**Introduction To Programming and Computation: Final Project**

**Introduction:**

The BE 1600 Project required analyzing a sample data set of vehicle data in Python. was tackled step by step, and the methods used to complete each step. Finally this will report identify the most important variables to predicting the fuel efficiencies of the vehicles provided, and demonstrate why using pair-wise correlation plots.

**Project Methodology:**

1. **Reading The Dataset:**

The initial dataset is stored in a .csv file. In order for python to be able to read .csv files, it must be converted into a format that the python programming language can understand. In this instance, a list was used. A function was defined for the purpose of converting the .csv file into a list in python. This function opened the file, converted each row into a string for each new line character indicator, then split the string into a list using the .split() function based on the commas present.

1. **Data Preparation:**
   1. **Identifying and Removing The Outlier Values:**

Outliers are values that represent extremes in a dataset. Outliers can exist due to errors in measurement and errors in data entry. These outliers can impact statistical analysis by creating distortions that make it difficult to determine if there is statistical significance. Outliers can be determined based upon the number of standard deviations a particular value is away from the mean of a data set. For this analysis, it was assumed that values with greater than 3 or less than -3 standard deviations from the mean were outliers.

A function was created in python to go through each column, ignoring null values, and calculating the standard deviation and mean of the values in that column. Outliers were then replaced with null values (These will later be replaced with the mean). The master list was then updated to remove the outliers.

* 1. **Identifying and Removing The Null Values:**

Null values were scanned column by column and replaced with the statistical mean for that column.

1. **Interpreting The Data Set Using Statistic Metrics:**

Python’s built in statistics module was used to calculate the mean, the median, and the standard deviation for each column. These were then exported to a .csv file for easy reading.

1. **Generating Plots For Pair Wise Correlation for Variables/Columns:**

A function was defined that generated subplots for every column of the master list against every other column inside of the master list. These plots were then printed in bulk.

1. **Providing a Pairwise Correlation Between Variables Using the Pearson Correlation Formula:**

A picture containing text

Description automatically generated

Figure 1. – This is the pairwise correlation formula for a sample.

A function was defined that calculated the pairwise correlation for each column of the master list using the above function. The resulting pairwise correlations were then exported to a list, which was then exported to a .csv file for easy reading.

1. **Defining and Plotting A New Set of Variables and Testing Them By Checking Their Correlation To MPG:**
   1. **Displacement on Power:**
   2. **Weight on Cylinder:**
   3. **Acceleration on Power:**
   4. **Acceleration On Cylinder:**

A function was defined that took an existing column of the master list, and compared it to an existing column in the master list. A new list was then returned of the first specified column divided by the second specified column. From this the desired values were calculated. These were then added to the master list via a new function that added existing lists to the master list. A pairwise correlation using the master list correlation function and master list plotting function was then ran.

1. **Implementing A Function To Mark Each Vehicles with a Low or High Fuel Efficiency”**

A function was defined that let the user determine the fuel efficiency, and then mark it as a high or low based upon the user’s benchmark fuel efficiency. A fuel efficiency of 30mpg was assumed as ‘high’ fuel efficiency, everything less than 30 was defined as low. This was then added back to the master list. The master list was then exported for easy reading.

**The Most Important Variables For Predicting Fuel Efficiency:**

Based on the pairwise correlation and the shape of the graphs, the strongest indications of low fuel efficiency was a high number of cylinders, a high displacement, greater horsepower, and high weight. Newer vehicles were associated with greater fuel efficiency, origin, and acceleration were also linked to higher fuel efficiency.

Application, table

Description automatically generated

Figure 2. – Correlation Coefficient as filtered in the CorrelationCoefficients.csv export file from the project.

From this we can conclude that the ideal fuel efficient has low weight, very low displacement, a very low number of cylinders, and very low horsepower, while having been built recently. It is therefore likely that the most fuel-efficient vehicle would be a new sailboat.

Icon

Description automatically generated

Figure 3. – The optimally fuel-efficient vehicle. Zero horsepower, no cylinders.