Homework 1

2022-09-29

Question 1 Supervised and unsupervised learning. In Supervised learning, there is an associated response variable for each of the predictors. In unsupervised learning, there is no associated response variable for each observation. Unsupervised learning is used to understand the relationships between the variables.

Question 2 A regression model uses quantitative data and quantitative response, while classification model uses qualitative data and response.

Question 3 Regression ML Problems: linear regression, multiple linear regression Classification ML Problems: logistic regression, neural networks

Question 4 Descriptive models: models that best visually emphasize a trend in data Inferential models: the goal is to predict the response variable with minimum error Predictive models: states relationship between response and predictors; test theories; asks which features are significant

Question 5 Mechanistic predictive models usually have assumptions about the relationship between the outcome and predictors, and are generally less flexible than empirically-driven models. Empirically driven models usually require a larger number of observations and are much more flexible. They are have no assumptions. A mechanistic model is generally easier to understand because we are predicting future events based on assumptions about the data.

Bias-variance tradeoff: Using more flexible statistical methods will result in higher variance and lower bias, and more inflexibility means low variance, and higher bias.

Question 6 A. This question is predictive since they are predicting which candidate a voter will choose based on the voter's data/characteristics. B. This question is inferential because its interested in the relationship between the outcome and predictors.

Exploratory Data Analysis

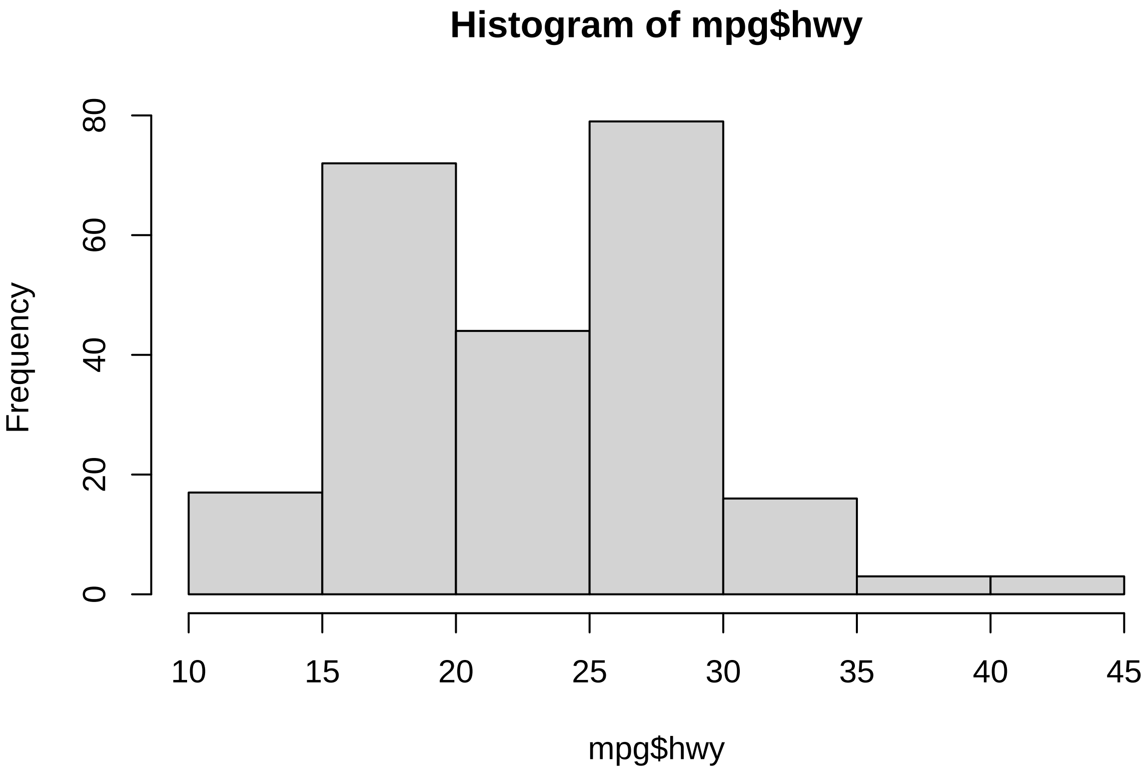
Downloading libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## 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
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

