Homework 1

Safiya Alavi

9/27/2022

Machine Learning Main Ideas

Question 1

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

Supervised learning involves building a statistical model for predicting or estimating an output based on one or more inputs. Unsupervised learning is similar in the sense that there are inputs, although there is no supervising output. So the difference between the two is due to the fact that one has an output based on the inputs, whereas the other involves analyzing the inputs to understand patterns within them.

Question 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, a regression model involves predicting a continuous or quantitative output value, whereas a classification model involves predicting a categorical or qualitative output. An example of a regression model is analyzing data about wages for males, and an example of a classification model is analyzing and modeling the stock market.

Question 3

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

Two commonly used metrics for regression machine learning problems are least squares linear regression and also logistic regression. Commonly used metrics for classification problems are logistic regression, K-nearst neighbors and boosting.

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: Descriptive models are used in statistics to provide useful and actionable results from a given data set. Descriptive models choose the best model that emphasizes a trend in the data.

Inferential models: Inferential models serve the purpose of understanding how the response variable changes as a function of X, or the predictors. In this model, we are more concerned with estimating the function f in order to

understand the way that Y is affected by X. The aim with inferential models is to test theories and state relationships between outcomes and predictors.

Predictive models: Predictive models involve making predictions about the response variable Y. In this case, we are not particularly interested in estimating f, but rather using f to yield accurate predictions for Y. The accuracy of the prediction model is dependent on both irreducible error given by the error terms and the reducible error, and our goal is to minimize the reducible error.

Question 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 means to relate theories to a phenomena. Whereas, empirically driven means to be based on or guided by experimentation. These two model types differ in the sense that Mechanistic models assume a parametric form for f and in definition will never match the true unknown f exactly. With mechanistic modeling, we can add parameters to improve model flexibility, but sometimes with too many, the model can be overfitting of the data. This means that it will follow noise or errors too closely yielding inaccurate results. On the other hand, empirically driven models have no assumptions about the form of f which is a positive since the model is naturally more flexible that mechanistic models. The negative is due to the fact that this model requires a much larger number of observations and, similarly to mechanistic models, can at times over fit the data.

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

In general, mechanistic models are easier to understand in comparison to empirically-driven models. Mechanistic models are boiled down to predicting a set of parameters for the model that fits the data best. While empirically-driven models can yield more accuracy, their interpretability is sacrificed due to the fact that the function f is an arbitrary function.

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

The bias-variance tradeoff is related to the use of mechanistic and empirically-driven models in the sense that both models are trying to simultaneously minimize the two sources of error, bias and variance, in order to prevent the supervised learning algorithms from generalizing beyond their training set (Wikipedia). Deciding on the most effective bias-variance tradeoff per model is an important decision that can affect the accuracy of the model.

Question 6

A political candidate’s campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions:

Given a voter’s profile/data, how likely is it that they will vote in favor of the candidate?
How would a voter’s likelihood of support for the candidate change if they had personal contact with the candidate?

Classify each question as either predictive or inferential. Explain your reasoning for each.

