Final Project – Data Science

Group 10

Documentation

1. Data management

As stated in the video presentation, the data came from two different *csv* files. The first one contains basic information from races from 1950 up until today (date, location, and identification number). The second dataset contains the detailed information for each race, like grid position, results, drivers, constructors, etc. Each race has a unique identification number that is invariant throughout the two datasets, which makes it easier for us to track them.

2. Data cleaning

In Formula 1, as in every sport of this kind, some drivers might not end a race, which is a problem when it comes to analyze positions and basically managing integers. So, to avoid that, we cleaned the whole dataset and used those retirements (NaN) to our advantage, to see what percentage of pole sitters would not finish the race in each circuit. Also, there were two circuits, Singapore and Sochi that have held a very small number of races, so we do not have enough data to apply any method, for that, we just plotted the data and let the user see for themselves.

3. Conclusion

After reviewing the data and the models we designed, we can determine that pole position is not as important in every circuit. The most important one seems to be Monaco because, eventhough the data is more distributed along the graph (pole sitters have finished in different positions, that is just because that circuit has had more races than the others, so it makes sense that more outliers appear.