we have checked F-test results with summary() function of R for deciding whether beta1 is equal to 0 or not. Also by checking the output of the summary() function we have decided on how much of the variation is explained by our model. Furthermore, in order to visualize our findings in analysis of research questions we have used some different libraries such as ggpubr, ggplot2 and functions such as plot(), boxplot(), ggboxplot().

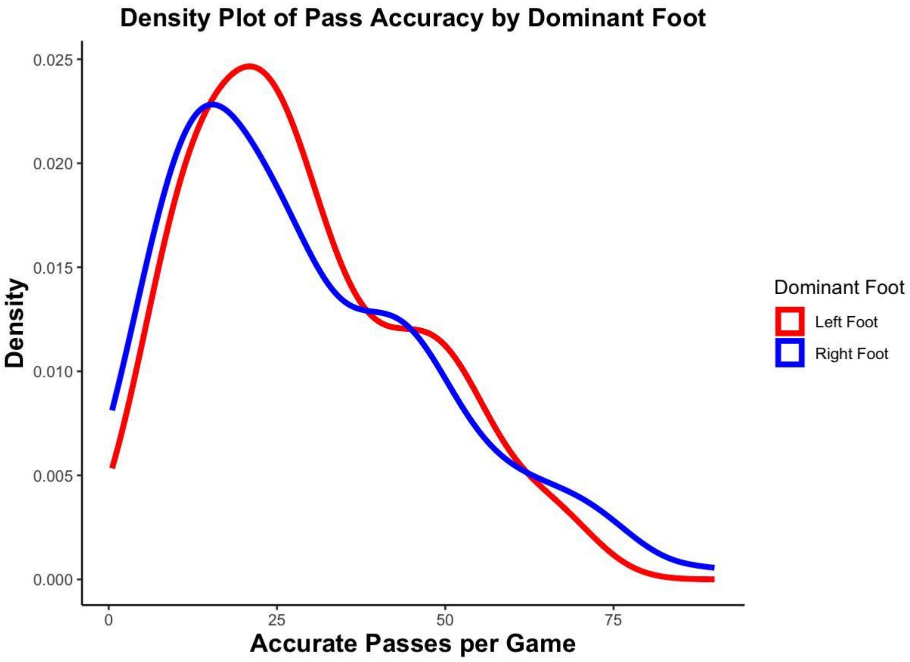
3. Results and Findings

3.1 - Is there a statistically significant difference between the mean goal expectations of players by leagues and mean goal expectations of players?



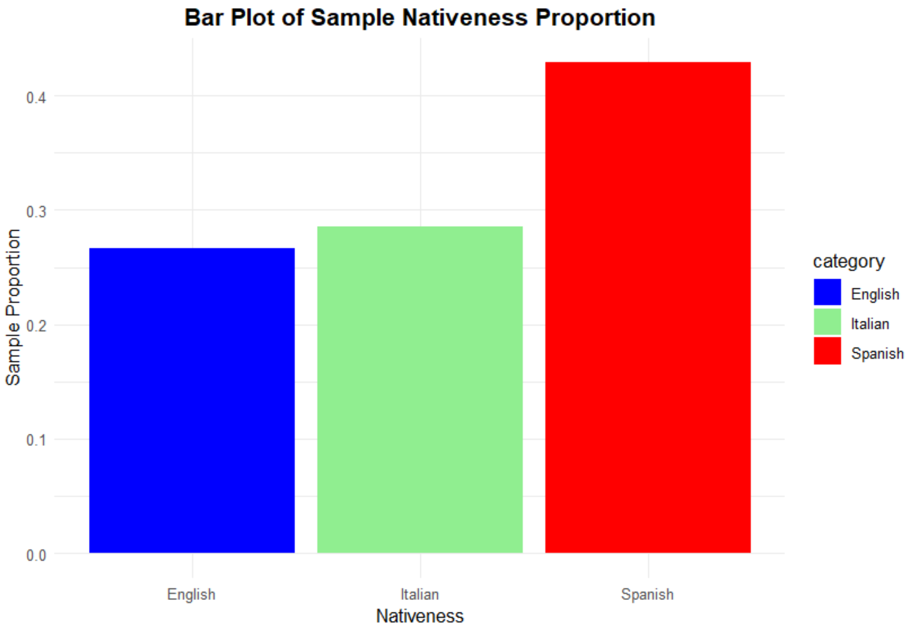
We did a hypothesis test for each league to determine if there is a statistically significant difference between players' goal expectation means according to their league and the population's goal expectation mean. According to the Shapiro test, the xG variable is not normal, but as n is greater than 30, we can say that the data is normal by Central Limit Theorem. The population mean is 4.517 and the variance is 22.43. The goal expectation means for the Premier League is 4.115; for La Liga, it is 4.831; and for the Serie A, the mean is 4.549. When we applied the t-test to each league, the p-value for all of them was less than 0.05. Therefore, we reject the null hypothesis, which means that there is a difference between players' goal expectations according to their league and the population's goal expectation mean.

