PSTAT 131 - Homework 1

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

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

```r
library(ggplot2)
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tune         1.0.0
## v infer        1.0.3     v workflows    1.1.0
## v modeldata    1.0.1     v workflowsets 1.0.0
## v parsnip      1.0.1     v yardstick    1.1.0
## v recipes      1.0.1
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Use suppressPackageStartupMessages() to eliminate package startup messages
```

Questions

1.

Supervised learning is a type of statistical machine learning in which a portion of the data set is used to fit the models (usually based on a number of different predictors) and then the accuracy of the model is determined, and testing data is used to see how well this model can predict other sample data. Unsupervised learning is using statistical machine learning in which there is no predictor variable, so many types of models cannot be done to test how accurate it may be. Overall, both types of learning are meant to provide insight to how a specific model can provide information about a population of which the sample is taken from, but supervised learning focuses on seeing how well the model fits other data and unsupervised learning does not work to try and determine its accuracy.

2.

A regression model is a supervised learning model in which the outcome variable is numeric and can thus be fit onto a sort of number line. A classification model is a supervised learning model in which the outcome variable is qualitative and thus fits into a certain category. A classification model can be more concrete when testing a model because it either falls into the correct category or not, where a regression model can show how accurate the model was for testing sets.

3.

Two commonly used metrics for regression ML problems would be MSE and R-squared. These metrics are used to measure how well the data fits the model that has been created from the data, i.e. how well one variable can predict the response. Two commonly used metrics for classification ML problems would be the Bayes error rate and K nearest neighbors. The Bayes error rate is based on conditional probability of one variable that can thus predict the category of which the response will fall under based on the previous observations from the sample. K nearest neighbors refers to the more realistic approach of estimating the response through identifying K points in the training data that are closest to x0 and then estimating the conditional probability for class j as the preaction of points in N0 whose response values equal j.

4.

a. Descriptive models are models generally used to show a trend in data to visualize these patterns. This tends to include data cleaning and then presenting the cleaned information with some functions to visualize the data for clearer understanding of their patterns and significance.
b. Inferential models are models that generally work to help test theories and provide insight to a relationship between variables. They work to find which features of the data are significant and thus provide information about the strength of patterns which we might see.
c. Predictive models are models that want to predict Y with minimum reducible error, meaning finding models that are as accurate as possible and cannot hopefully be accurate beyond random error.

5.

a. Mechanistic refers to a model type which assumes a parametric form of the model f, meaning that there are set standards of which this function will follow to predict or assess the data set. Empirically-driven models are models which have no strict assumptions about f and are much more flexible than parametric models, but therefore also require a greater number of observations. These models are similar in that they both aim to assess the given data in some way and fit the data into some created function. They differ, however, in their goals for what the data will tell us about inferences we have and how much flexibility we can have in terms of our model.
b. In general, a parametric model is easier to understand because it creates a much more simple way of thinking about the data: it follows a pattern based on a few numbers and equations. Empirically-driven models work more to our own brains and are more realistic in their assumptions about the data, but it then makes it much more difficult to understand in terms of mathematics.

c. The bias-variance tradeoff is the idea that you want to balance MSE, bias, and variance of a model when finding a way to fit the data. If you have high variance, that means that you will likely have low bias, and vice versa. A good model will aim to balance these two or be explicit in the tendency of the model to have more variability or bias. In mechanistic models, it is fairly easy to calculate these with different tests and see if tweaks to the model can bring the balance closer together. In empirically-driven statistics with less strict rules for f, then you will not be able to calculate bias and variance as well but it is still important to understand that bias and variance exist and balance between them is still important to avoid overfitting the model.

