EFFECTIVE BALL PLAY TIMES IN PREMIER LEAGUE AND TURKISH SUPER LEAGUE

Ömer Faruk Kolçak - 2024776054

1. INTRODUCTION

In this project, I analyzed **effective ball play times** in the **Premier League (PL)** and the **Turkish Super League (TSL)**. Effective ball play time refers to the actual duration during a football match when the ball is actively in play, without counting the stoppages such as fouls, goal celebrations, corner kicks, injuries, substitutions, and time-wasting. Unlike the total match time, which is fixed at 90 minutes plus added time, effective play time often adds up to much less, sometimes only 50–60 minutes. This metric is directly linked to the tempo, intensity and fluidity of a game. Higher effective ball play time mostly leads to a more enjoyable and engaging viewer experience. It also shows the quality and the level of competitiveness in play, which can influence various tactical strategies, match tempo and overall quality of the game. As football continues to empower with the collected data, understanding and improving effective ball play time has become a key focus for leagues and broadcasters.

2. METHODOLOGY

Generally, PL is considered to be the best and the most enjoyable football league by the football community. In this work, I examined if there is a statistical difference in effective ball play times between PL and TSL. Also, I study how different statistics in a game affects the ball play time. This section will explain different parts of the study such as: **data collection**, **exploratory data analysis**, **underlying distributions**, **goodness of fit, hypothesis testing**, and **regression analysis**.

2.1 Data Collection

The games are collected for both leagues for 23-24 and 24-25 seasons. Games and corresponding match statistics are scraped from the website FBREF. The following game statistics are collected.

- Number of goals
- Number of corners

- Number of yellow cards
- Number of red cards
- Number of fouls
- Number of crosses
- Number of interceptions
- Number of offsides

These statistics are collected for both home and away teams. However, FBREF does not contain the effective ball play time statistic. Therefore, 40 games for each league are sampled, 20 samples for each season, from the dataset and effective ball play times are manually inserted to the dataset from Mackolik mobile application. In total, there are 80 games in the final dataset. The overall data collection diagram can be inspected in Figure 1.

