602 Semester Project – Notes on Part 3

Overview

For my semester project I wrote a program to analyze the impact of precipitation on the quantity of goals scored in English professional soccer games. Specifically, I focused on teams based in London as that allowed me to use the same precipitation data point for a given day across all the teams.

Data & Process

I used two datasets for my analysis. The soccer dataset was a csv file and it contained 188,060 rows of results stretching back to the 1888 season. It was sourced from https://github.com/jalapic/engsoccerdata . However, the specific date of the games was only populated in the dataset from the 1993 season onwards.

The weather data was sourced from the NCDC website: http://www7.ncdc.noaa.gov . I managed to download 18 years of data (from April 1992 to February 2010) into a csv file containing 6,503 rows, with each row corresponding to one day of weather data measured by the London Weather Station in Bloomsbury.

While I could have amended the two csv files in excel, and used perfectly clean csv files for my regressions, I decided it was more appropriate to do all of the clean up within the python code using pandas data frames. Therefore, I read in the csv files using the pandas read_csv() function, rather than creating a file picker via Tkinter.

I refined the files by removing all the rows I didn't need from the soccer dataset (pre 1993 season and post Feb 2010, along with the non-London clubs as the home team) and only pulling in the precipitation field from the weather dataset. The precipitation field needed some cleaning to extract the actual precipitation quantity from the alpha-numeric field. Also, the rows with missing data needed to be removed (missing data was represented by 99.99).

The final data frame contained data at game level, specifically the date, home team, total goals in the game and the precipitation in inches for that date. From this base final dataset I was able to create the data frames I needed to run the regressions. For the overall regression, I obtained the average number of goals scored (dependent variable) per precipitation amount (independent variable). I also took 4 subsets using this same structure for the main London clubs (Arsenal, Chelsea, Tottenham Hotspur & West Ham).

The regressions were run using the ols function from the statsmodels package. The summary function was useful for printing the output to the console. The plots were created using matplotlib and the regression lines were constructed using some numpy functions.

Findings

The program produces two pieces of output. Firstly, it prints out the regression summaries for all the games, as well as the home games of the four main London clubs. Secondly, it creates five corresponding plots of the regressions, which appear in one output window.

The overall finding is that there is a positive correlation between precipitation and goals. There is an increase in the average number of goals scored associated with larger precipitation amounts in all cases aside from the West Ham home games.