**Automobile Selector: Analysis on Automobile Attributes and features to demonstrate comparison between vehicle manufacturers.**

Hsiaoan Wang\*, Priyanka Annapureddy\*\*, Rani Sebastian\*\*\*

[\*Hsiaoan.wang@marquette.edu](mailto:*Hsiaoan.wang@marquette.edu), [\*\*priyanka.annapureddy@marquette.edu](mailto:**priyanka.annapureddy@marquette.edu), [\*\*\*rani.sebastian@marquette.edu](mailto:***rani.sebastian@marquette.edu)

**Abstract---**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**1 Introduction**

Shopping for an automobile has proven to be challenging for many due to the sheer number of options we have nowadays. Comparing fuel consumption has been the most common tactic that buyers use to determine which vehicle best suits their needs. However, we wanted to ensure that the fuel consumption of the automobile is not the sole attribute the buyer considers when making such a long-term investment. Therefore we will be assisting the buyers in making an automobile investment by considering more features and attributes of the vehicle manufacturer. The historical data for each car manufacturer along with their vehicle model information will be used to help us perform a comprehensive analysis on the relationship between the fuel consumption of the vehicle and other attributes such as vehicle horse power, vehicle acceleration speed, the number of cylinders the vehicle has and the displacement of the vehicle to help the end user better understand each car manufacturer and identify the dream car manufacture and model that they should invest in based on their needs.

**2 Related Works**

**3 Dataset**

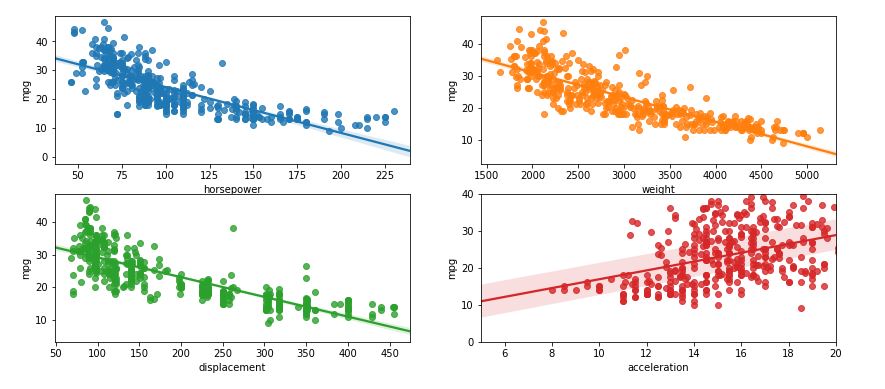
The dataset that we will be performing the analysis on was taken from the StatLib library maintained by Carnegie Mellon University. The dataset itself is now publicly available and can be found in the UCI machine learning repository. It contains a total of 398 instances of various vehicles from various manufacturer. Each data instance consists of nine different attributes, namely –MPG (miles per gallon or fuel consumption per gallon), cylinder numbers within vehicle, displacement, horsepower, weight, acceleration, model year, origin and car name. The attributes are either continuous, multi-valued discrete or are strings. The dataset contains vehicles that are of the same models but were manufactured in a different model year and contains data between the model years of 1970 through 1982. Although it has plentiful information regarding car manufacturer and attributes, the dataset comes with several missing values which may require some cleaning and replacement work.

**4 Methods**

**4.1 Exploratory Data Analysis**

We began with exploratory data analysis, allowing us to get familiar with the data and find out which variables have a strong correlation with mpg that will be good predictors for our model.

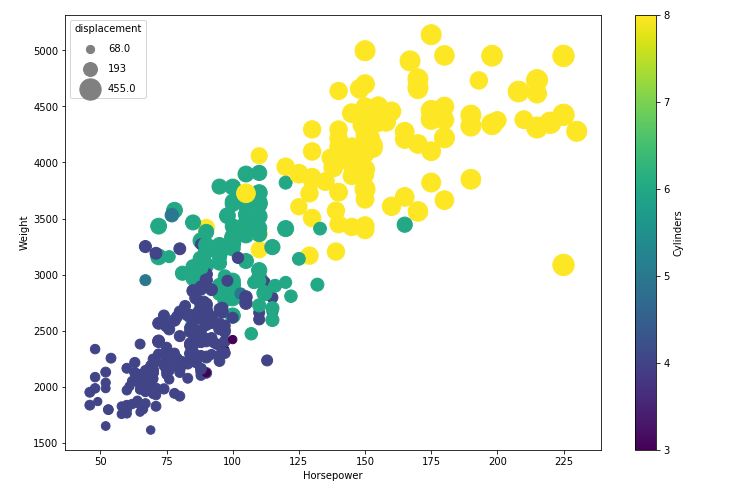
Investigating on *mpg*, we found that mpg has clear linear relationships with *horsepower, weight, displacement* (inversely proportional) and *acceleration* (directly proportional). Also, the dataset had some missing values for horsepower; which we dealt with using linear regression to impute the missing values using *weight*. These linear relationships are clearly visualized in the plot shown in Figure 1.1.

