**Statistical Modelling Proposal**

*Group #4*

*Alex Zakrzeski, Aditya Nayak, Purvi Jain, Tracie Robinson, and Alejandra Mejia*

To further understand the patterns within the data and be able to unmask the factors that truly have an influence in our dependent variable, we propose to use the following statistical modelling techniques:

* **ANOVA:** We propose to build three ANOVA models to determine the influence that other variables have in CO2 emissions.
  1. A two-way ANOVA with CO2 Emissions as a function of Number of Cylinders and Fuel Type
  2. A two-way ANOVA that displays CO2 Emissions as a function of the interaction between Number of Cylinders and Fuel Type
  3. A two-way ANOVA model with the same dependent variable and independent variables from Model 2 but excluding CO2 Emissions outliers.
* **Linear regression:** In other to include other predictors aside of the categorical ones used in ANOVA, we propose to build two linear models
  1. A simple linear regression model with no control variables, where the dependent variable is CO2 Emissions, and the independent variable is Engine Size.
  2. A multivariate linear regression model, using CO2 emissions as the dependent variable. As independent variables we propose using Engine Size, and five control variables that will include three new binary variables that we will create based on Premium Gas, Four Cylinders and if the car is an SUV, and two dummy variables representing the location where the vehicle was manufactured.

* **Logistic regression:** We also want to be able to answer the following question: What vehicle features are associated with CO2 emissions greater than 245 g/km? To do so, we propose running two logistic regressions:
  1. In our first model the dependent variable will be a binary variable we create that indicates whether a vehicle’s CO2 emissions are above 245 g/km and our independent will only be engine size.
  2. On the second model we propose to use the same dependent and independent variables as in model 1, but including control variables such as Fuel\_Consumption, Prem\_Gas, SUV, and USA.
* **K-nearest neighbors:** Finally, we propose using two K-nearest neighbors’ models to help predict CO2 emissions.
  1. Under the first model, we will predict CO2 emissions using engine size and fuel consumption as predictors. Under this model we propose not using any special resampling or multiple-fold validation.
  2. Under the second model, we will still predict CO2 emissions but we will include an additional predictor, engine cylinders, to try to improve the model fit.