Smith School of Business Queen's University | Master of Management Analytics *Toronto*

Course Number: MMA 860
Course Name: Acquisition and Management of Data

Assignment Name: Project
Due Date: July 12, 2024 12pm

Team Name: Team Gordon

Student Name	Student Number
Alisha Sahota	20497348
Anthony Ramelo	20499391
Chris Wu	10182394
Elizabeth Zhang	20161231
Emily Zhao	10096273
Sam Hossain	20466500

Order of files:

Filename	Pages	Comments and/or Instructions
MMA 860 Project Outline_Team Gordon	4	

Additional Comments:

Formula 1: Factors Impacting Race Outcome

Data source Ergast Formula 1 API

Summary

Formula 1 is the pinnacle of motorsports and one the world's most prestigious motor racing competition. There are 10 teams and 20 drivers racing, and 24 races that take place all over the world. Driver's go as fast as 370km/h and decisions are made in a blink of an eye.

Formula 1 is a data driven sport where each millisecond makes a difference to the race outcome. Over the last decade, teams have collected large amounts data to better understand where they can improve race performance. While the driver's skill and experience are a large part of the race outcome there are other factors such as, number of pitstops, starting position in the race, lap times that influence the race. In this analysis, we are going to be looking at each of these factors to determine what teams should focus on to get the best race outcome.

Methodology

1. **Data Cleaning and Treatment**
   For an accurate and meaningful analysis, we will need to treat for any missing/incomplete values and incompatible datatypes in the dataset.
   We will need to convert all time values into one type (seconds). These columns will be pitstop duration, fastest lap and lap speed. For any missing values from time columns, we will be setting as zero, this is due to a driver or team may have not participated in the race.

2. **Data Exploration**
   This is to get an understanding of the dataset to discover the relationships, patterns and potential problems:
   1. Perform a correlation analysis to identify relationships between the different variables and how they influence race outcomes, by analyzing the T-test results. As part of the analysis, also evaluate the initial regression models' ability to have predictive power on race times by performing the F-Test.
   2. Re-running the regression model to capture the correlated variables identified in Step 1 above, and further evaluating their impact on the model using T-Tests.

   3. Based on the updated regression model, we will be performing visualizations such as, box plots, histograms, scatter plots with the intention of identifying any regression assumptions that were violated (i.e. linearity of data, independent vs. dependent variables correctly identified, normality of residuals). Lastly, assess whether the r2 is producing sufficient level of accuracy for our objective.

3. **Problems we expect and how to handle:**
   Due to races being held around the world, lap times may differ depending on the circuit. This could cause our predictive model to have inaccuracies. The way we plan on handling this is by splitting the races into two types of circuits, 'street' and 'race'. As the two would have similar lap times.

We also expect our initial model will not capture all the variables that may influence race times, including any uncertainties that may sway the behaviors of specific variables. We plan to use the Chow Test to detect any structural changes or breaks within our model, and any sub-categories of the population that were not initially considered.

Lastly, we expect our model will not be perfect as it will contain a degree of error. We will be testing our model for heteroskedasticity to ensure the error rate is consistent with our expectations.

4. **Race Strategy Analysis/ Visualization**
   Analyze the relationship between qualifying positions and the race outcome to understand the importance of starting grid position.

5. **Predictive Modelling**
   Create a prediction model using the historical dataset, evaluate the model's prediction power using the F-Test. Review the model against the race strategy developed above.

Expected Outcomes

Based on our analysis, we expect to build a linear regression model that will allow us to predict race times and design a winning strategy that can be applied to F1 races. The model will capture all the relevant factors and variables that may influence race outcomes.

`