3.2 - Is there a significant difference in average of accurate passes per game between right- and left-footed players?



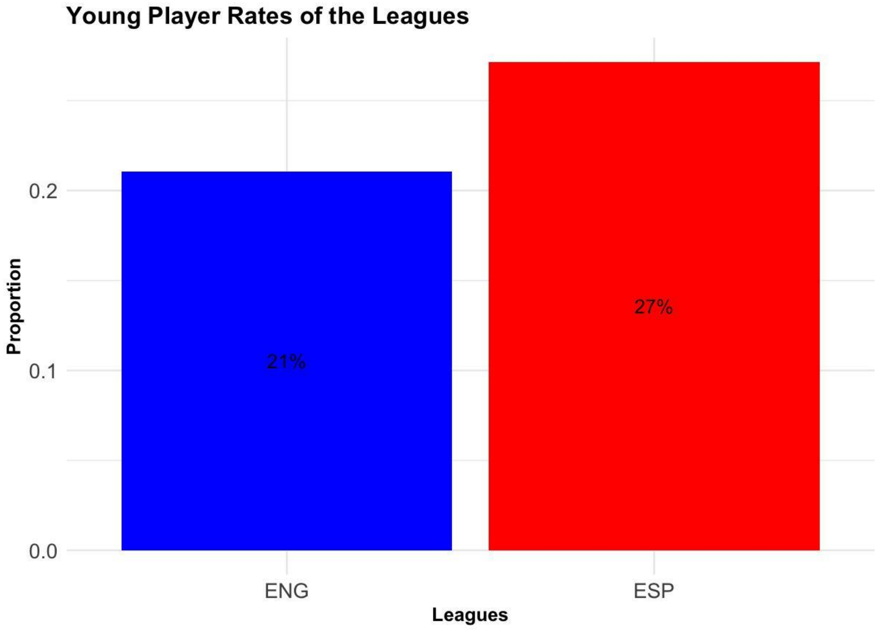
We know that variances are unknown, populations are independent, populations are normally distributed according to Shapiro - Wilk Normality test and we assume that population variances are different. Thus, according to the Welch Two Sample t-test results, there is no significant difference in pass accuracy between left-footed and right-footed players. The p value of 0.9788 indicates that any observed difference in passing accuracy is likely due to random chance rather. From this, we can conclude that the foot used by football players does not positively or negatively affect the accuracy of their pass.

3.3 - Is the proportion mean of nativeness by leagues are significantly higher than 0.4?



First, we did a proportion hypothesis test and assumed that significance level was 0.10 to find if the proportions were higher than 0.40. According to the Shapiro test, the Nation variable is not normal, but as n is greater than 30, we can say that the data is normal according to the Central Limit Theorem. The proportion value for the Premier League is 0.27; for La Liga, it is 0.43; and for Serie A, it is 0.29. Then, by using the p-test, we found that Premier League's p-value was 0.05. For LaLiga p-value is 0.43, and for Serie A, it is 0.07. Only the p-value of LaLiga is smaller than alpha, so we reject Premier League's and Serie A's null hypothesis and fail to reject LaLiga's null hypothesis. Therefore, only LaLiga’s proportion mean of nativeness by leagues are significantly higher than 0.4.

3.4 - Is there a significant difference according to young player proportions between English and Spanish League?

