PSTAT 131 HW1

Jenny Do (5669841)

April 03, 2022

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 describes the situation in which for each observation of the predictor measurement(s) x_i, $i = 1, ..., n$ there is an associated response measurement y_i. You know the true value of the outcome, therefore you can calculate an error rate. This allows us to do prediction, estimation, model selection and inference. Moreover, in supervised learning, cluster analysis is more clear as we know certain observations belong to a certain groups.

Unsupervised learning describes the situation in which for every observation $i = 1, ..., n$, we observe a vector of measurements x_i but no associated response y_i. In contrast to supervised learning, you don't know what the true value of the outcome is, resulting in a more uncertain error rate. This makes unsupervised learning more challenging as the goal is more ambiguous. In addition, regarding clusters it is also more challenging (especially difficult for overlapping points). As we don't know which observations belong to which groups,there is no viable way to identify clusters.

Other differences include supervised learning to be associated with linear regression, logistic regression, k-nearest neighbors, decision trees, random forests and support vector machine(s) while unsupervised learning is associated with PCA, k-means clustering and hierarchical clustering.

(Reference: ISLR pg. 37, Lecture 2 Slides 1-4)

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 refers to a quantitative response while a classification model involves a qualitative response. Note that whether the predictors are qualitative or quantitative is generally considered less important. It's all based on the outcome.

(Reference: ISLR pg. 28)

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 ML problems are the training and test mean squared error (MSE). Note that lowest training MSE does not equal lowest test MSE.

Two commonly used metrics for classification ML problems are the training and test error rate.

(Reference: Lecture 2 Slide 53)

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 chosen to best visually emphasize a trend in the data. For example, we can fit a line on a scatterplot.
- Inferential models: Inferential models asks the question of "what features are significant?". It aims to test theories, look at (possibly) casual claims, and state relationship between outcome and predictor(s).
- Predictive models: Predictive models as the question "what combo of features fits best?". It aims to predict Y with minimum reducible error. In addition, the focus with this model is not on hypothesis tests.

(Reference: Lecture 2, Slide 7)

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, also known as parametric, assumes a parametric form for f(ie., $\beta_0 + \beta_1 + ...$). Note that this will not match true unknown f. In contrast, empirically-driven, also known as non-parametric, has no assumptions about f and requires a larger number of observations. The similarities that these two include the fact that they can be overfitted and are more flexible (mechanistic = we add more parameters to make more flexible, empirically-driven = by default is more flexible).

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

In general, I believe that the mechanistic model is easier to understand. This is because we have parameters and we have made an assumption. Empirically-driven is more ambiguous as have we have no assumptions on the structure, denoted by f.

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

Mechanistic models tend to be generalizable which means they have a higher bias and a lower variance. In contrast, empirically-driven models will have a lower bias and a higher variance. Therefore, with both types of models, this bias-variance tradeoff is always present. The challenge is finding a method for which both the variance and bias are low (this is our goal, want this to be achieved simultaneously).

Question 6:

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

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?**

This question could be classified as predictive. This is because we are looking at all the features we have–the combination of features that produces the best fit. We look at all the data (all features for a single voter) and predict if those factors make a voter more likely to vote in favor of a candidate.

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

This question could be classified as inferential. This is because we are testing one feature (isolate one feature and try to see what happens when it changes): if a voter had personal contact with the candidate. The question is posed in a way that aims to test the theory that voters who have had personal contact with the candidate might be more or less likely to support the candidate.

(Reference: Lecture 2 Slide 6)

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.

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.

```r
# Load the needed data, mpg data set
data(mpg)
```

```
## Warning in data(mpg): data set 'mpg' not found
```

```r
# Load tidyverse, ggplot2 and corrplot (needed in problem 5)
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()
```

