Research Idea 1: Mini Project

Elsa Hritz

Research Question:

Do the most influential variables in F1 crashes change for different track conditions? This project hopes to explore the most influential variables in predicting crashes for numerous weather conditions. These would include combinations of wind conditions, precipitation, and temperature.

Data

The data comes from OpenF1–a large free and open-source API database. This database collects and stores real-time and historical Formula 1 data but is not maintained by Formula 1 companies. This database contains multiple data frames including: Car data, Drivers, Intervals, Laps, Location, Pit, Position, Race control, Sessions, Stints, Team radio, and Weather. This project would use columns from the Car data, Drivers, Laps, Location, Position, Race control, Sessions, Stints, and Weather data frames. The data is relatively clean, but each variable has a large set of categories, which may need cut down to increase model speed. This data is extensive and offers information on many different aspects of F1.

Data Retrieval

Since this data is stored in an API, each data frame must be pulled into R through an HTTP query. These requests will be written and run in R to pull as many different races and practices that are stored in the API as csv files. The specific columns and conditions will not be specified in the API request itself but coded in R afterward. Once these files are pulled off the API, a data frame object will be created for each in a new script. At this point, the 9 data frames will be cleaned to only include the variables that may have an impact on crashes, the crash data, and all weather-related columns/variables. Inner joins will be run using the keys to properly align the columns in order to connect all nine data frames. Using the event column in the Race control data frame, a new binary column will be added to the entire data frame indicating whether a crash occurred for that driver during that race.

Model

Once the main data frame is created-- it will be divided into smaller data frames using a filter for n different weather conditions. Each of these n data frames will then be divided into training and testing sets and used to train and test a logistic regression model to predict if a crash will occur. The weights of the factors for each of the n conditions will be compared to determine if the condition affects the factors that are most influential in predicting a crash.

Stakeholder Implications

Stakeholders would be interested in this project because it affects the driver’s safety, and future success of Formula 1, could reduce crash rates, and is impactful for those who bet on the races. Knowing what factors are most likely to cause crashes

for each set of conditions would potentially help the trainers prepare the drivers for safer practices in a multitude of conditions. Additionally, the mechanics can focus on creating more effective tires or a body on a car that performs better in high winds if necessary. Regardless of the results, ideally, this project allows the proper people to make the changes needed to prevent even more crashes from occurring. This would overall improve Formula 1’s statistics and reputation even more while providing greater peace of mind for the drivers and their families involved.

Ethical Considerations

It is important to consider that this data did not come directly from Formula 1. It is not guaranteed to be as correct as necessary to provide the accurate level of analysis required to make adjustments in such a large industry. If it was used and someone changed their tires or training style, and the data was inaccurate, it could unfortunately result in more crashes rather than fewer–putting even more lives at risk. Additionally, Formula 1 may not have permitted this data to be used for analysis. Therefore, if anyone uses this analysis for profit, such as trying to predict a crash for betting purposes, it could be considered unethical, and someone could be sued. This type of analysis could also further encourage illegal betting and malicious action to be taken.