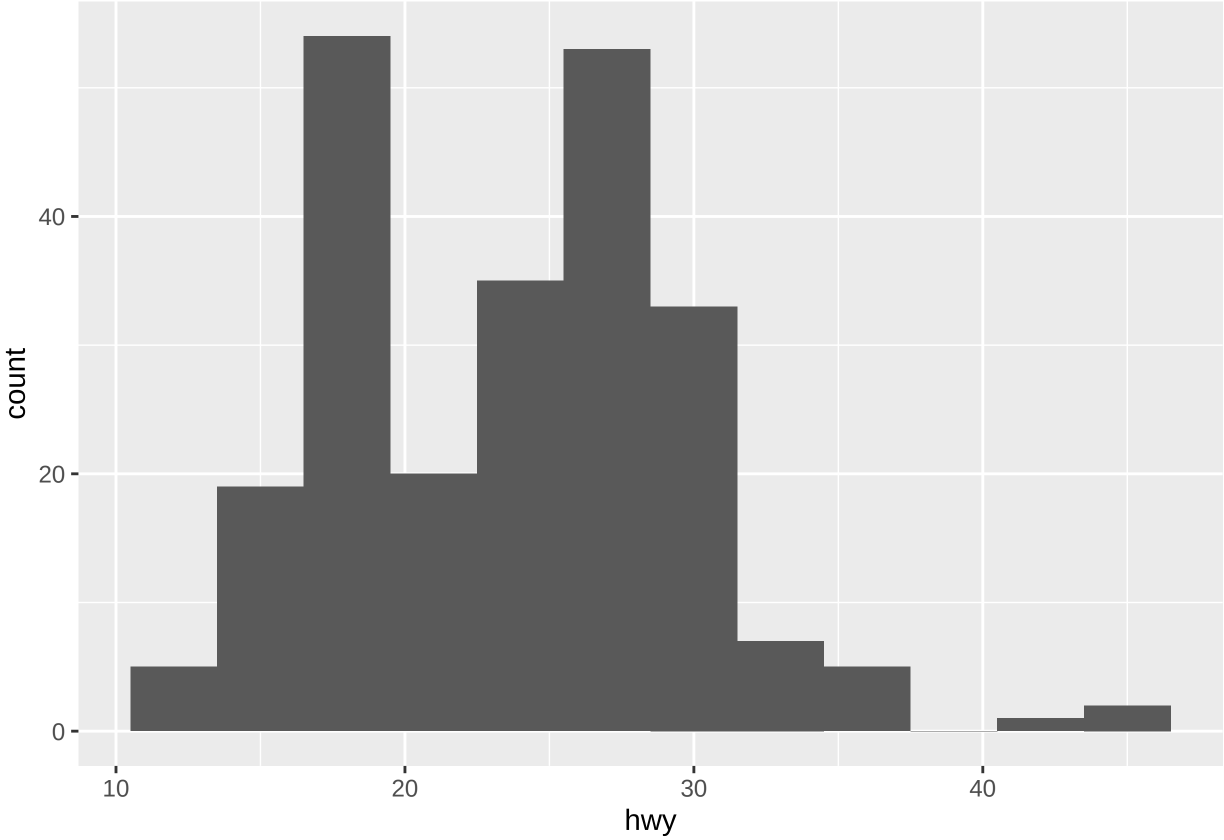
6.

To see how likely a constituent would be to vote in favor of a candidate, we would consider that to be predictive, as we aren't testing a theory of the model but instead predicting the behavior of a new observation based on the data which we currently have. To see how a personal relationship would impact the likelihood of a voter voting a certain way would be inferential, as we are testing the significance of a feature of the observation, i.e. is their relationship with the politician significant in their voting behavior.

