AUTOMOBILE MPG PREDICTION

Intelligence of Mathematic System

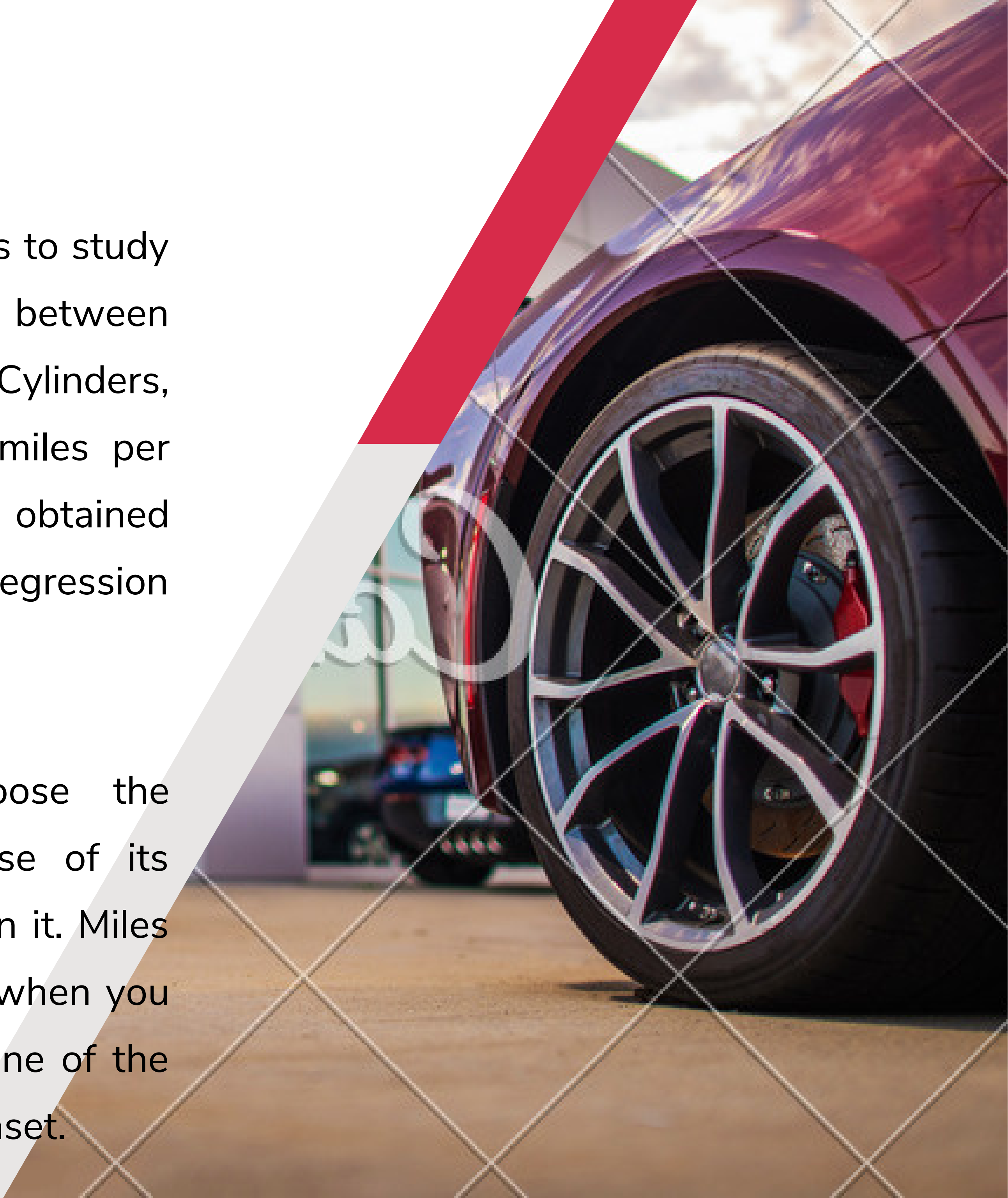
Group 8

INTRODUCTION

- The objective of this Project is to study the relative relationship between Horsepower, Displacement, Cylinders, Acceleration and weight on miles per gallon(MPG). The dataset was obtained from the UCI Website and Regression Analysis was conducted.

Reason

- The reason why we choose the particular dataset was because of its practical applications involved in it. Miles per Gallon(mpg) will be useful when you purchase a car and that was one of the reason why we choose this dataset.

