# PSTAT 131 Homework 1

Ezra Torio

2022-10-01

## Machine Learning Main Ideas

### Question 1

Supervised learning is used to accurately predict future responses given a labeled training data set. Unsupervised learning is used to gain insight on unlabeled data sets where the response variables are unknown. Supervised learning requires more human intervention than unsupervised learning. With unsupervised learning, you don't have the answer key of knowing the response variable. Examples of supervised learning include linear regression, random forests, and k-nearest neighbors. Examples of unsupervised learning include PCA, k-means clustering, and hierarchical clustering.

### Question 2

The main difference between regression and classification is that regression relates to quantitative responses and classification relates to categorical or qualitative responses. When testing the accuracy of these models, regression uses mean squared error while classification uses error rate.

### Question 3

Two Common Regression Metrics: Mean Squared Error, Mean Absolute Error

Two Common Classification Metrics: Accuracy, Precision

### Question 4

Descriptive Models: Choose a model that can effectively emphasize a trend in the data.

Predictive Models: Choose a model that can effectively combine features to predict the response variable.

Inferential Models: Choose a model to determine which features are significant. Models can state relationships between predictors and the outcome.

### Question 5

In Mechanistic Modelling, we make assumptions about the outcome in order to make our prediction. For Mechanistic Modelling, we assume forms for f. In Empirical Modelling, we make no assumptions about the outcome and instead learn by experimenting. For Empirical Modelling, we make no assumptions about f.

I believe that Empirically Driven Models are easier to understand. In Empirical Models, you don't need to bring in any external assumptions and instead just learn straight from the data that you are given.

A flexible model like the Empirical Model is likely to have higher variance but lower bias. On the other hand, a simpler model like the Mechanistic Model is likely to have lower variance but higher bias.

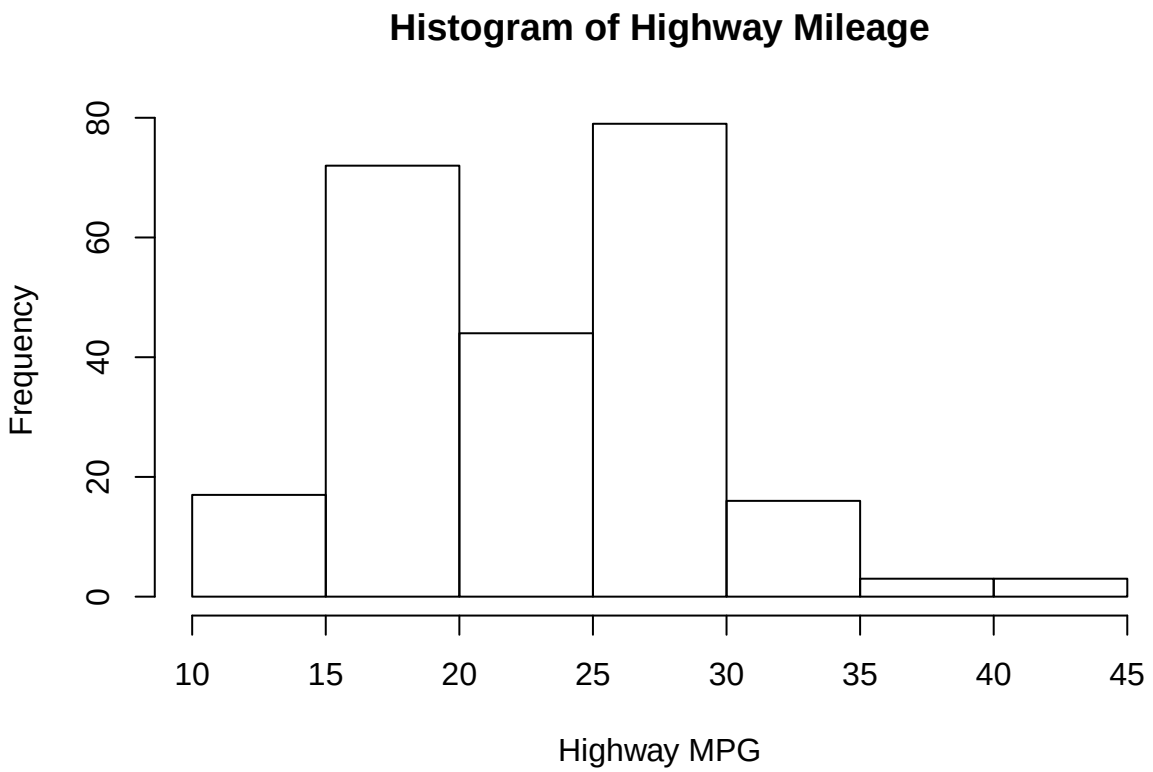
## Question 6

The first question is inferential. This question is prompting us to look at past data in order to draw a relationship about predictors and the outcome. The second question is predictive. This question is prompting us to make a prediction about something that has not yet happened.

