Homework 1

PSTAT 131/231

Machine Learning Main Ideas

Please answer the following questions. Be sure that your solutions are clearly marked and that your document is neatly formatted.

You don't have to rephrase everything in your own words, but if you quote directly, you should cite whatever materials you use (this can be as simple as "from the lecture/page # of book").

Question 1:

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

- Supervised learning: It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross validation process (From https://www.ibm.com/cloud/learn/supervised-learning)
- Unsupervised learning: also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets.(From https://www.ibm.com/cloud/learn/unsupervised-learning)
- Supervised learning algorithms require input-output pairs (i.e. they require the output), unsupervised learning requires only the input data (not the outputs).

Question 2:

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

- Regression model: Regression analysis is a fundamental concept in the field of machine learning. It falls under supervised learning wherein the algorithm is trained with both input features and output labels. It helps in establishing a relationship among the variables by estimating how one variable affects the other. Especially, The Y is quantitative which is numerical values.(https://builtin.com/data-science/regression-machine-learning)
- Classification model: Classification is a task that requires the use of machine learning algorithms that learn how to assign a class label to examples from the problem domain. An easy to understand example is classifying emails as "spam" or "not spam." The Y is qualitative which is categorical values.(https://machinelearningmastery.com/types-of-classification-in-machine-learning/)

Question 3:

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

None.

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: Choose model to best visually emphasize a trend in data i.e., using a line on a scatterplot
- Inferential models: What features are significant? Aim is to test theories; (Possibly) causal claims;State relationship between outcome & predictor(s) (From 131 lecture_day2)
- Predictive models: What combo of features fits best? Aim is to predict Y with minimum reducible error; Not focused on hypothesis tests

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: Assume a parametric form for f(i.e. beta_0 + beta_1 + . . .); Won't match true unknown f; Can add parameters = more flexibility;Too many = overfitting.
  Empirically-driven: No assumptions about f; Require a larger # of observations;Much more flexible by default; Overfitting.
  The Mechanistic can be more flexibility by adding more parameters and empirically-driven more flexible by default; The Mechanistic will cause overfitting by adding more parameters and empirically-driven is overfitting by default.
- In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice.
  Mechanistic model is easier. A mechanistic model uses a theory to predict what will happen in the real world. The empirical modeling, studies real-world events to develop a theory. We can just use theory to do the calculation.
- Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models.
  None.

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.

- Given a voter's profile/data, how likely is it that they will vote in favor of the candidate? *This is predictive model. Since we just predict the outcome and without hypothesis test.*
- How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate? *This is inferential model.Since this is a causal claim which the candidate whether has personal contact.*

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.

```
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

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

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

## Warning: package 'dplyr' was built under R version 4.0.5

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