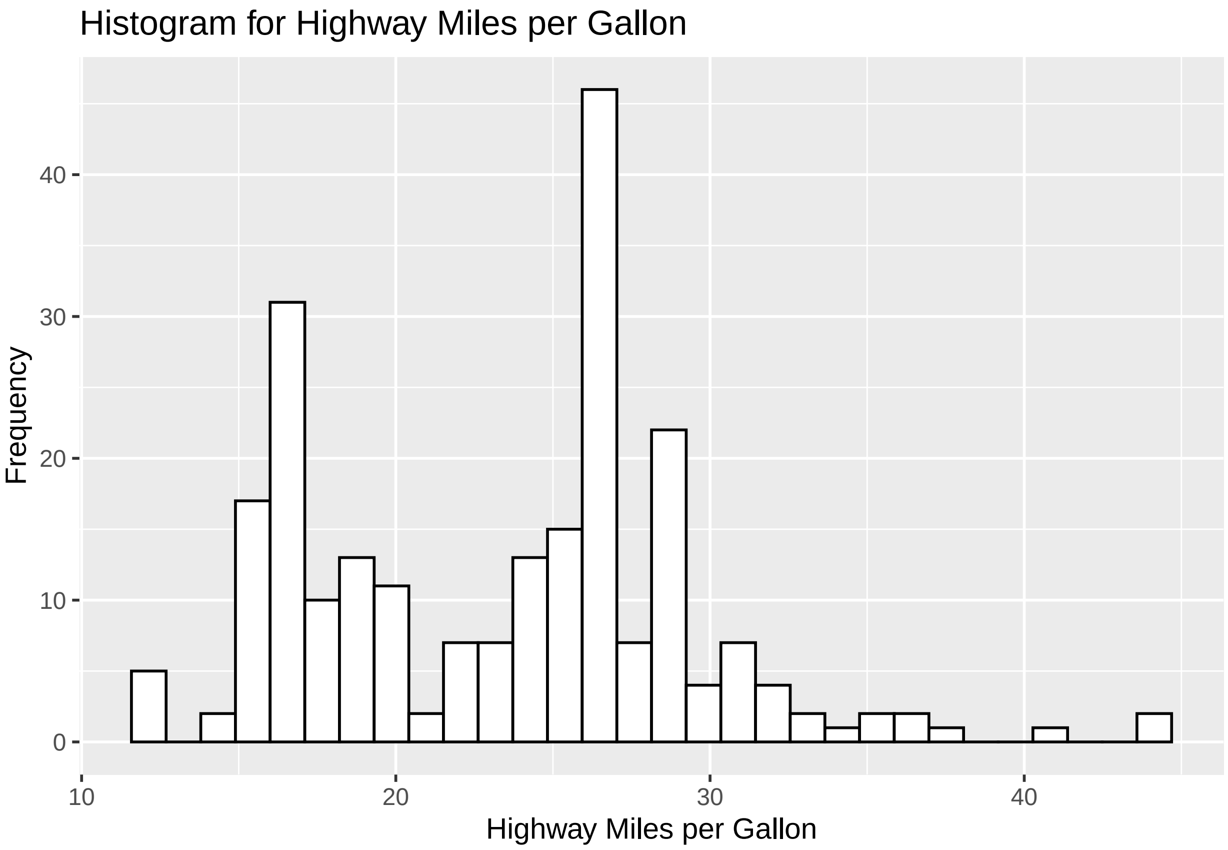
```r
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

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.

