Assignment 1

Luis Dale Gascon
Computer Science
Towson University
lgascon1@students.towson.edu

I. Introduction

II. Dataset

The dataset is a modified version from Carnegie Mellon University's StatLib dataset. Each row contains the following information about a vehicle:

- horsepower: Describes the power output of engines
- weight: Describes how heavy something is
- acceleration: Rate at which velocity changes over time
- displacement: Total volume of air and fuel an engine can displace
- model_year: A vehicle's production period

These features are stated to have an endgoal of predicting a vehicle's miles per gallon.

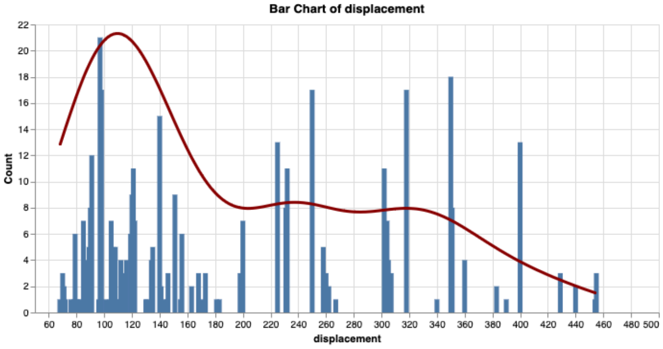
III. Exploratory Data Analysis

A. Missing Values

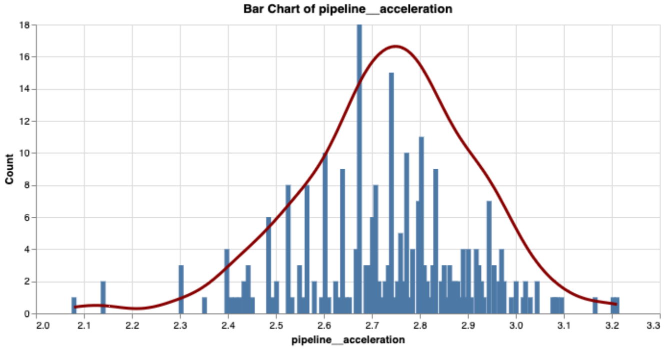
There are a total of 6 rows that were missing values for their horsepower feature. Since we're working with 398 points, we've decided to simply remove those 6 points as they only account for about 1% of the data.

B. A feature with non-normal distribution

Looking at the histogram for the feature of displacement, the data is positevly skewed.



Applying a log transformation



IV. Data Preprocessing

We want to set a standard value for numerical values to allow features with different magnitudes to contribute the same impact. I took features of type float such as displacement, horsepower and acceleration. I also included weight.

A. Data Binning

We bin the data with clustering-based binnning by using the k-means clustering algorithm to group the data based on similarities.

V. Regression

VI. Conclusion

References