ADDIS ABABA INSTITUTE OF TECHNOLOGY

DONE BY: NAHOM SENAY
ID: GSR/4848/17
TO: DR. FANTAHUN
DATE: DECEMBER 2/2024 G.C

Problem Definition

The main aim of this project is to be capable of predicting the miles per gallon (MPG) provided that different parameters are provided. These parameters are vehicle characteristics like cylinders, horsepower. Being able to predict the miles per gallon of a car allows policy makers to be able to ratify environmentally friendly decisions based on precise definitions placed by international institutions. In addition, it will assist vehicle manufacturers to design their cars that are economized to the fuel based economy.

This model development process is planned to produce an MSE < 15 as an acceptable range. This means on average it is strived to achieve an error less 4 mpg on a car. This might seem big but due to the less availability of data it is thought to be fair.

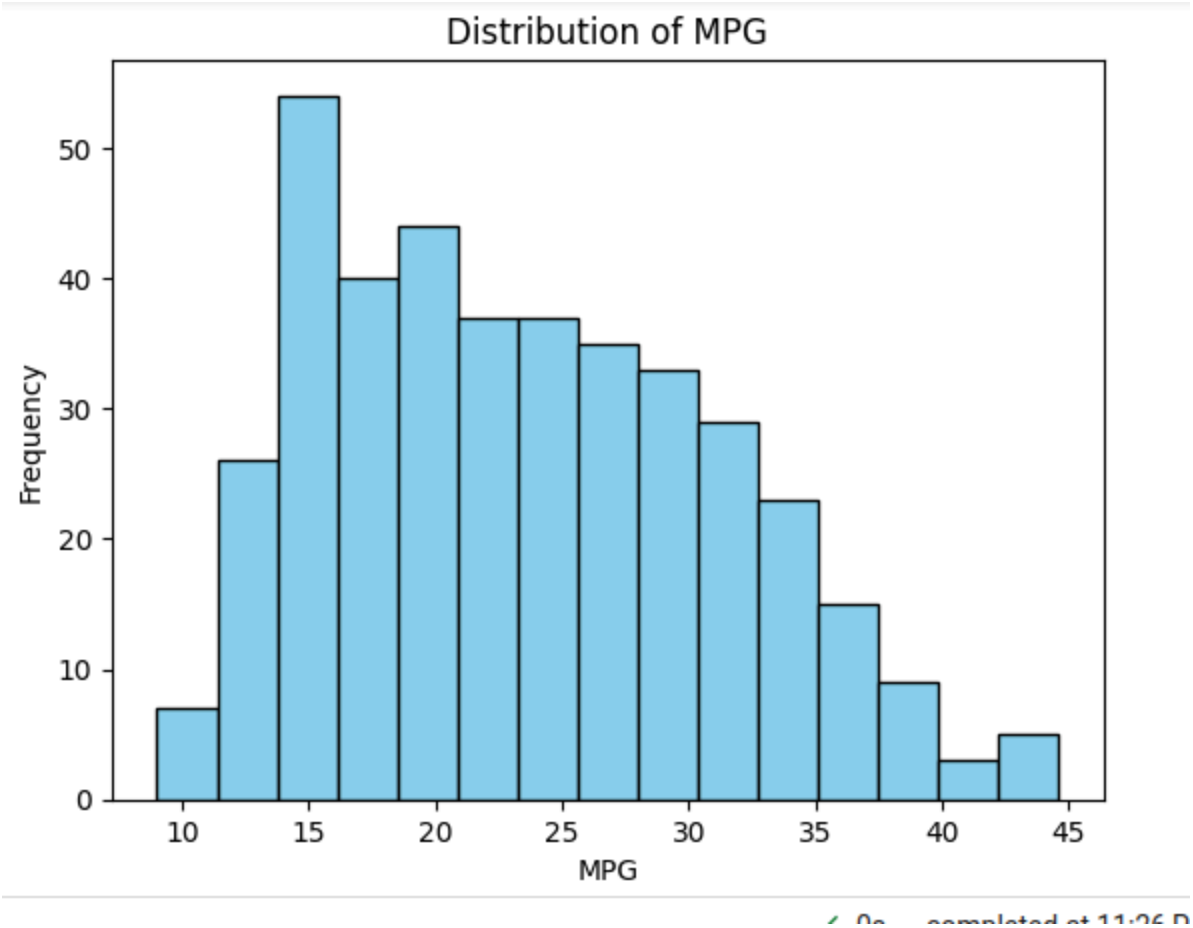
Data Collection

The dataset used in the model training process is found from kaggle. I have used the famous Auto-mpg dataset. The link for the dataset can be found [here]. It has 8 independent variables with one target variable. In addition, it has 398 rows. This is a linear regression problem whereby the miles per gallon is predicted out of the model.