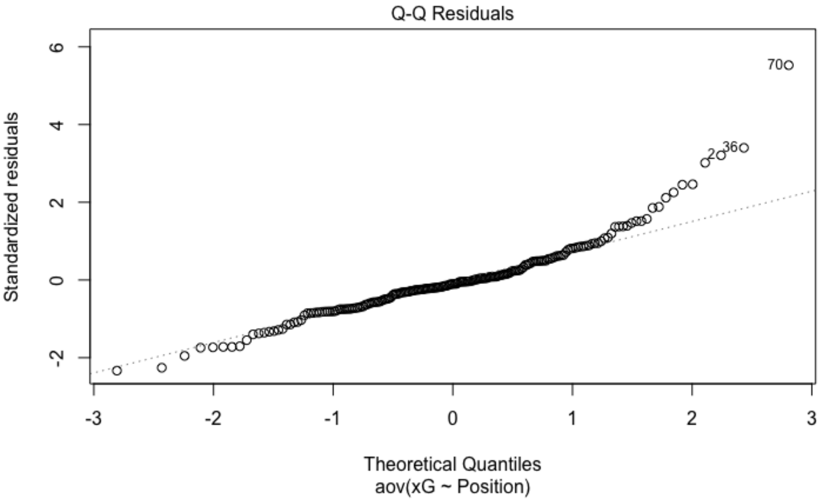


In order to evaluate this research question, we must apply hypothesis testing for two population proportions. We have obtained that while youth player rate of the sample of English league is 0.21, same proportion for Spanish league is 0.27. We have used this obtained proportions later in order to apply hypothesis testing. Before we start to hypothesis testing, we have verified the assumptions to apply test that populations are independent and normal by CLT. Then, we found that p-value of the proportion test is. 0.5572. Thus, the null hypothesis represents that there is no difference between the proportions of young players in England and Spain. Additionally, alternative hypothesis: two-sided which represents that proportions are not equal. Since the p-value is 0.5572, which is greater than the typical significance level of 0.05, we fail to reject the null hypothesis. Therefore, we do not have enough evidence to conclude that there is a significant difference in the proportions of young players between England and Spain.

3.5 - Is there a significant difference in mean goal expectations according to positions of players?

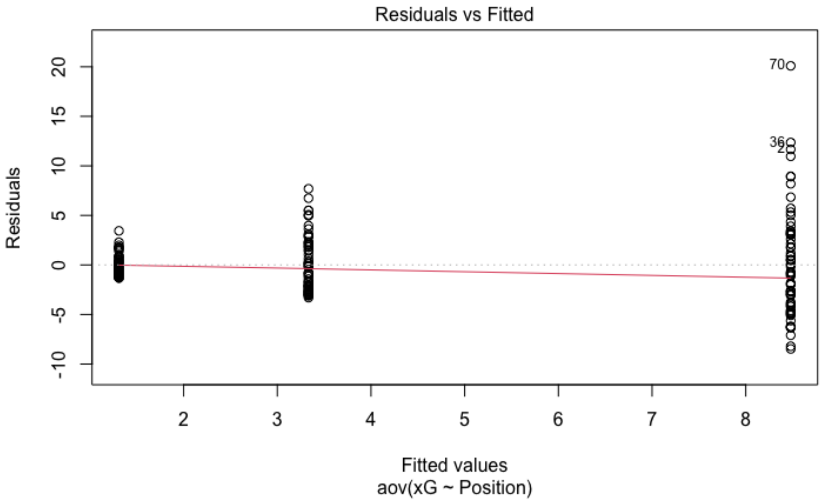
To evaluate this question, we can conduct hypothesis testing that whether there is difference between mean goal expectations of positions. Since we have more than 2 factors and means which are positions of attackers, midfielders, and defenders; we can test this with ANOVA method. In order to apply ANOVA test method, we need to check assumptions that normality, equality of variances and independence. We know that samples are independent, so we have checked other two conditions by tests.