```r
# use ggplot2 to create a histogram for hwy
p1<-ggplot(mpg, aes(x=hwy)) +
  geom_histogram(bins = 30L, color="black", fill="white") +
  labs(x ="Highway Miles per Gallon", y = "Frequency") +
 ggtitle("Histogram for Highway Miles per Gallon")
p1
```



Answer(Exercise 1): After creating a histogram using the variable hwy (denoting highway miles per gallon), we can see that this is a bimodal (two peaks), asymmetrical histogram.

In addition, we can see that there are a couple observations (greater than 40 highway miles per gallon) that are higher than all the other observations. These could be seen as outliers, fuel efficient cars that equate to a higher amount of highway miles per gallon.

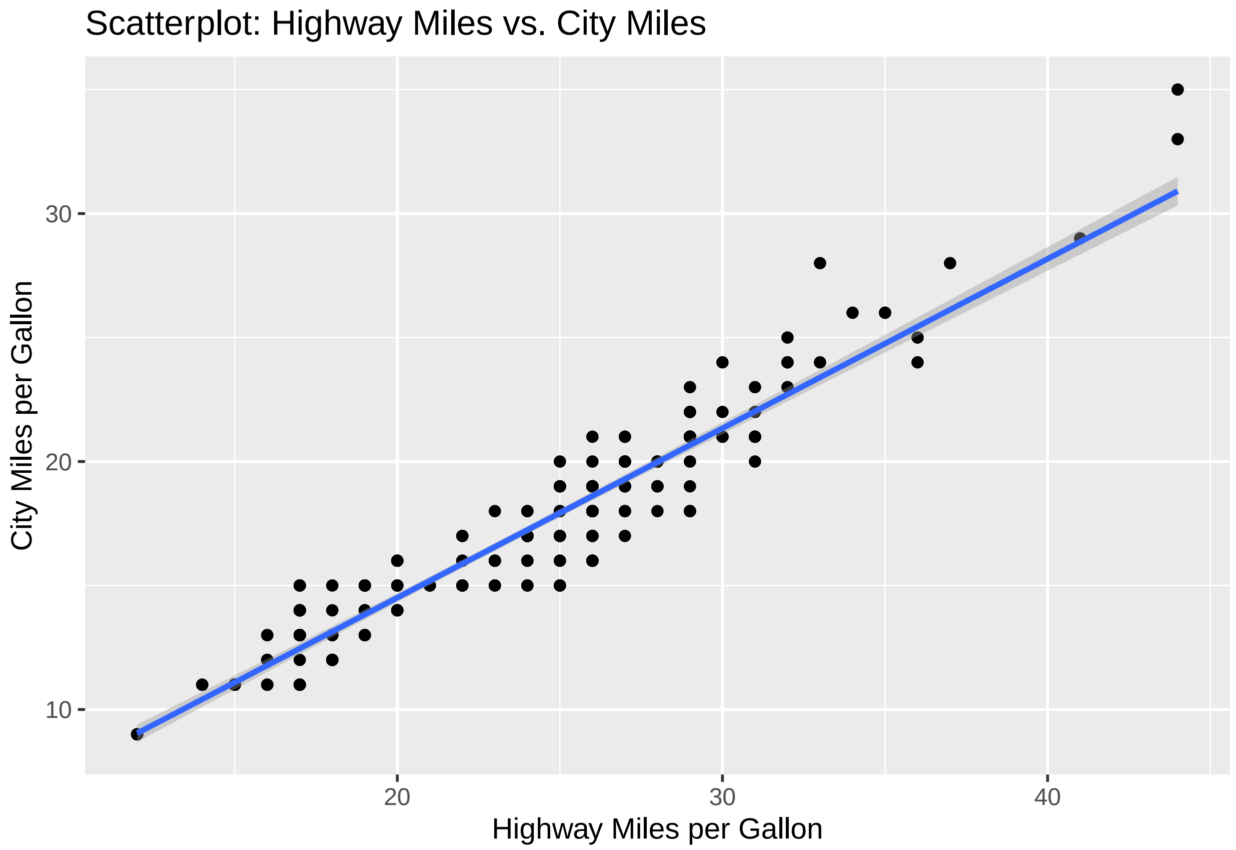
On the other hand, we can see one observation (on the lower end of the 10 to 20 highway mpg range) that are lower than all the other observations. This could also be seen as an outlier, fuel inefficient cars that equate to a lower amount of highway miles per gallon.

(Reference: Describing the Shapes of Histograms - https://www.youtube.com/watch?v=boVJzMXjcV8)

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?