# Exploratory Data Analysis

## Excercise 1

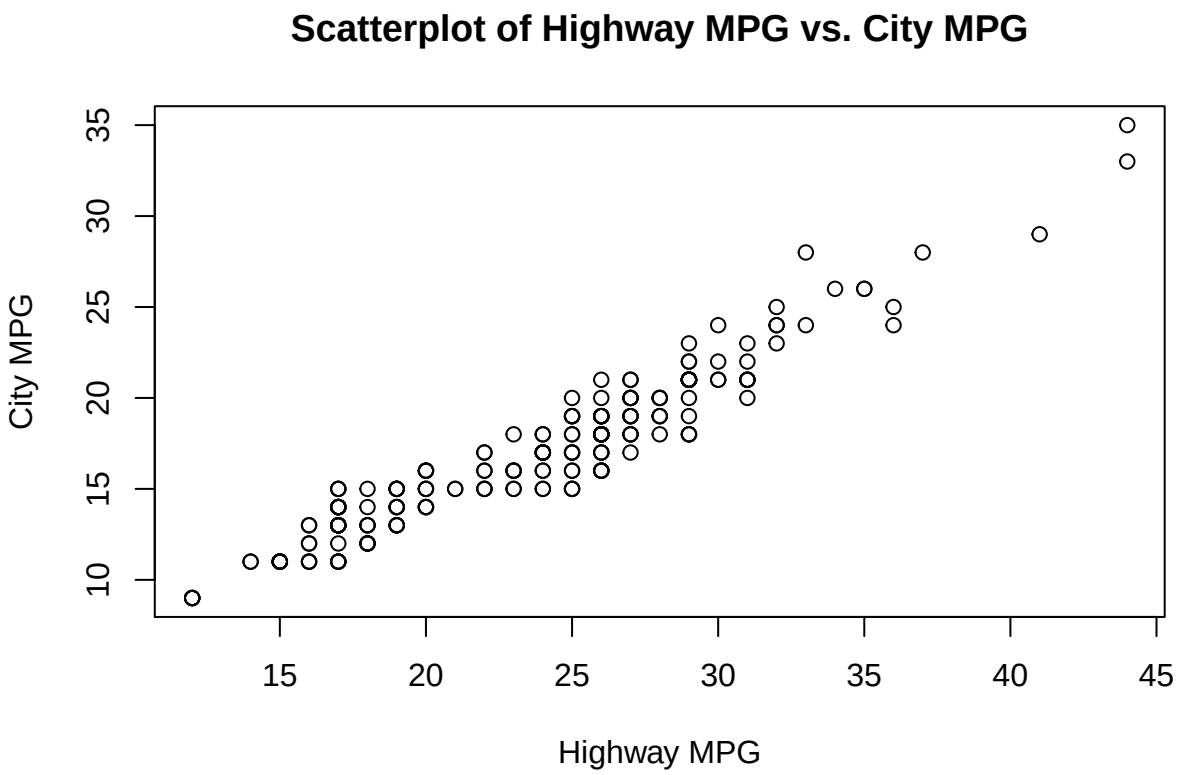
```
data <- mpg
hist(mpg$hwy, main = "Histogram of Highway Mileage", xlab = "Highway MPG")
```




In this histogram, we can see that most cars get between 15 and 30 miles per gallon on the highway.