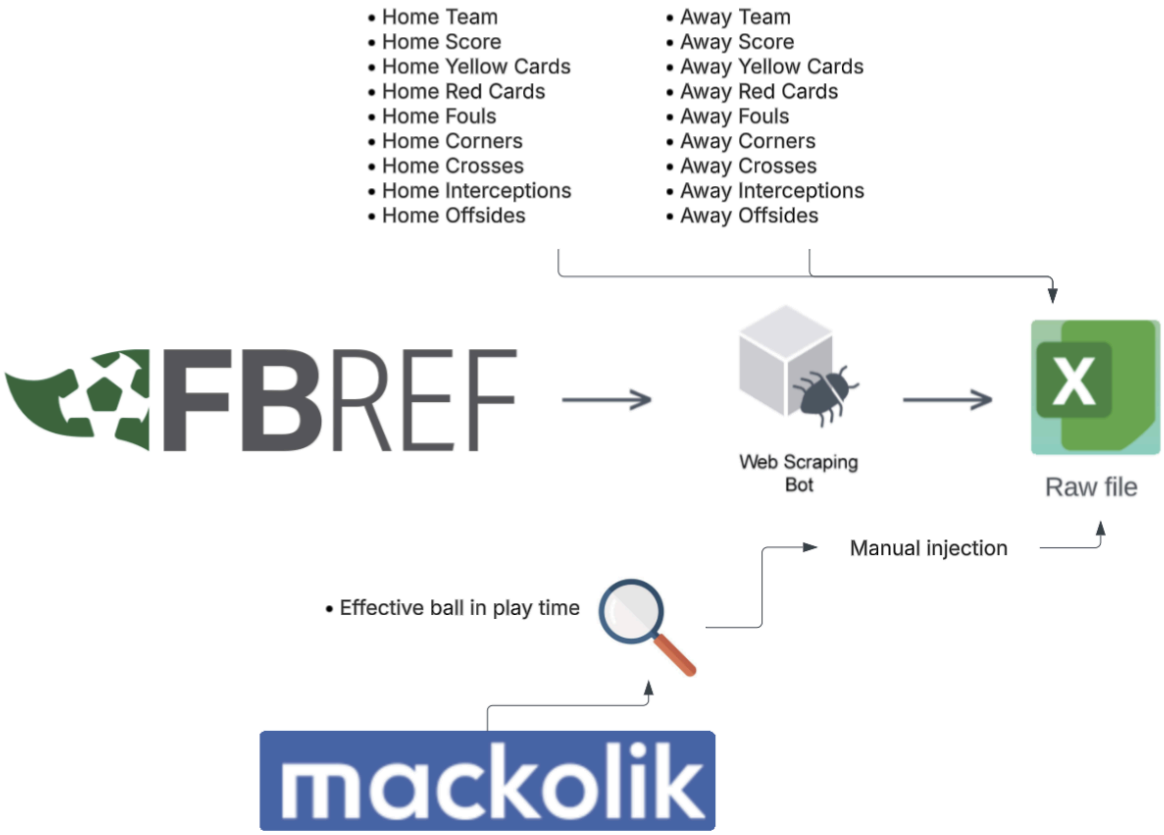
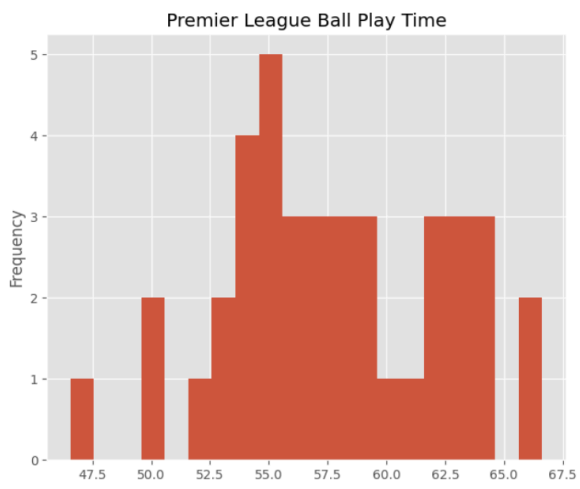
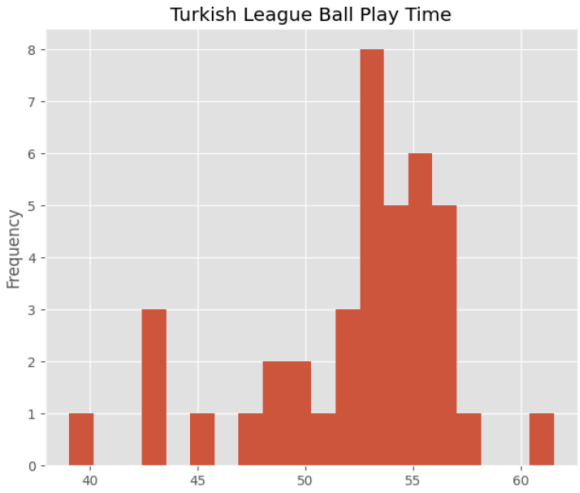


Figure 1: Data collection diagram

2.2 Exploratory Data Analysis

In this section, various sample statistics and visualizations are demonstrated to understand the data better. Histograms and sample statistics of the both distributions can be observed in Figure 2.






```
Premier League Sample Statistics:
---------------
Mean:      57.70
Median:    57.34
Std:       4.64
Skewness: -0.08
Kurtosis: -0.46
```

```
Turkish League Sample Statistics:
---------------
Mean:      52.25
Median:    53.19
Std:       4.59
Skewness: -1.00
Kurtosis: 0.87
```

Figure 2: Sample statistics of both distributions

Sample statistics of the PL data suggests that it is fairly a symmetric curve and it can be approximated by a normal distribution. However, TSL data shows that it is a left-skewed data with a value of -1 skewness coefficient. Additionally, Figure 3 presents the distribution of in-play times as a box plot, indicating that the PL generally exhibits higher in-play durations which is to be proved in **Section 2.5 (Hypothesis Testing)**.