****

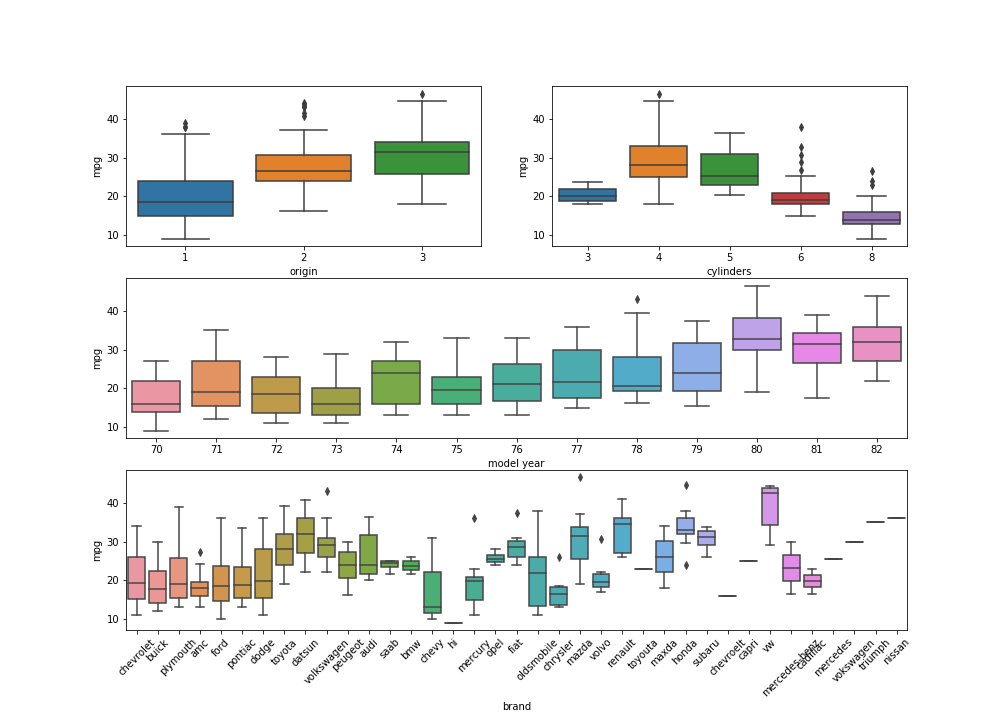
**Figure 1.1 Comparison of mpg against various attributes.**

The number of *cylinders* have a clear impact on mpg too. Our data shows us that increasing the number of cylinders may give the vehicle more power, but it does not result in a better mpg value. Figure 1.2 highlights this – the larger the cylinder count, the heavier and more powerful the vehicle will be, with displacement also generally higher. The ideal cylinder count for the optimal mpg value is found to be 4 as shown in Figure 1.3.

*Origin* is another interesting field shown to have some correlation with mpg. Origin is 1 for USA, 2 for Europe and 3 for Japan – we find that US cars are the worst among the three and Japanese automobiles have the best mpg to offer among the three. The *model year* highlights another trend of the automobiles in general – average mpg values of the later model year (80s) vehicles are significantly higher than the mpg values of the former years (70s). These trends are shown in the Figure 1.3. It also shows the various brand names.