Excercises

Exercise 1: Creating a histogram of hwy

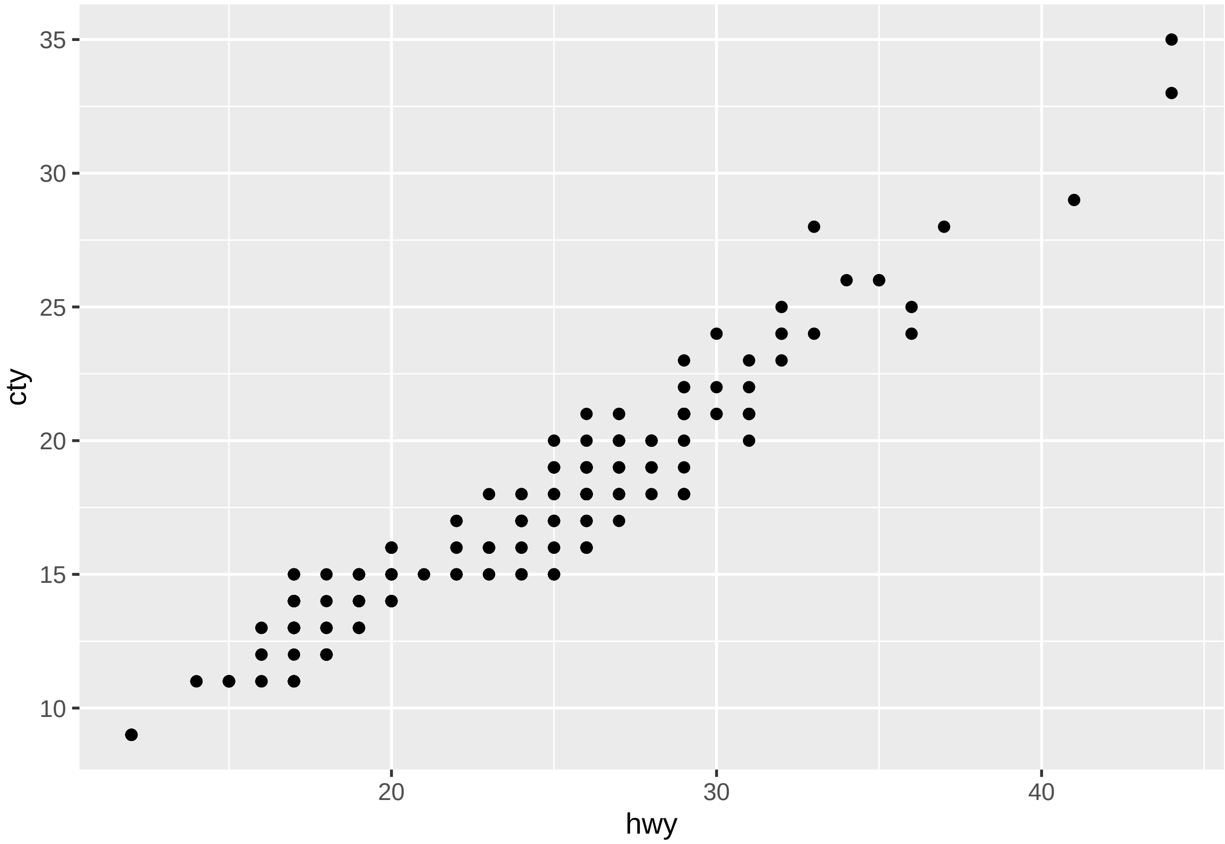
```
?mpg
ggplot(mpg, aes(x=hwy)) + geom_histogram(binwidth=3)
```



Based on the created histogram, we can see that there is a right-skew to the data and it is bimodal. Because we are only counting the counts of different highway miles per gallon. We can assume that the two counts with high miles per gallon would likely be for newer, smaller, more fuel-efficient cars and the bimodal peaks would be a distinction between compact cars and SUVs, with compact cars being more fuel-efficient.