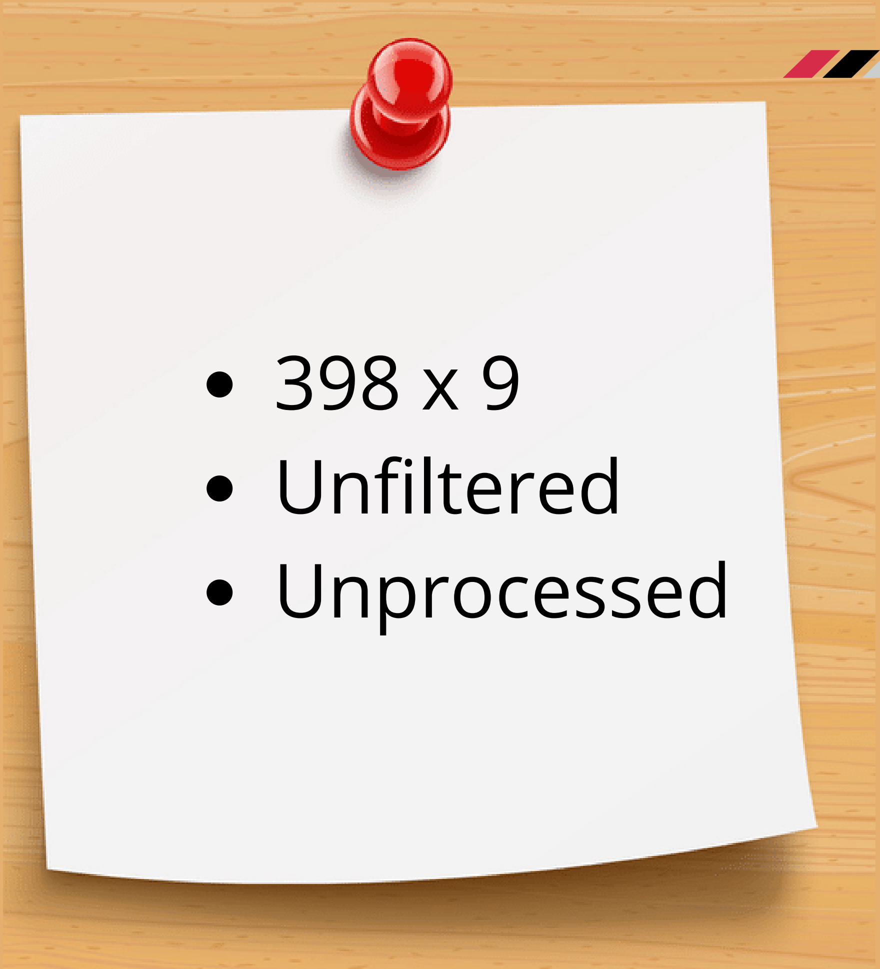


DATA SOURCING

- __This dataset is a slightly modified version of the dataset provided in the StatLib library. In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg", 8 of the original instances were removed because they had unknown values for the "mpg" attribute.

- __It has been extensively used by students, educators and researchers all over the world and is the primary source for the regression dataset Analysis.

LINK TO DATASET:https://archive.ics.uci.edu/ml/datasets/auto+mpg



- 398 x 9
- Unfiltered
- Unprocessed

- The Automobile MPG prediction data set contains the list of cars and their miles distances.
- We analysis the data into several sub class.
- Dataset is 398*9 dimensions and their properties.

METHODOLOGY

- The model that we have used to perform regression analysis is multivariable. It has more than two variables and therefore Multiple Regression Analysis is Conducted.

- The variable here to predict is called the dependent variable. The variables here to predict the dependent variable are called the independent variable

- The automotive industry is extremely competitive. With increasing fuel prices and picky consumers, automobile makers are constantly optimizing their processes to increase fuel efficiency.
- But, what if you could have a reliable estimator for a car’s mpg given some known specifications about the vehicle?
- Then, you could beat a competitor to market by both having a more desirable vehicle that is also more efficient, reducing wasted R&D costs and gaining large chunks of the market.

