PSTAT231 HW1 CHENG YE

2022-10-02

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

##From the page 26 of book An Introduction to Statistical Learning with Applications in R.Supervised learning is a machine learning approach that's defined by its use of labeled datasets. These datasets are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately.Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention. ##Difference:Supervised learning uses labeled input and output data, while an unsupervised learning algorithm does not. In supervised learning, the goal is to predict the value of an outcome measure based on a number of input measures.In unsupervised learning, there is no outcome measure, and the goal is to describe the associations and patterns among a set of input measures. (Such conclusion is derived from the page 26 of book An Introduction to Statistical Learning with Applications in R)

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

##Difference:Problems with a quantitative response are classified as regression problems, while those involving a qualitative response are often referred to as classification problems.(Such conclusion is derived from the page 28 of book An Introduction to Statistical Learning with Applications in R)A regression model could help to predict a continuous quantity, while a classification model tends to predict discrete class labels.In the regression model, Y is quantitative. In the Classification model, Y is qualitative. (This idea is concluded from the first lecture, slide 33)

#Question 3:Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.

##For regression ML problems, Mean Square Error(MSE) and Mean Absolute Error(MAE) are two commonly used metrics ##For classification ML problems, Error Rate and K-nearest neighbors (KNN) classifier are two commonly used metrics(Such conclusion is derived from the page 37 of book An Introduction to Statistical Learning with Applications in R)

#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 (This idea is concluded from the first lecture, slide 39) ##Inferential models:Aim is to test theories,(Possibly)test causal claims, test state relationship between outcome & predictor(s) (This idea is concluded from the first lecture, slide 39) ##Predictive models:Aim is to predict Y with minimum reducible error,not focused on hypothesis tests (This idea is concluded from the first lecture, slide 39)

#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? ## The mechanistic predictive models predict the future based on a theory, while the empirically-driven predictive models would develop a theory through studying real world events. As for differences, mechanistic predictive models assume a parametric form for f (i.e. $\beta 0 + \beta 1" + \cdots$), won't match true unknown f, and can add parameters which means more flexibility. empirically-driven models make no assumptions about f, require a larger number of observations, and is much more flexible by default. Both of them share the similarity of overfitting.(This idea is derived from the first lecture, slide 38)

#In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice. ## I think the mechanistic model is easier to understand in general. Because mechanistic models are based on theories while empirically-driven models require real world data to establish a theory. In addition, a mechanistic model assume a parametric form for f, making it obivous to model (and is more direct to perceive than that of empirical-driven models).

#Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models. ##The bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters (This idea is derived from Wikipedia). Both the use of mechanistic/ empirically-driven models would produce errors/ inaccurate predictions. By bias-variance tradeoff, we need to choose the model that both accurately captures the regularities in its training data, but also generalizes well to unseen data.

#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.

