PSTAT 131 - Homework Assignment 1

Akshat Ataliwala (7924145)

April 01, 2022

Machine Learning Main Ideas

1 - Define supervised and unsupervised learning. What are the difference(s) between them?

Supervised learning and unsupervised learning are both categories of machine learning problems. In supervised learning, our dataset contains the true response values we are trying to predict, meaning that we are able to see our prediction accuracy on a testing set because we know the actual values. On the other hand, datasets in unsupervised learning do not contain the true response values. This difference makes either problem quite different, as supervised learning uses the features of a specific observation to find the corresponding response, and unsupervised learning clusters like observations together and then assigns them a value based on a chosen distance measure.

2 - Explain the difference between a regression model and a classification model, specifically in the context of machine learning.

In the context of machine learning, the difference between regression classification comes from the type of response value we are trying to predict. Both are supervised problems, but regression predicts numerical/continuous/quantitative values, and classification is predicts categorical/qualitative values.

3 - Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.

Regression ML: Training MSE and Test MSE

Classification ML: Training Error Rate and Test Error Rate

4 - As discussed, statistical models can be used for different purposes. These purposes can generally be classified into the following three categories. Provide a brief description of each.

Descriptive Models - Models that analyze our current data for patterns and trends

Inferential Models - determine the accuracy of our predictions/estimations

Predictive Models - predicting/forecasting future response values given our data

5 - Predictive models are frequently used in machine learning, and they can usually be described as either mechanistic or empirically-driven. Answer the following questions.

Define mechanistic. Define empirically-driven. How do these model types differ? How are they similar?

Mechanistic is when we assume the form of a model, which means that we only have to estimate parameters instead of the function f. Empirically-driven is when we do not assume the form of a model, and therefore have to estimate the function f. These model types differ in the sense that the mechanistic approach has the possibility that the form used to estimate f is not actually right, resulting in very poor model fit. On the other hand, an empirically-driven approach avoids this issue, but this typically requires more data to try and accurately estimate f.

In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice.

Mechanistic is generally easier to understand, because if we assume the form of a model, it makes it gives us more clarity on what types of relationships we might be dealing with right from the jump.

Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models.

Mechanistic models tend to have higher bias, because we are assuming the form of our model, and depending on whether or not we are right our performance will be excellent or extremely poor. More flexible models like those that are empirically-driven tend to have low bias as we do not assume any form and are capable of molding to the data, but higher variance across the board as they are less strict.

6 - A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions. Classify each question as either predictive or inferential. Explain your reasoning for each.

Given a voter's profile/data, how likely is it that they will vote in favor of the candidate?

This is a prediction question because we are using predictor variables (voter data) to determine a likelihood of a vote, which could take the form of a 1-10 score, the label of the most likely vote, etc.

How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?

This is an inference question because we are essentially testing whether or not personal contact with the candidate is actually associated with the vote that was given.

Exploratory Data Analysis

This section will ask you to complete several exercises. For this homework assignment, we'll be working with the mpg data set that is loaded when you load the tidyverse. Make sure you load the tidyverse and any other packages you need.

Exploratory data analysis (or EDA) is not based on a specific set of rules or formulas. It is more of a state of curiosity about data. It's an iterative process of:

- generating questions about data
- visualize and transform your data as necessary to get answers
- use what you learned to generate more questions

A couple questions are always useful when you start out. These are "what variation occurs within the variables," and "what covariation occurs between the variables."

