Homework 1

Jonathan Palada Rosal

2022-04-03

(1). Define Supervised and Unsupervised learning. What are the difference(s) between them?

Supervised Learning - In this type of learning, the actual data Y is the supervisor, and it also needs to give the model observed output and input. Unsupervised Learning - It learns without a supervisor and we never see the answer key. Some differences between them are Supervised learning consists of: Linear/Logistic Regression, k-th nearest neighbors, decision trees, random forests, support vector machine(s), and neural networks. Unsupervised Learning consists of Principal Component Analysis (PCA), k-means clustering, Hierarchical clustering, and Neural Networks.

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

A classification model has the task of predicting a discrete class label vs. a regression model has the task of predicting a continuous quantity.

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

R^2 and Adjusted R-squared.

(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: A model to best visually emphasize a trend in data (Using a line on scatterplot)

Inferential models: Tries to predict Y with minimum reducible error. This model is not focused on hypothesis tests. (Features that fits best)

Predictive models: Tries to test theories. Maybe casual claims. States relationships between the outcome and predictors. (Features that are significant)

(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? Answer: Mechanistic - Assumes a parametric form for f, won't match true unknown f, can add parameter

(more flexibility), and too many means there will be overfitting. Empirically-Driven - There are no assumptions about f (different to mechanistic), requires a larger number of observations (different to mechanistic), flexible by default (similar to mechanistic because they both could be flexible), and overfitting (similar to mechanistic because they can both contain overfitting).

In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice. Answer: I think that empirically-driven is easier for me to understand. The reason being that there are no assumptions on f, and it is already flexible by default without adding parameters.

Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models. Answer: Bias-Variance is when we choose to lower bias by increasing variance or vice-versa. Based on the model we choose, one trade-off would be better than the other.

(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? Answer: Predictive because based on the data we are trying to predict their choice of a candidate. I think of the voter's profile/data as features. Based on the combination of their features we will predict Y or the candidate.

How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate? Answer: Inferential because we are trying to test a theory. The theory for this is if the voter had personal contact with the candidate, would they be more likely to vote for them?

```
library(tidyverse)
```

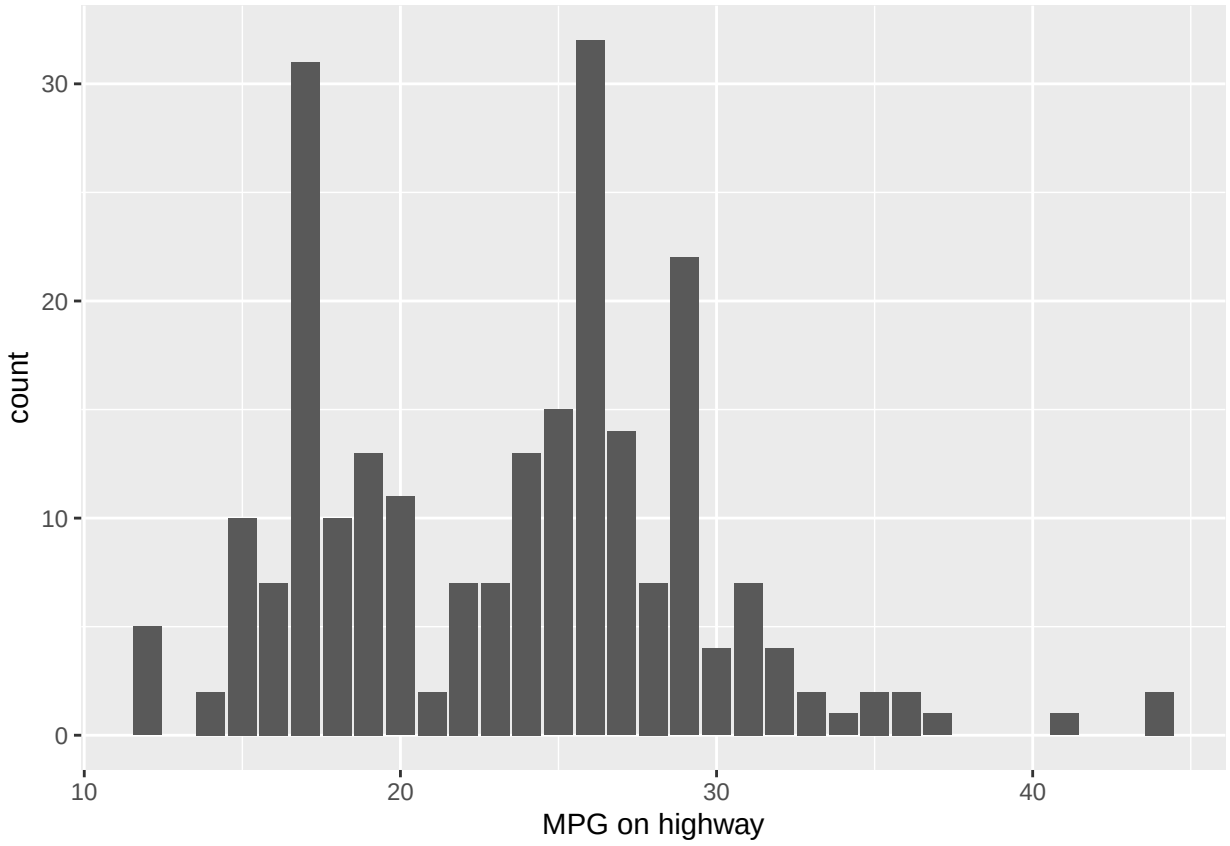
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## 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()
```

