James Bloor

Boston University Metropolitan College

MET CS 555

Term Project

Summer 2 2021

BU ID: U45448564

Project Question:

Being a soccer supporter, data in the sports world has always intrigued me. Learning what I have over the past few weeks in this class has made me sought out some soccer data and brainstorm possibilities. I have found a data set of the top 20 goal scores from five major soccer leagues around the world, year over year. I am interested in seeing if there is any correlation between time played and goals scored for these players.

Dataset Preparation:

I have carved the original data set to only show the name of each player, their time played, and goals scored. I have also filtered for five leagues around the world, MLS, La Liga, EPL, Bundesliga, and Serie A while only looking at the year 2019. The original dataset can be found in the link below.

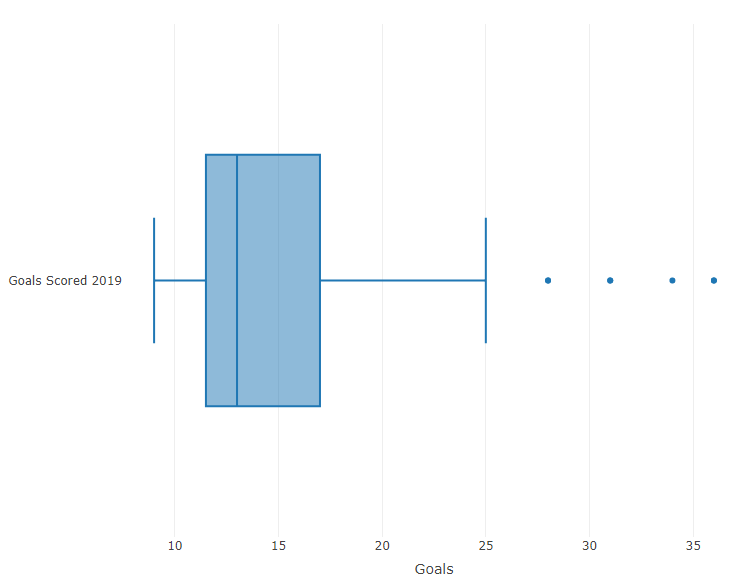
<https://www.kaggle.com/mohamedhanyyy/top-football-leagues-scorers>

Method Plan:

The question I want to answer, is there a correlation between time played and goals scored for the most prolific goals scorers in the world? I will formally test the data I have and see if there are outliers, a correlation, and if there is a possible simple linear regression. I will test the hypothesis that Beta1 = 0 at the alpha = 0.05. If H0: Beta1 = 0, there is no linear association. If H1: Beta1 != 0, there is linear association.

The first step I took, after carving my data from the original data set, was to identify outliers by creating a box plot. Then I will be able to see what data points need to be extracted.





Here we can see that there are a handful outliers, and these points can be points of influence, so I am going to pull these points from my data. After doing this, I am left with 93 observations.



The test I will use for this project is F = MS Reg/ MS Res, there are 92 degrees of freedom.

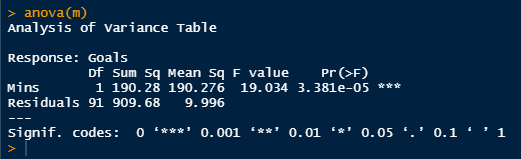
Results:

By running the statement qf(.95, df1 = 1, df2 = 92) in R, we receive the output of 3.945. Our decision rule then becomes the following, Reject H0 if F>=3.945, otherwise, do not Reject H0.



After running the necessary functions, I have computer the F statistic to be 19.034, which is larger then 3.945. Because of this we can reject the H0 hypothesis and say that there is a linear relationship between minutes played and goals scored for our players.

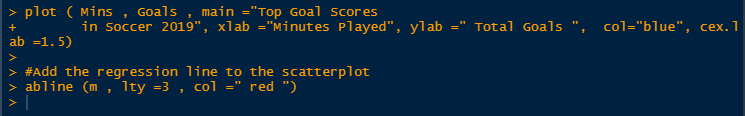


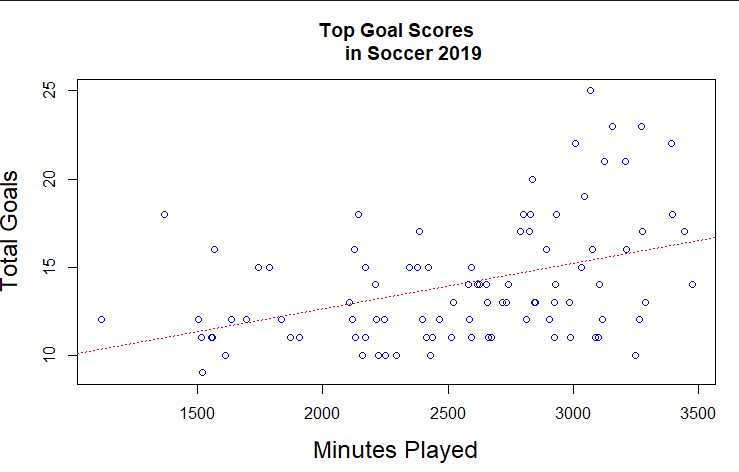


I was also interested in the correlation coefficient, so I ran some more code in R to find it to be .416.



To get a visual sense of the data, I also created a scatter plot and plated the regression line as well.





Above we see each of our players as a dot and the red line visibly showing the relationship between minutes player and goals scored.

Conclusion & limitations:

In conclusion, I did someone that there is a positive linear relationship between minutes played and goals scored for the players we have data on. By using the methods, I have learned this past semester, I was able to calculate this conclusion as well as visualize it using R.

There are a few limitations to this data. First, this data is not representative of all soccer players, looking at this list can quickly tell us that these are majority attacking position players. Other player positions and other attackers are not taken into consideration of this data. I am sure that if we brough in all players that played in these five leagues our results would look very different. The work I have done here is just for the most proven goal scorers. A few added steps can be taking into consideration in the future to pair along with this data as well. I could add more years of data into my work here to see how it effects my results. I could also add in other soccer metrics and run multiple linear regression to see what other metrics can help determine a goal scorers’ true level. Expected goals, assist, expected assist, could just be a few I can use in the future to rerun my work above and bring me to more advanced conclusions.