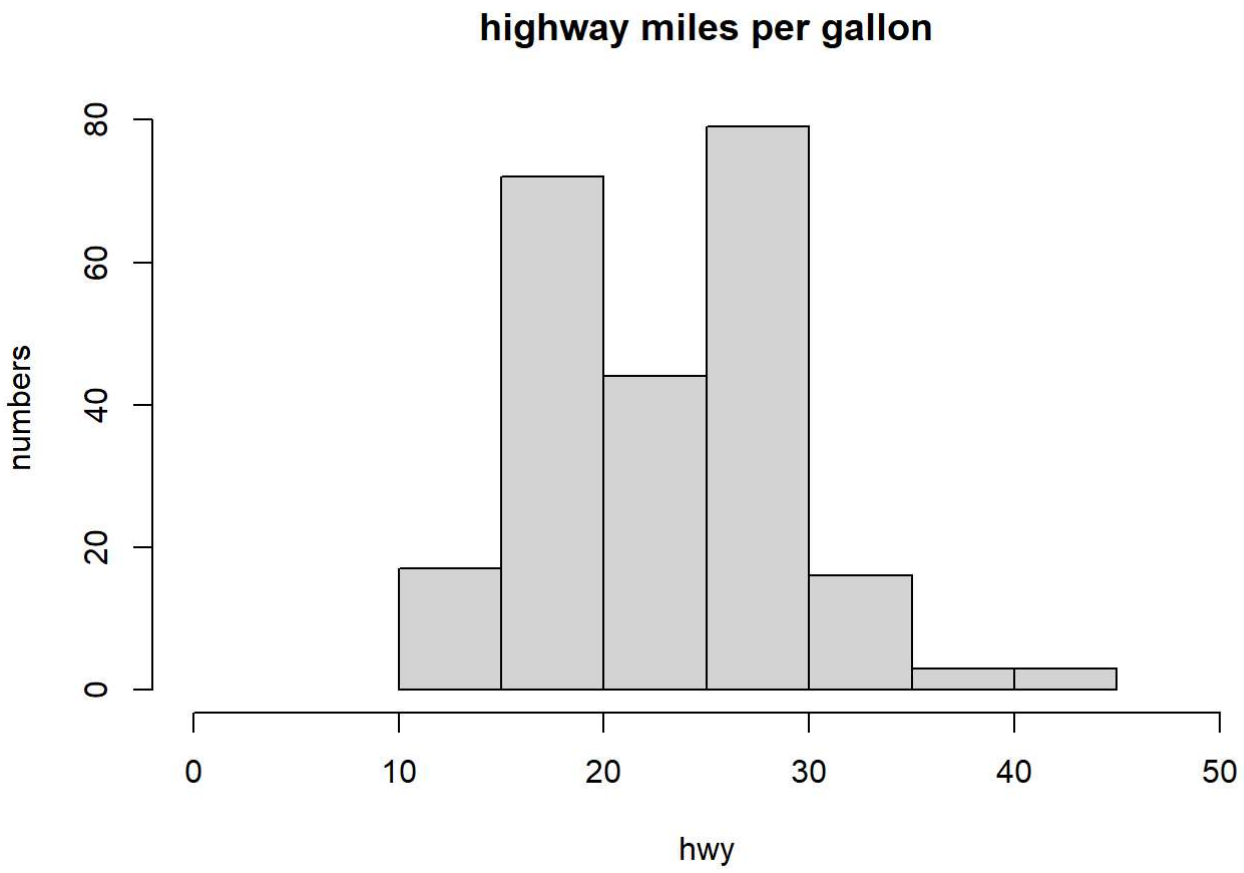
##By definition, predictive models: Aim is to predict Y with minimum reducible error,not focused on hypothesis tests. Inferential models:Aim is to test theories,(Possibly)test causal claims, test state relationship between outcome & predictor(s). Hence the first question is predictive as it wishes to predict the voter's choice of candidate. The second question is inferential as it aims to test whether voter's likelihood to support the candidate would be altered by personal contact with the candidate.

Exploratory Data Analysis #Exercise 1:

```
library(ggplot2)
data("mpg")
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~
```

```
hist(mpg$hwy, main="highway miles per gallon", breaks=7, xlim = range(0:50), xlab="hwy",ylab= "n
umbers")
```