```r
# Use ggplot
# Want to add a trend line for better interpretation
p2 <- ggplot(mpg, aes(x=hwy, y=cty)) + geom_point() +
  labs(x ="Highway Miles per Gallon", y = "City Miles per Gallon") +
  ggtitle("Scatterplot: Highway Miles vs. City Miles")+
  geom_smooth(method=’lm’, formula= y~x)
p2
```



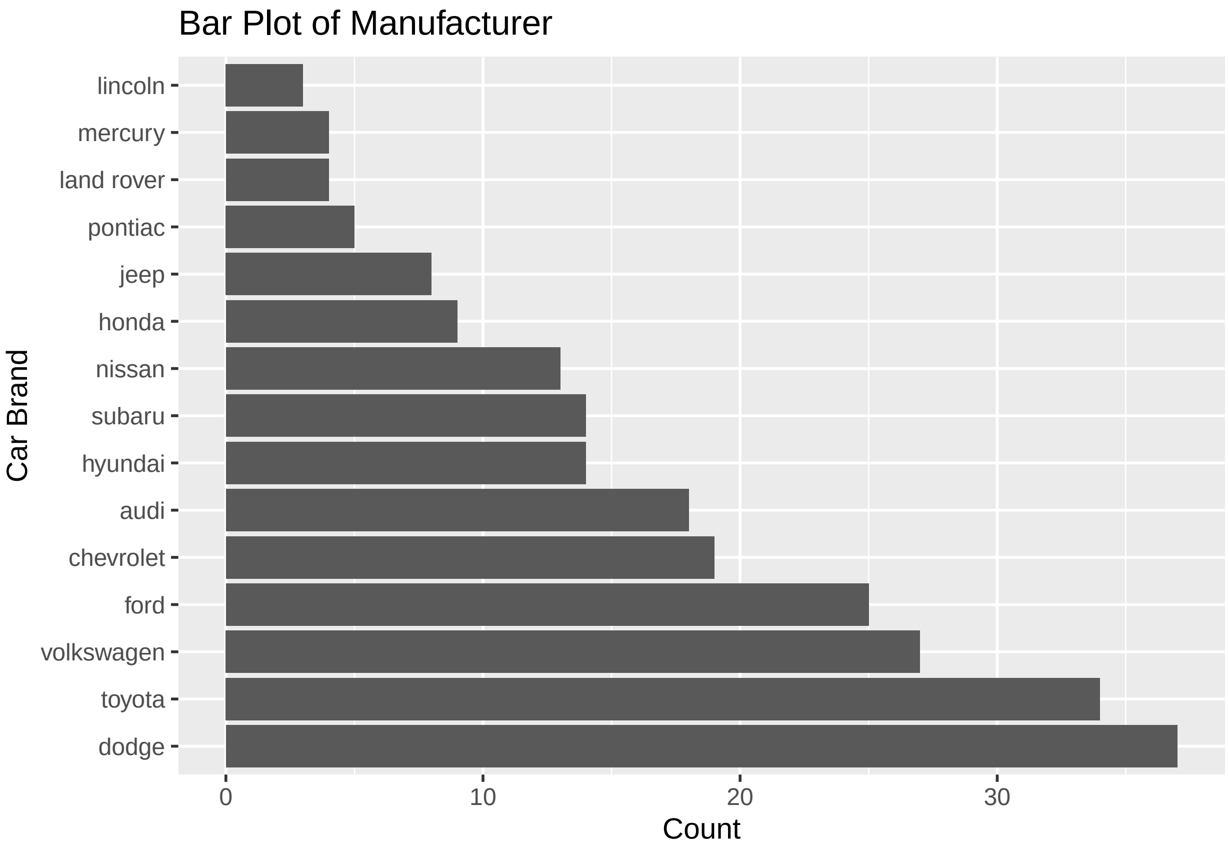
Answer(Exercise 2): After creating a scatterplot (hwy on the x-axis and cty on the y-axis), we can see that there is a relationship (positive correlation) between the two variables. This means that as highway miles per gallon increases, city miles per gallon increases. In addition, we note that there might be some outliers (past 40 and on the lower end of 20 - this matches what we saw in Exercise 1 above).

(Reference: https://www.texasgateway.org/resource/interpreting-scatterplots)

##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?**

Note: As the prompt did not specify if order ascending or descending, we will be ordering the bars by increasing height.