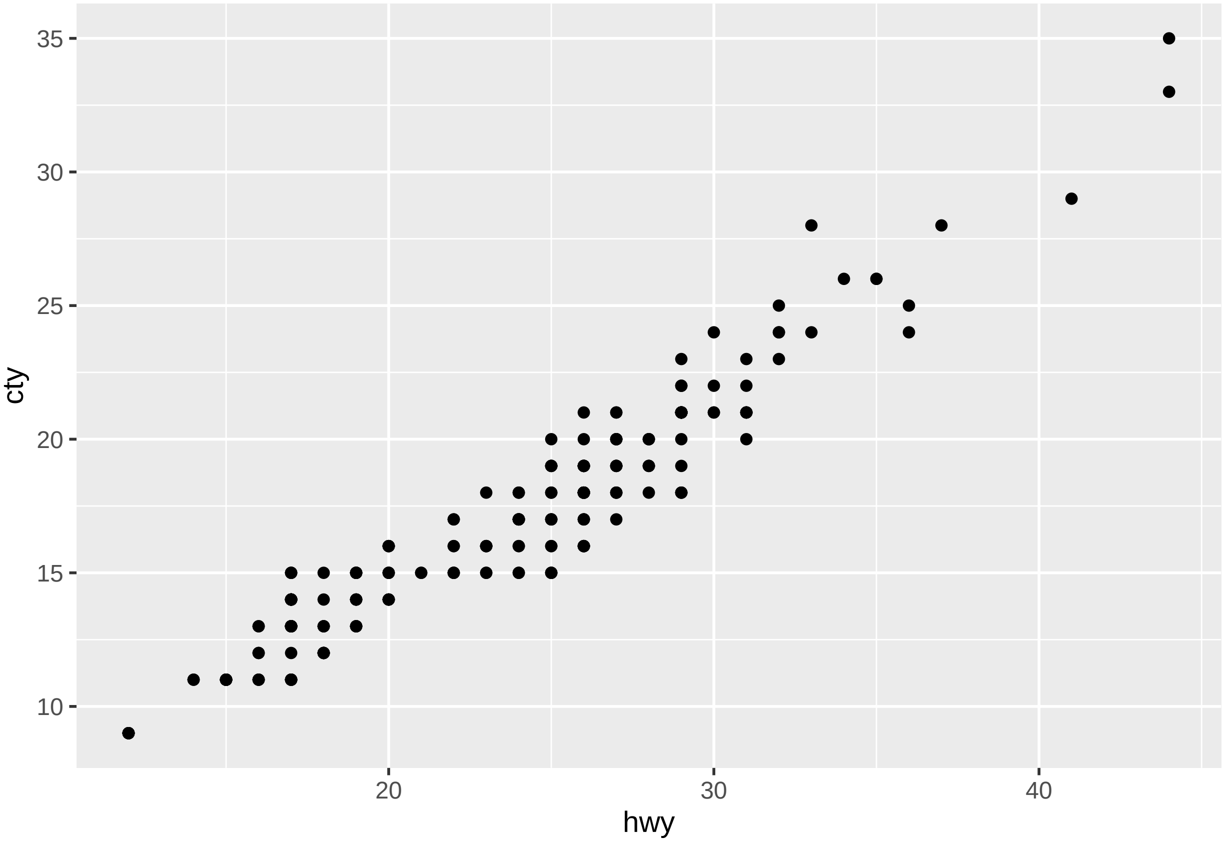
```
hist(mpg$hwy)
```



Question 1

The histogram seems to be rightly skewed, which means that there are higher frequencies of lower highway mileage. Highway mileage between 25-30 mpg have the highest frequency.