```
head(mpg)
```

```
## # A tibble: 6 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~
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa~
```

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.

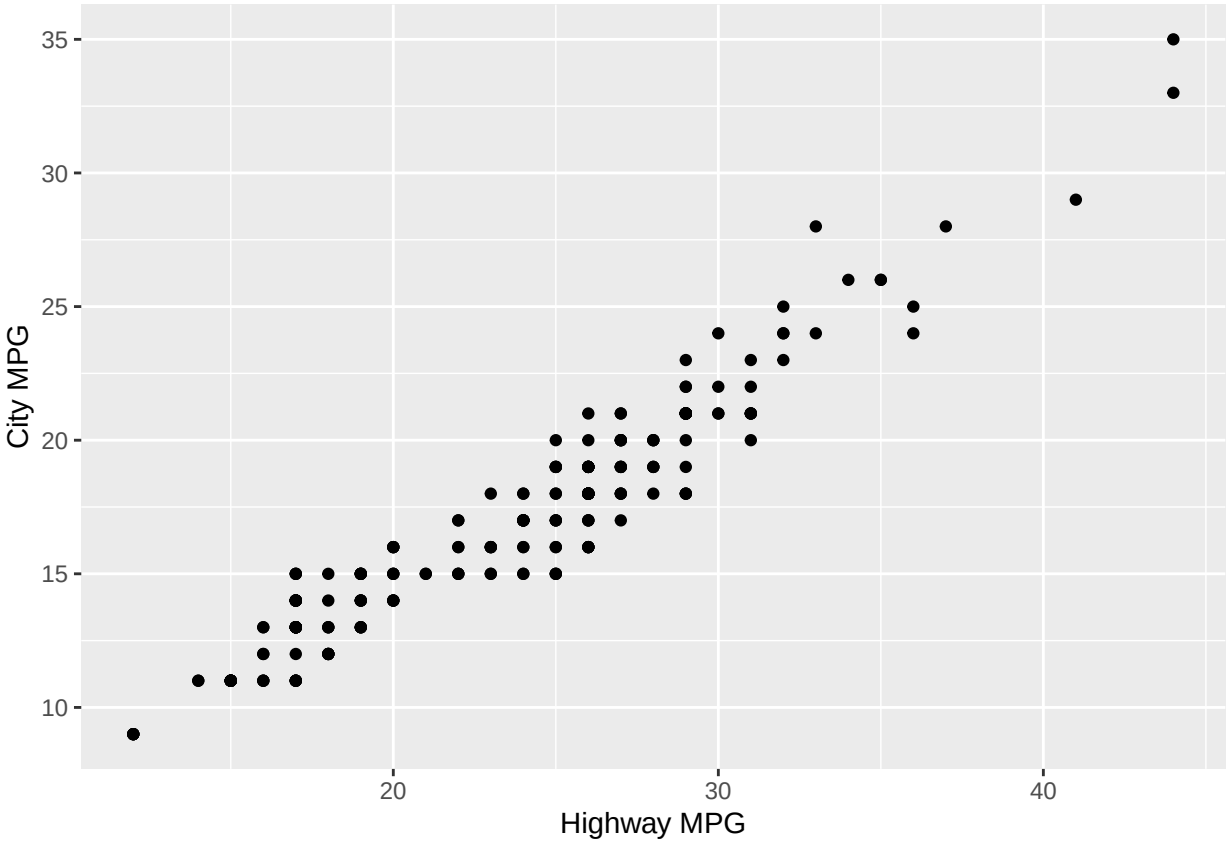
```
library(ggplot2)
ggplot(mpg,aes(hwy,)) + geom_bar() + xlab("MPG on highway")
```



Based on my histogram, most of the cars in this dataset get 26 mpg 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?