1. Checking normality:



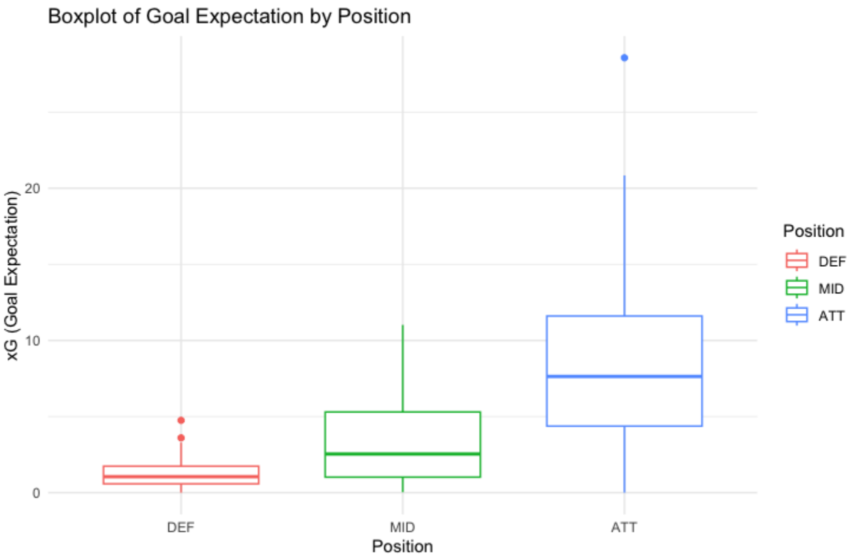
According to the normality plot of residuals and Shapiro-Wilk normality test, we have satisfied the condition of normality. Since the majority of points are approximately along this reference line and p-value of the normality test is 2.655e-09 which is less than our significance level 0.05.

2. Checking homogeneity of variance:

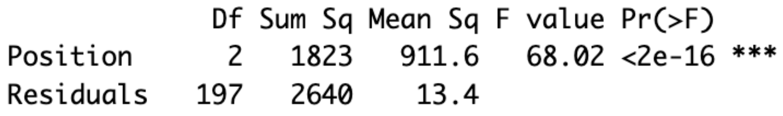


Even though we could not prove that variances of samples are not equal by the p-value (2.2e-16) of bartlett test of homogeneity of variances which is less than significance level 0.05, we have assumed homogeneity in order to apply ANOVA.

Before considering the ANOVA, we have visualized our data in order to have an idea about the averages of each group.



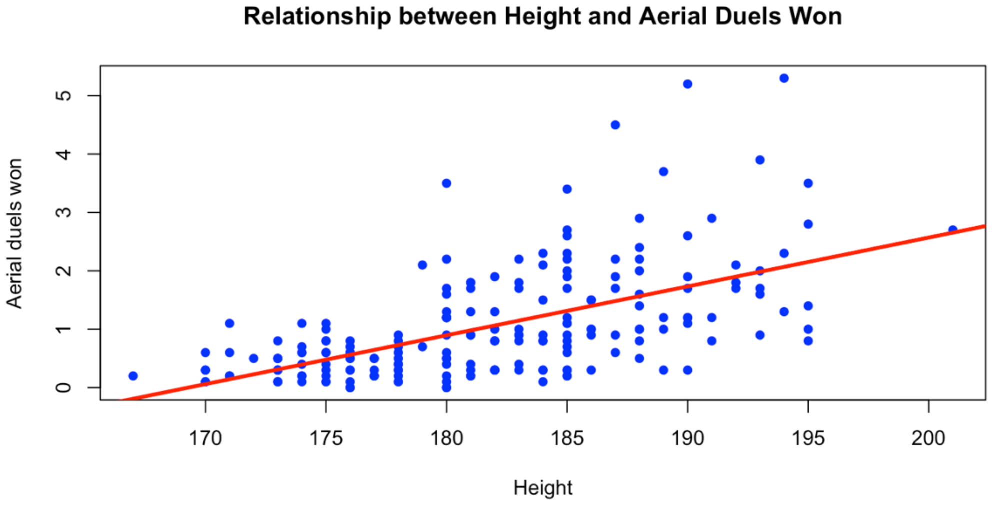
It is possible observe from boxplots that there may be difference between Goal expectations of positions. However, we must apply ANOVA to get precise results.



```
             Df Sum Sq Mean Sq F value Pr(>F)    
Position      2   1823   911.6   68.02 <2e-16 ***
Residuals   197   2640    13.4                   
```

Considering the results of ANOVA test, we have rejected null hypothesis that mean goal expectations of all positions are equal at the 0.001 and more significance levels. There is enough evidence to claim that mean goal expectations differ by positions of players.