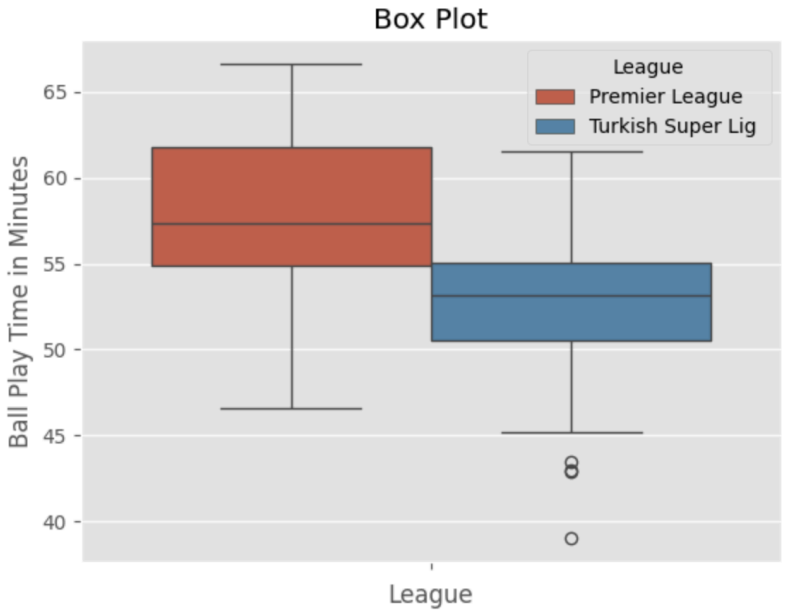


Figure 3: Box plot of ball play in times

2.3 Determining the Underlying Distributions

Based on the findings from the previous section, given the near-zero skewness, moderate spread, and visual similarity to a bell-shaped pattern, the PL data can be approximated by a normal distribution. In contrast, the distribution of data from the TSL more closely resembles a left-skewed version of the normal distribution. This asymmetry suggests that the data may be better modeled using a skew-normal distribution, which introduces an additional shape parameter to the probability density function (PDF) of the normal distribution to account for skewness. However, this brings additional complexity for the statistical inference, so that it is going to be approximated as a normal distribution for simplifying the process.

2.3.1 Method of Moments (MME)

In the Method of Moments Estimation (MME), the mean and variance of a normal distribution can be approximated using the following equations.

$$\mu_{MME} = \frac{1}{N}\Sigma x_i \qquad (1)$$

$$\sigma_{MME}^{\ 2} = \frac{1}{N}\Sigma(x_i^{\ 2} - \bar{x}^{2}) \qquad (2)$$

Based on these, estimated distribution parameters for the both distributions are shown in Table 1.

	Mean MME & MLE	Variance MME & MLE
Premier League	57.70	20.98
Turkish Super League	52.25	20.53

Table 1: MME and MLE estimations

2.3.2 Maximum Likelihood Estimation (MLE)

In the MLE, the mean and variance of a normal distribution can be approximated using the following equations 2 and 3. Actually, for normal distribution MME and MLE are equal point estimations.

$$\mu_{MLE} = \frac{1}{N}\Sigma x_i \qquad (2)$$

$$\sigma_{MLE}^{\ 2} = \frac{1}{N}\Sigma(x_i - \bar{x})^2 \qquad (3)$$

2.3.3 Bayesian Inference

When the mean is unknown and the variance is known the both prior and the posterior distribution follows normal distribution. Posterior hyperparameters are calculated as in Equation 4.

$$\sigma_1^{\ 2} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^{\ 2}}}\ ,\ \mu_1 = \left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^{\ 2}}\right)\sigma_1^{\ 2} \qquad (4)$$

Based on the conducted research, the known variance of PL effective play times is set to 25, and relatively a weak prior is set with mean 65 and variance 20. For TSL, the known variance is set to 30, and again a relatively weak prior is set with mean 50 and variance 20. Prior and posterior distributions can be observed in Figure 4. In Bayesian inference, a common point estimate is the mean of the posterior distribution. Point estimates obtained through Bayesian inference and Maximum Likelihood Estimation are also compared in Figure 4.