Exercise 2: Creating a scatterplot of hwy and cty

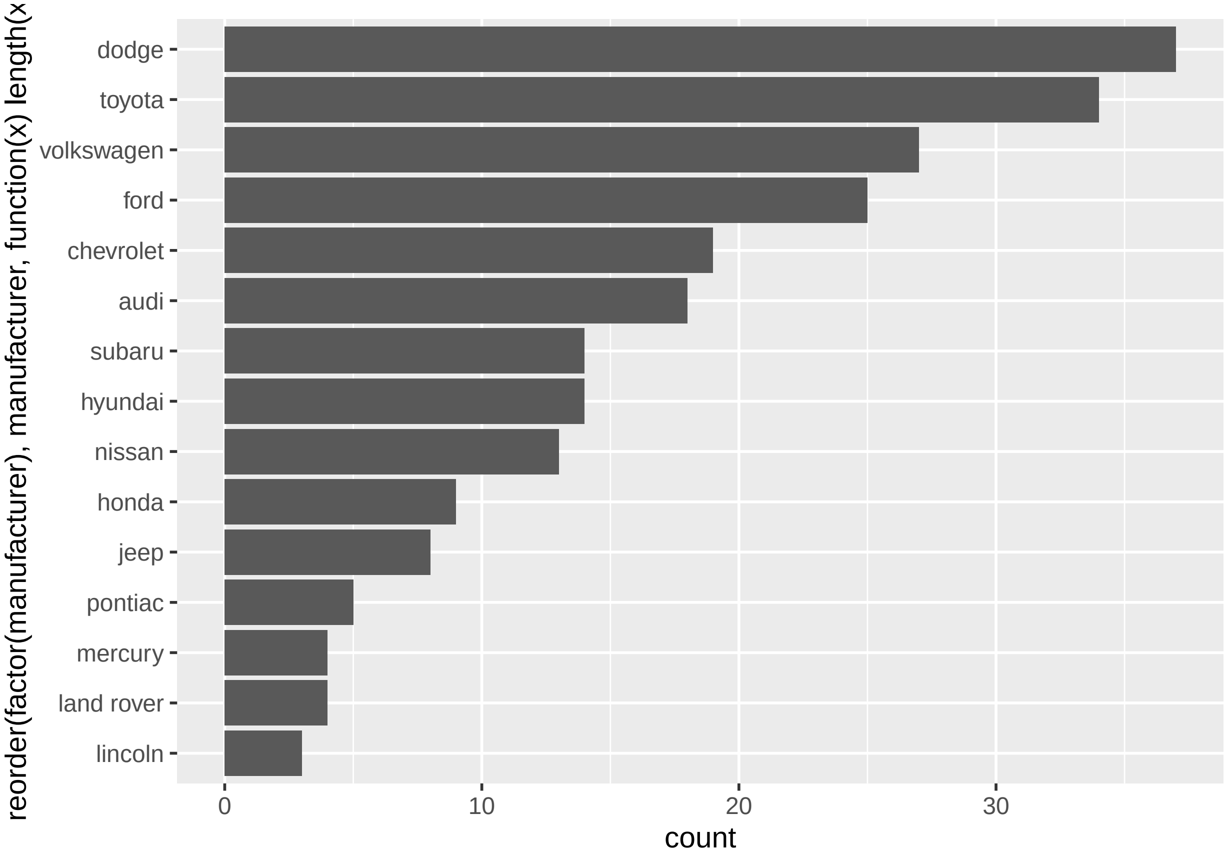
```
ggplot(data = mpg, aes(x = hwy, y = cty)) + geom_point()
```



The relationship between hwy and cty shows a fairly strong positive linear relationship. This isn't very shocking as they both are measures of miles per gallon, just changing on where they are measured and thus the fuel efficiency. In general the highway mpg is higher than the corresponding city mpg, but this makes sense as city driving takes more acceleration between stops and generally has a slower speed limit.