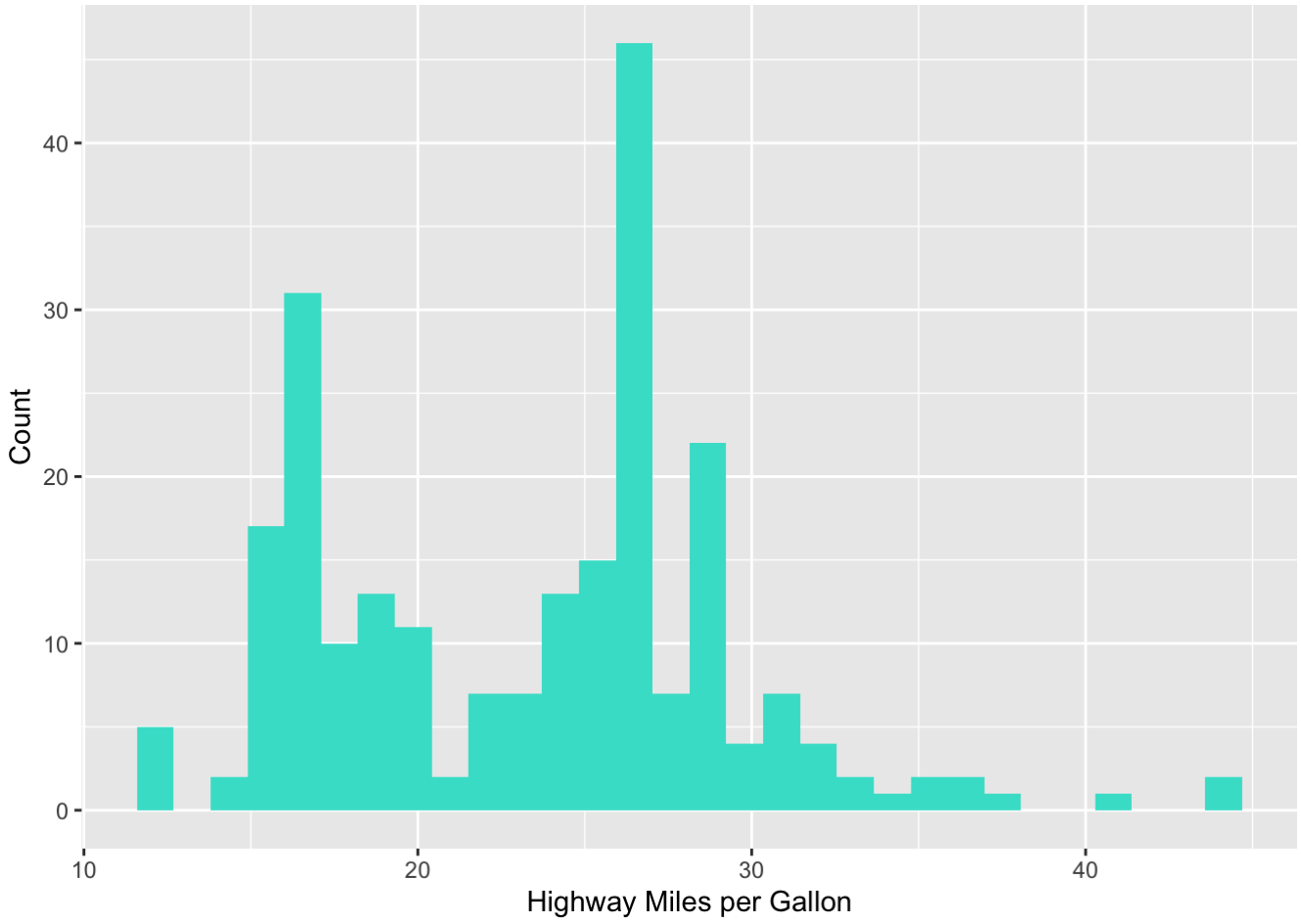
Question 1 is a prediction problem because we are given a set of variables for X, and asked to predict what the outcome Y will be based on the given information. On the other hand, Question 2 is an inferential question because we are interested in how an individual input variable affects the overall outcome Y, or in other words, we are interested in how Y changes as a function of X.

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.

