Cameron Gridley - Capstone 1 Proposal

Motivating Questions

Most sports have a home base where they routinely practice and compete. As such, competitors get to know the nuances of their home turf more than their rivals, giving them the "home field advantage". Formula 1 is quite different as all of the drivers have equal time on all of the circuits from year to year. Still, the home field advantage is frequently referenced. So I wanted to see if this was supported by the data based on a circuit's location and the driver's nationality.

- Ho: avg finishing position at home = avg finishing position away
- Ha: avg finishing position at home > avg finishing position away

The Data

I am using a collection of csv files containing extensive data on all Formula 1 races from it's inaugural season in 1950 through the end of last year. To start, I am merging 6 CSVs obtained from Kaggle.com that contain data about finishing results, races, drivers, circuits (race tracks), teams and car status.

This merged data contains 24,620 rows and 46 columns of data. There are 32 object cols, 11 int cols and 3 float cols. It is quite a clean dataset with minimal NaNs. Each file contains a primary key col to help join them together accurately. From initial inspection, some of the columns are object dtypes when it would make more sense to be an integer (e.g. the finishing position col, which will only ever be a number between 1-22, is of dtype object) so several cols may need to be converted from object to int.

MVP/Goals for Capstone 1

- Practice pandas data manipulation, especially groupy, filters, and aggregate functions
- Practice data visualization and making informative and attractive plots with MatPlotLib
- Make it object oriented so I could use this model with data from other auto racing series, such as NASCAR, Indy and IMSA.

MVP+

- Basyesian Hypothesis testing

MVP++

- Teenagers are becoming F1 drivers. Does age/experience matter? How has average age changed over time and what does the age:performance ratio look like over time?
- Get and merge weather data to see if there is a correlation with driver performance and temperature