```
library(ggplot2)
library(dplyr)
tidyverse_packages()
```

```
##  [1] "broom"         "cli"           "crayon"        "dbplyr"
##  [5] "dplyr"         "dtplyr"        "forcats"       "googledrive"
##  [9] "googlesheets4" "ggplot2"       "haven"         "hms"
## [13] "httr"          "jsonlite"      "lubridate"     "magrittr"
## [17] "modelr"        "pillar"        "purrr"         "readr"
## [21] "readxl"        "reprex"        "rlang"         "rstudioapi"
## [25] "rvest"         "stringr"       "tibble"        "tidyr"
## [29] "xml2"          "tidyverse"
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(stsm)
```

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."

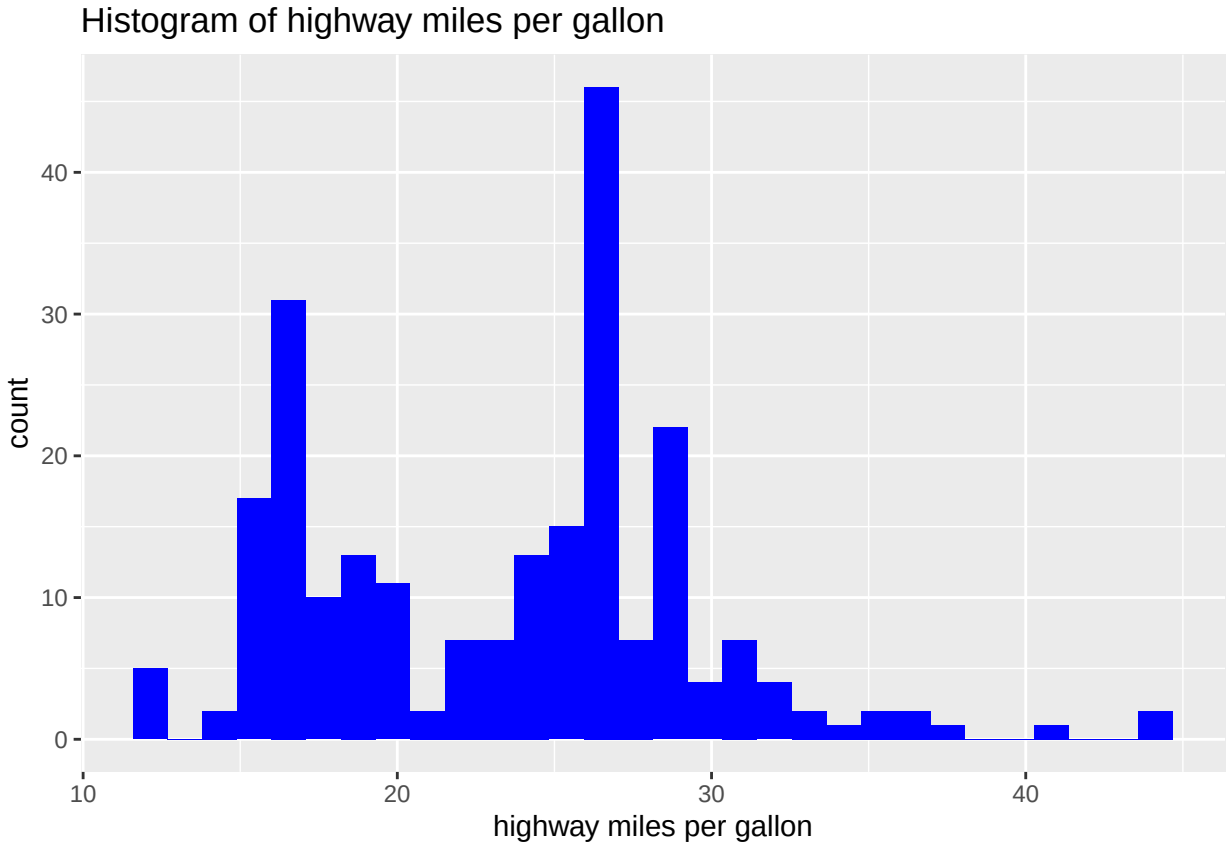
You should use the tidyverse and ggplot2 for these exercises.

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(x=hwy)) + geom_histogram(fill="blue") +
  xlab("highway miles per gallon") +
  ggtitle("Histogram of highway miles per gallon")
```