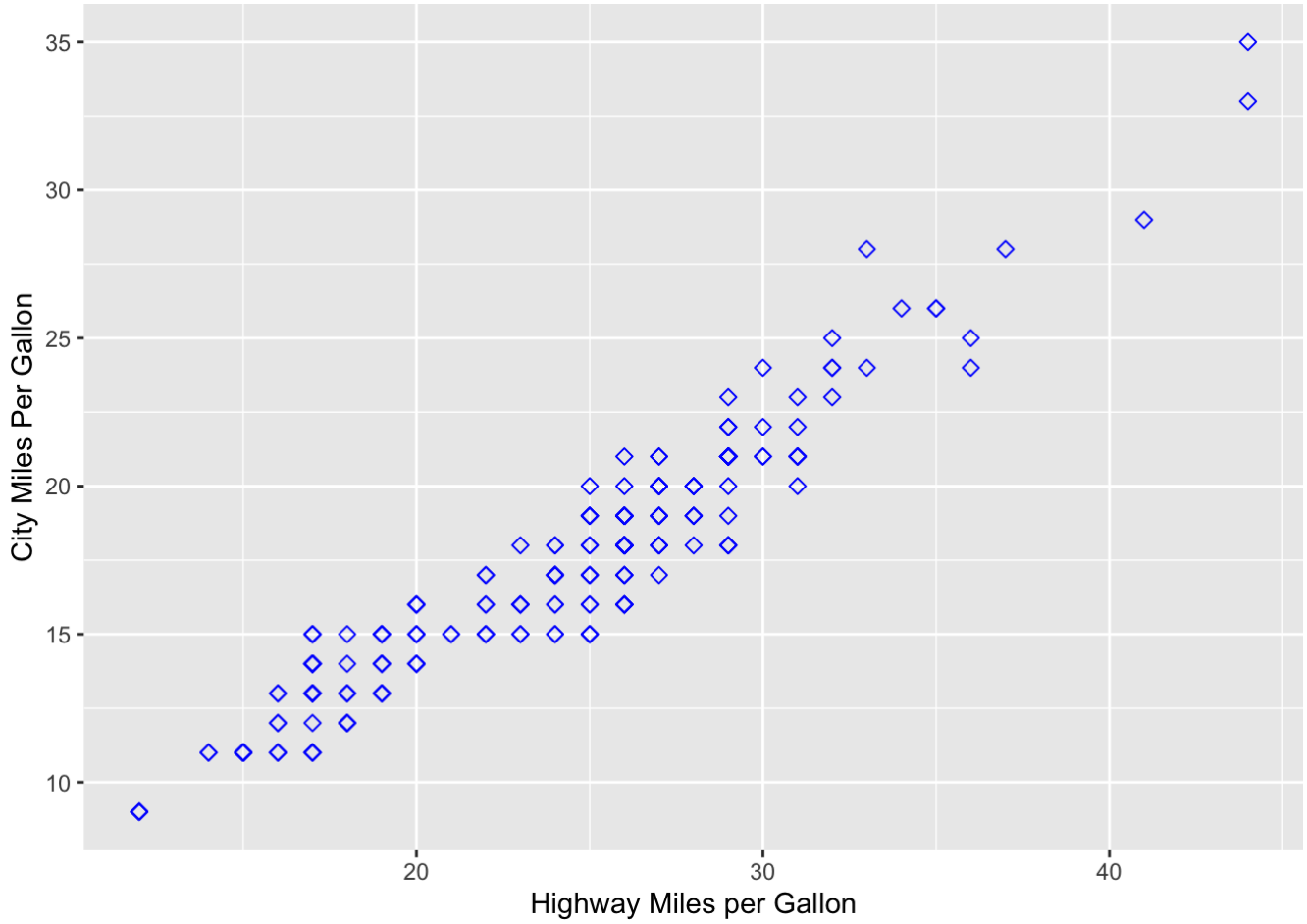
```
ggplot(mpg,aes(hwy))+geom_histogram(aes(color = hwy), fill = "turquoise")+labs(y= "Coun
t", x = "Highway Miles per Gallon")
```



From this histogram, I observe that there are one large spike in the data at around 26 highway miles per gallon. There are additionally two more smaller spikes in the data around 17 highway miles per gallon and 29 highway miles per gallon. The count for the rest of the highway miles per gallon is less than 20. To me, it looks like the data has two bumps in it, or in other words, two normal distributions right next to each other. The data trails off with a very low count of highway miles per gallon exceeding 32.

