**FUEL CONSUMPTION RATING**

The fuel consumption rating report aims to categorise the vehicles based on categorical and numerical criteria and predict the CO2 emission of the vehicles. Several data science approaches were used to make this prediction, including data cleaning, categorical data encoding, separating the dataset into training and testing sets, training the model, and finally forecasting the test set outcomes.

The data was cleaned to make sure it was accurate, consistent, and usable for creating our model and generating predictions afterwards. This procedure entails locating data mistakes, fixing them, and switching the data types of the columns to the proper data types. The Jupyter Notebook that is attached to this report has a detailed description of each data cleaning procedure.

**Question 1:**

Describe the steps required to train a model.

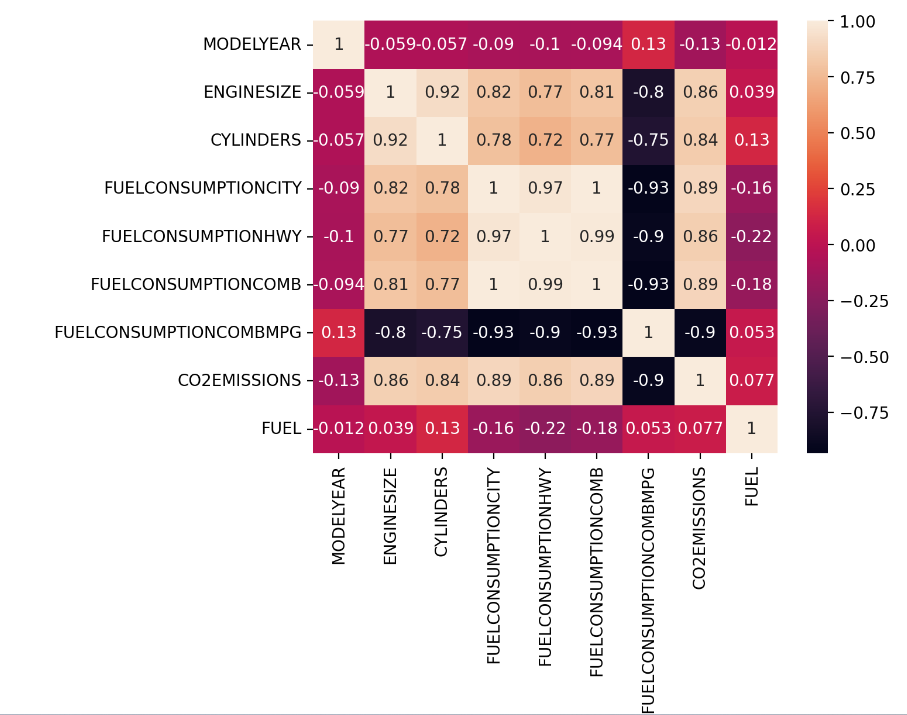
* First step is to import the necessary libraries
* Read in the dataset
* Clean the dataset by checking for missing values, null values and any inconsistent data
* Encode the categorical data
* Split the dataset into training, test and validation
* The model must then be trained by testing various parameters in order to produce accurate predictions.
* Train the model using Linear Regression
* Analyze its performance to determine how effectively it can foresee future data. Typically, to do this, you must divide your data into training and test sets, train the model on the training set, and then assess its performance on the test set.
* Obtaining the final linear regression equation and its coefficient values Create predictions based on new data using it. This entails supplying the model with data and using the model's output to decide what to do or how to do it in your application.
* Use a given set of independent variables to predict using the linear regression equation

**Question 2:**

Perform exploratory data analysis to select a subset of the variables and repeat the procedure. Compare your models and report if there are any differences in the models’ performances?

Explain your findings.

1. We first plot a correlation matrix of all the data features in order to study the dataset, and we next view the pair plots. Additionally, a few pair plots were displayed.



1. With the exception of FUELCONSUMPTION COMB MPG, the heatmap figures demonstrate a substantial connection between the dependent and independent variables. Except for CO2EMMISSIONS which is a dependent variable, all numerical variables are independent variables.
2. The numerical variables in the dataset are; ENGINESIZE, FUELCONSUMPTIONCITY, FUELCONSUMPTIONHWY, FUELCONSUMPTIONCOMB, FUELCONSUMPTIONCOMBMPG and CO2EMMISSIONS.
3. We scaled the characteristics of the independent variables in order for the predictors to have a mean of 0 as is frequently advised in order to develop our multiple linear regression utilising numerical data.

**Subset 1**:

After creating the regression model, we used metrics such as MeanSquared Error, Mean Absolute Error, and R-Squared or Coefficient of Determination to assess the model's performance in the regression study.

From the Multiple Linear Regression analysis:

* The mean absolute error is: 15.03
* The result of the root mean squared error is: 21.51
* The result of the coefficient of determination is: 0.89

**Subset 2:**

This subset utilises all features for multiple linear regression with the exception of 'CO2EMISSIONS', 'FUELCONSUMPTIONCITY', 'FUELCONSUMPTIONHWY', and 'FUELCONSUMPTIONCOMB.

From the Multiple Linear Regression analysis:

* The result of the mean\_absolute\_error is: 15.51
* The result of the root mean squared error is : 22.64
* The result of the coefficient of determination is: 0.88

**Overfitting**

A typical issue in machine learning is overfitting, where a model performs well on training data but struggles to generalise to new data (test data). When a model is overly complicated—for example, when there are too many parameters in relation to the quantity of observations—this happens.