You should use the tidyverse and ggplot2 for these exercises.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model      displ  year   cyl trans drv     cty   hwy fl    class
##    <chr>        <chr>      <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
##  1 audi         a4           1.8  1999     4 auto~ f        18    29 p     comp~
##  2 audi         a4           1.8  1999     4 manu~ f        21    29 p     comp~
##  3 audi         a4           2    2008     4 manu~ f        20    31 p     comp~
##  4 audi         a4           2    2008     4 auto~ f        21    30 p     comp~
##  5 audi         a4           2.8  1999     6 auto~ f        16    26 p     comp~
##  6 audi         a4           2.8  1999     6 manu~ f        18    26 p     comp~
##  7 audi         a4           3.1  2008     6 auto~ f        18    27 p     comp~
##  8 audi         a4 quattro   1.8  1999     4 manu~ 4        18    26 p     comp~
##  9 audi         a4 quattro   1.8  1999     4 auto~ 4        16    25 p     comp~
## 10 audi         a4 quattro   2    2008     4 manu~ 4        20    28 p     comp~
## # ... with 224 more rows
```

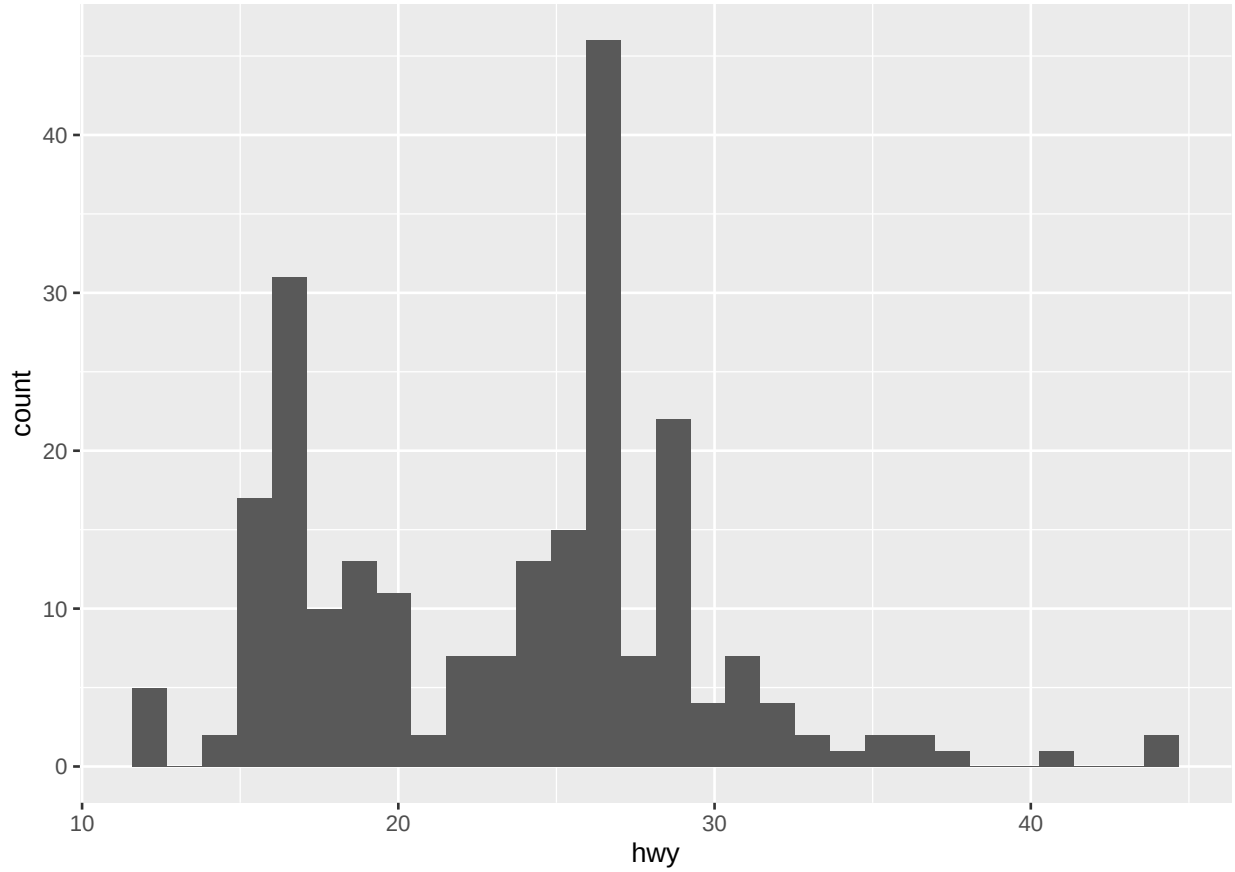
```
head(mpg, 5)
```

```
## # A tibble: 5 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
```

1. We are interested in highway miles per gallon, or the hwy variable. Create a histogram of this variable. Describe what you see/learn.