### Exercise 2

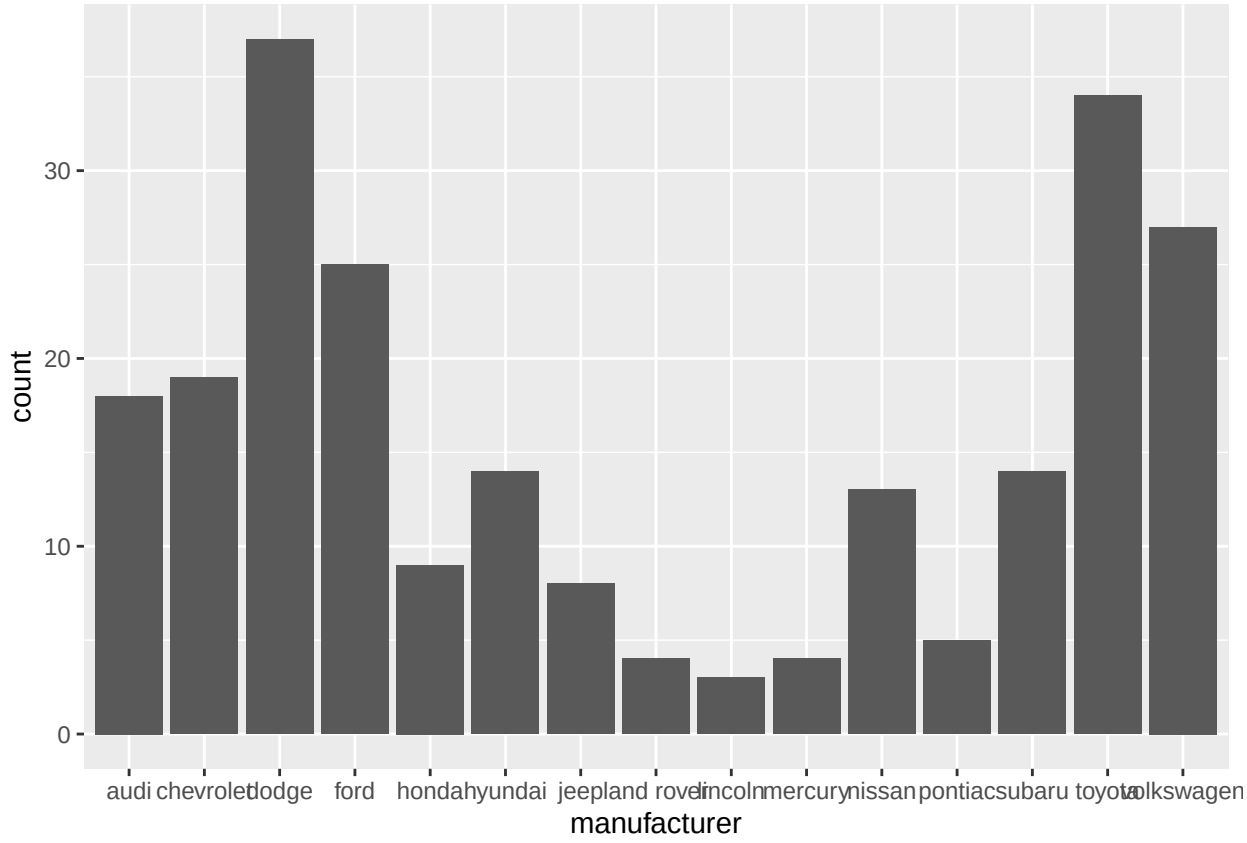
```
data <- mpg
plot(mpg$hwy, mpg$cty, main = "Scatterplot of Highway MPG vs. City MPG",
     xlab = "Highway MPG", ylab = "City MPG")
```



