Samuel Amoabeng

PROJECT 1 REPORT

DATA PREPARATION

Data preparation is made using different steps in data processing, first, all the libraries needed to perform this process. Installation of these libraries is very important for this stage, it gives all the necessary packages for the preparation stage. The installation helps in calculating statistical descriptions and plots for the virtualization of distributed key variables to identify patterns and potential outliers for the model.

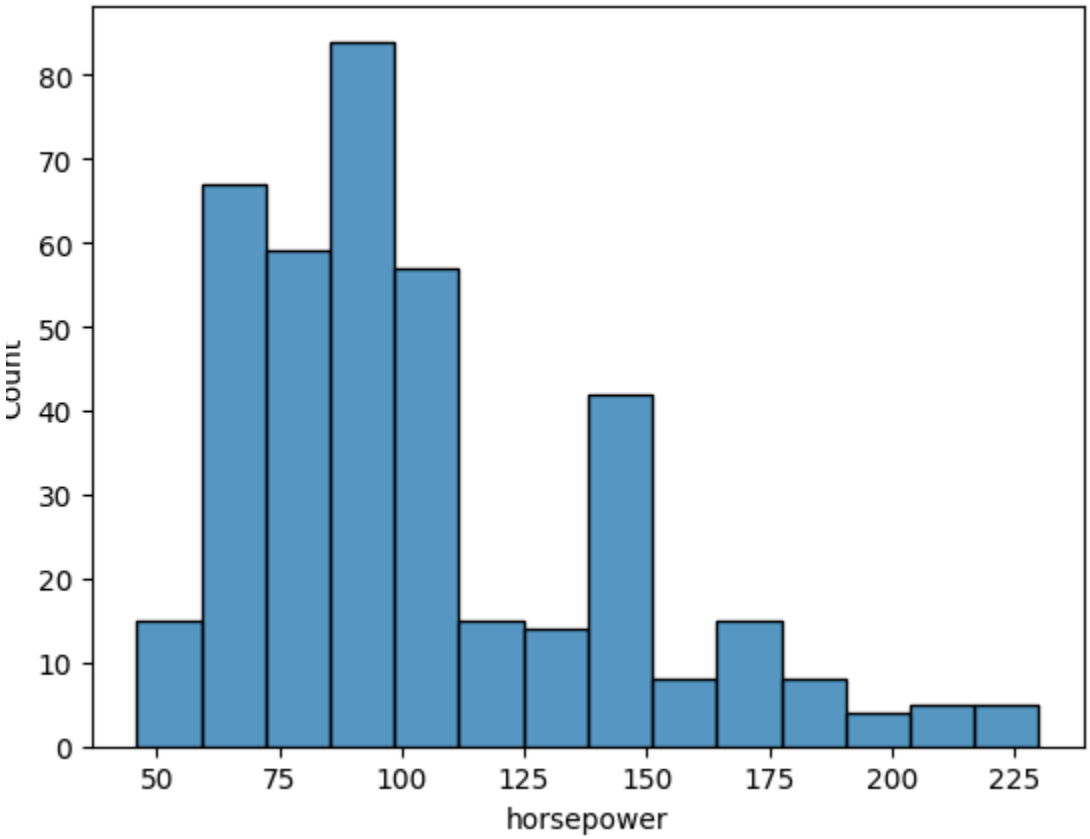
Cleaning the data is one of the integral processes by checking missing, and handling duplicates. It is also important to check the relationships, distribution of the data, and what part of the data is needed for the modeling. Creating new features or modifying the existing dataset to improve modeling performance. This can be done by converting categorical variables into formats that best fit the model, for this data process, horsepower was converted from an object to float. Descriptive statistics, such as mean, standard deviation, minimum, maximum, quartile, and inter-quartile ranges were calculated for each numeric column in other to understand the dataset.