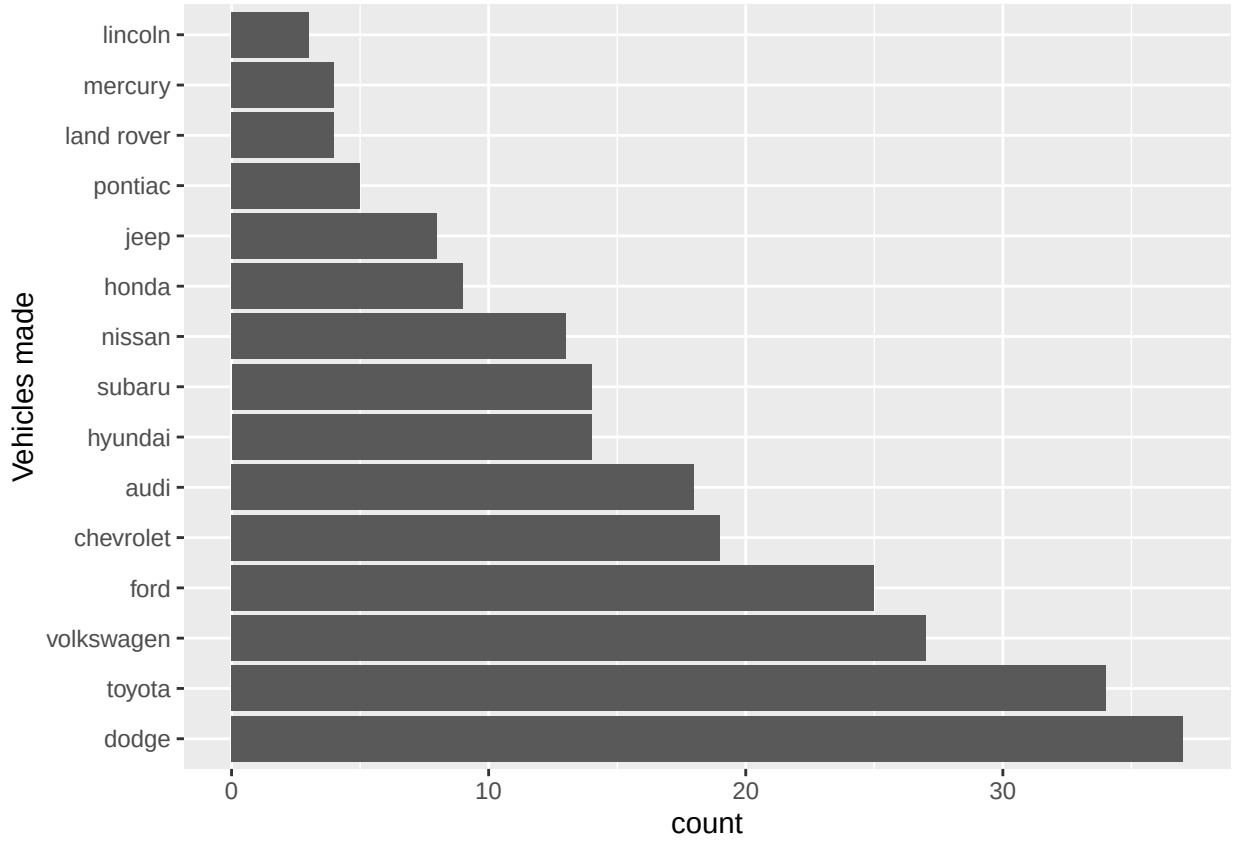
```
ggplot(mpg, aes(hwy, cty,), ) + geom_point() + xlab("Highway MPG") + ylab("City MPG")
```



There does seem to be a relationship between city MPG and highway MPG. There is a positive correlation.

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=reorder(manufacturer,manufacturer,function(x)-length(x)),), horizontal = TRUE,) + geom
```