Introduction:

- This notebook contains the following columns: mpg, cylinders, horsepower, weight, acceleration, etc., which should all be self-explanatory.
- Displacement volume of the car’s engine, usually expressed in liters or cubic centimeters.
- Origin-It is a discrete value from 1 to 3.
- Model year - It is decimal number representing the last two digits of the 4-digit year

Data Ingestion:

According to others using this dataset, some of the mpg values for the cars are incorrect, meaning that some of our predictions will be off by a large amount, but we shouldn’t always trust the listed mpg value.
There are also unknown mpg values in the dataset, marked with a ‘?’. We will need to manually replace these with the correct mpg value.

DATA COLLECTION AND FILTERING

CAR NAME

- Ford turino and Ford turiono st are same car. So , change both to Ford turino
- Correct incorrect car names

DUPLICATES AND CONVERSION

- Remove duplicates
- Convert all string values to float to be fit in the modal

Data processing

The purpose of the data preprocessing stage is to minimize potential errors in the model as much as possible.
Generally, a model is only as good as the data passed into it, and the data preprocessing we do ensures that the model has as accurate a dataset as possible.
While we cannot perfectly clean the dataset, we can at least follow some basics steps to ensure that our dataset has the best possible chance of generating a good model.

4. EDA

The purpose of EDA is to enhance our understanding of trends in the dataset without involving complicated machine learning models. Oftentimes, we can see obvious traits using graphs and charts just from plotting columns of the dataset against each other.

5.Model training

List of packages we have used:

MODAL CREATION

TRAINING SET

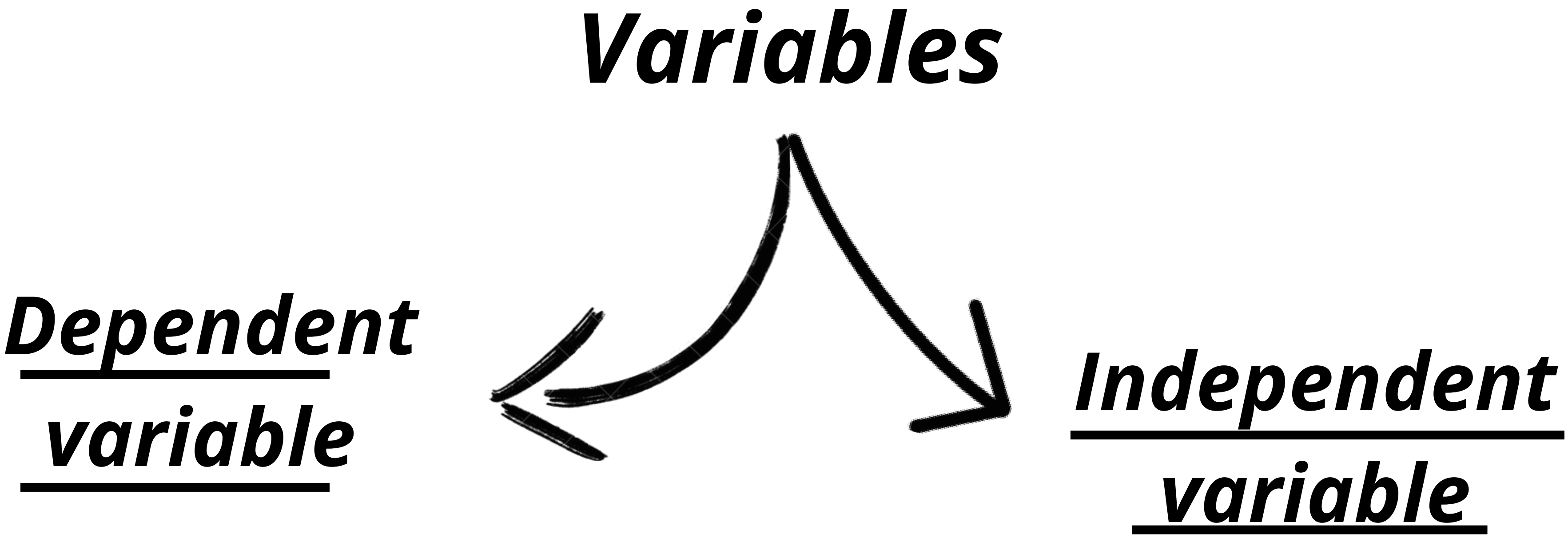
- We take feature(s) into the model to get the predicted values.
- We get a set of predicted values that has to passed on to the testing set