Exercise 3: Constructing a barplot of manufacturer

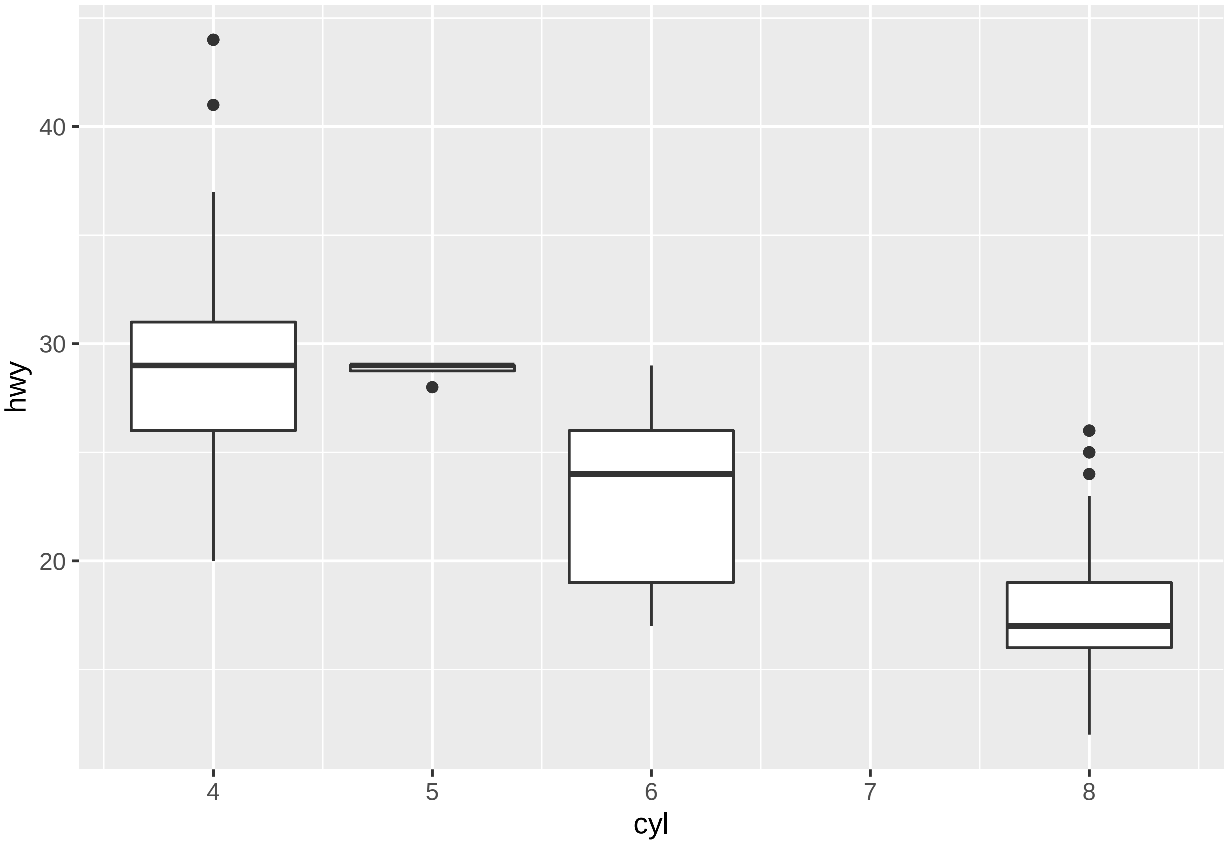
```
ggplot(data = mpg, aes(x = reorder(factor(manufacturer), manufacturer, function(x) length(x)))) + geom_
```



From the sorted bar plot, we notice that the manufacturer who produced the most cars in the data set is Dodge, and the manufacturer with the least cars is Lincoln.

Exercise 4: Creating a boxplot of hwy grouped by cyl

```
ggplot(data = mpg, aes(y = hwy, x = cyl, group = cyl)) + geom_boxplot()
```



From these grouped boxplots, I can see that a majority of the cylinders are divisible by 2 and that the lower the number of cylinders, the higher the highway miles per gallon. This makes sense in the fact that bigger cars would require more cylinders and that they would also have worse mpg because of their size. Also, there are a few outliers in the data, with most showing higher fuel efficiency than the others in their respective cylinder groups.