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?

```
ggplot(mpg, aes(x=hwy, y=cty)) + geom_point(size=2, shape=23, color = "blue")+labs(y= "C
ity Miles Per Gallon", x = "Highway Miles per Gallon")
```

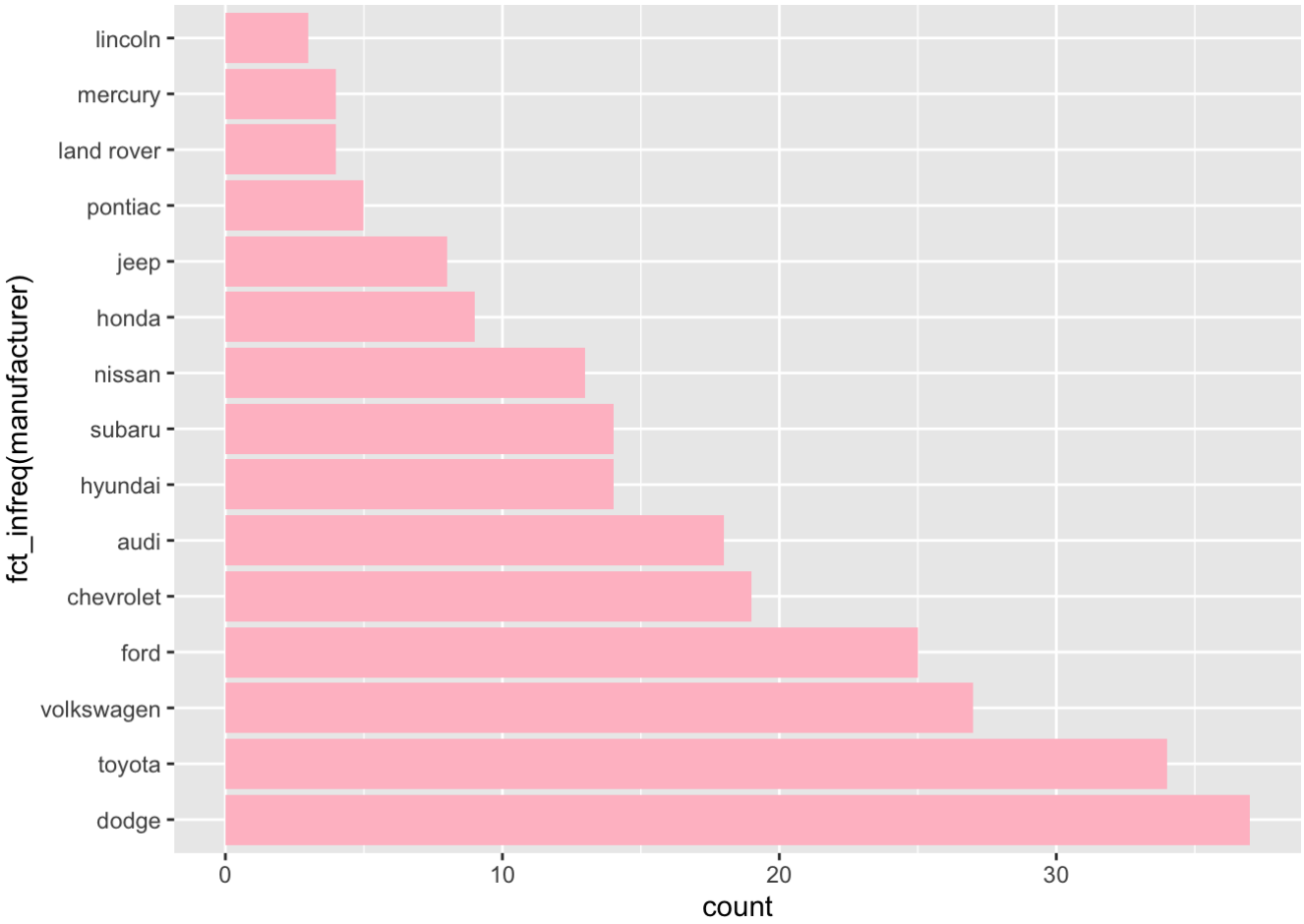


I notice that the points lie on the plot in a very linear fashion. Generally, as the highway miles per gallon increases, the city miles per gallon also increases. Another observation is that the range of values city miles per gallon is for every highway miles per gallon slowly increases and then decreases. For example, when highway miles gallon is about 5, the city miles per gallon is also equal to about 5. As the highway miles per gallon increases to 25, the range of value of city gallons in between 15 and 21. It is very safe to say that highway miles per gallon and city miles per gallon is highly correlated, which also makes sense intuitively. The highway miles per gallon is also generally slightly higher than city miles per gallon.

