PSTAT 131 Homework 1

Luke Fields (8385924)

March 31, 2022

```
## corrplot 0.92 loaded

## -- 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()
## x dplyr::select() masks MASS::select()

## -- Attaching packages --------------------------------------- tidymodels 0.2.0 --

## v broom        0.7.12     v rsample      0.1.1
## v dials        0.1.0      v tune         0.2.0
## v infer        1.0.0      v workflows    0.2.6
## v modeldata    0.1.1      v workflowsets 0.2.1
## v parsnip      0.2.1      v yardstick    0.0.9
## v recipes      0.2.0

## -- 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 dplyr::select()   masks MASS::select()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Use suppressPackageStartupMessages() to eliminate package startup messages
```

Machine Learning Main Ideas

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

Supervised learning is when we know the outcome of the data, whereas unsupervised learning is when we only know the input data. Not only are there many different methods of machine learning that fall into

each category, like linear regression being supervised and hierarchial clustering being unsupervised, but we can actually check our error and have an "answer key" when we are using supervised learning.

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

While regression models predict continuous values, like quantitative numbers, classification models predict categorical values, like a qualitative class.

Question 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

Question 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 best emphasize a previous/past trend in data.

Inferential models: Models that evaluate the condition of the estimation and predictors.

Predictive models: Models that predict an exact outcome with the least amount of error with a combo of predictors that work bet.

Question 5: Predictive models are frequently used in machine learning, and they can usually be described as either mechanistic or empirically-driven. Define mechanistic. Define empirically-driven. How do these model types differ? How are they similar?

Mechanistic: When we assume or identify a form of the model, and then proceed to fit our model accordingly.

Empirically-driven: When we do not make any assessments about the form of our model.

Similarities: Both model types are trying to correctly assess the model fit and form in the best way.

Differences: Mechanistic is usually a simpler method to fit our model, but it can sometimes be far off from the actual model form, whereas empirically-driven requires more observations to get closer to the true form of the model, but usually is more flexible and closer to the true form when it succeeds.

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

Mechanistic is generally easier to understand because it requires a lot less observations, and can be in a simpler form like a linear model.

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

Mechanistic models will be of higher bias because we are assuming our model will form in a certain way and less variance if it the model does take that form, while more flexible methods like empirically-driven models result in less bias because they mold to the true form of the data better, but more variance because it might not perform as well across all models.

Question 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?

Predictive, as we are trying to directly predict a likelihood, whether that be an actual probability value or a scale like low to high or 1-10, with our voter's information (predictors).