```
hwy_mpg_hist <-ggplot(mpg, aes(x=hwy)) + geom_histogram()
hwy_mpg_hist
```

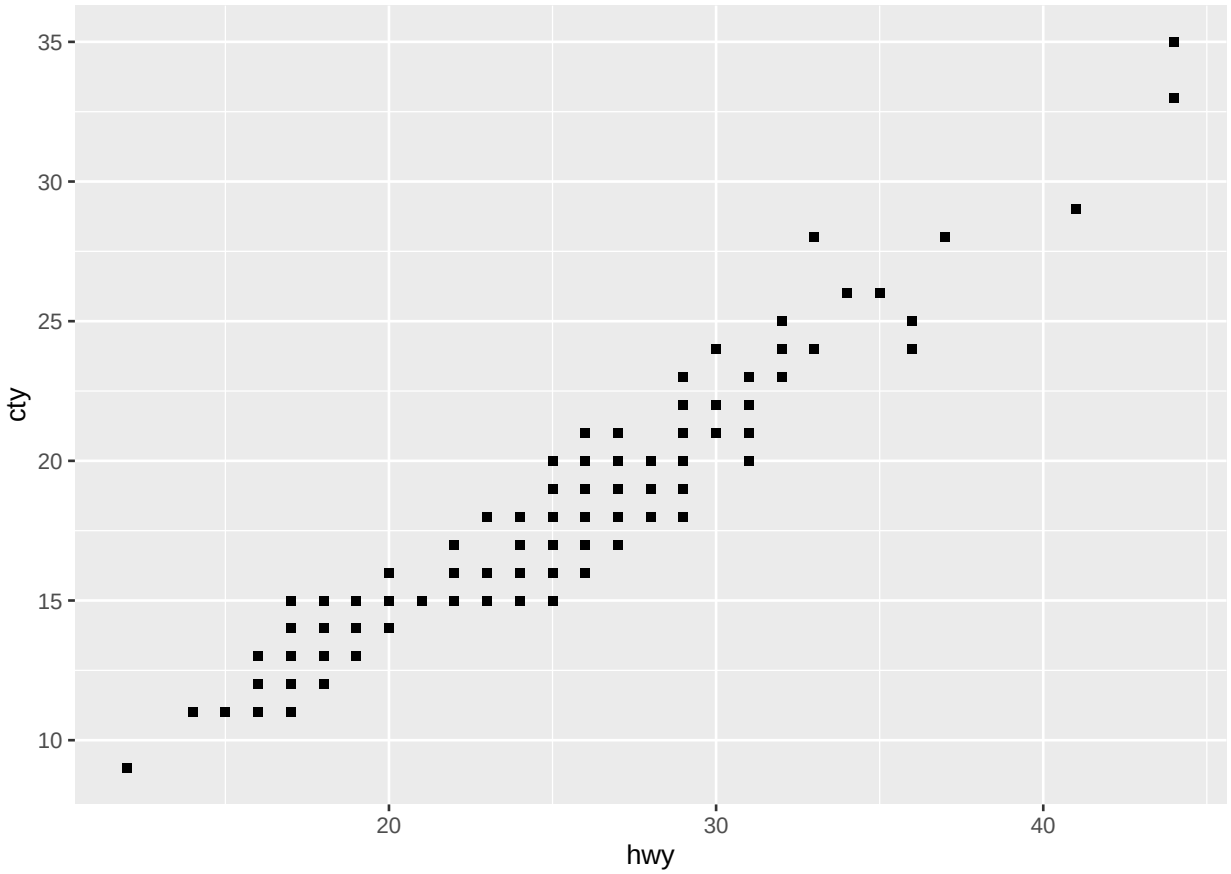
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Based on the histogram, we can see that hwy is bimodal with 2 major peaks at 13 and 27. This is probably because there are usually 2 types of cars, with powerful gas guzzlers that get low mileage and more economical vehicles. There are very few outliers, which indicate cars that could be either extremely inefficient or very economical.

2. Create a scatterplot. Put hwy on the x-axis and cty on the y-axis. Describe what you notice. Is there a relationship between hwy and cty? What does this mean?

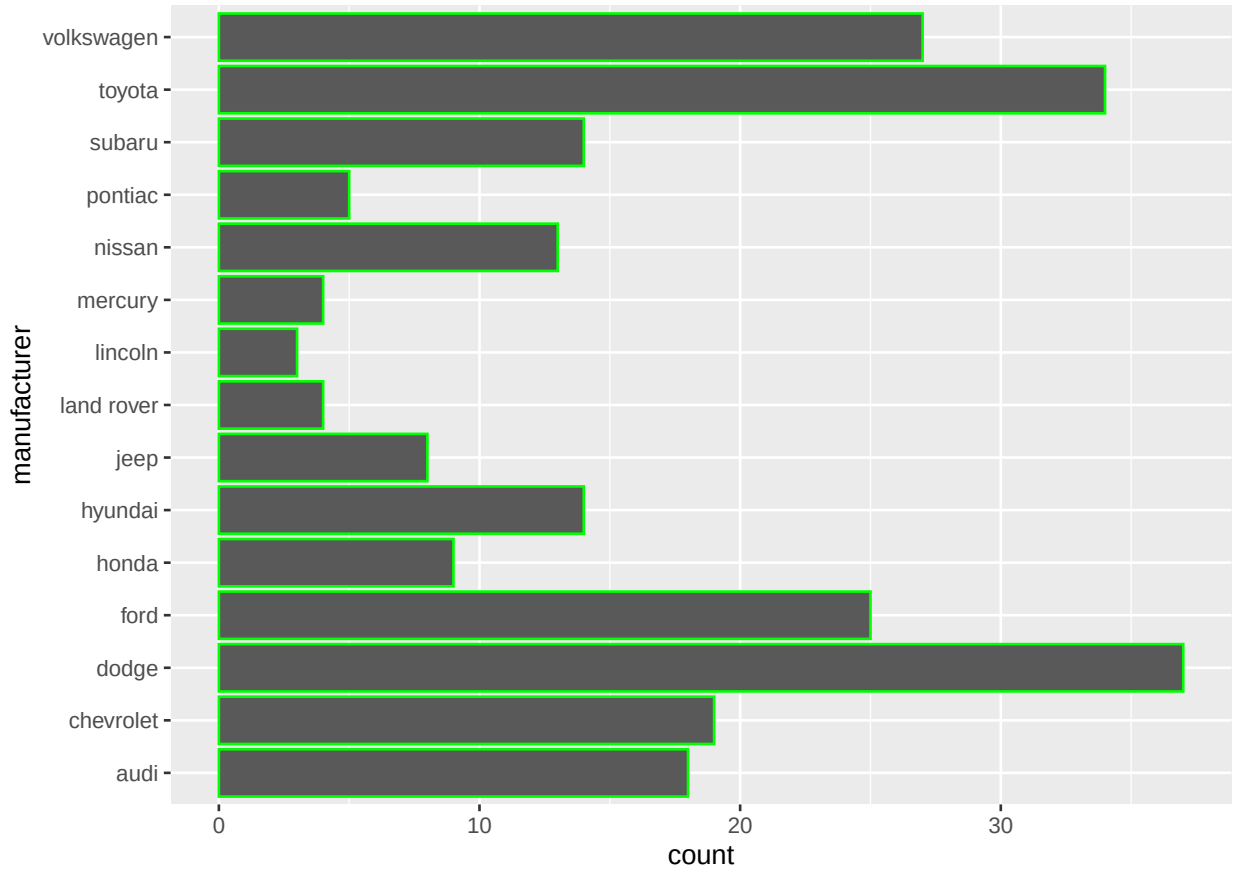
```
scatter <- ggplot(mpg, aes(x = hwy, y = cty)) + geom_point(shape=15)