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

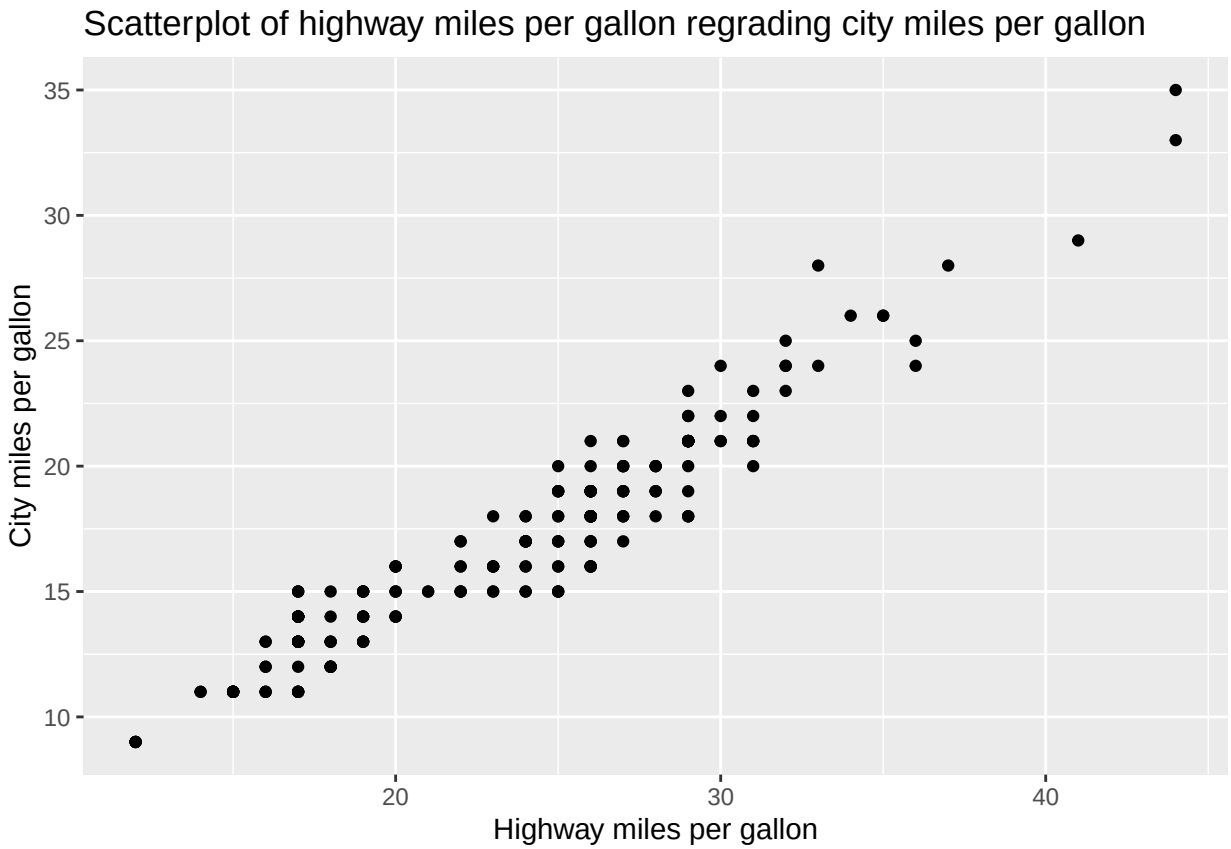


As we can see, The data is not evenly distributed. Most data between 16-20 hwy & 21-29 hwy. The biggest count is nearly 26 hwy and the least count is nearly 41.

