# Report on Data Manipulation and Predictive Modelling

**1. Handling Missing Values:** The dataset was loaded from the 'automobile.csv' file.

- NaN values were used to replace the dataset's '?' values.

- Numeric columns with medians and categorical columns with modes were used to fill in the gaps left by missing values.

- The 'new_automobile.csv' file contains the cleaned dataset.

**2. Examining and Handling Data Types**: The columns with missing values' data types were examined.

- Since all columns with missing values had the same data type, no data type conversions were required.

**3. Identifying Correlated Features:** The cleaned dataset from 'new_automobile.csv' was loaded.

- To find characteristics connected with "Price," the correlation matrix was generated.

- The attributes that were most closely associated with "Price" were printed.

## 4. Creating a Predictive Model with the Independent Variable "engine size"

- The target variable "price" and the independent variable "engine size" were chosen.

- Data sets for training and testing were divided into two groups: training (70% of the data, and testing 30%).

- The training data were used to initialize and train a linear regression model.

- Predictions were produced on the testing set, and the model's effectiveness was evaluated using the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2).

## 5. Creating a Predictive Model with the Independent Variable "horsepower"

- The goal variable "price" and the independent variable "horsepower" were chosen.

- Data sets for training and testing were divided into two groups: training (70% of the data, and testing 30%).

- The training data were used to initialize and train a linear regression model.

- Predictions were produced on the testing set, and the model's effectiveness was evaluated using the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2).

## Results

- The dataset was saved as 'new_automobile.csv' after addressing the missing values.

- The correlation matrix was used to identify the attributes that were connected to "Price".

**- The performance metrics produced by the prediction model with "engine size" as the independent variable were as follows:**

- MSE (Mean Squared Error): [9683687.279391937]

- [3111.862349043083] is the Root Mean Squared Error.

R2 is calculated as [0.615672224689068].

**- The prediction model with 'horsepower' as the independent variable produced the following performance metrics:**

- MSE (Mean Squared Error): [29897790.25119074]

- [5467.887183473224] is the Root Mean Squared Error (RMSE).

- R2 = [0.18658842261524633]

The data wrangling procedures, correlation analysis, and predictive modelling employing two different independent variables are all described in this study.