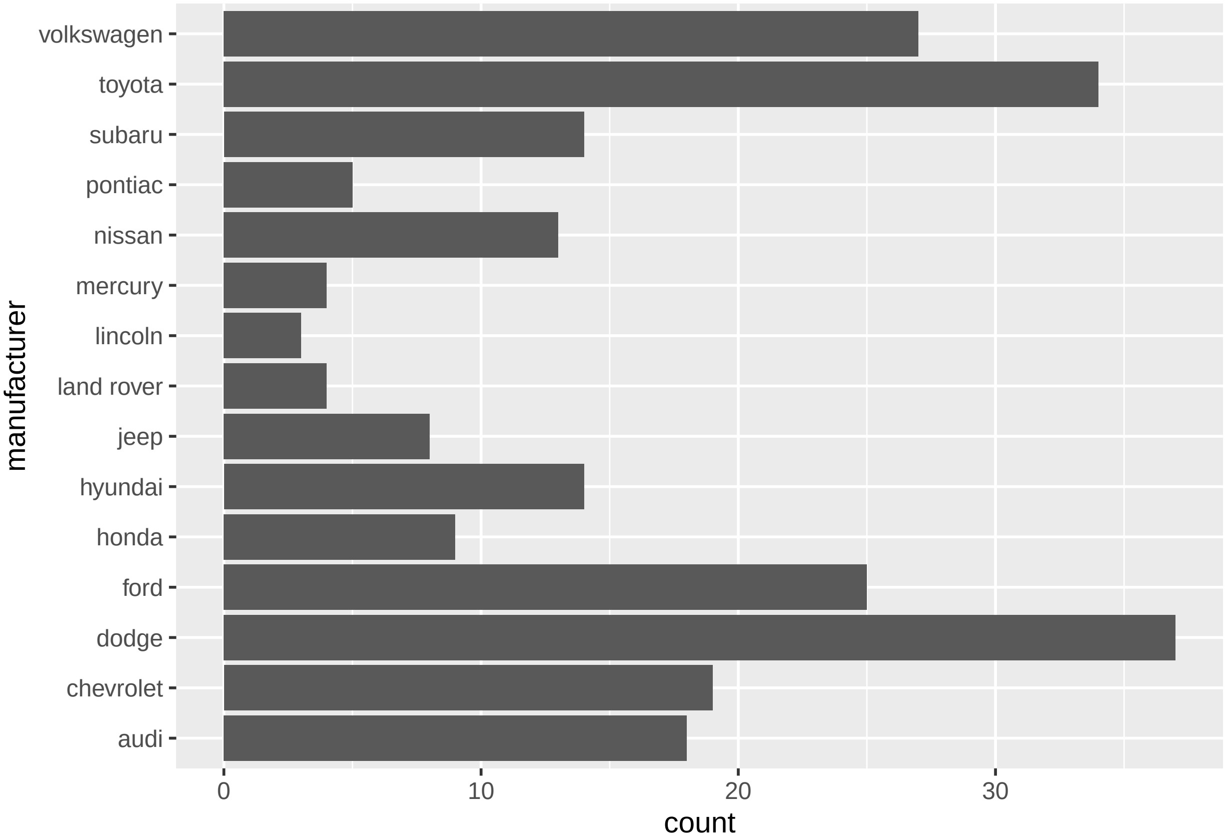
```
ggplot(mpg, aes(x=hwy, y=cty)) + geom_point()
```



Question 2

There seems to be a positive linear relationship between hwy and cty. As the highway mileage increases, city mileage also increases.

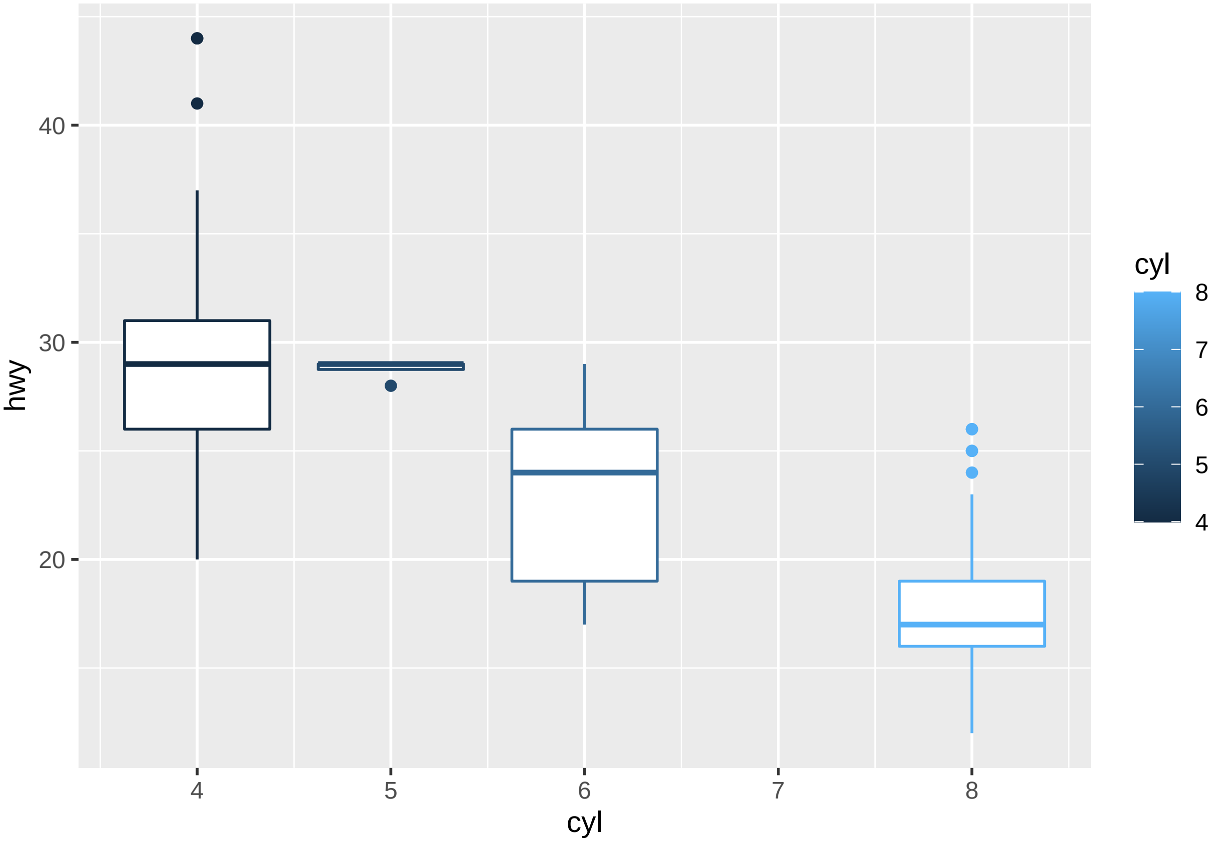
```
ggplot(mpg, aes(manufacturer)) + geom_bar(stat="count") + coord_flip()
```