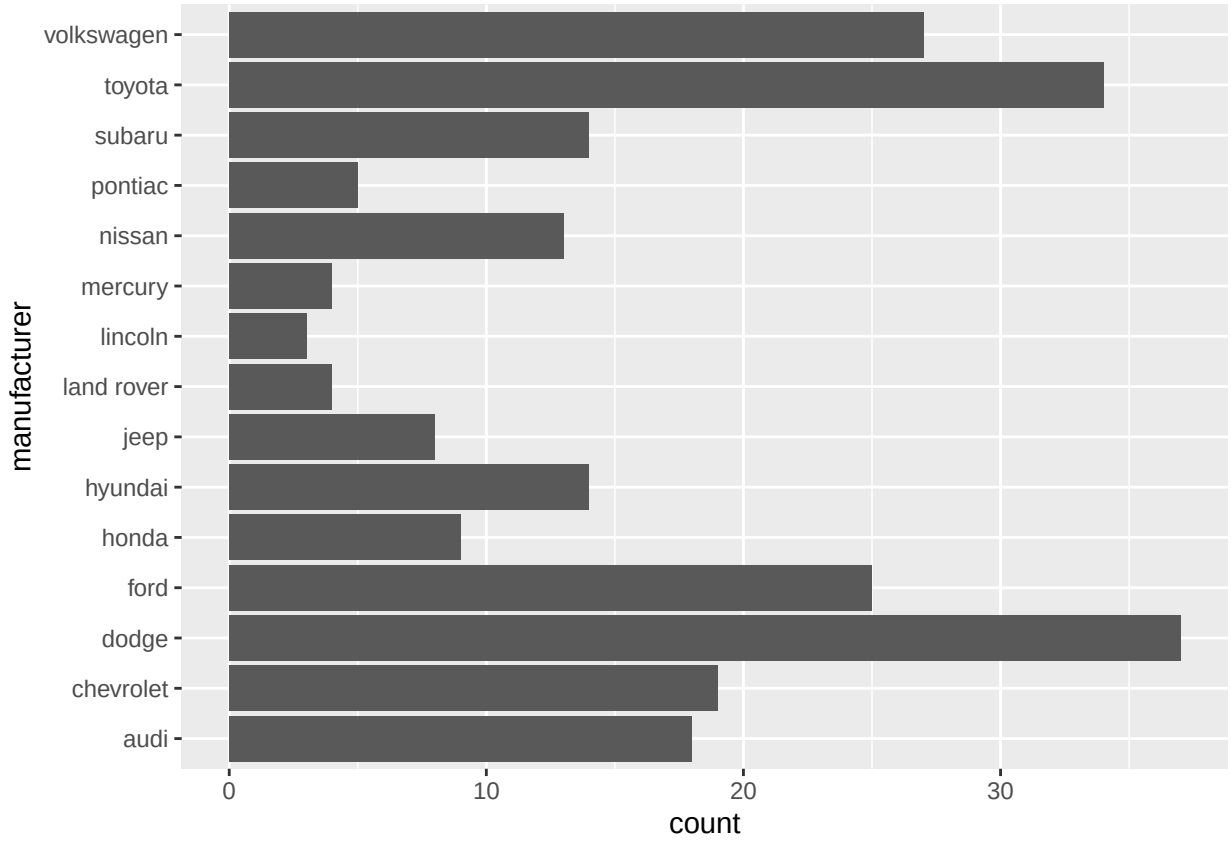

From this scatter plot, we can conclude that as highway miles per gallon increases, city miles per gallon increases and vice versa.