Exercise 5: Using the corrplot package

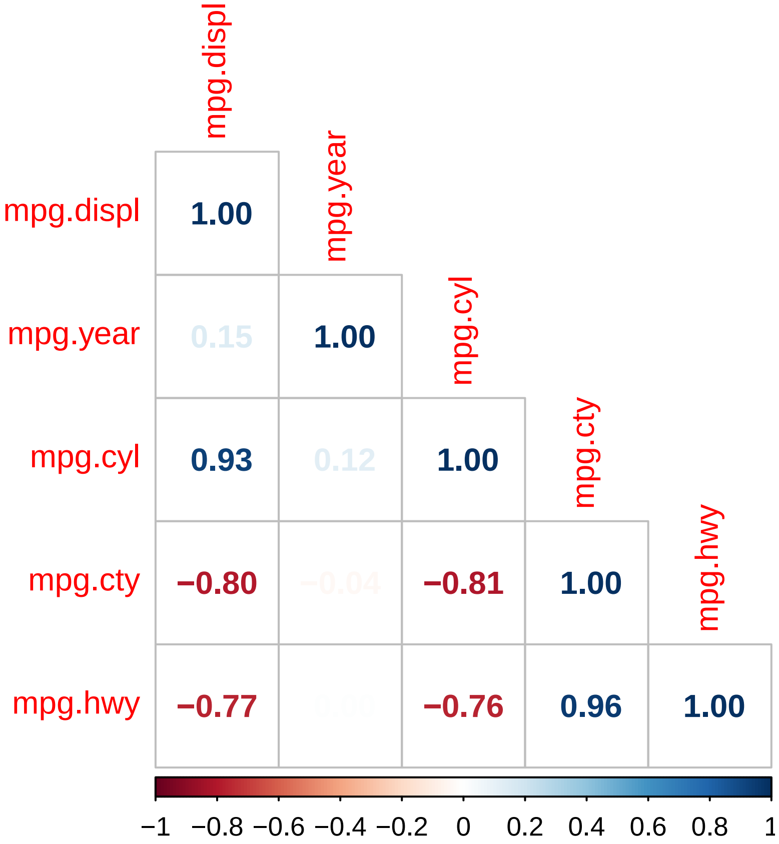
```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
mpgNumeric <- data.frame(mpg$displ, mpg$year, mpg$cyl, mpg$cty, mpg$hwy)
mpgNumeric
```

```
##     mpg.displ mpg.year mpg.cyl mpg.cty mpg.hwy
## 1         1.8     1999       4      18      29
## 2         1.8     1999       4      21      29
## 3         2.0     2008       4      20      31
## 4         2.0     2008       4      21      30
## 5         2.8     1999       6      16      26
## 6         2.8     1999       6      18      26
## 7         3.1     2008       6      18      27
## 8         1.8     1999       4      18      26
## 9         1.8     1999       4      16      25
## 10        2.0     2008       4      20      28
## 11        2.0     2008       4      19      27
## 12        2.8     1999       6      15      25
## 13        2.8     1999       6      17      25
## 14        3.1     2008       6      17      25
## 15        3.1     2008       6      15      25
## 16        2.8     1999       6      15      24
## 17        3.1     2008       6      17      25
## 18        4.2     2008       8      16      23
## 19        5.3     2008       8      14      20
## 20        5.3     2008       8      11      15
## 21        5.3     2008       8      14      20
## 22        5.7     1999       8      13      17
## 23        6.0     2008       8      12      17
## 24        5.7     1999       8      16      26
## 25        5.7     1999       8      15      23
## 26        6.2     2008       8      16      26
## 27        6.2     2008       8      15      25
## 28        7.0     2008       8      15      24
## 29        5.3     2008       8      14      19
## 30        5.3     2008       8      11      14
## 31        5.7     1999       8      11      15
## 32        6.5     1999       8      14      17
## 33        2.4     1999       4      19      27
## 34        2.4     2008       4      22      30
## 35        3.1     1999       6      18      26
## 36        3.5     2008       6      18      29
## 37        3.6     2008       6      17      26
## 38        2.4     1999       4      18      24
## 39        3.0     1999       6      17      24
## 40        3.3     1999       6      16      22
## 41        3.3     1999       6      16      22
## 42        3.3     2008       6      17      24
## 43        3.3     2008       6      17      24
## 44        3.3     2008       6      11      17
## 45        3.8     1999       6      15      22
## 46        3.8     1999       6      15      21
## 47        3.8     2008       6      16      23
## 48        4.0     2008       6      16      23
## 49        3.7     2008       6      15      19
## 50        3.7     2008       6      14      18
## 51        3.9     1999       6      13      17
## 52        3.9     1999       6      14      17
## 53        4.7     2008       8      14      19
## 54        4.7     2008       8      14      19
## 55        4.7     2008       8       9      12
## 56        5.2     1999       8      11      17
## 57        5.2     1999       8      11      15
## 58        3.9     1999       6      13      17
## 59        4.7     2008       8      13      17
## 60        4.7     2008       8       9      12
## 61        4.7     2008       8      13      17
## 62        5.2     1999       8      11      16
## 63        5.7     2008       8     13     18
## 64        5.9     1999       8     11     15
## 65        4.7     2008       8     12     16
## 66        4.7     2008       8      9     12
## 67        4.7     2008       8     13     17
## 68        4.7     2008       8     13     17
## 69        4.7     2008       8     12     16
## 70        4.7     2008       8      9     12
## 71        5.2     1999       8     11     15
## 72        5.2     1999       8     11     16
## 73        5.7     2008       8     13     17
## 74        5.9     1999       8     11     15
## 75        4.6     1999       8     11     17
## 76        5.4     1999       8     11     17
## 77        5.4     2008       8     12     18
## 78        4.0     1999       6     14     17
## 79        4.0     1999       6     15     19
## 80        4.0     1999       6     