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

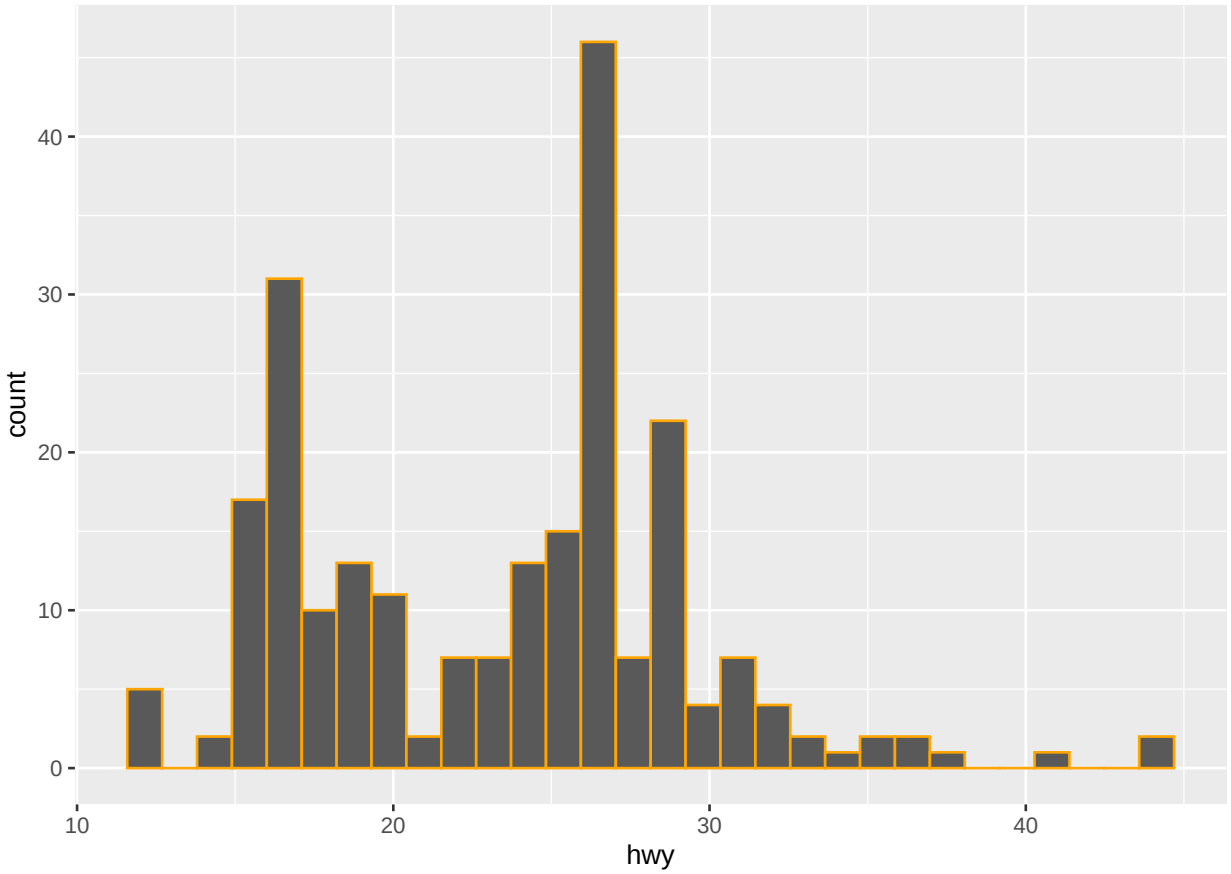
Inferential, as we are testing whether there is an association between the voter knowing the candidate personally and their support for them, instead of predicting an outcome.

Exploratory Data Analysis

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

```
# creating histogram plot of highway miles per gallon
hwy_hist <- ggplot(mpg, aes(x = hwy)) + geom_histogram(color = "orange")
hwy_hist
```