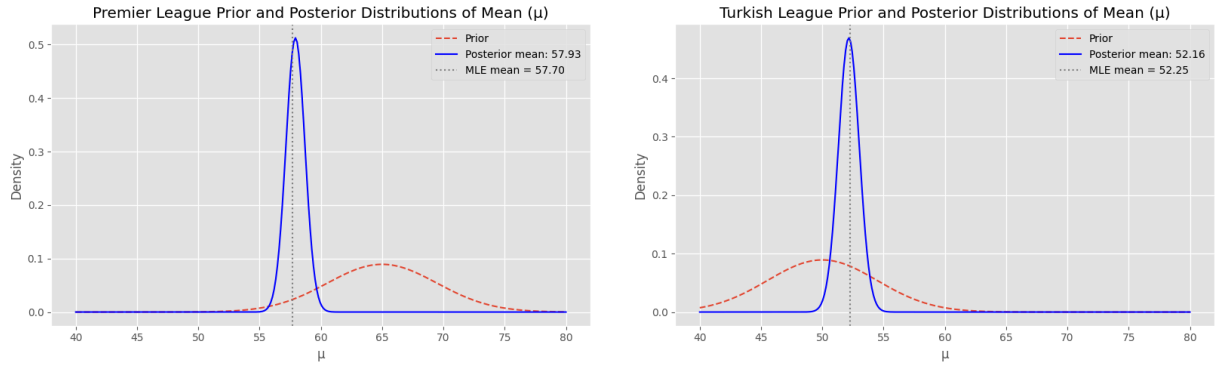


Figure 4: Comparison of posterior and prior distributions of Premier League and Turkish Super League. Additionally, MLE for mean can be compared with the bayesian point estimation.

2.4 Goodness-of-Fit

In this section, goodness-of-fit tests are conducted using Chi-square test. The experiment setup is as follows:

- **Hypothesis**
  - H0: The data follows a normal distribution.
  - H1: The data does not follow a normal distribution
- **Bin the data**
  - The data is binned in 5 minute intervals to convert the continuous distribution to categorical, since Chi-square distribution is for categorical distributions.
- **Calculate Expected Frequencies**
  - By assuming the normality with the parameter estimations from sample data, expected frequencies (counts) for each bin is calculated by using normal cumulative distribution function.
- **Calculate Chi-square Score**
  - $\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i}$ where O_i is the actual frequency of the related bin and E_i expected frequency based on the normality assumption.
- **Degrees of Freedom and Critical Value**
  - Degrees of freedom is calculated as the "**#bins - 1**".

For PL data details of the test are depicted in Figure 5. Null hypothesis can not be rejected, meaning that there is not enough evidence to say that this data does not follow a normal distribution with mean 57.70 and std 4.58. The p-value is calculated as 0.818 which is a very large value.

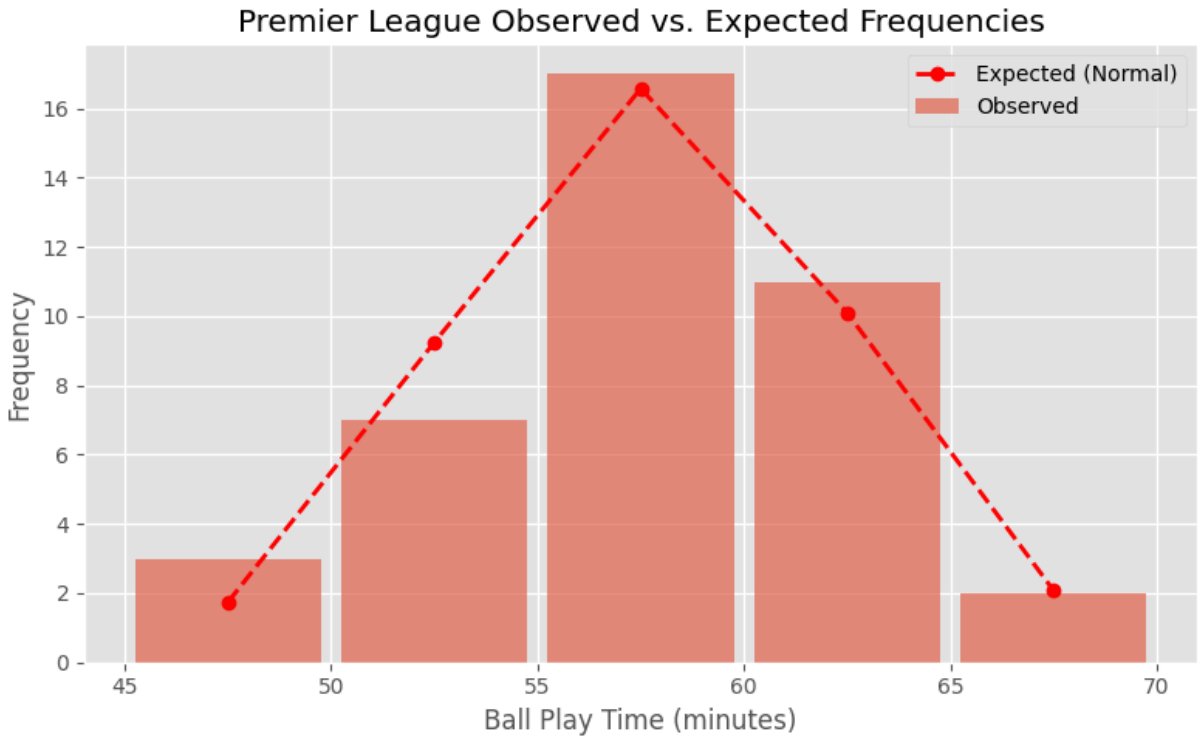


Figure 5: Actual frequencies and the expected frequencies from normal distribution are compared for Premier League effective ball play times. Chi-square score is calculated as 1.55, and the critical value is 7.779 for degrees of freedom 4 (#bins - 1). Therefore, we can not reject the null hypothesis.