14     17
## 81        4.0     2008       6     13     19
## 82        4.6     2008       8     13     19
## 83        5.0     1999       8     13     17
## 84        4.2     1999       6     14     17
## 85        4.2     1999       6     14     17
## 86        4.6     1999       8     13     16
## 87        4.6     1999       8     13     16
## 88        4.6     2008       8     13     17
## 89        5.4     1999       8     11     15
## 90        5.4     2008       8     13     17
## 91        3.8     1999       6     18     26
## 92        3.8     1999       6     18     25
## 93        4.0     2008       6     17     26
## 94        4.0     2008       6     16     24
## 95        4.6     1999       8     15     21
## 96        4.6     1999       8     15     22
## 97        4.6     2008       8     15     23
## 98        4.6     2008       8     15     22
## 99        5.4     2008       8     14     20
## 100       1.6     1999       4     28     33
## 101       1.6     1999       4     24     32
## 102       1.6     1999       4     25     32
## 103       1.6     1999       4     23     29
## 104       1.6     1999       4     24     32
## 105       1.8     2008       4     26     34
## 106       1.8     2008       4     25     36
## 107       1.8     2008       4     24     36
## 108       2.0     2008       4     21     29
## 109       2.4     1999       4     18     26
## 110       2.4     1999       4     18     27
## 111       2.4     2008       4     21     30
## 112       2.4     2008       4     21     31
## 113       2.5     1999       6     18     26
## 114       2.5     1999       6     18     26
## 115       3.3     2008       6     19     28
## 116       2.0     1999       4     19     26
## 117       2.0     1999       4      19      29
## 118       2.0     2008       4      20      28
## 119       2.0     2008       4      20      27
## 120       2.7     2008       6      17      24
## 121       2.7     2008       6      16      24
## 122       2.7     2008       6      17      24
## 123       3.0     2008       6      17      22
## 124       3.7     2008       6      15      19
## 125       4.0     1999       6      15      20
## 126       4.7     1999       8      14      17
## 127       4.7     2008       8       9      12
## 128       4.7     2008       8      14      19
## 129       5.7     2008       8      13      18
## 130       6.1     2008       8      11      14
## 131       4.0     1999       8      11      15
## 132       4.2     2008       8      12      18
## 133       4.4     2008       8      12      18
## 134       4.6     1999       8      11      15
## 135       5.4     1999       8      11      17
## 136       5.4     1999       8      11      16
## 137       5.4     2008       8      12      18
## 138       4.0     1999       6      14      17
## 139       4.0     2008       6      13      19
## 140       4.6     2008       8      13      19
## 141       5.0     1999       8      13      17
## 142       2.4     1999       4      21      29
## 143       2.4     1999       4      19      27
## 144       2.5     2008       4      23      31
## 145       2.5     2008       4      23      32
## 146       3.5     2008       6      19      27
## 147       3.5     2008       6      19      26
## 148       3.0     1999       6      18      26
## 149       3.0     1999       6      19      25
## 150       3.5     2008       6      19      25
## 151       3.3     1999       6      14      17
## 152       3.3     1999       6      15      17
## 153       4.0     2008       6      14      20
## 154       5.6     2008       8      12      18
## 155       3.1     1999       6      18      26
## 156       3.8     1999       6      16      26
## 157       3.8     1999       6      17      27
## 158       3.8     2008       6      18      28
## 159       5.3     2008       8      16      25
## 160       2.5     1999       4      18      25
## 161       2.5     1999       4      18      24
## 162       2.5     2008       4      20      27
## 163       2.5     2008       4      19      25
## 164       2.5     2008       4      20      26
## 165       2.5     2008       4      18      23
## 166       2.2     1999       4      21      26
## 