Jax Lehmann, Chibuzo Ifezue, Vincent Guminski

STAT-230

Adriana Picoral

3/25/2025

Final Project - Premiere League Analysis

For the final project, we have chosen to conduct a statistical analysis on an English League soccer data set. We have asked the questions, “What is the average number of goals scored each match in the Premiere League for the Home teams? As well as “How does the average number of fouls committed differ between Home and Away teams/does being Home or Away impact the number of fouls?” The average number of goals scored is a numeric discrete variable. The home and away teams are categorical variables.

Question 1 analysis and solution: To find this answer, we began by piping a histogram of the total goals. However, this did not give us any valuable input on the question we were asking. A better approach that we decided on was to simply calculate the mean of the total home goals. We got the result that the average number of goals for home teams was roughly 1.5 per game. Ultimately, this question was a little weak and is a good basis for future questions and research, such as “Is there a home field advantage based on average home goals and average away goals?” This could be something we add in the future, but the second question has a lot of elaborative research that exceeded this assignment’s requirements.

Question 2 analysis and solution: To find this answer, we began by attributing a box plot. However, since both variables (home and away fouls) are numerically discrete, a better approach was actually to conduct a scatter plot. The scatter plot showed a trend that the more fouls committed each game by any one team would result in more fouls than the other. However, it was very cluttered and was not indicative of one side having significantly more fouls than the other. We did also draw another question from this scatter plot that could be an interesting question to ask which is, “Are teams more aggressive or passive in conjunction with the opposition’s nature?” We could then conduct linear testing for an answer to this question. Firstly, we need to have a null and alternate hypothesis. Our null hypothesis was that fouls were not impacted by the team category. Our alternative hypothesis is that fouls are directly correlated to team type. To get our answer for question 2, we decided to run ANOVA and Linear Modeling. The ANOVA showed a P-value of <2e-16 (less than 0.05). This means we can reject the null hypothesis. From this, the ANOVA did not tell us the team fouls and which had more. We then ran the means and compared (home had 11.3 fouls on average and away had 11.8 fouls on average ) After seeing this results, we scratched our head as 11.8 is not much greater than 11.3 However, we realized our data set has thousands and thousands of variables, so the data was more sensitive. We came to the conclusion that statistically significant DOES NOT always mean practically significant. The Linear Modeling showed the same P-value as well as an R^2 value of 0.003844, showing a .38 variance in fouls.