For TSL data details of the test are shown in Figure 6. Also, for this data, we do not reject the null hypothesis meaning that there is not enough evidence to say that this data does not follow a normal distribution. The p-value for the corresponding test is calculated as 0.126.

However, these results are highly sensitive to the number of bins defined. In order to show this, the experiment is repeated for different numbers of bins and the results are depicted in Table 2. As we can clearly see, for the TSL data, as the number of bins increases it is easier for us to reject the null hypothesis which perfectly makes sense. In Section 2.3, a left-skewed distribution was observed in the data. However, for the sake of simplification, the data was approximated using a normal distribution rather than a left-skewed normal distribution. Now we observe the effect of such simplification. Although we do not reject the null hypothesis for TSL in Figure 6, Table 2 clearly shows that this is not the case, and the normal approximation is not fully appropriate for this data.

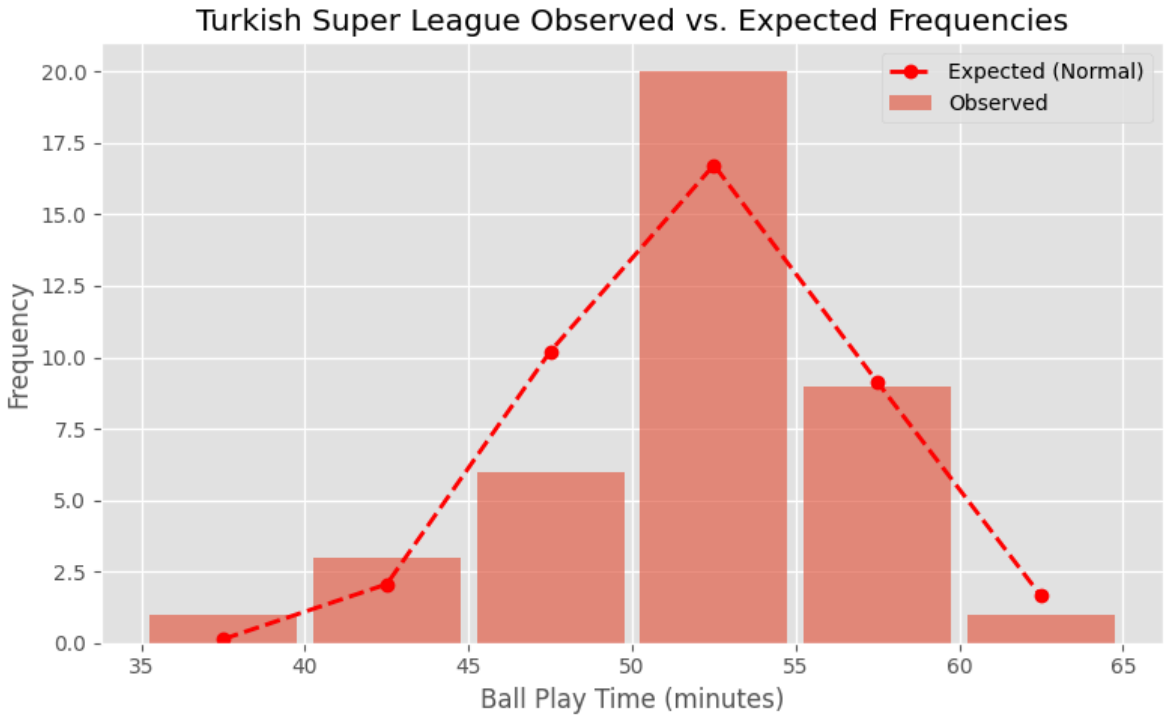


Figure 6: Actual frequencies and the expected frequencies from normal distribution are compared for Turkish Super League effective ball play times. Chi-square score is calculated as 8.61, and the critical value is 9.236 for degrees of freedom 5 (#bins - 1). Therefore, we reject the null hypothesis in favor of the alternative one.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
PL	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	R	A
TSL	A	A	A	R	A	R	A	A	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R

Table 2: Columns show the number of bins, rows show the league and 'A' means accept the null hypothesis and 'R' means reject the null hypothesis for the corresponding bin number. For PL, normal approximation seems very good because we only reject the null hypothesis in bin number 29. However, for TSL normal approximation seems not appropriate since we start always rejecting the null hypothesis as the number of bins increases.