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?

```
ggplot(mpg, aes(x=fct_infreq(manufacturer))) + geom_bar(stat = "count", fill = "pink")+
 coord_flip()
```

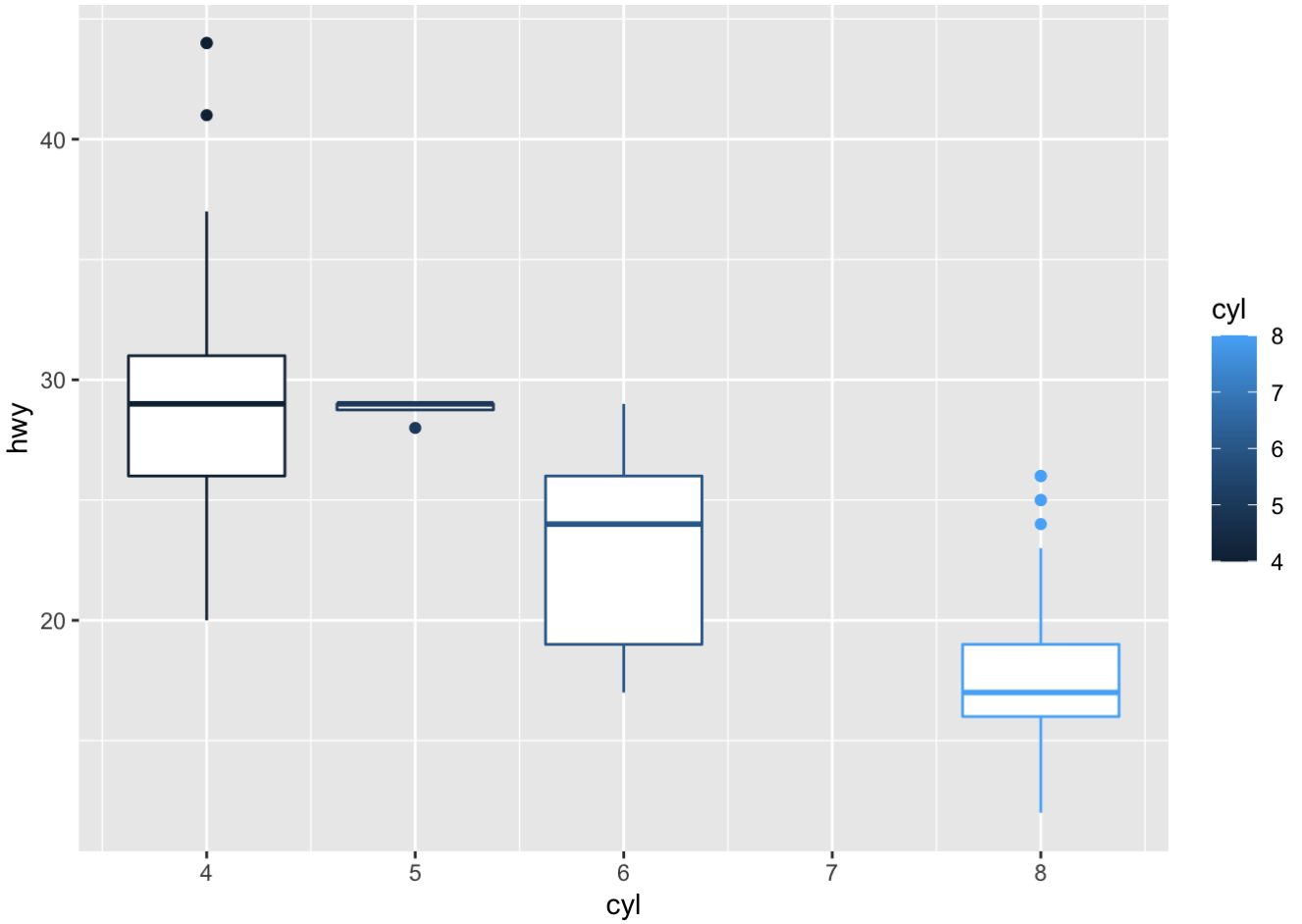


Dodge produced the most cars from 1999 to 2008, whereas Lincoln produced the least.