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() +
  xlab("Highway miles per gallon") +
  ylab("City miles per gallon") +
  ggtitle("Scatterplot of highway miles per gallon regrading city miles per gallon")
```

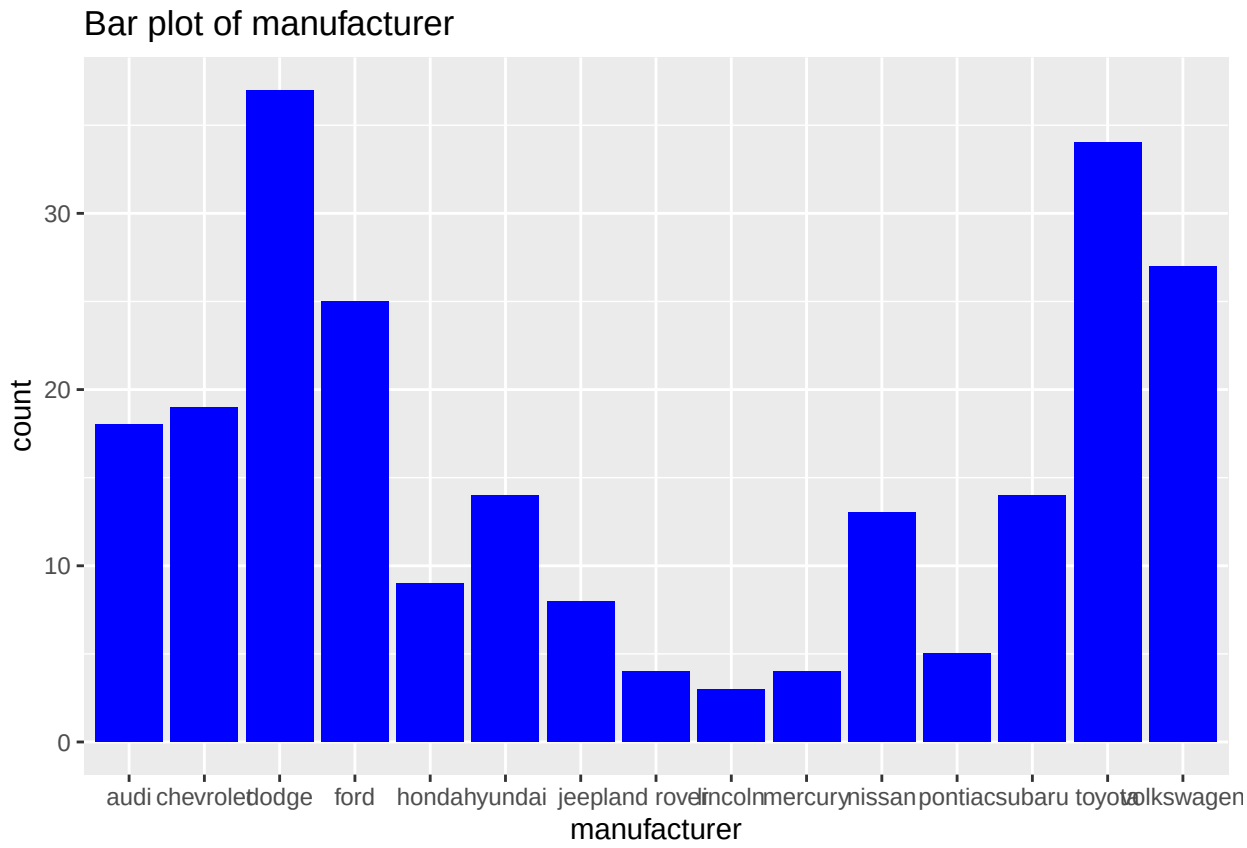


From the plot, we can notice that the hwy has a positive correlation to the cty which imply a car has higher highway miles per gallon also has a high city miles per gallon.

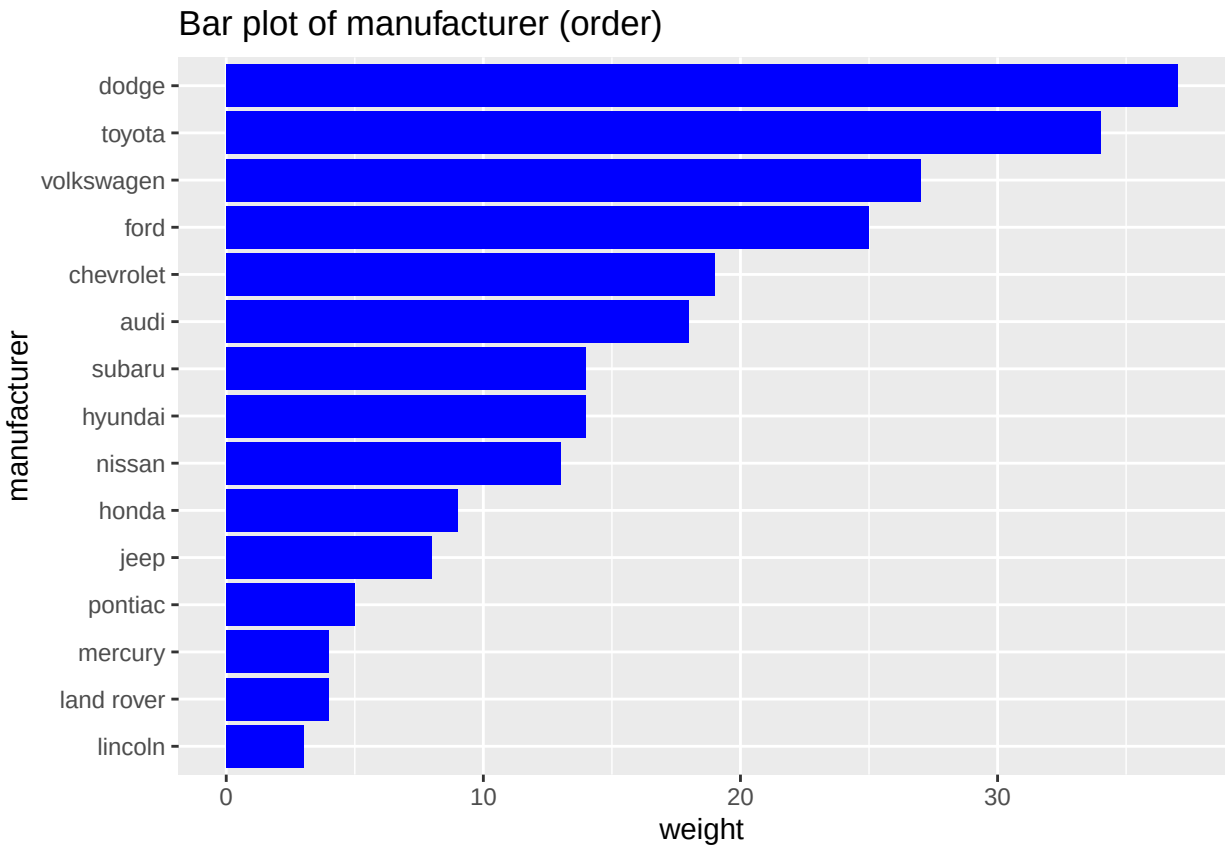
Exercise 3:

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?