```
# Use ggplot to make a bar plot
# use reorder to order the bars by height
p3 <- ggplot(data = mpg,
             aes(x=reorder(manufacturer, -table(manufacturer)[manufacturer])))+
  coord_flip()+ geom_bar()+ labs(x ="Car Brand", y = "Count") +
  ggtitle("Bar Plot of Manufacturer")
p3
```



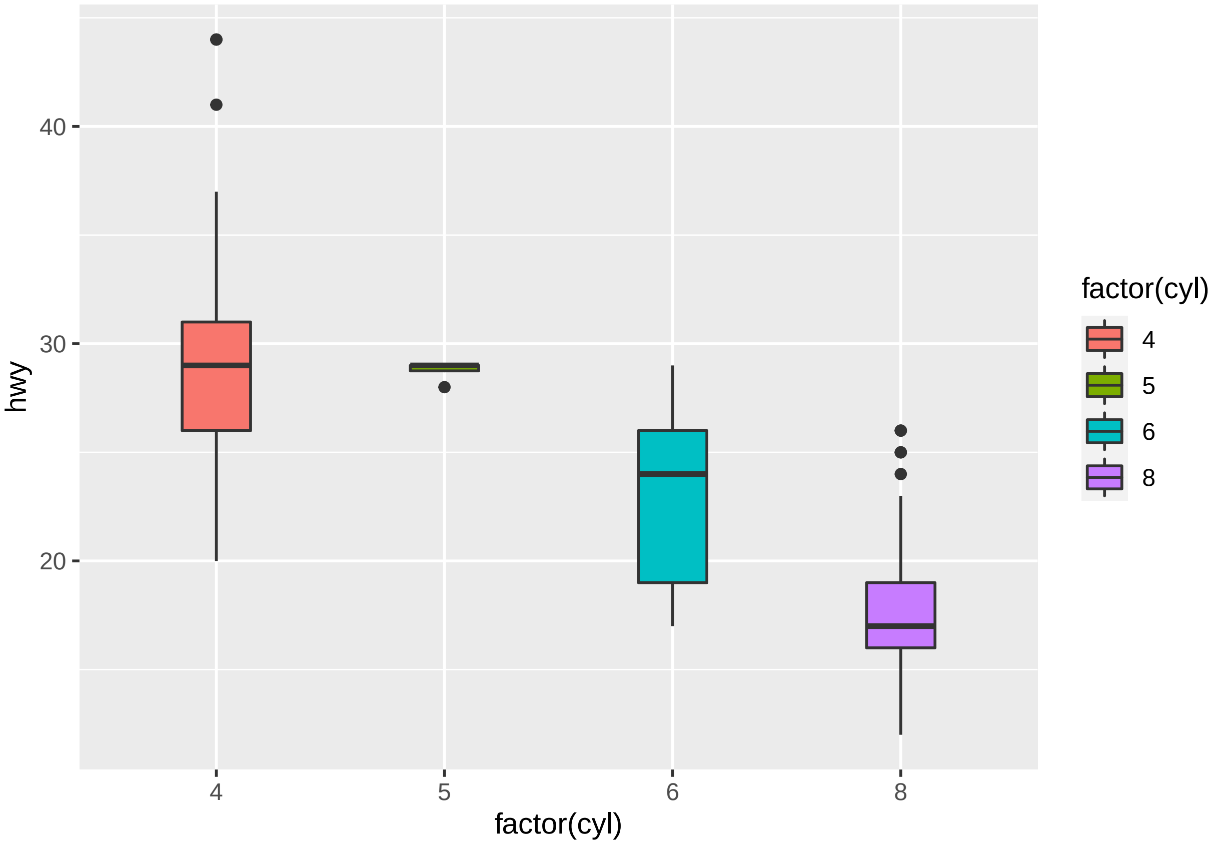
Answer(Exercise 3): After ordering the bars by height, we can see that manufacturer Dodge produced the most cars while manufacturer Lincoln produced the least cars.

(Reference (for how to order the bars): https://stackoverflow.com/questions/5208679/order-bars-in-ggplot2-bar-graph)

Exercise 4:

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

```
# Note there are no cars with 7 cylinders
p4 <- ggplot(mpg, aes(x = factor(cyl), y = hwy, fill = factor(cyl))) +
  geom_boxplot(width = 0.3)
p4
```



Answer(Exercise 4): After creating a box plot of hwy, grouped by cyl, we do see a pattern. As we increase the amount of cylinders in the car, there is a decrease in highway miles per gallon. This makes sense; when a car has more power (more cylinders), it tends to be less fuel efficient (a decrease in miles per gallon). In addition, we note that there are some outliers seen in 4 cylinder, 5 cylinder and 8 cylinder.

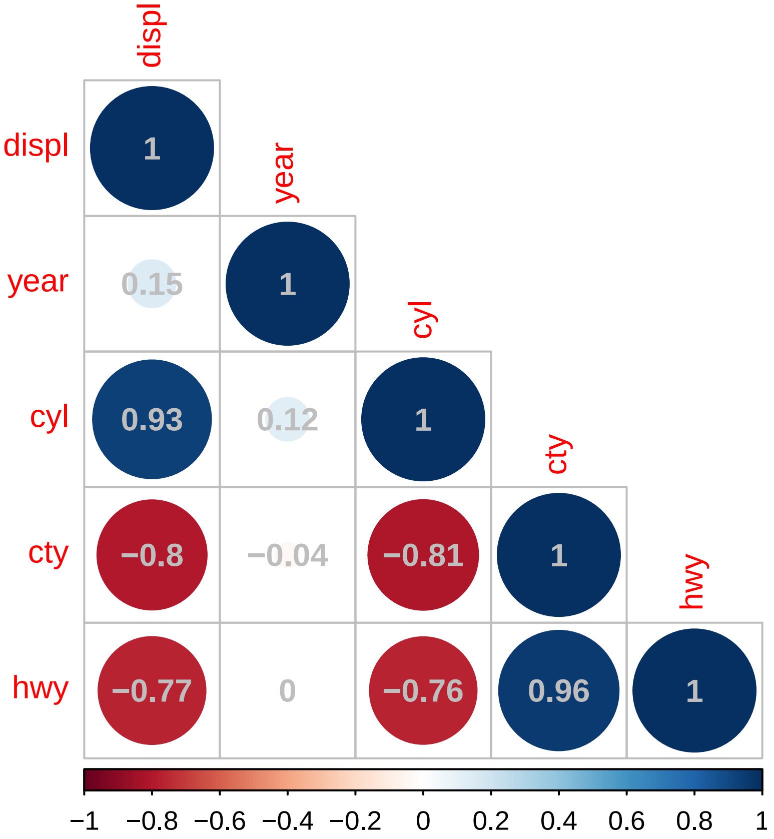
(Reference: https://r-coder.com/boxplot-r/)

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?