167       2.2     1999       4      19      26
## 168       2.5     1999       4      19      26
## 169       2.5     1999       4      19      26
## 170       2.5     2008       4      20      25
## 171       2.5     2008       4      20      27
## 172       2.5     2008       4      19      25
## 173       2.5     2008       4      20      27
## 174       2.7     1999       4      15      20
## 175       2.7     1999       4      16      20
## 176       3.4     1999       6      15      19
## 177       3.4     1999       6      15      17
## 178       4.0     2008       6      16      20
## 179       4.7     2008       8      14      17
## 180       2.2     1999       4      21      29
## 181       2.2     1999       4      21      27
## 182       2.4     2008       4      21      31
## 183       2.4     2008       4      21      31
## 184       3.0     1999       6      18      26
## 185       3.0     1999       6      18      26
## 186       3.5     2008       6      19      28
## 187       2.2     1999       4      21      27
## 188       2.2     1999       4      21      29
## 189       2.4     2008       4      21      31
## 190       2.4     2008       4      22      31
## 191       3.0     1999       6      18      26
## 192       3.0     1999       6      18      26
## 193       3.3     2008       6      18      27
## 194       1.8     1999       4      24      30
## 195       1.8     1999       4      24      33
## 196       1.8     1999       4      26      35
## 197       1.8     2008       4      28      37
## 198       1.8     2008       4      26      35
## 199       4.7     1999       8      11      15
## 200       5.7     2008       8      13      18
## 201       2.7     1999       4      15      20
## 202       2.7     1999       4      16      20
## 203       2.7     2008       4      17      22
## 204       3.4     1999       6      15      17
## 205       3.4     1999       6      15      19
## 206       4.0     2008       6      15      18
## 207       4.0     2008       6      16      20
## 208       2.0     1999       4      21      29
## 209       2.0     1999       4      19      26
## 210       2.0     2008       4      21      29
## 211       2.0     2008       4      22      29
## 212       2.8     1999       6      17      24
## 213       1.9     1999       4      33      44
## 214       2.0     1999       4      21      29
## 215       2.0     1999       4      19      26
## 216       2.0     2008       4      22      29
## 217       2.0     2008       4      21      29
## 218       2.5     2008       5      21      29
## 219       2.5     2008       5      21      29
## 220       2.8     1999       6      16      23
## 221       2.8     1999       6      17      24
## 222       1.9     1999       4      35      44
## 223       1.9     1999       4      29      41
## 224       2.0     1999       4      21      29
## 225         2.0      1999       4      19      26
## 226         2.5      2008       5      20      28
## 227         2.5      2008       5      20      29
## 228         1.8      1999       4      21      29
## 229         1.8      1999       4      18      29
## 230         2.0      2008       4      19      28
## 231         2.0      2008       4      21      29
## 232         2.8      1999       6      16      26
## 233         2.8      1999       6      18      26
## 234         3.6      2008       6      17      26
```

```
mpg_cor <- cor(mpgNumeric)
corrplot(mpg_cor, type = 'lower', method = 'number')
```



The positive correlations which the table shows are between highway and city mpg, cylinder and year, and displacement and year. The negative correlations exist between city mpg and displacement, highway mpg and displacement, city mpg and year, city mpg and cylinders, and highway mpg and cylinders. The strongest relationships here are between displacement and all other variables besides year, and the city/highway mpgs and all other variables besides year. I expected there to be a stronger relationship between year and other variables, but upon rereading the information about the data set i realized the data only spanned 9 years and technology for cars only could have improved so much between 1999 and 2008.