```
# original plot
ggplot(mpg, aes(x= manufacturer)) +
  geom_bar(stat = "count", fill="blue") +
  ggtitle("Bar plot of manufacturer")
```




```
# Filp and order plot
mpg %>%
    group_by(manufacturer) %>%
    summarise(count = n()) %>%
    ggplot(aes(y = reorder(manufacturer,(count)), x = count)) +
        geom_bar(stat = 'identity',fill="blue" )+
    xlab("weight")+ylab("manufacturer") +
    ggtitle("Bar plot of manufacturer (order)")
```

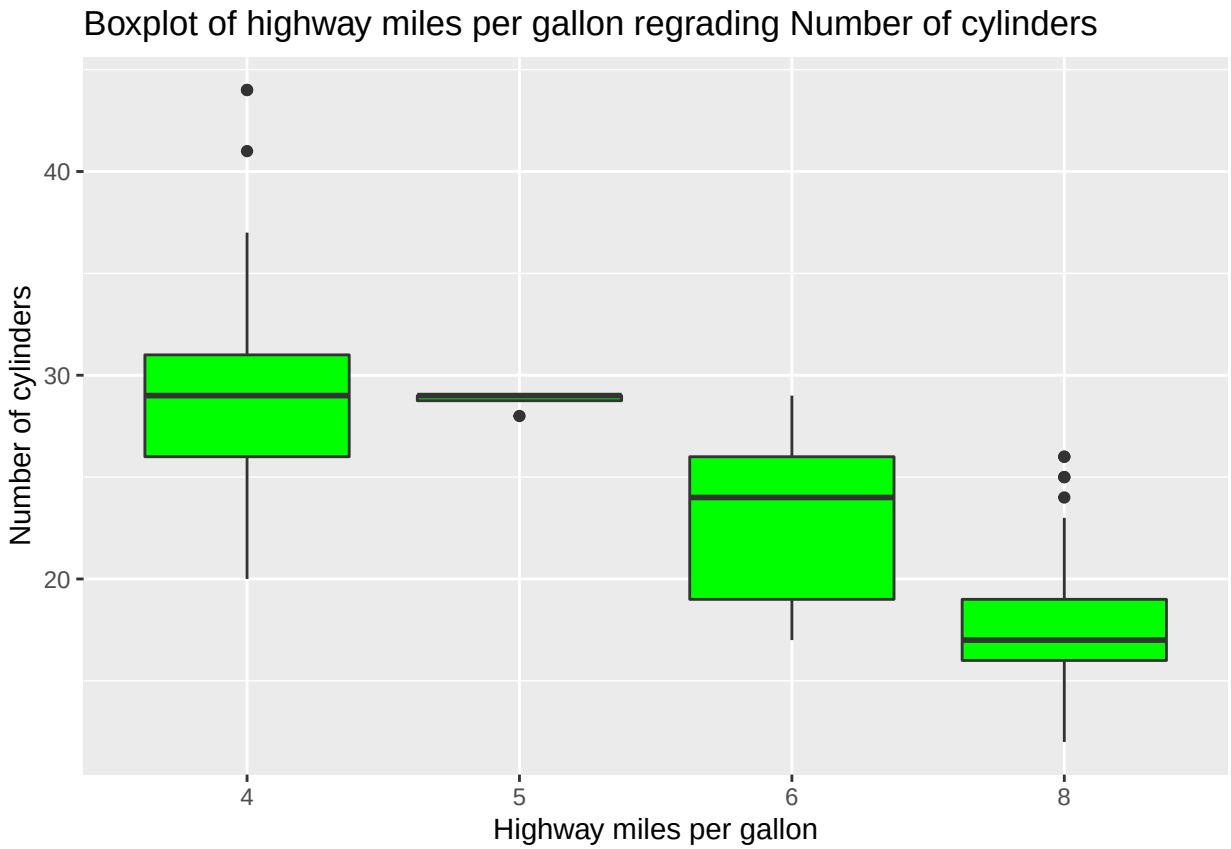


As we can see, the dodge produced the most cars and lincoin produced the least cars.

Exercise 4:

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

