Stat 4355 Project Proposal

Team Members: Jonathan Serrano, Umar Ali-Salaam, Daniel Li, Yaseen Mohammad
Team Name: JUDY's Carbon Crusaders
Data: https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles

INTRODUCTION TO THE DATASET:

We found our dataset on Kaggle, however the origin of the data is from the Canadian Government official open data website. Since the data originates from a government institution, we believe that the data is reliable. The data contains 7385 observations of vehicles from a period of 7 years, 6282 of them being unique. There are 12 columns that describe certain features of the car, statistics of the fuel, and CO_2 emissions.

ANALYSIS GOAL:

The goal of our analysis is to determine which factors of a car, such as fuel consumption, the amount of cylinders, fuel type, and engine size, are best at modeling the amount of correlation to the CO_2 emission of each car.

DATA VARIABLES:

Our dataset contains 12 variables, where the CO_2 emission will be our response variable. We believe that there are two variables, namely the Make and Fuel Consumption Combined, that are either redundant or may have minimal or no impact on predicting CO_2 emission. By eliminating those 2 variables and using the 10 other variables as our predictor variables, we will create our linear regression models and check which predictors have a positive or negative correlation with CO_2 emissions. These are the following variables we chose:

1. **CO2 Emissions(g/km) (Ranges from 96 to 522):** The tailpipe emissions of carbon dioxide (in grams per kilometer) for combined city and highway driving.
2. **Model (Discrete):** Model of car with details of type of drive.
3. **Vehicle Class (Discrete):** Class of vehicle depending on their utility, capacity and weight.
4. **Engine Size (Ranges from 0.9 to 8.4):** Size of engine used in Litre.
5. **Cylinders (Discrete):** Number of cylinders.
6. **Transmission (Discrete):** Transmission type with number of gears.
7. **Fuel Type (Discrete):** Type of Fuel used.
8. **Fuel Consumption City (Ranges from 4.2 to 30.6):** Fuel consumption in city roads (L/100 km).

9. **Fuel Consumption Hwy (Ranges from 4 to 20.6):** Fuel consumption in highways (L/100 km).
10. **Fuel Consumption Comb (Ranges from 4.1 to 26.1):** The combined fuel consumption (55% city, 45% highway) is shown in L/100 km.

ANALYSIS PLAN:

1. **Data Cleaning -** Initially we should clean our dataset by removing any duplicate data points, unneeded variables, and all unnecessary NA values. This should help reduce bias and leave us with a clean dataset with unique data points.
2. **Explore the Data -** Then we'll make different graphs showing the predictor variables against the response variable, and calculate the amount of correlation between each predictor and the response variable.
3. **Multicollinearity Test -** We would like to have predictor variables that do not correlate highly with each other to make our model more reliable. Therefore, we will conduct a variance inflation factor (VIF) test to measure the collinearity between each predictor variable. Then correct for the variables with high collinearity.
4. **Multilinear Regression Models -** We will choose the predictor variables with high correlation to the response variable and then make a couple of multilinear regression models with permutations of the predictor variables.
5. **ANOVA analysis -** Using models made in step 4, we use the ANOVA analysis to find the multilinear regression model that best fits the data.
6. **Conclusion -** Quantify the impact of each of the predictors on the CO_2 Emissions (response), and draft a final conclusion about the model and each predictor.

RESPONSIBILITIES:

We'll determine how to clean the data as a group; then everyone will be delegated a set of predictor variables to analyze the significance of correlation between them and the response variable. Each member will then compare the strength of correlation of each predictor variable to the response variable and conduct a linear regression model. Finally, we'll come together to compare each model, decide which model is the best, and prepare a conclusion, presentation, and report.