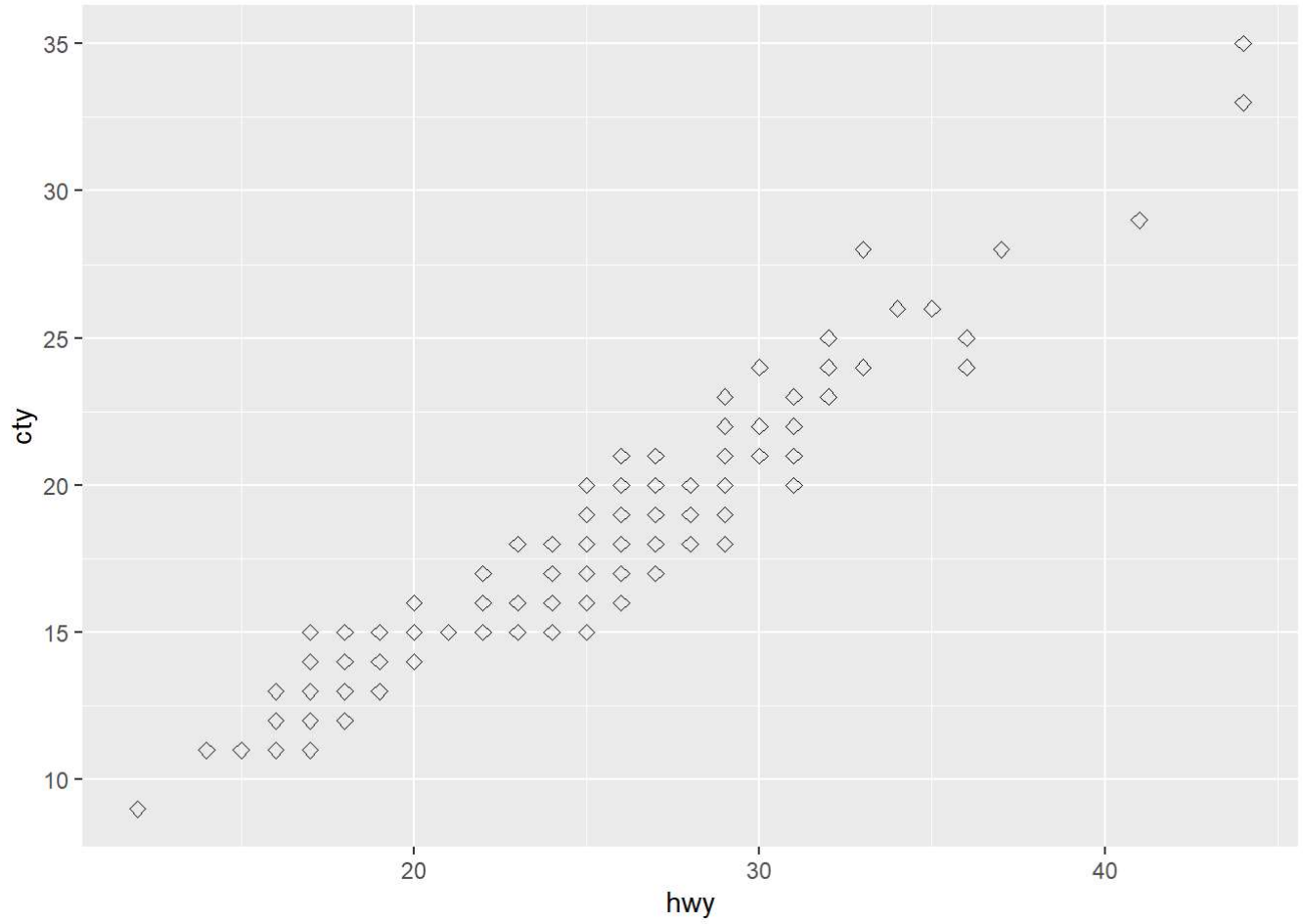

```
# From the dataset and graph, we could see that the majority of cars are within the 15~30 highwa
y miles per gallon range and a few cars are within the 10~15 highway miles per gallon range as w
ell as the 30~45 highway miles per gallon range.
```

#Exercise 2:

```
library(ggplot2)
data("mpg")
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~
```

```
ggplot(mpg, aes(x = hwy, y = cty)) + geom_point(size = 2, shape = 23)
```