Data Exploration and Preparation

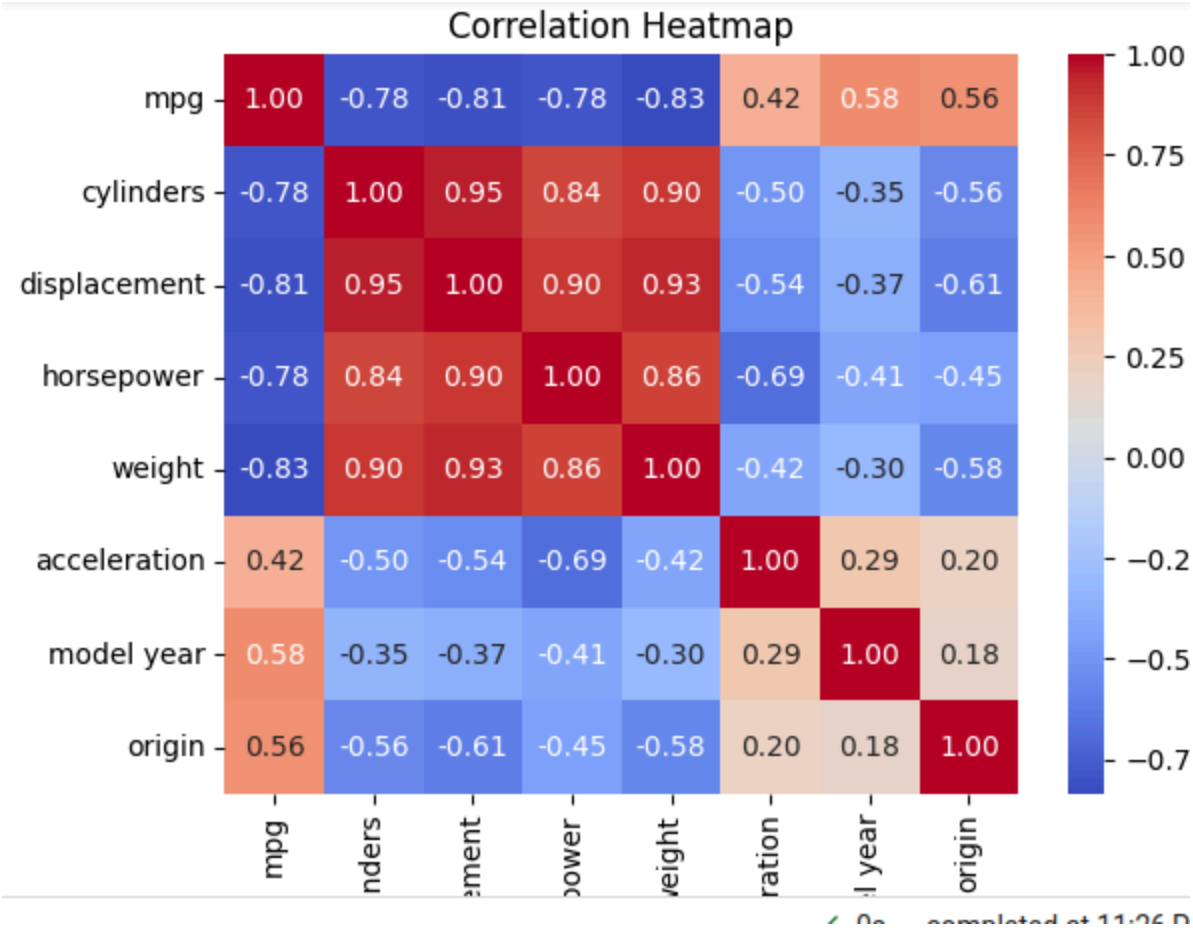
The dataset initially has irrelevant values. Hence the problem was curtailed by dropping irrelevant columns. For instance, it had a ‘car name’ property which had to be removed since it was computationally irrelevant.

Different MPGs were represented as a histogram with there respective frequency and the output was the following:



As shown in the above, the most common type of distribution appears for cars having around 15 miles per gallon. This according to the dataset in which the model fitting will be done on. But It can be inferred that most cars are between 15 miles per gallon to 30 miles per gallon.

In addition, correlation among variables is also studied using a correlation matrix between the 9 variables here is the result:



From the above displacement and cylinder have shown the biggest correlation. Followed by cylinders and weights. The biggest negative correlation has been shown between weight and mpg. Followed by mpg and horsepower.

In addition, further data cleaning to remove outliers was made using the inter quartile range. The interquartile range was used to get the upper and lower bounds so that it can be used as a boolean mask to filter the data that is in between the upper and lower bound.

Algorithm Selection

For this data set, linear regression was selected because the data was relatively small as compared to datasets required by connectionist approaches and implementations. It is also relatively computationally inexpensive as compared to other models. Furthermore, it is suitable for continuous data.

Model Training

During the model training phase the LinearRegression implementation from scikit learn was used coupled by the standard scaler. The standard scaler was used to avoid bias of different

components of a vector to result in an output. The output was found to be 10.16 MSE. Which was around 3 miles per galon. An acceptable result as compared to the goal mentioned.

Model Deployment

The model was deployed using the free feature of streamlit. The data was saved as a pickle file and accessed via a custom ui made for streamlit.