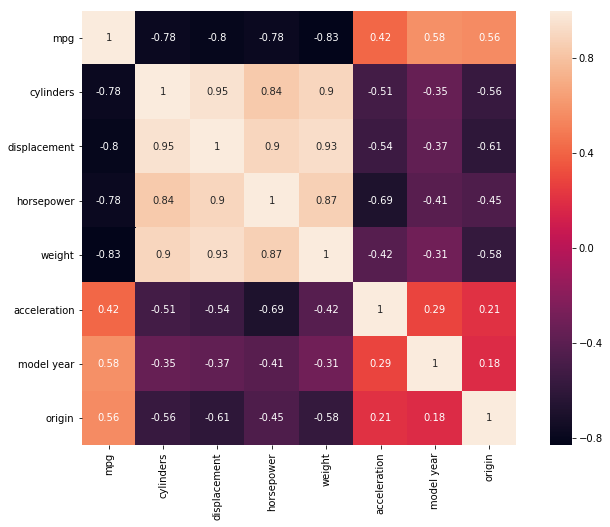


**Figure 1.2 The relationships between cylinders, horsepower, weight and displacement**

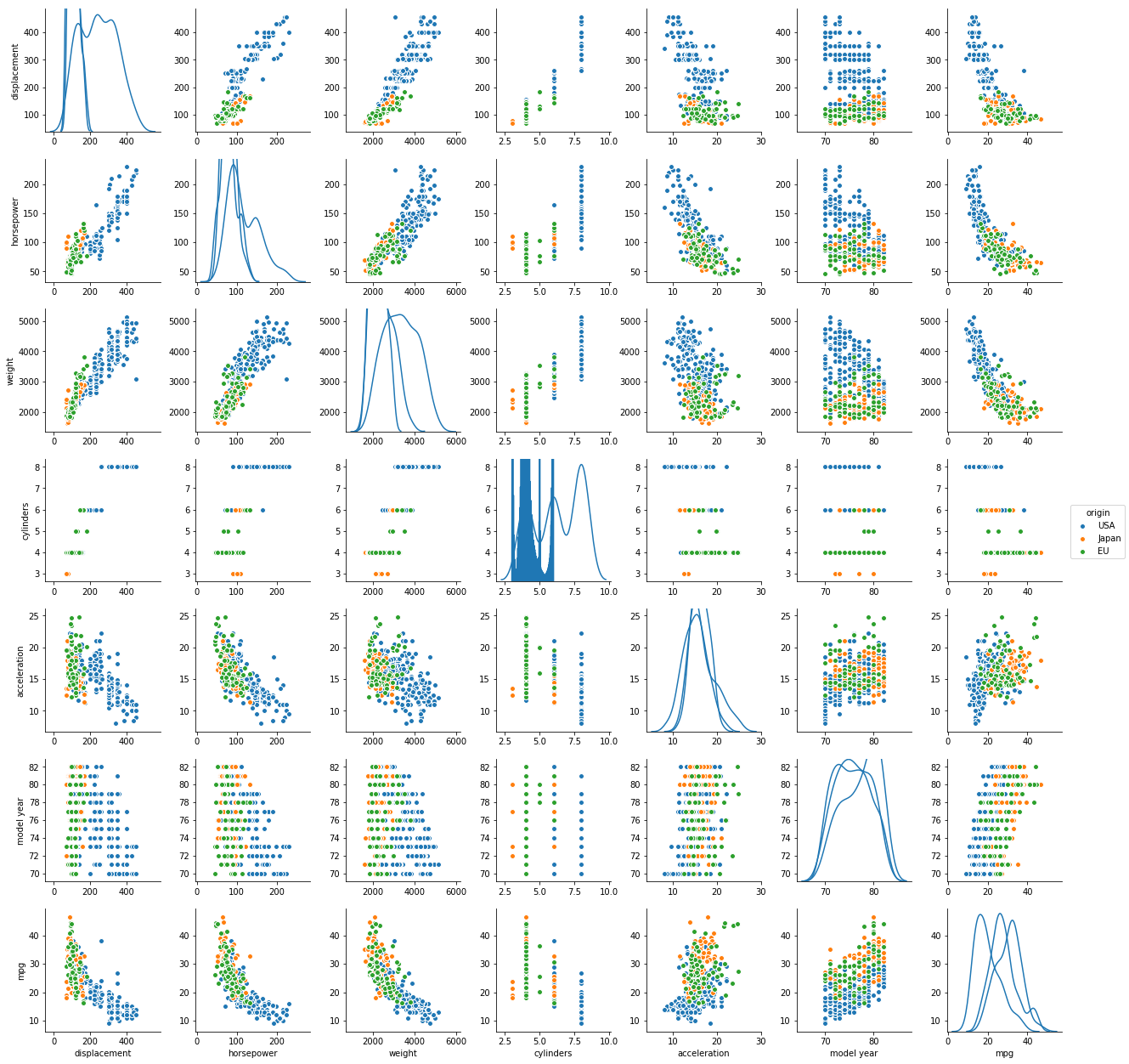
****

**Figure 1.3 mpg VS origin, cylinders, model year and brand**

The correlation between all the variables can be summed up as shown in Figure 1.4 using a heat map and Figure 1.5 using the pair plot:



**Figure 1.4 Heat Map**



**Figure 1.5 mpg Pair plot**

**4.2**

**4.3**

**5 Results**

**6 Discussion**

**7 Conclusion**

**Appendix A team member contributions**

**Priyanka:** Data fetching, data analysis scripting, mid-way report.

**Hsiaoan:** Interpretation of data plots and data attribute relationships. Composition of mid-way report.

**Rani:** Interpretation of data plots and data attribute relationships. Composition of mid-way report, slides.