DATA INSIGHTS

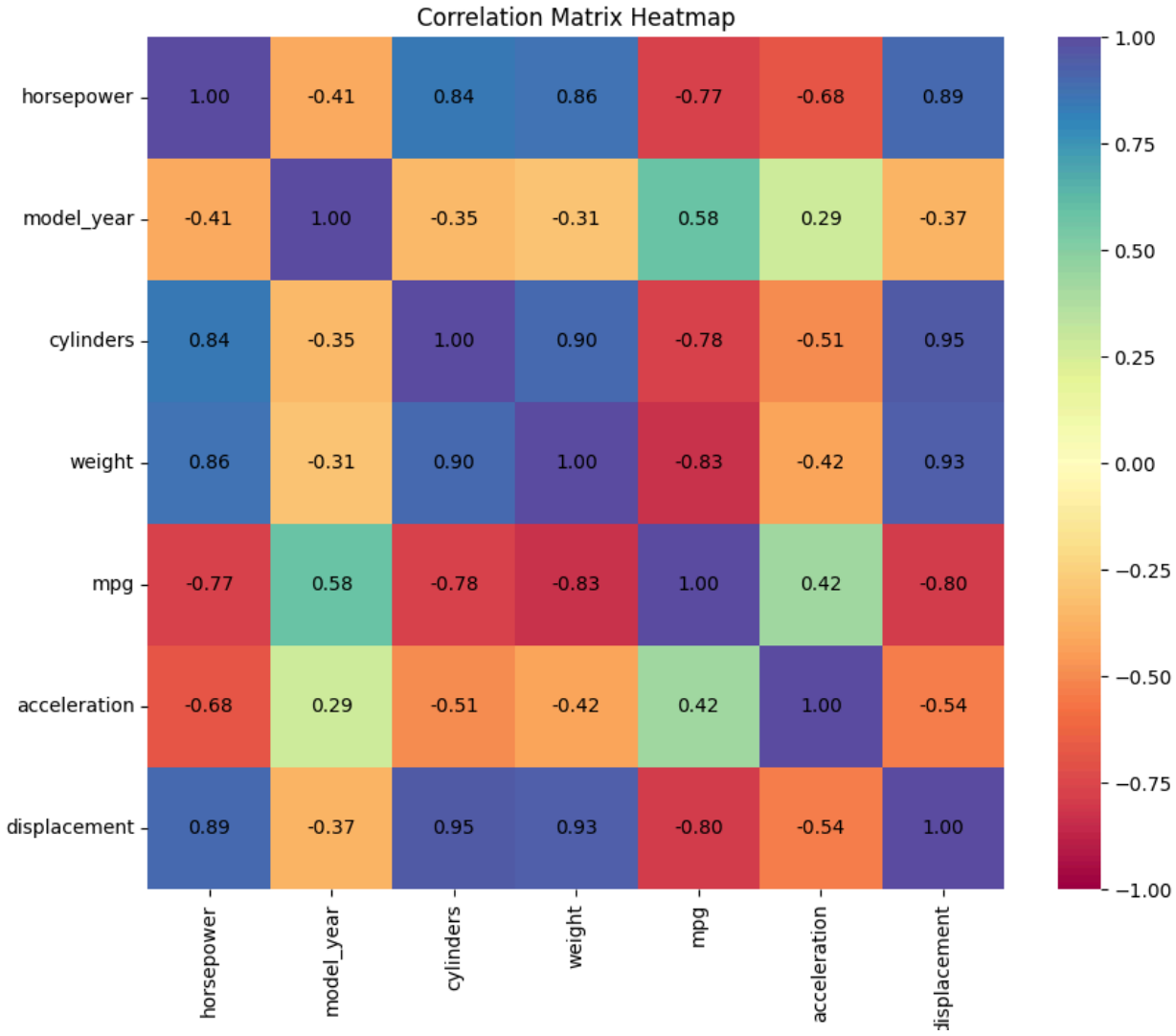
The data set contains information on 398 vehicles with 9 columns. The rows for the dataset represent the names of specific automobiles. The columns include number of cylinders, engine displacements, origin, model, acceleration, car name, and mpg. The data includes both numeric and which is in floats and integers and objects in strings. The horsepower columns datatype is in objects(strings), which needed to be converted to float. There was another character in the data

set that was replaced by the mean. In creating a heat map for the dataset I used a chatGPT to help get the code that provided the missing correlation matrix from my table.

DATA VIRTUALIZATION



Histogram plot.



This heatmap shows the correlations

PROCEDURE FOR TRAINING MODEL

A linear regression process was used to train the model, for predicting ‘mpg’ based on other features of different categories of automobiles in the dataset. The mean squared error was used to assess the model’s performance on the test data. The data split was 70% of the data used in training and 30% for testing. The process creates a linear regression model which helps the training model to make predictions. The MSE of the data was 8.95, this can be improved by reducing the value in other to get a well-trained model.

Based on the model MSE of 8.95 can be classified as a little no desirable for training a model, but the model can also be improved by achieving a more desirable value in this case a lower MSE for the model to work at its optimum best.