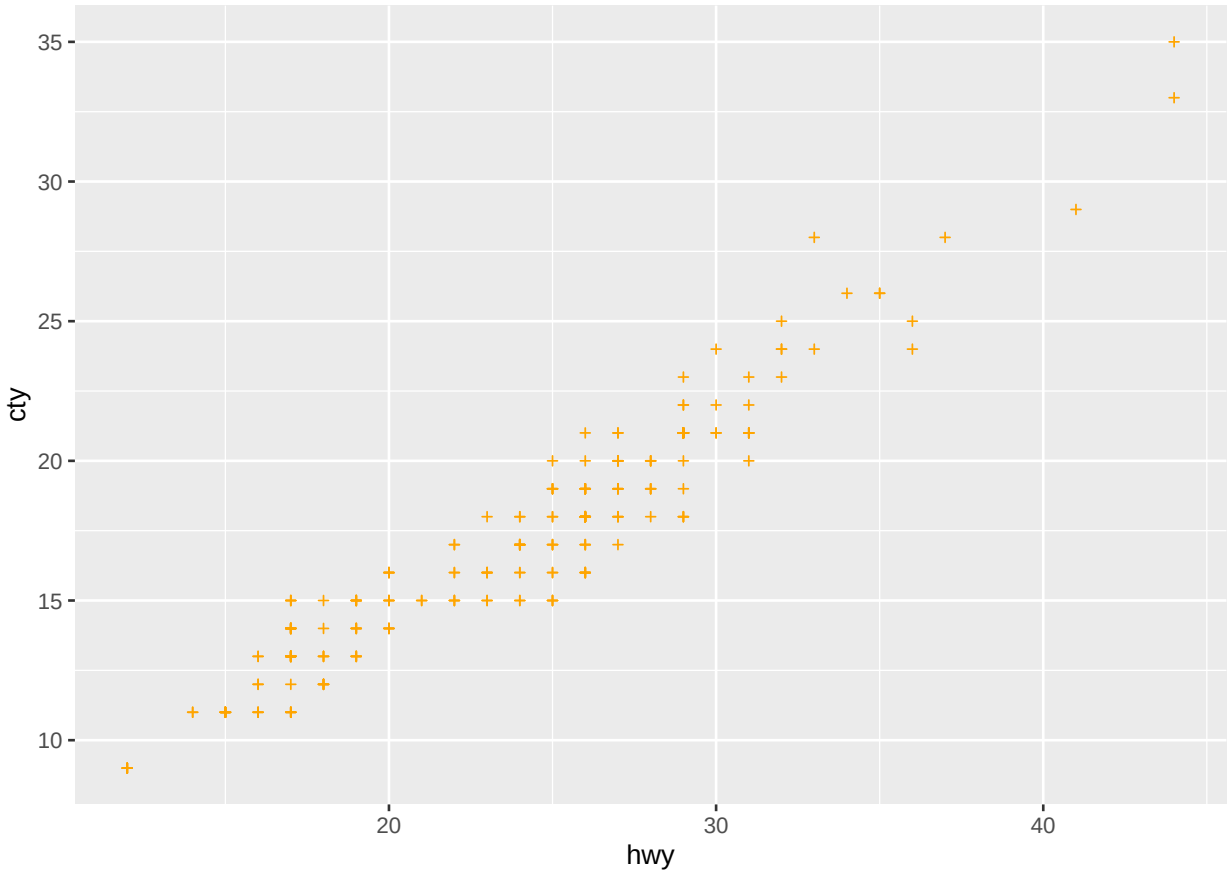
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From this histogram, we can see that cars have high spikes in highway mpg at about 15 miles per gallon and about 25 miles per gallon, and there are very few cars whose highway mpg is over 30. So, most cars have an average mpg between 15 and 30 on the highway.

Exercise 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?