### Exercise 3

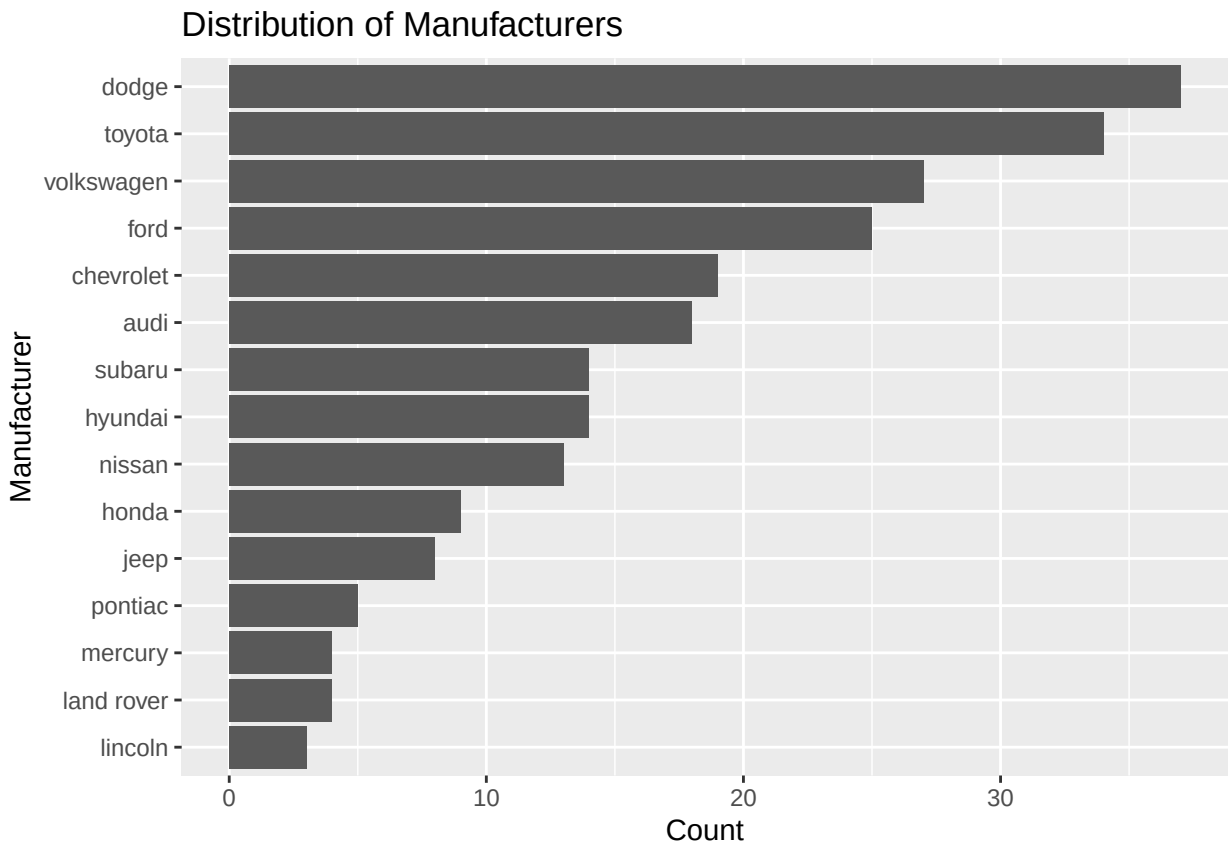
```
mpg %>% group_by(manufacturer) %>%
  summarise(count = n()) %>%
  ggplot(aes(x = manufacturer, y = count)) +
  geom_bar(stat = "identity")
```



```r
mpg %>% group_by(manufacturer) %>%
  summarise(count = n()) %>%
  ggplot(aes(x = manufacturer, y = count)) +
  geom_bar(stat = "identity") + coord_flip()
```



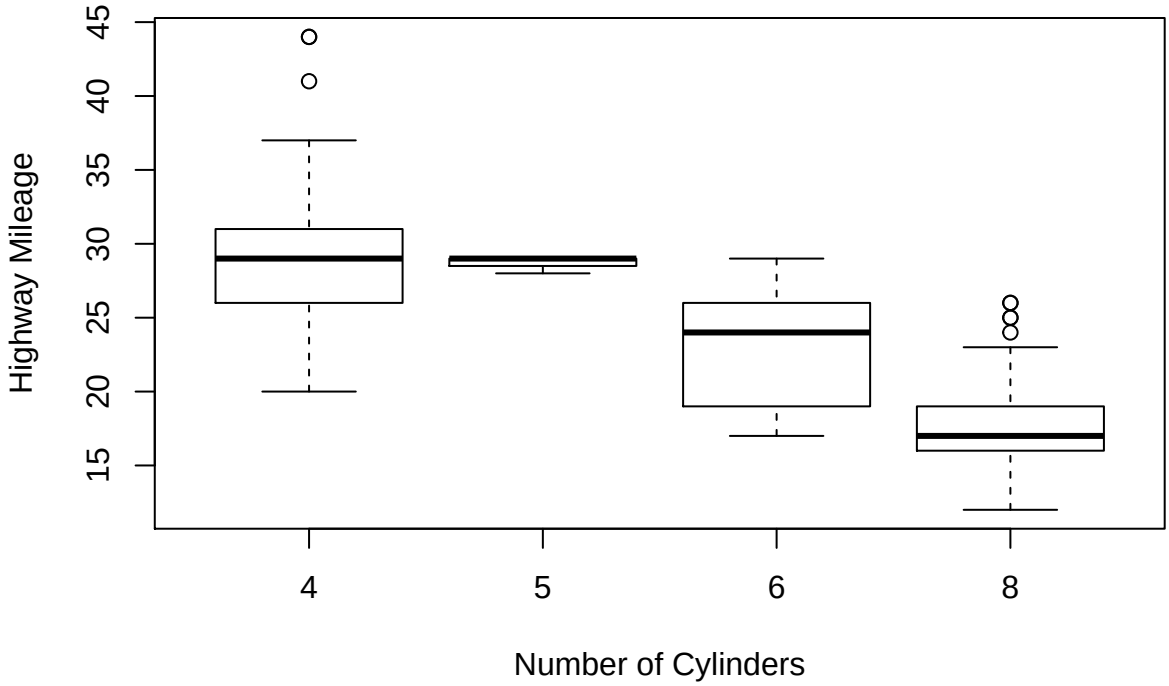
```r
mpg %>% group_by(manufacturer) %>%
  summarise(count = n()) %>%
  ggplot(aes(x = reorder(manufacturer, +count), y = count)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Distribution of Manufacturers", y = "Count", x = "Manufacturer")
```