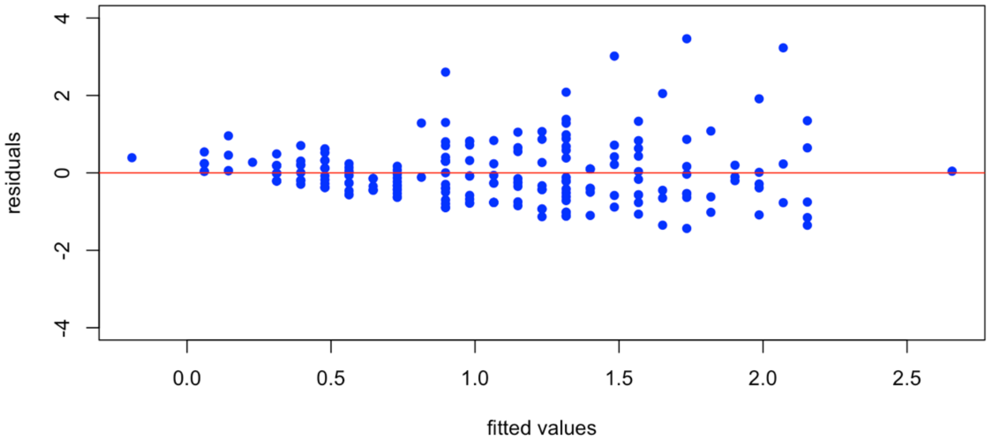
3.6 - Is there a significant simple linear relationship between heights of players and aerial duels won?



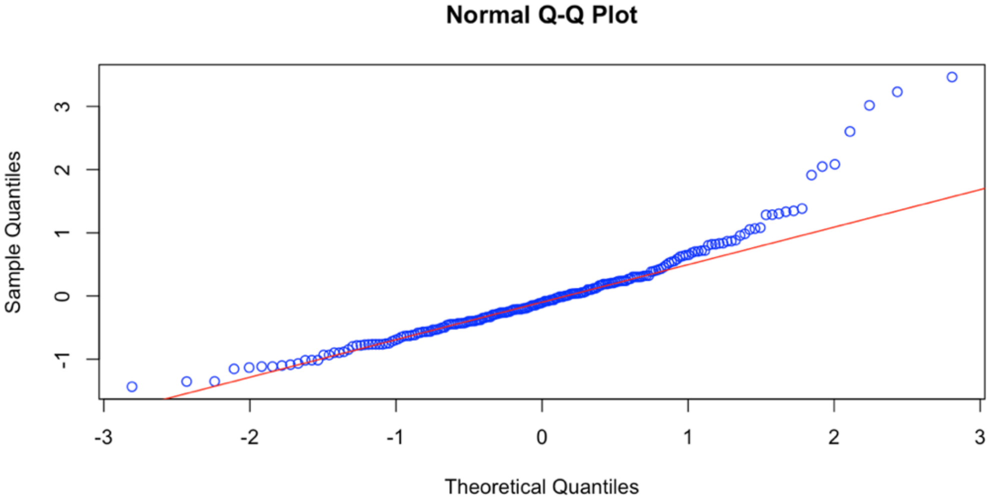
The relationship looks roughly positive linear, so we can use this linear model.

We can detect a linear relationship. But we will check the assumptions by using residuals. Linear regression makes several assumptions about the data, such as:

1. Linearity of the data: The relationship between x and y is assumed to be linear.
2. Normality of residuals: The residual errors are assumed to be normally distributed.
3. Homogeneity of residuals variance: The residuals are assumed to have a constant variance (homoscedasticity)
4. Independence of residuals error terms.