scatter
```



There seems to be a strong positive linear relationship between hwy and cty. This means that as cars tend to have higher hwy mileage, they also tend to have higher cty mileage.

3. Make a bar plot of manufacturer. Flip it so that the manufacturers are on the y-axis. Order the bars by height. Which manufacturer produced the most cars? Which produced the least?

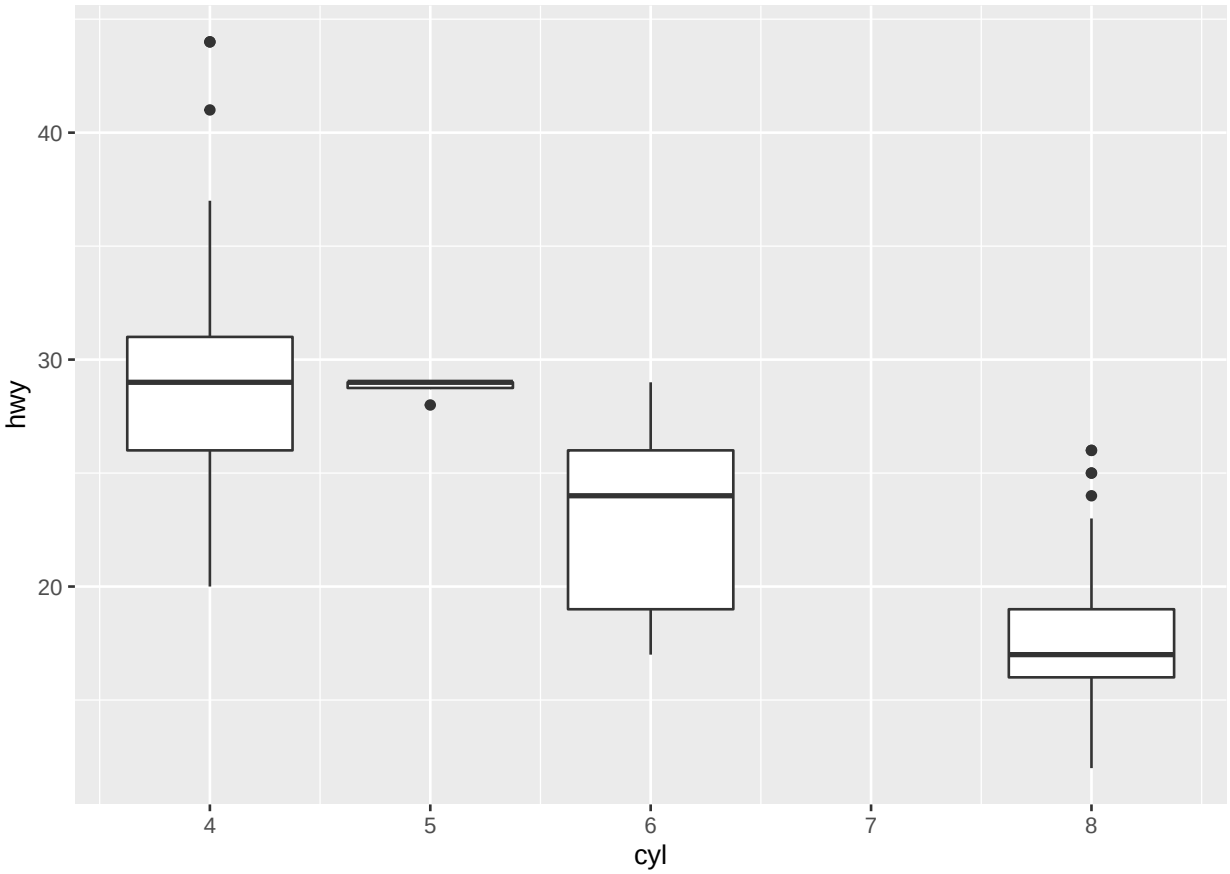
```
manufacturers <- ggplot(mpg, aes(x = manufacturer)) + geom_bar(color = "green", stat = "count") + coord_
manufacturers
```



Dodge produced the most cars, and Lincoln produced the least.

4. Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?

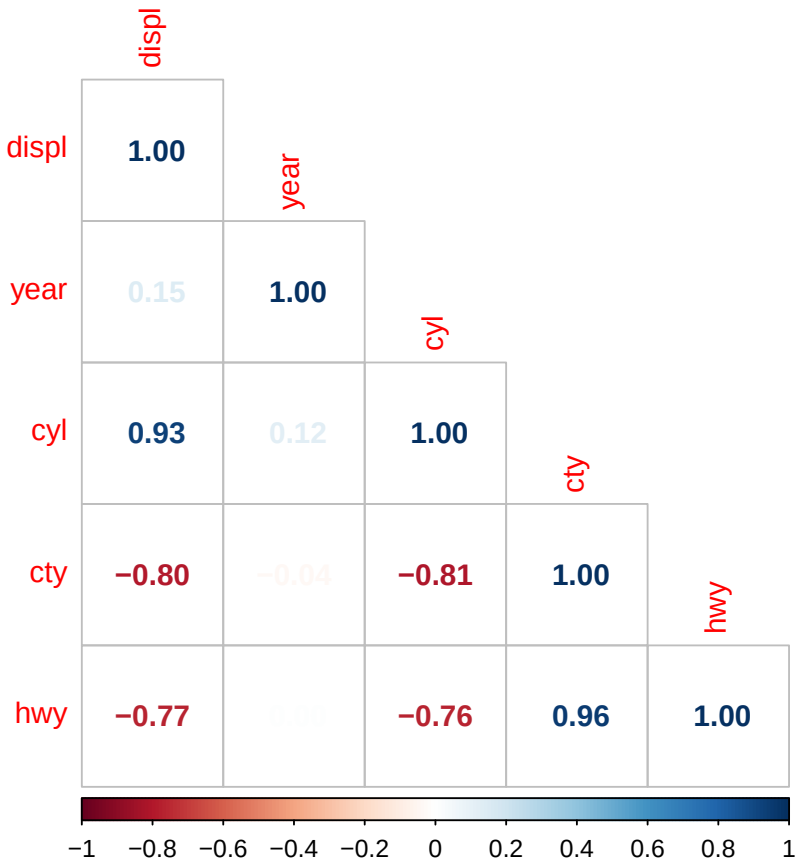
```