```
# as cor function requires values to be numeric:
# use subset function to remove all the variables that are not of numeric type
mpg_drop <- subset(mpg, select = -c(manufacturer, model, trans, drv, fl, class))
# Define corr
corr_val <- cor(mpg_drop)
# Use the corrplot function to make a lower triangle correlation matrix of mpg
# with variables manufacturer, model, trans, drv, fl, class excluded
corrplot(corr_val, type = "lower", addCoef.col ="gray")
```



Answer(Exercise 5): After using the corrplot package to make a lower triangle correlation matrix of mpg (with all non numeric variables excluded), we note the following:

- Correlation coefficients along the diagonal are all equal to 1. This makes sense because each variable is perfectly correlated to itself.
- Positively correlated:
  - The correlation between variable displ (denoting engine displacement, in litres) and cyl (denoting number of cylinders) is 0.93, which indicates that they are strongly positively correlated. This makes sense, engine displacement means an engine can move more air and fuel giving it the potential to make more power. The more cylinders an engine possesses, the faster power can be generated. These two factors are directly related.
  - The correlation between variable cty (denoting city miles per gallon) and hwy (denoting highway miles per gallon) is 0.96, which indicates that they are strongly positively correlated. More city miles are strongly related to higher amounts of highway miles. This makes sense; there is constant stopping and starting on city roads and frequent changes of speeds. If this figure is high, the car should have better miles on the highway which has less stopping and re-acceleration.
- Negatively correlated:
  - The correlation between variable displ (denoting engine displacement, in litres) and cty (denoting city miles per gallon) is -0.8, which indicates that they are strongly negatively correlated. More engine displacement is associated with less city miles. This is makes sense; engine displacement means an engine can move more air and fuel giving it the potential to make more power. A higher engine displacement equates to a higher fuel consumption, meaning less city miles per gallon.
  - The correlation between variable displ (denoting engine displacement, in litres) and hwy (denoting highway miles per gallon) is -0.77, which indicates that they are strongly negatively correlated.

    More engine displacement is associated with less highway miles. Similar to the result above, a higher engine displacement equates to a higher fuel consumption, meaning less highway miles per gallon.
  - The correlation between variable cyl (denoting number of cylinders) and cty (denoting city miles per gallon) is -0.81, which indicates that they are strongly negatively correlated. More cylinders are associated with less city miles. This makes sense; the more cylinders an engine has, the faster power can be generated in the engine. More cylinders in an engine equates to a higher fuel consumption, meaning less city miles per gallon.
  - The correlation between variable cyl (denoting number of cylinders) and hwy (denoting highway miles per gallon) is -0.76, which indicates that they are strongly negatively correlated. More cylinders are associated with less highway miles. Similar to the result above, more cylinders in an engine equates to a higher fuel consumption, meaning less highway miles per gallon.

- I found it surprising that engine displacement, number of cylinders, city miles per gallon and highway miles per gallon has little to no correlation to the year of manufacture. I thought that as the years progressed, there would be a more engine displacement, more cylinders added to new cars or the cars would become more fuel efficient with time (resulting in more city and highway miles per gallon).

(Reference: https://www.statology.org/how-to-read-a-correlation-matrix/)