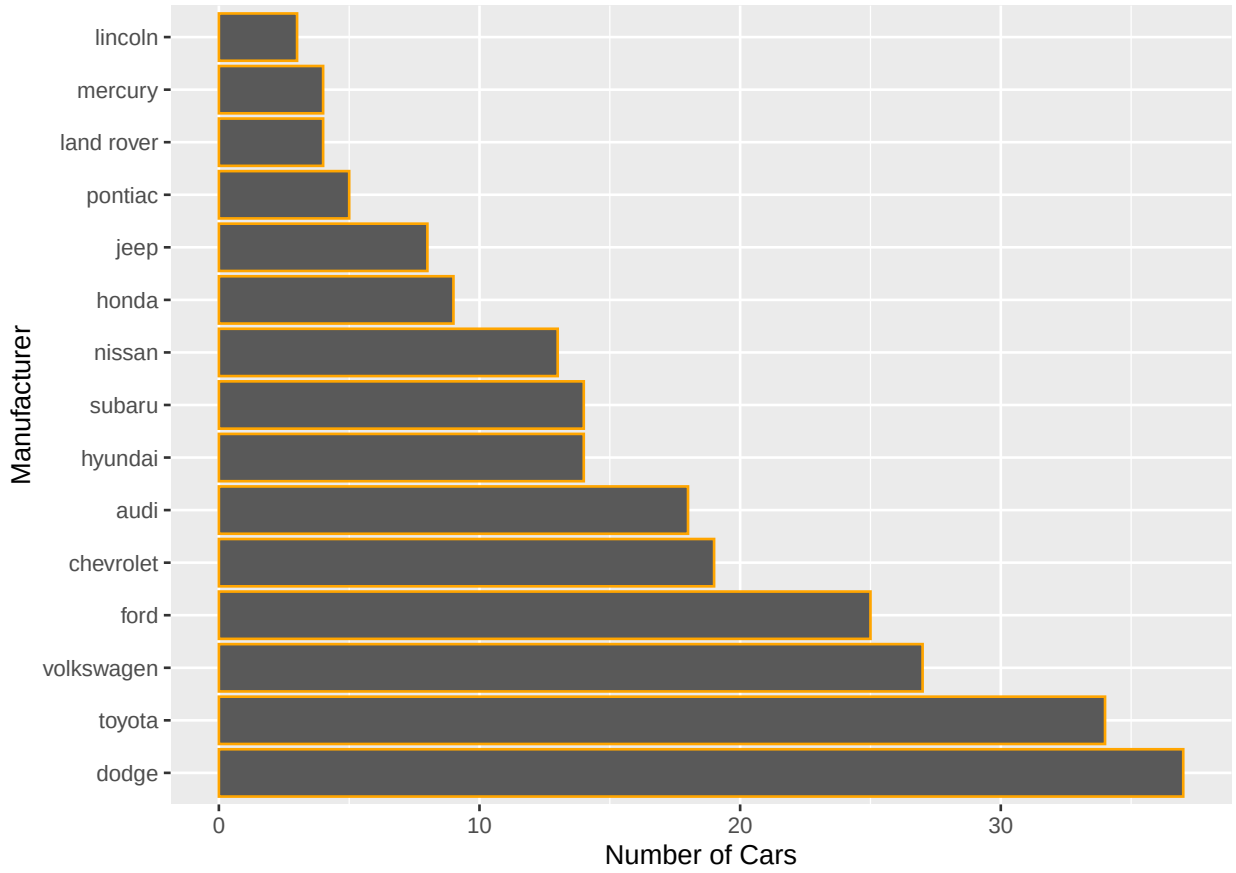
```
# creating scatter plot of highway mpg and city mpg
hwy_cty_scatter <- ggplot(mpg, aes(x = hwy, y = cty)) + geom_point(size = 1, shape = 3, color = "orange
hwy_cty_scatter
```



From this scatter plot, we can see that there is a high positive correlation between cars' highway mpg and cars' city mpg. In other words, cars that have a lower highway mpg will most likely have a lower city mpg, and cars that have a higher highway mpg will most likely have a higher city mpg, which makes sense in the real world.

Exercise 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?