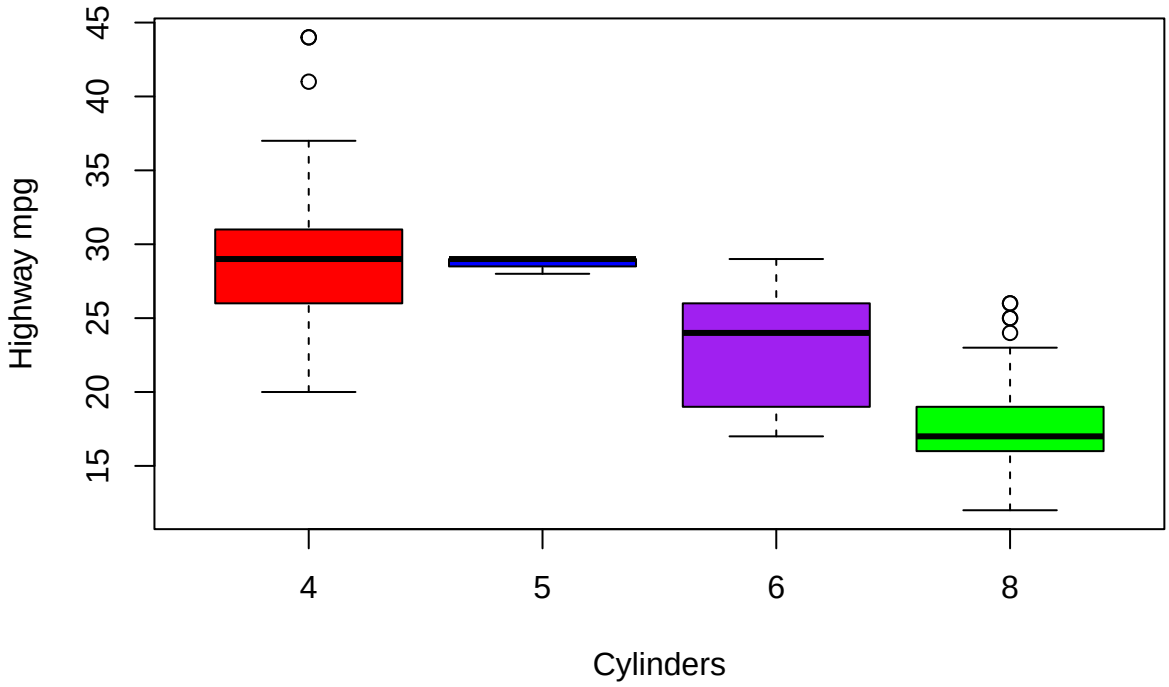



```
# source for how to reorder- (https://www.geeksforgeeks.org/how-to-reorder-boxplots-in-r-with-ggplot2/#
```

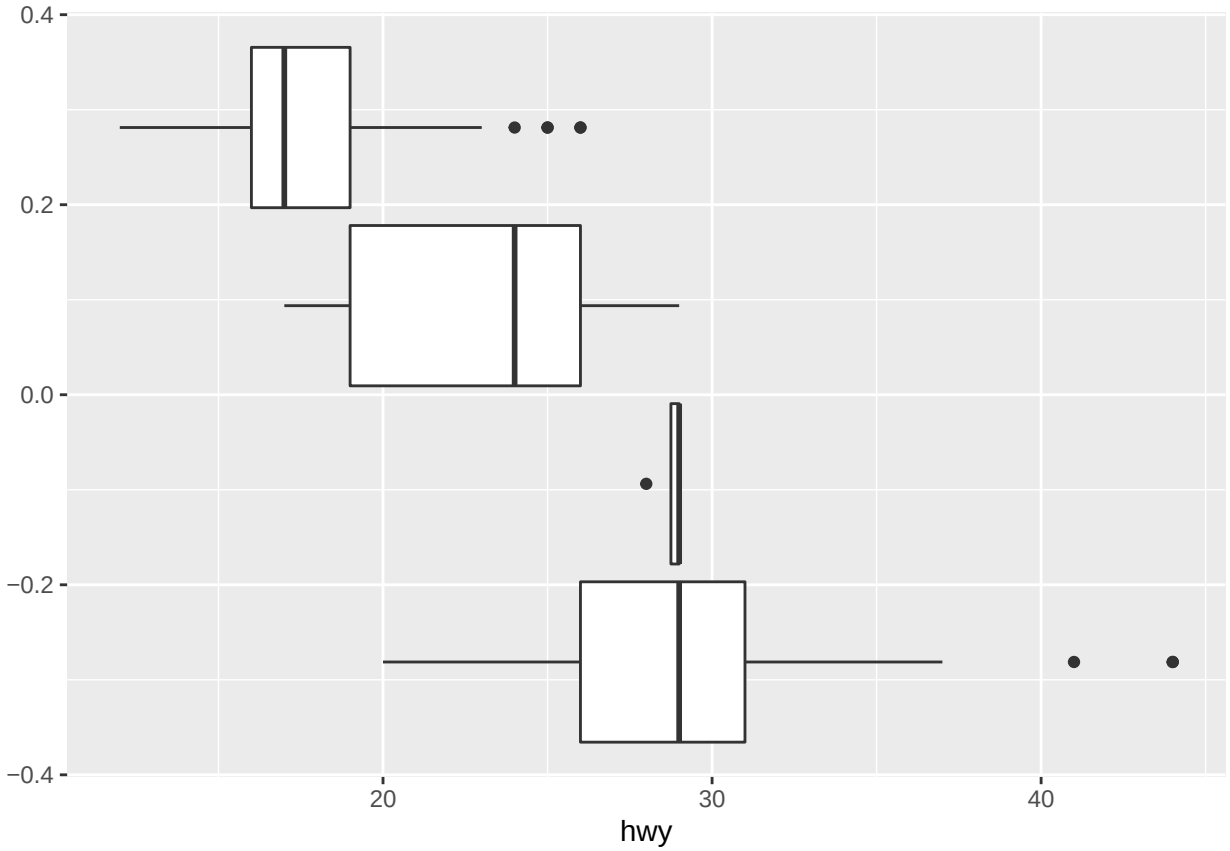
The manufacturer that produced the most vehicles was Dodge. The manufacturer that produces the least vehicles was Lincoln.