box_plot <- ggplot(mpg, aes(x = cyl, y = hwy)) + geom_boxplot(aes(group = cyl))
box_plot
```



The most common cylinder counts are 4, 6, 8, with 5 being an extremely rare occurrence. 4-cylinder cars tend to have a median hwy of just under 30, 6-cylinder cars tend to have a median hwy of just under 25, and 8-cylinder cars have a median of around 17. This generally means that as the number of cylinders in a car increases, the worst gas mileage it tends to have. 4 and 6 cylinder cars are also more common, which is why their plots are much larger than the 8 cylinder plot that is narrower in range.

5. Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset. Which variables are positively or negatively correlated with which others? Do these relationships make sense to you? Are there any that surprise you?

```
correlations <- cor(mpg[, c("displ", "year", "cyl", "cty", "hwy")])
corrplot(correlations, method = "number", type = "lower")
```



The variables that have the strongest positive correlations are hwy and cty with 0.96, cyl and displ with 0.93. On the other hand, the strongest negative correlations are cty and displ with -0.80 and cty and cyl with -0.81. In general there is a very weak relationship between year and other predictor variables. Most of these relationships make sense to me, as larger engines (displ) would generally be in more powerful cars, which is why it is strongly correlated to cyl, for example. This would also explain the very strong negative correlation between cty mileage and cyl + displ, as larger cars are usually less efficient.