Dodge makes the most amount of cars while Lincoln makes the least amount of cars.

### Exercise 4

```
boxplot(mpg$hwy~mpg$cyl, xlab = "Number of Cylinders", ylab = "Highway Mileage")
```



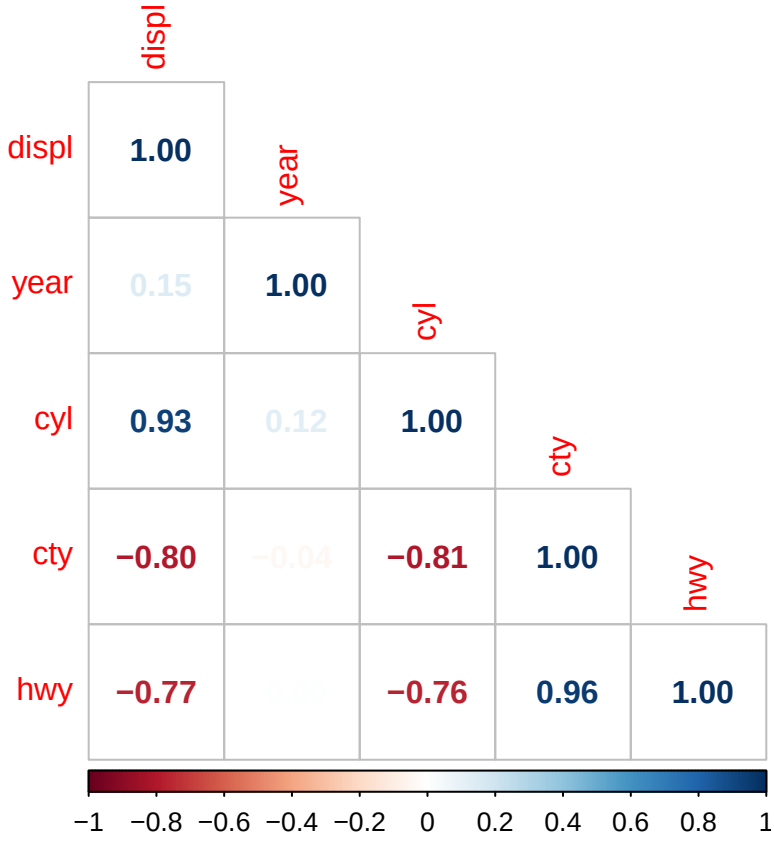

From the box plot, we can conclude that having less cylinders is correlated with better highway mileage.

### Exercise 5

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
mpg_filtered <- mpg[,c(-1,-2,-6,-7,-10,-11)]
M = cor(mpg_filtered)
corrplot(M, method = "number", type = "lower")
```




Highway mileage and city mileage are positively correlated with each other. This is the relationship that I would expect given that both metrics are measures of mileage. Number of cylinders and engine displacement are positively correlated with each other. This is the relationship that I would expect given that engine displacement essentially measures the swept volume of the pistons in a cylinder. Therefore, it makes sense that more cylinders would lead to a higher volume.

Both city and highway mileages are negatively correlated with both number of cylinders and engine displacement. The general rule with automobiles is that engines with less cylinders offer better fuel economy. Therefore, I would expect number of cylinders to have a negative relationship with mileage. Furthermore, we have already touched on the idea that more cylinders relates to more engine displacement, thus by extension, I would expect engine displacement to have a negative relationship with mileage.

Year ultimately has no relationships with any other variable which is to be expected because year has very minimal effect on physical attributes of automobiles, especially given our short range of years.