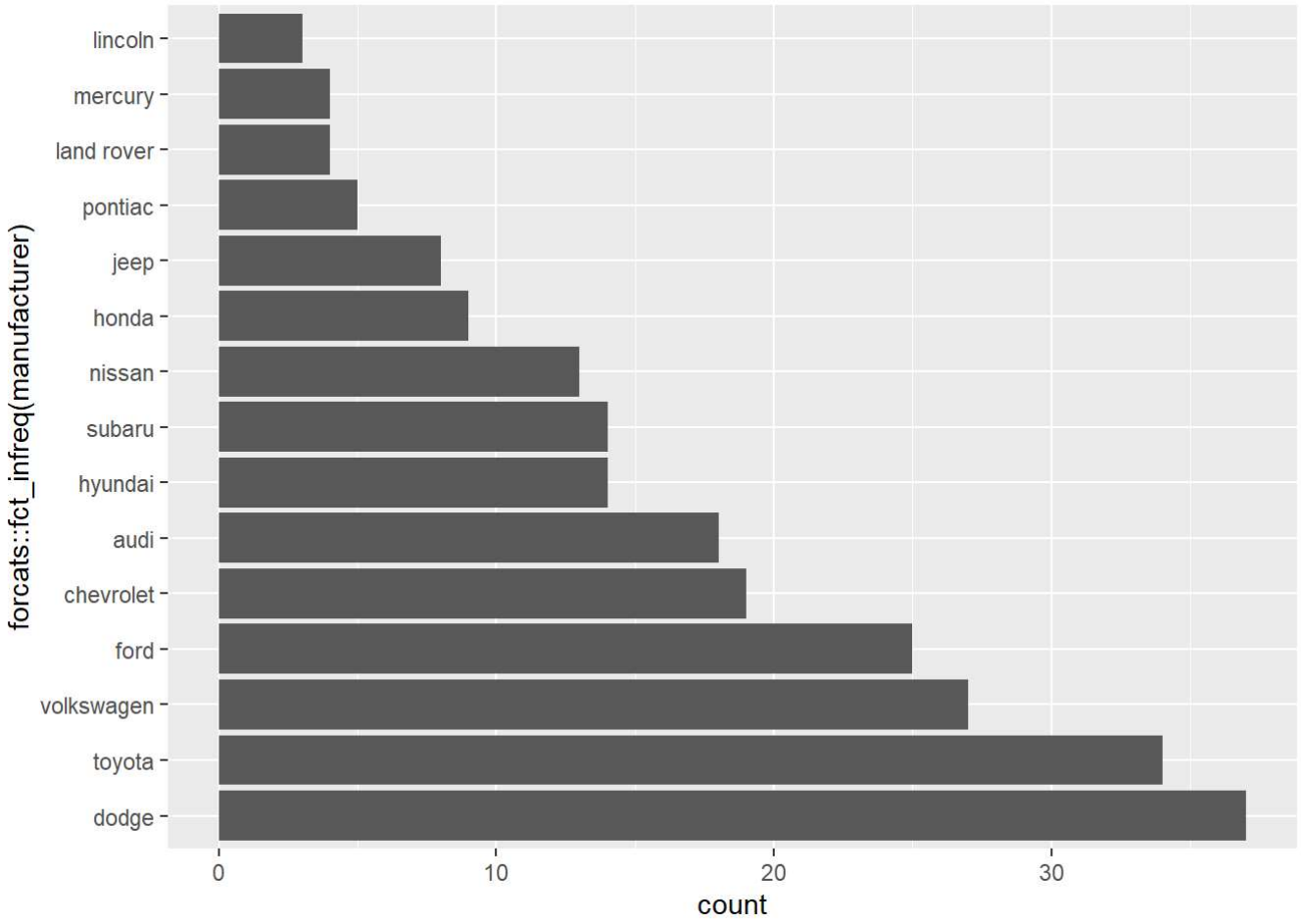

```
# From the dataset and graph, we could observe that as hwy increases, so does cty increases. Hen
ce there is a relationship between hwy and cty. This means that when hwy is big, cty is also bi
g; when hwy is smaill, so is cty.
```

#Exercise 3:

```
library(ggplot2)
data("mpg")
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~
```

```
p <- ggplot(mpg,aes(x = forcats::fct_infreq(manufacturer)))
p+ geom_bar() + coord_flip()
```