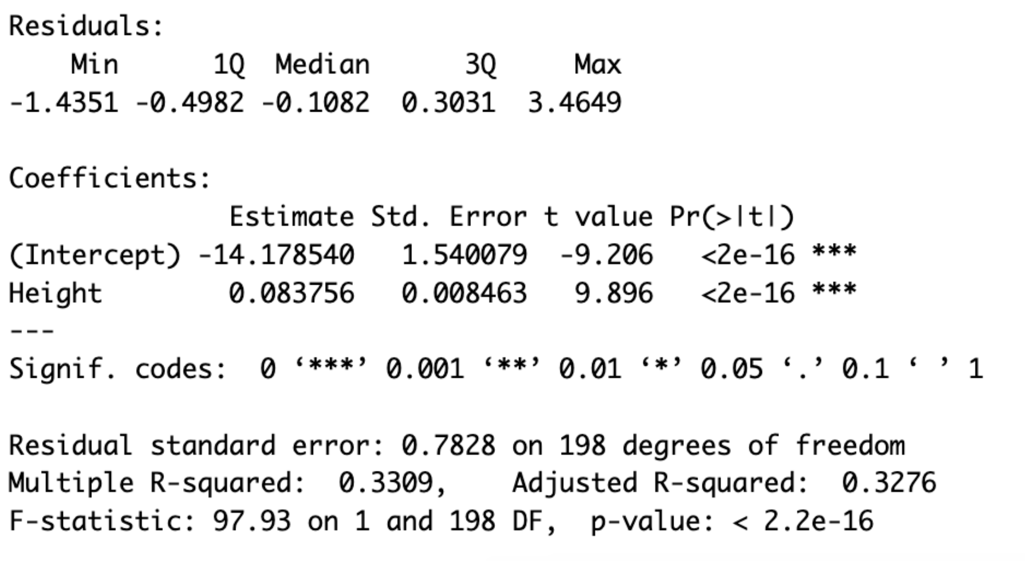


The residuals are mostly clustered around the 0 line, indicating that the variances are constant.



The percentiles mostly follow a straight line, which indicates that the residuals are normally distributed. Since we only have one independent variable and one dependent variable, we can say that the independence assumption is satisfied.

After checking the assumptions, let’s comment the model. If we look at the summary statistics using R, the results are as follows.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-1.4351 -0.4982 -0.1082  0.3031  3.4649 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -14.178540   1.540079  -9.206   <2e-16 ***
Height        0.083756   0.008463   9.896   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7828 on 198 degrees of freedom
Multiple R-squared:  0.3309,    Adjusted R-squared:  0.3276 
F-statistic: 97.93 on 1 and 198 DF,  p-value: < 2.2e-16
```

The symmetry of the residuals and the fact that the median is close to zero indicate that the residuals are approximately normally distributed.

Both p values are 2e-16 and are less than 0.05. This shows that both the intercept and slope are statistically significant. (Intercept: -14.178540 Slope: 0.083756) This shows that player heights positively affect the number of aerial duels won.

The F-statistic is 97.93 and the p-value is 2.2e-16. This shows that the model is generally significant. The model explains approximately 33.09% of the variance in aerial duels won. (Multiple R-squared: 0.3309)

4. Conclusion

We have made conclusions related to our main goal and research questions. In general, we have seen that while most of the factors related to footballers’ performance have statistically significant effects on it, even some are not have. On the other hand, we have observed in the demographic analysis of leagues that while the youth player ratios do not differ from each league, domestic player rates of leagues differ. In order to recall and conclude our findings about research questions, we can focus on their results separately and in short.

In the analysis of first research question, we have seen that mean goal expectations of players by leagues are not different from the population including players from all leagues.

In the analysis of second research question, we have detected that there is not any difference in mean pass accuracy according to preferred foot whether is left or right.

In the analysis of third research question, we have observed that while the domestic player rates of Italian and English leagues are significantly less than 0.40, rate of Spanish league is significantly more than 0.4. Thus, we have concluded that having high native player rates is significantly important for Spanish league.

In the analysis of fourth research question, we have noticed that there is no significant difference in the young player rates between the Spanish and English leagues.

In the analysis of fifth research question, we have recognized that, mean goal expectations of attacking, defending, and midfield positions of players are significantly different by ANOVA test.

In the analysis of sixth research question, we have realized that the simple linear model about effect of players' heights on aerial duels won is statistically significant model. Considering the results of questions, may provide comprehensive perception about factors affecting performances of players in football.

References

Erkaya, P. (2024). *Recitation 5*

Erkaya, P. (2024). *Recitation 9*

Erkaya, P. (2024). *Recitation 11*

One-Sample T-test in R - Easy Guides - Wiki - STHDA. (n.d.). http://www.sthda.com/english/wiki/one-sample-t-test-in-r

Palit, A. (2024). Stat250_project. GitHub. https://github.com/ArdaPalit/stat250_project

Sofascore. (2024). Football scores. Sofascore. https://www.sofascore.com

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., & RStudio. (2024). *ggplot2: Elegant Graphics for Data Analysis* (3rd ed.). Springer. https://ggplot2-book.org/