```
ggplot(mpg, aes( y=hwy, x=factor(cyl))) +
  geom_boxplot(fill="green") + xlab("Highway miles per gallon") +
  ylab("Number of cylinders") +
  ggtitle("Boxplot of highway miles per gallon regrading Number of cylinders")
```



From the plot we can general say the Highway miles per gallon has a negative correlation to the number of cylinders which imply a car has less number of cylinders will have a high highway miles per gallon.

Exercise 5:

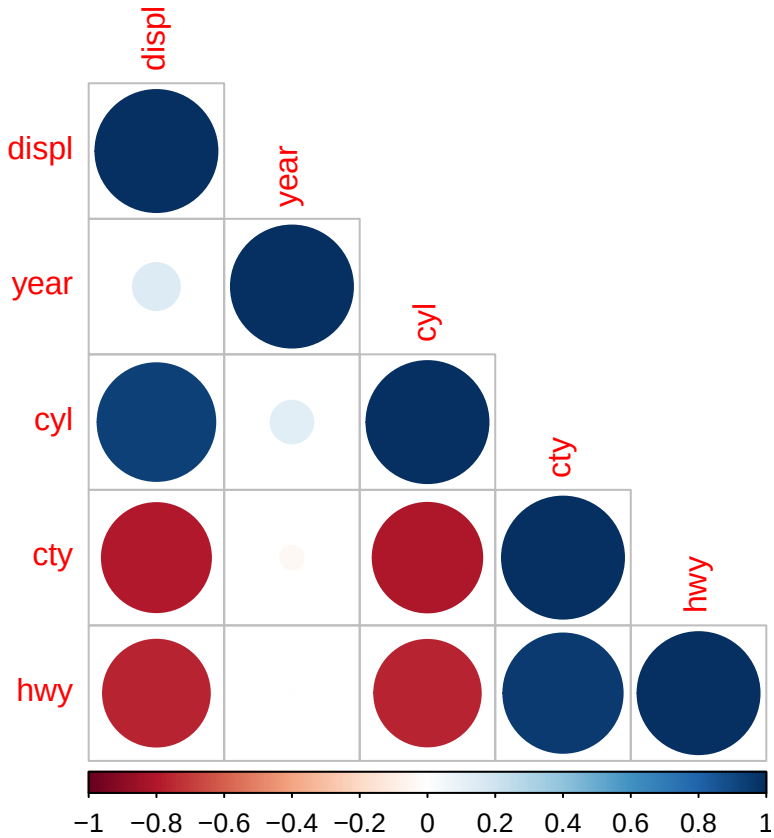
Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset. (Hint: You can find information on the package here.)

```
data <- subset(mpg, select = -c(manufacturer,model,trans,drv,fl,class))
data
```

```
## # A tibble: 234 x 5
##    displ  year   cyl   cty   hwy
##    <dbl> <int> <int> <int> <int>
##  1   1.8  1999     4    18    29
##  2   1.8  1999     4    21    29
##  3   2    2008     4    20    31
##  4   2    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    2008     4    20    28
## # ... with 224 more rows
```

```
M <- cor(data)

corrplot(M, type="lower")
```



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

- The highway miles per gallon and city miles per gallon are have negative relationship with displ. This is make sense since the higher displ will use more oil and decrease the miles per gallon.
- The highway miles per gallon and city miles per gallon are have negative relationship with cyl (the number of cylinders). This is make sense since the higher number of cylinders will use more oil and decrease the miles per gallon.
- The highway miles per gallon has positive relationship with city miles per gallon which is make sense.
- The number of cylinders has positive relationship with disp which is also make sense, more sylinders can provide higher power and cause higher disp.