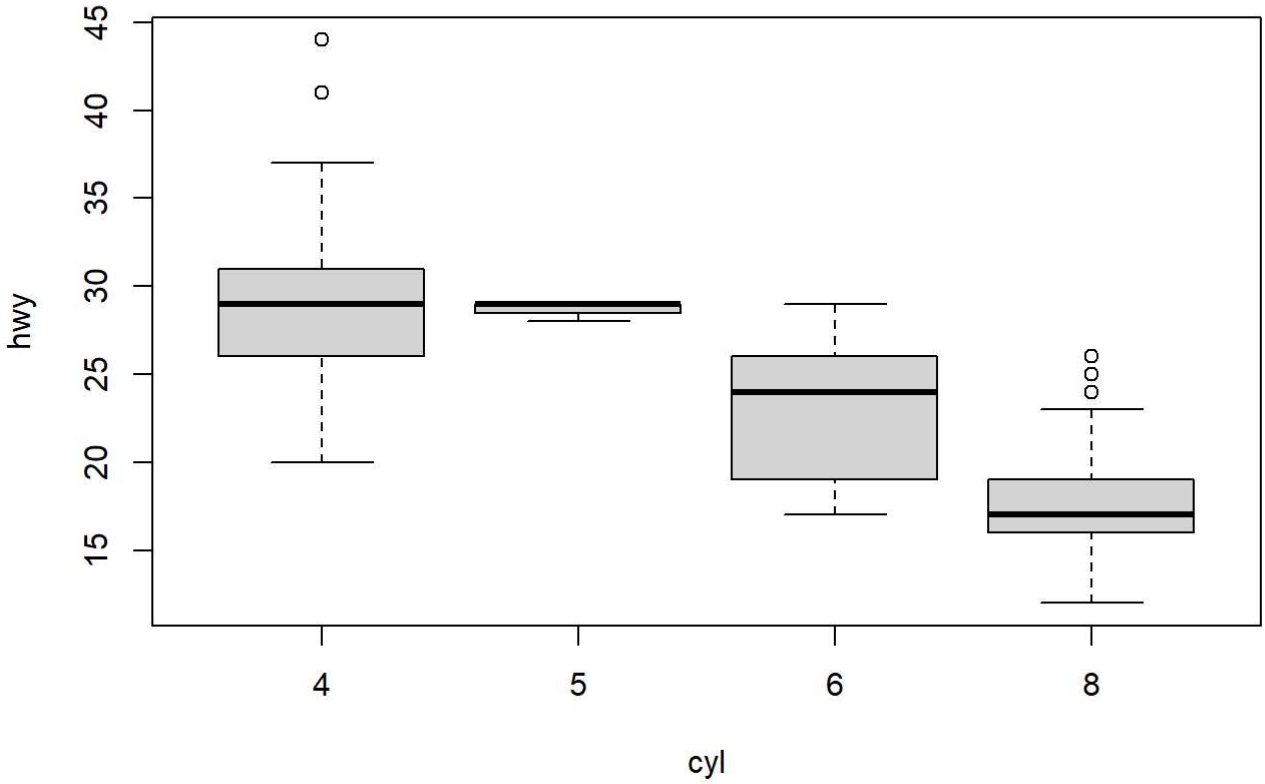


#Exercise 4:

```
library(ggplot2)
data("mpg")
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~
```

```
boxplot(hwy ~ cyl, data = mpg)
```




```
# From the dataset and graph we could see a obvious pattern: as cyl increases, hwy decreases.
```

#Exercise 5:

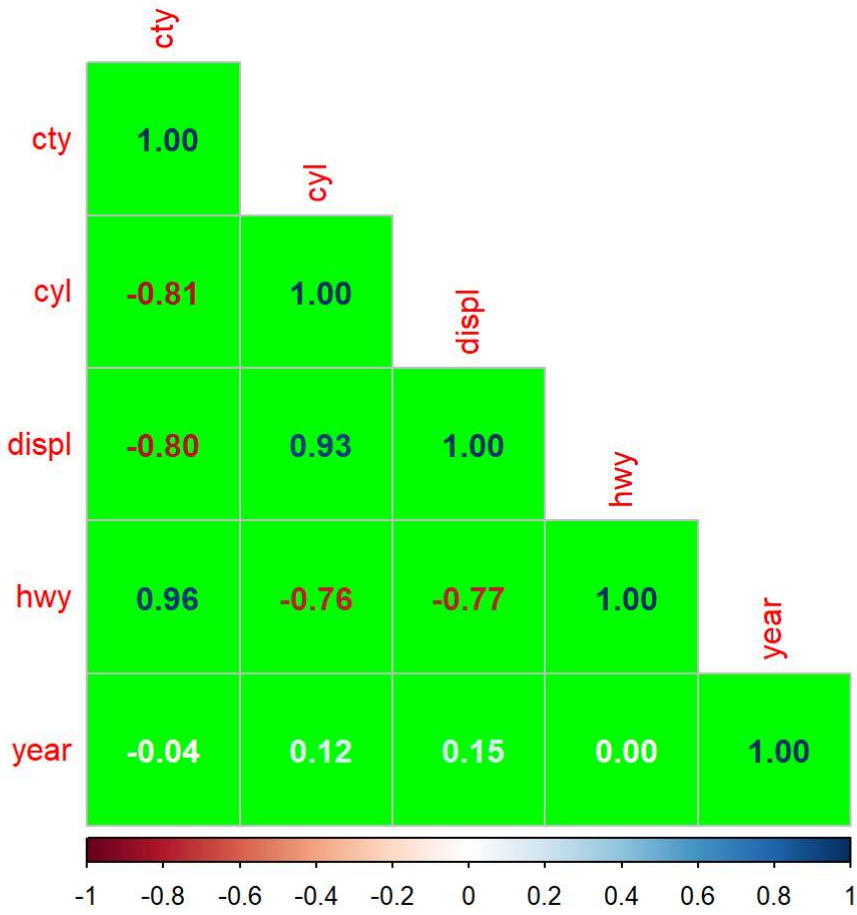
```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

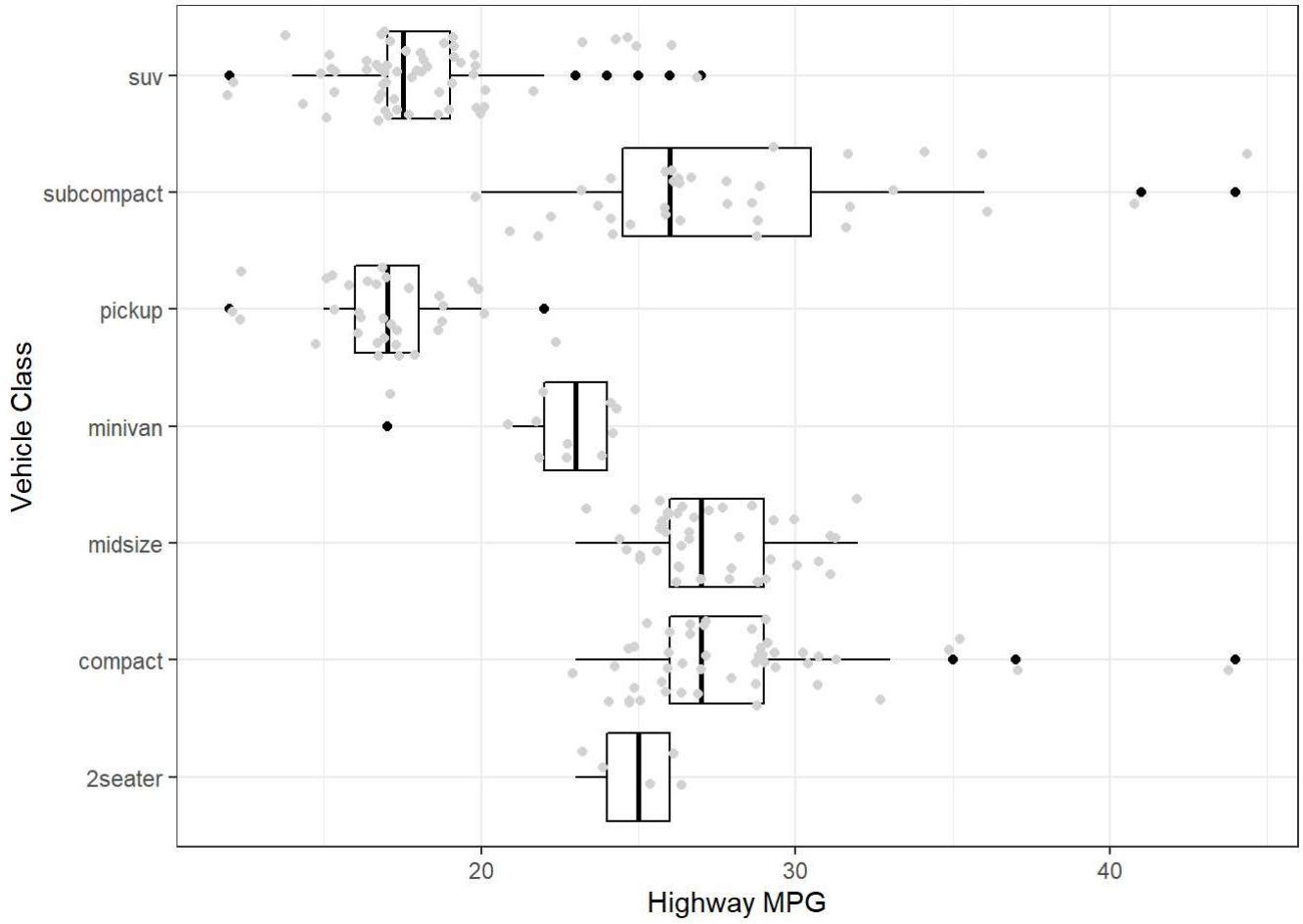
```
M<- ggplot2::mpg %>% select_if(is.numeric) %>% cor(.)
corrplot(M, method = 'number',bg = "green", order = 'alphabet', type = 'lower')
```