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

```
boxplot(mpg$hwy ~ mpg$cyl ,xlab = "Cylinders",ylab = "Highway mpg", col=c("red","blue","purple","green"
```




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



The pattern I notice is that the lower the cylinders, the higher the highway mpg. It has a negative correlation.

Exercise 5: Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset. (Hint: You can find information on the package here.) Which variables are positively or negatively correlated with which others? Do these relationships make sense to you? Are there any that surprise you?

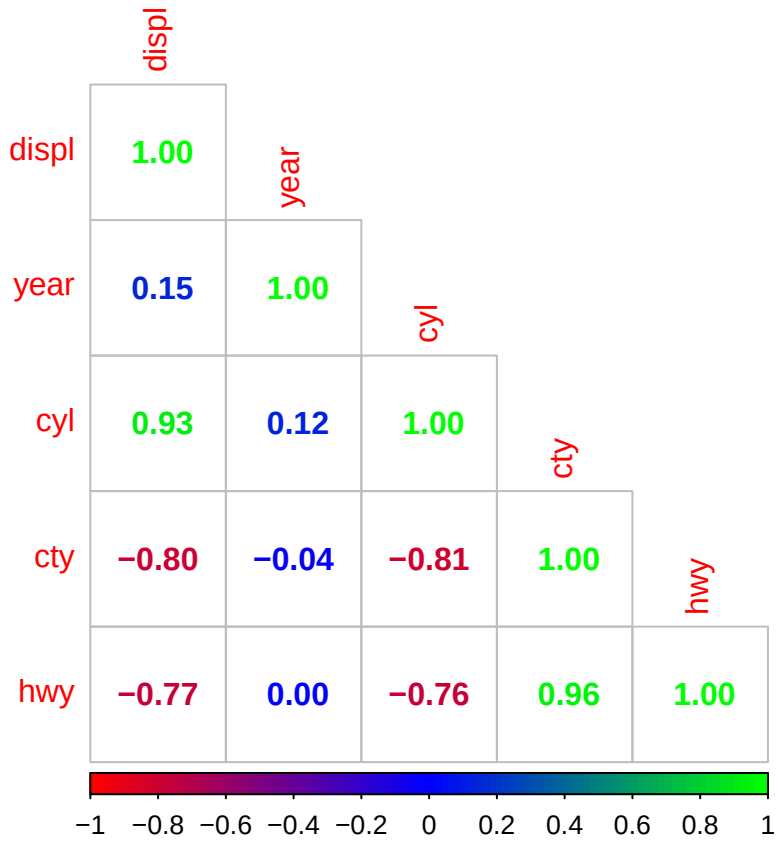
```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(dplyr)
mpg_matrix <- mpg %>%
  select_if(is.numeric) %>%
  cor(.)
mpg_matrix
```

```
##            displ         year        cyl         cty          hwy
## displ  1.0000000  0.147842816  0.9302271 -0.79852397 -0.766020021
## year   0.1478428  1.000000000  0.1222453 -0.03723229  0.002157643
## cyl    0.9302271  0.122245347  1.0000000 -0.80577141 -0.761912354
## cty   -0.7985240 -0.037232291 -0.8057714  1.00000000  0.955915914
## hwy   -0.7660200  0.002157643 -0.7619124  0.95591591  1.000000000
```

```
corrplot(mpg_matrix, method = 'number', col = colorRampPalette(c("red","blue","green"))(100), type = 'lo
```



City per gallon is positively correlated with highway mpg but is negatively correlated everywhere else. Highway mpg is negatively correlated with engine displacement and cylinders. Year has a positive correlation with engine displacement and cylinders. Engine displacement has a positive correlation with cylinders.