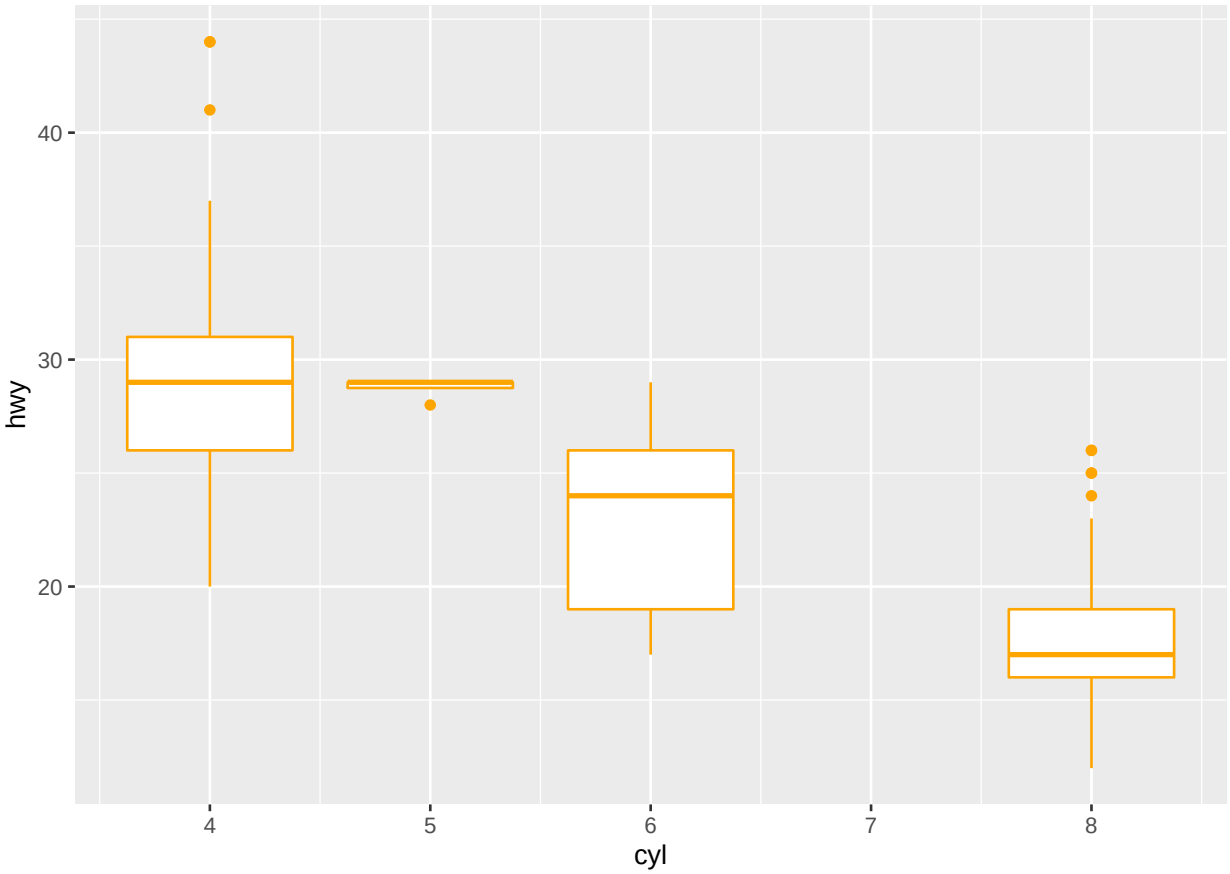
```
# creating bar plot of manufacturer count
manuf_bar <- ggplot(mpg, aes(x=reorder(manufacturer, manufacturer, function(x)-length(x)))) + geom_bar(
manuf_bar + coord_flip()
```



From this bar graph, we can see the amount of cars manufactured by brand, ordered in least amount to most. Lincoln has less than 5 cars manfacturers at the minimum, while dodger has over 30 cars manufactured at the maximum. The amount of cars manufactured by region (e.g. USA, Japan, etc.) is fairly even in this graph.

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

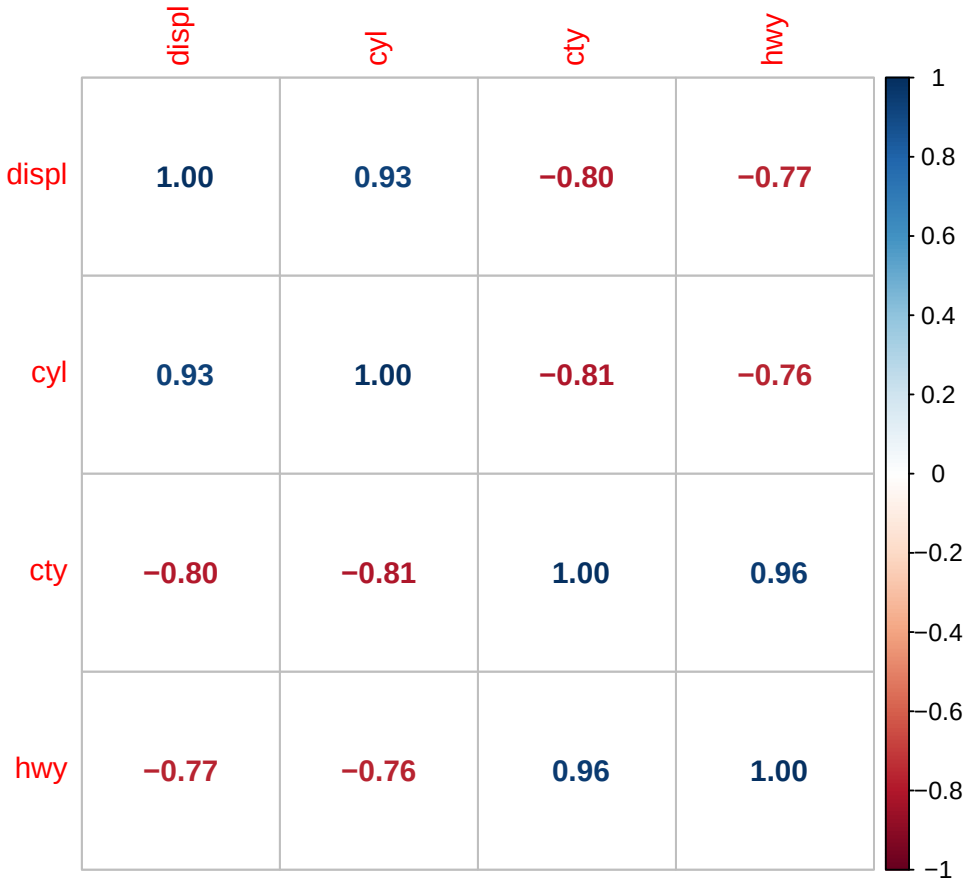
```
# creating box plot for highway mpg grouped by cyl
hwy_cyl_box <- ggplot(mpg, aes(x = cyl, y = hwy)) + geom_boxplot(aes(group = cyl), color = "orange")
hwy_cyl_box
```



