Eldon Tsoi

SI 330

Project Pitch

This project is based on the inspiration from my personal experience in watching Formula 1. Formula 1 (also known as F1 or Formula One), is a single seater automobile racing sport, arguably the highest class amongst other single seater racing categories. The season starts in March and ends in late November. Each team has 2 drivers and a reserve driver. Both drivers are required to participate in each race, and reserve drivers are available if accidents happened to the two official drivers. Each race happens throughout the weekend, which consists of three “Free Practice” Sessions, also known as FP1, FP2, and FP3. Then, there are three qualifying sessions, where the bottom 5 cars are eliminated from each qualifying (Q1, Q2, Q3). If the cars are eliminated during qualifying, they have to start wherever they care qualified on race day. The race consists of number of laps, ranging from 50 laps to 70 laps, depending on the length of the track. Each race has a maximum of two hours in length, and the race will stop if the race goes over the two hour limit. The race also consist of pit stops and safety cars. Pit stops are used when the tire degradation is too high, and loses grip such that the driver needs to come in and change tires. Pit stops used to be able to do refueling as well. Safety cars are deployed if there are accidents on track, or when the rain is too big during the start. There are different flags that means different things as well, including a virtual safety car. Each driver wins 25 points if they win a race, and from 2nd place to 10th place will each get a lower number of points as they descend down the rank. Drivers that are placed 11th place or below will not get any points.

As I watch Formula 1, I am always interested in knowing how these drivers did in previous years, and predict how they’ll perform in coming races. One prediction is the number of times that this driver has won on each circuit, or qualified pole position during qualifying. These are strong indicators that the driver will win in this circuit. Other measurements include how often the driver is qualified top 10, finished in top 10, finished in the podium, or finished in the podium with certain conditions (such as qualified out of top 10 and finished in the podium). Another interesting measurement that shows how good a driver is is the number of championships that the driver won. To win a championship title, the driver has to win the most number of points compared to all the other drivers on the grid during that particular season. If you have the most points, you win the championship title for that year. Similarly, each team can also win a constructors championship title, which means the two drivers’ points add up and see which team has the highest number of points. To show each driver and the team’s strength, people often look at the consecutive titles these drivers win back to back, which is very challenging. Some drivers win races back to back, or qualified pole position back to back. It is very hard to win races and win the title.

There are many websites that list raw statistics for motor racing. Hence, I plan to pull that data with python and correlate it in a way that is easy for a user to see side by side comparisons of certain statistics between my favorite drivers.

To present the idea, this code will essentially demonstrate its ability to plot graphs, plot charts, intelligently detecting outliers and redundant information from our excel spreadsheet we gathered. As such, the code focuses on building the fundamentals for a code that can potentially used in many more areas.

I also plan to use interactive elements in Jupyter notebook by using different libraries, such that the data frames can be chosen by the user minimal typing, but instead can use sliders and other functions to interact with the code. Since the data on the internet is very likely to be messy, I anticipate using regex to parse through the data and extract useful information. I will most likely join data frames as well because there will be scenarios where I have to join different drivers’ data and compare them by column or a specific statistic. To organize data frames, I will use pandas. I will also use different APIs and web scrapping methodology to get data from the internet, hopefully converting them to CSVs, and present the idea by visualizing the data into graphs and charts.