TESTING SET

- We have a set of actual values ans set of predicted values. Now we have to calculate the MSE between those sets. Then we find the accuracy using score

6.Testing the model

ATTRIBUTE INFORMATION



Variables

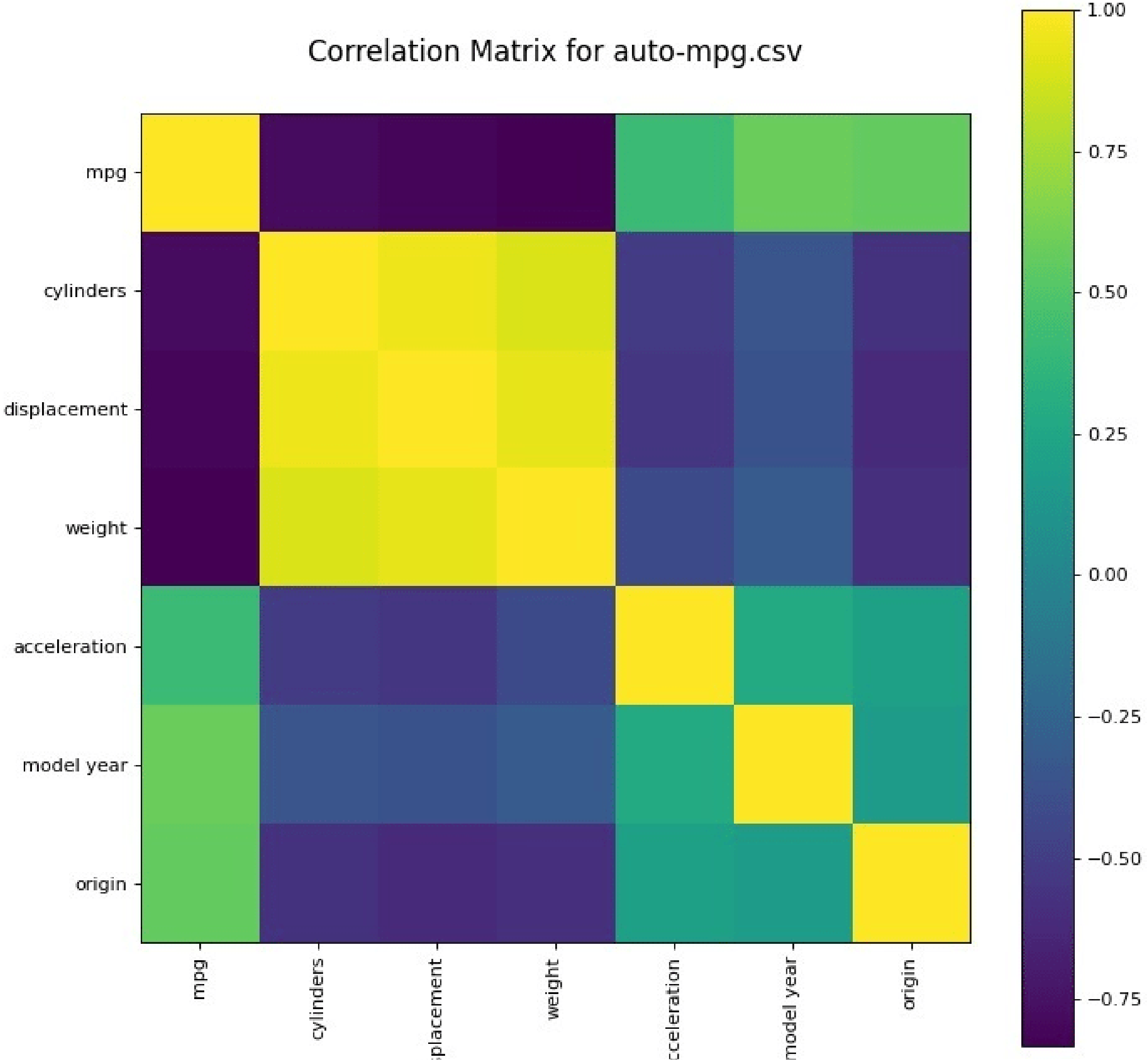
Dependent variable

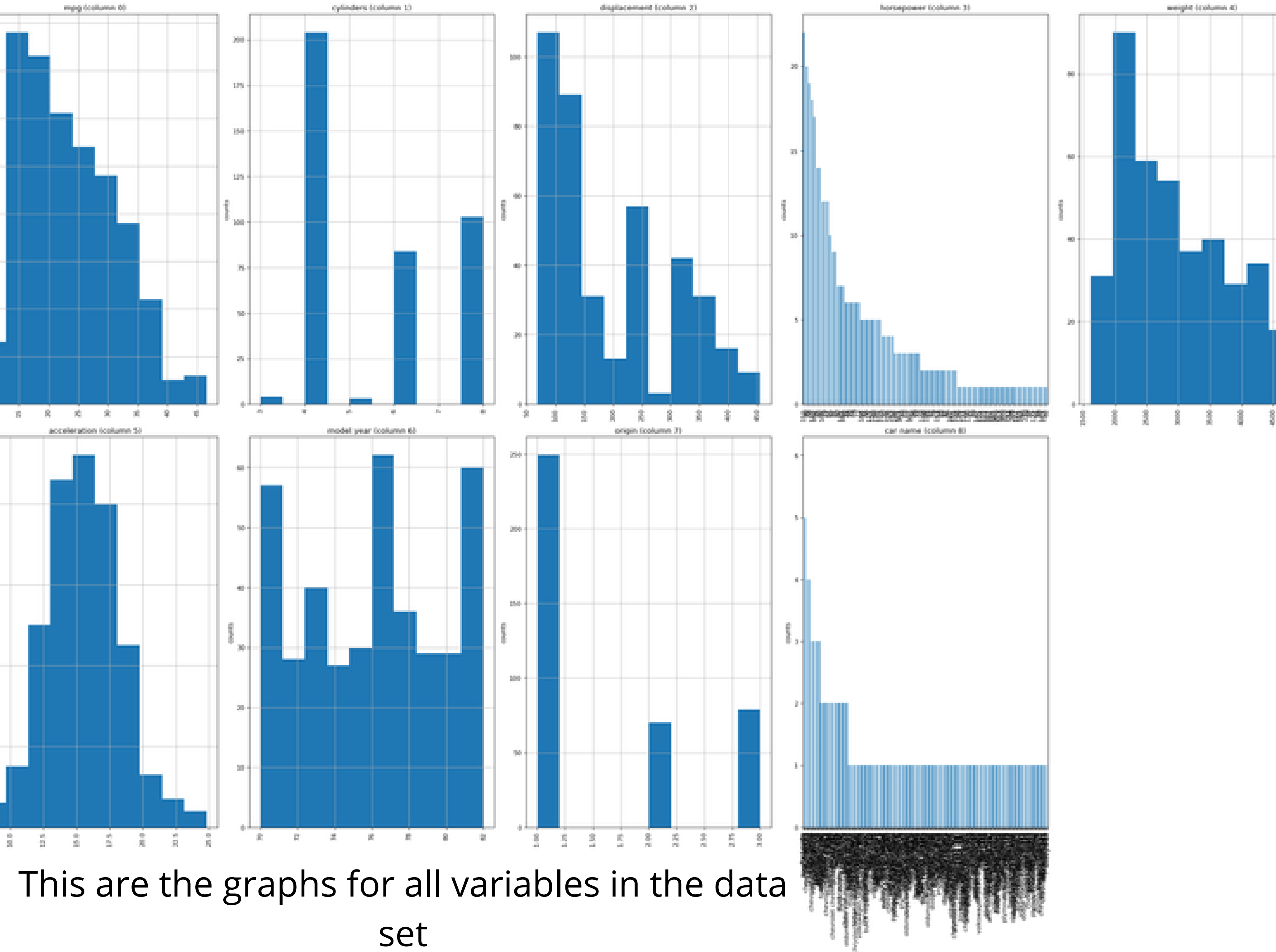
Independent variable

- Miles per Gallon(Mpg)-Continuous

- Cylinders-Multi valued discrete
- Displacement-Continuous
- Horse power-continuous
- Weight-Continuous
- Acceleration-Continuous.

DATA VISUALISATION (Correlation Matrix)





This are the graphs for all variables in the data set

THANK YOU

DONE BY-

- ABHISHEK ASHOKKUMAR
- GURU ASWINI DATH
- ERIC OOMMEN MATHEW
- VISHNUJITH MANU
- MANAV MANIPRASAD
- ANUPA SAJIKUMAR

Thank you