From this box-plot, we can see that the mean and median highway mpg all continue to decrease with an increase in cylinders within the car. So, less cylinders within a car mean the car will have higher highway mpg, and a car with more cylinders with have a smaller highway mpg.

Exercise 5: Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset.

```
# Creating Correlation Plot For Continuoius Variables
mpg_numeric <- mpg[sapply(mpg, is.numeric)]
mpg_continuous <- subset(mpg_numeric, select = -c(year))
mpg_continuous_cor <- cor(mpg_continuous)
mpg_low_tri_cor <- corrplot(mpg_continuous_cor, method = "number")
```




```
mpg_low_tri_cor
```

```
## $corr
##         displ     cyl     cty     hwy
## displ  1.0000  0.9302 -0.7985 -0.7660
## cyl    0.9302  1.0000 -0.8058 -0.7619
## cty   -0.7985 -0.8058  1.0000  0.9559
## hwy   -0.7660 -0.7619  0.9559  1.0000
##
## $corrPos
##    xName yName x y    corr
## 1  displ displ 1 4  1.0000
## 2  displ   cyl 1 3  0.9302
## 3  displ   cty 1 2 -0.7985
## 4  displ   hwy 1 1 -0.7660
## 5    cyl displ 2 4  0.9302
## 6    cyl   cyl 2 3  1.0000
## 7    cyl   cty 2 2 -0.8058
## 8    cyl   hwy 2 1 -0.7619
## 9    cty displ 3 4 -0.7985
## 10   cty   cyl 3 3 -0.8058
## 11   cty   cty 3 2  1.0000
## 12   cty   hwy 3 1  0.9559
## 13   hwy displ 4 4 -0.7660
## 14   hwy   cyl 4 3 -0.7619
```

```
## 15   hwy   cty 4 2  0.9559
## 16   hwy   hwy 4 1  1.0000
##
## $arg
## $arg$type
## [1] "full"
```

Which variables are positively or negatively correlated with which others? Do these relationships make sense to you? Are there any that surprise you?

From this correlation plot that only takes into account continuous variables, we can see displacement and highway mpg are each highly positively correlated with cylinder are highly correlated at 0.93 and 0.96, respectively, and displacement and cylinder are highly negatively correlated with city mpg at -0.8 and -0.81, respectively. The surprising things to me are that cylinder can be so strongly postitvely correlated and so strongly negatively correlated with two different types of mpg, but the idea that more cylinders impacts any form of mpg makes sense to me.

END OF HOMEWORK 1