Exercise 4

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

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

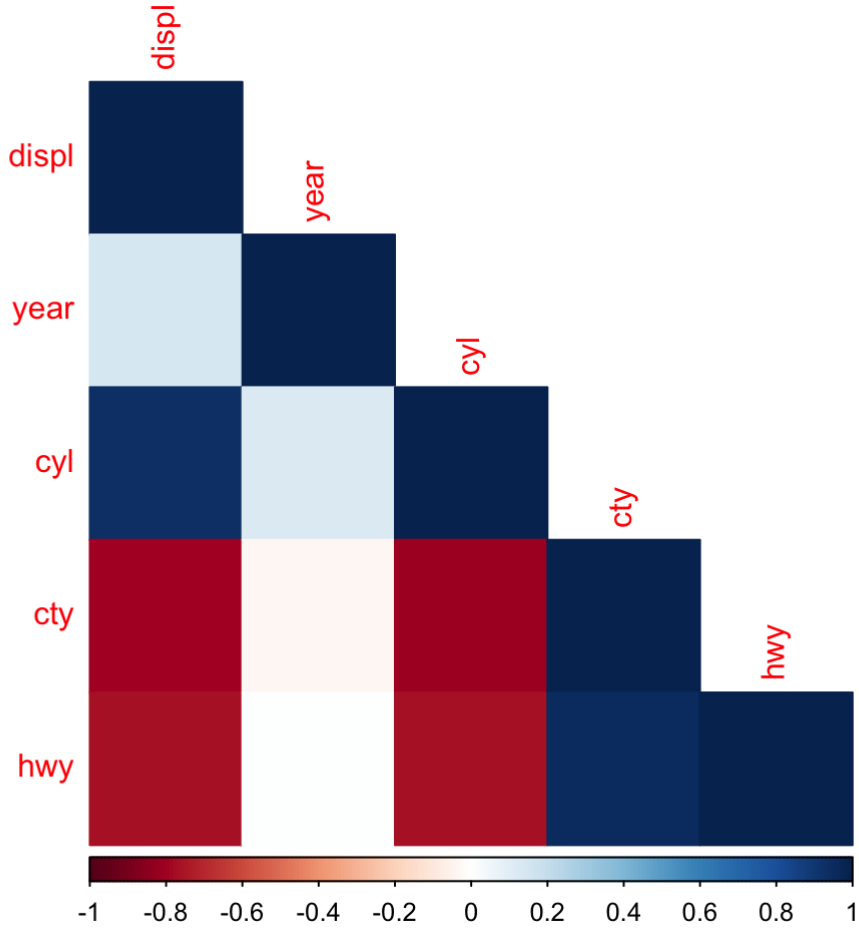


Majority of number of cylinders are even numbers 4,6,8. I see a pattern that generally as the number of cylinders increases, the highway mile per gallon decreases.

Exercise 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?

```
M = cor(mpg[,sapply(mpg,is.numeric)],use="complete.obs",method="pearson")
corrplot(M, method = 'color', type = "lower")
```



Based on this correlation plot, we can see that cty/hwy and displ/cyl are positively correlated, whereas cyl/cty and cyl/hwy and displ/hwy and cty/displ are negatively correlated. In general, displ/year, cyl/year, year/cty and year/hwy have almost to no correlation with eachother. In general, there relationships make sense to me, especially the high correlation between hwy/cty and displ/cyl. I am surprised by the little correlation between year/cty and year/hwy because I would have inferred that the higher the year, the better the city and highway mileage is.