```
# From the dataset and graph we could see that displ and cyl,  hwy and cty,  year and cyl, year
 and displ have positive relationships; cyl and cty, displ and cty, hwy and cyl, hwy and displ,
 year and cty have negative relationships. These relationships make sense to me as we observed e
arlier that hwy and cty has some correlation. The data sets that surprised me is that year and c
ty, year and cyl, and year and displ have slight relationships compared to other relationships,
 and year has no relationship with hwy.
```

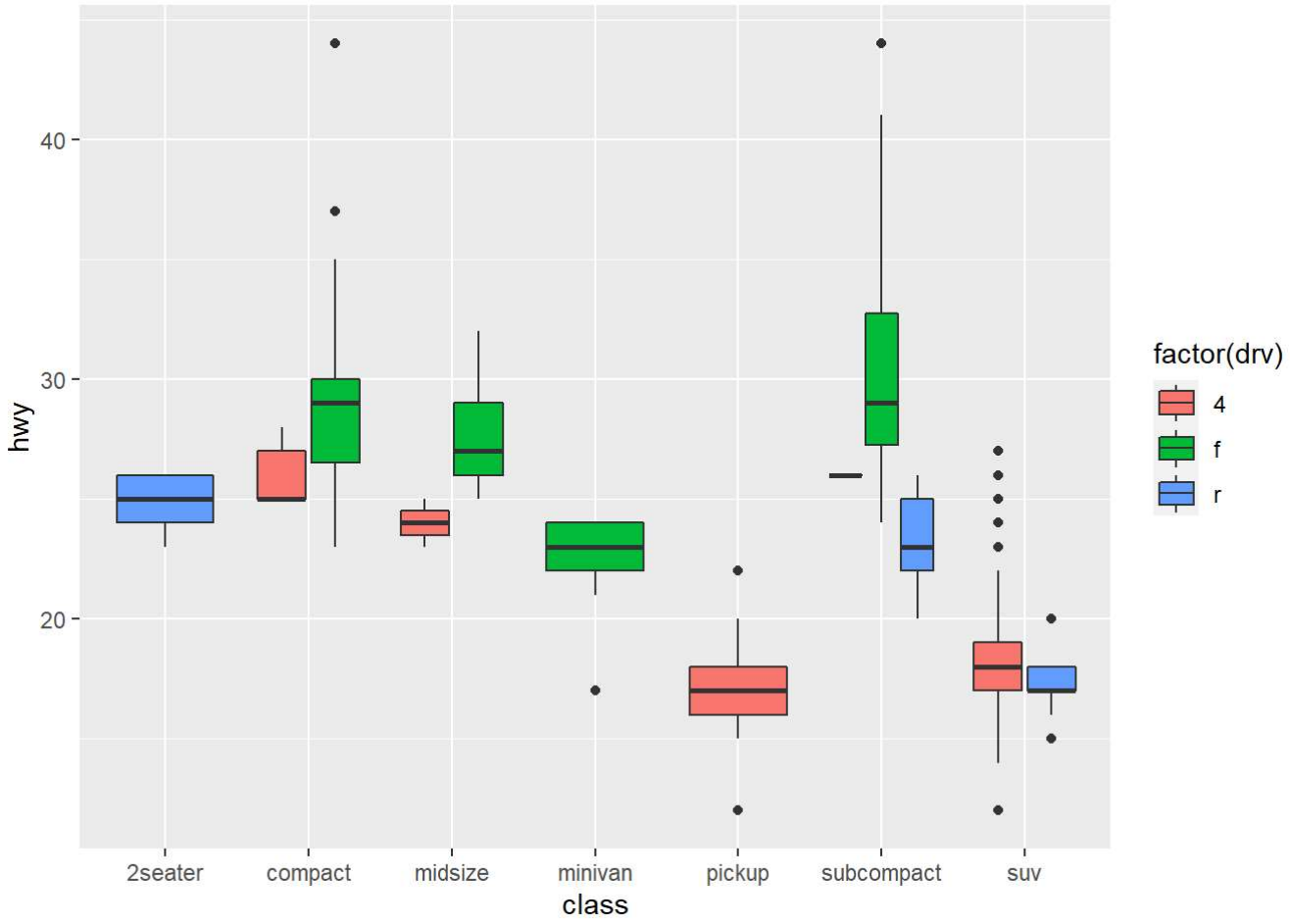
#Exercise 6:

```
library(ggthemes)
library(ggplot2)
ggplot(data = mpg, aes(x=hwy, y=class))+
  geom_boxplot(color = "black")+geom_point(position = "jitter", color = "light gray") + theme_bw
() + labs(x = "Highway MPG", y = "Vehicle Class")
```



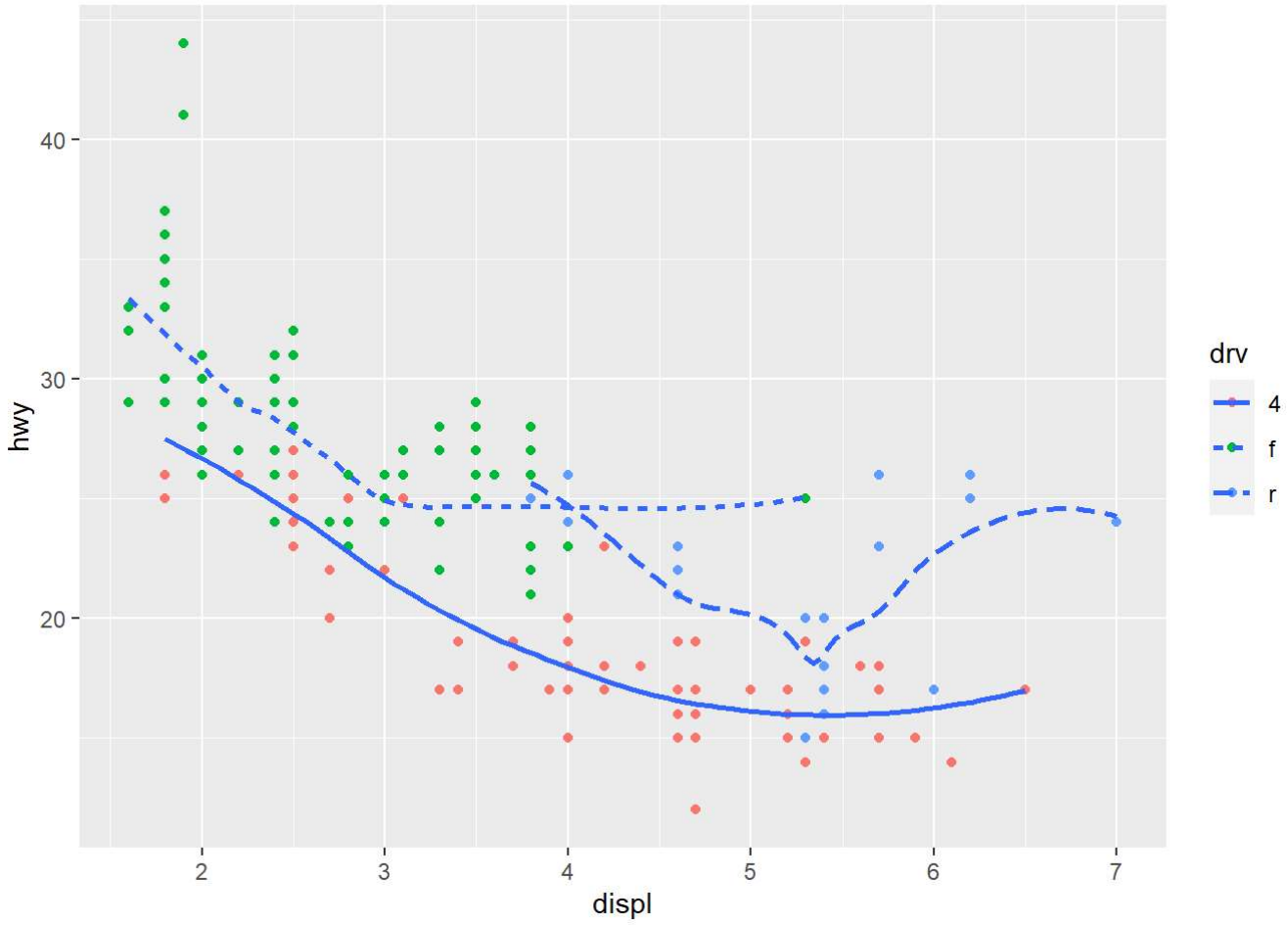
#Exercise 7:

```
library(ggthemes)
library(ggplot2)
k <-ggplot(mpg, aes(x = class, y = hwy, fill = factor(drv))) +geom_boxplot()
k
```



#Exercise 8:

```
library(ggthemes)
library(ggplot2)
k <-ggplot(mpg, aes(x = displ, y = hwy))
k + geom_point(aes(color = drv))+geom_smooth(formula = y~x, method = "loess", se = FALSE, aes(li
netype = drv))
```


hwy
40
30
20
2
3
4
5
6
7
displ
drv
4
f
r