2.5 Hypothesis Testing

This section tests whether the effective ball-in-play times in the PL are significantly higher than those in the TSL. The following hypothesis testing is conducted using t-test with a calculated degrees of freedom 77 and alpha 0.01.

- H_0: $\mu_{PL} - \mu_{TL} = 0$
- H_1: $\mu_{PL} - \mu_{TL} > 0$

- $t_{calc} = \frac{\bar{x}_{PL} - \bar{x}_{TL}}{\sqrt{\frac{{s_{PL}}^2}{n} + \frac{{s_{TL}}^2}{m}}}$, where n and m is the number of samples for Premier League and Turkish Super League respectively.

- $v = \frac{({s_{PL}}^2 + {s_{TL}}^2)^2}{\frac{{s_{PL}}^2}{n} + \frac{{s_{TL}}^2}{m}}$ is the degree of freedom calculation.

Based on these, t-calc is 5.308 and the critical value is 2.3758 which the null hypothesis is rejected in favor of the alternative hypothesis. This suggests that Premier League effective ball play times are significantly higher than the Turkish Super League.

2.6 Regression Analysis

Multivariate linear regression analysis is experimented for both Turkish Super League and Premier League dataset separately. Figure 7 and 8 show the results of the analysis.

Parameter Estimates for Turkish Super League

Term	Estimate	Std Error	t-Calc	Significance
intercept	70.91	inf	0.00	Not Significant
total_score	-0.47	0.36	-1.31	Not Significant
total_yellow_cards	-0.43	0.28	-1.54	Not Significant
total_red_cards	0.80	1.59	0.51	Not Significant
total_fouls	-0.29	0.15	-1.94	Significant
total_corners	-0.39	0.21	-1.88	Significant
total_crosses	-0.11	0.09	-1.25	Not Significant
total_interceptions	0.05	0.14	0.40	Not Significant
total_offsides	-0.32	0.38	-0.84	Not Significant

Figure 7: Multivariate linear regression analysis for Turkish Super League.

Parameter Estimates for Premier League

Term	Estimate	Std Error	t-Calc	Significance
intercept	73.30	inf	0.00	Not Significant
total_score	-0.11	0.43	-1.10	Not Significant
total_yellow_cards	0.09	0.31	-1.35	Not Significant
total_red_cards	-0.29	1.46	0.55	Not Significant
total_fouls	-0.45	0.13	-2.20	Significant
total_corners	-0.69	0.23	-1.73	Significant
total_crosses	0.01	0.08	-1.45	Not Significant
total_interceptions	0.07	0.17	0.32	Not Significant
total_offsides	-0.30	0.36	-0.89	Not Significant

Figure 8: Multivariate linear regression analysis for Premier League

Both results suggest that the most important variables to predict effective ball play times are the **total fouls** and **total corners,** and they affect negatively which perfectly makes sense. Also, **total offsides** decreases the in play time but the variable is not found to be statistically significant for the regression analysis. Although the **total number of yellow cards** in the Turkish Super League is associated with a decrease in in-play time, indicated by a high t-value, this relationship is not statistically significant. In contrast, in the Premier League, yellow cards are found to have a positive effect on in-play time, which is an unexpected result. Another factor negatively affecting in-play time is the **number of goals scored**, due to the time spent on goal celebrations and potentially on Video Assistant Referee (VAR) reviews. Finally, **total crosses** and **total interceptions** parameters are very close to 0 suggesting they do not have huge impacts on in play time which is an expected result.

3. CONCLUSION

This project's main findings demonstrate that the **effective playing times of the Premier League were significantly higher than the Turkish Super League effective playing times**. Also, the most important factors that reduce the effective playing times were found to be **the number of fouls** and **the number of corner kicks**. However, there are a variety of additional possibly important candidate variables that are not examined in this project such as, substitution of players, team doctor interventions, possession percentage—where close possession between home and away teams may indicate a continuous contest for the ball, penalty kicks, goalkeeper kicks etc. Additionally, this study could have examined potentially positively correlated variables such as the number of passes in a game and game tempo etc. These mentioned variables can be examined in the future work.