Question 3

Dodge produced the most cars and Lincoln has produced the least.

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

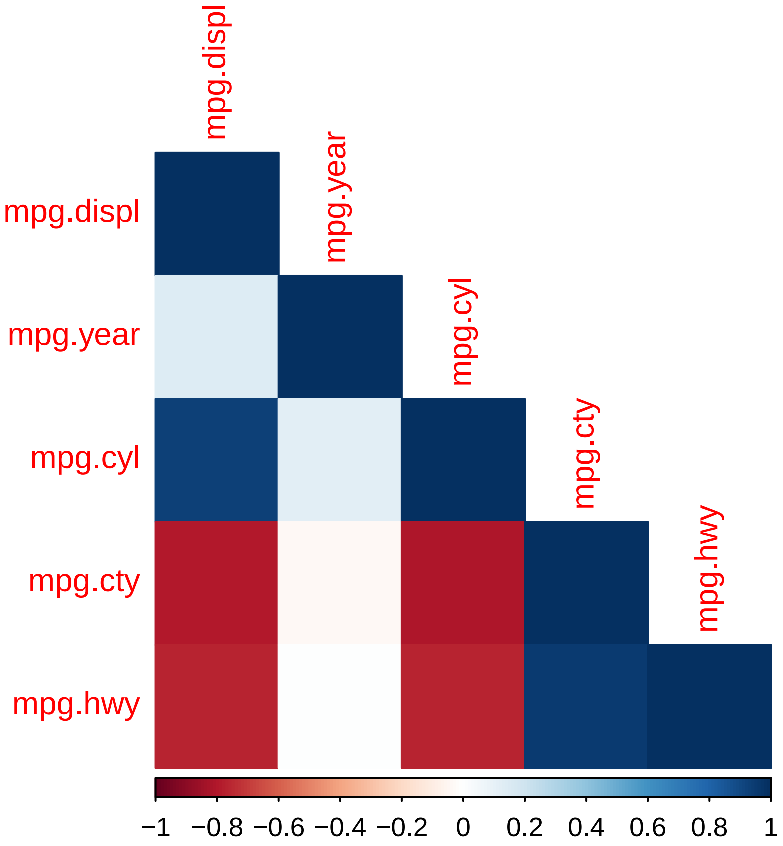


Question 4
Yes it seems that less highway mileage is being used as the number od cylinders increase. There is also no data for cars with 7 cyls.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
df <- data.frame(mpg$displ, mpg$year, mpg$cyl, mpg$cty, mpg$hwy)
m <- cor(df)
corrplot(m, method = 'color', type = 'lower')
```



Highly positively correlated: hwy & cty, cyl & displ

Highly negatively correlated: hwy & displ, cty & displ, hwy & cyl, cty & cyl These relationships make sense to me. Hwy & cty, and hwy & cyl relationships make sense since we have visualized these relationships in #1 and 4.