Pstat 131 HW1

Hongxu Ma

2022-10-01

- Question 1

Supervised learning accurately predict future response by giving the model observed output and input. Unsupervised machine learning to analyze unlabeled data with no response.

The major difference between them is that unsupervised machine learning model learn without a supervisor,because unsupervised learning model does need to discover hidden patterns itself. the supervised learning is created to measure its accuracy.

*from lecture slides page #31-34

- Question 2

The Y of regression is quantitative and the Y of Classification is qualitative, which meas that the former predict a continuous values, classification predicts categorical values.

*from lecture slides page #33

- Question 3

Regression model: Mean Squared Error (MSE), Mean Absolute Error (MAE)

Classification model: Accuracy, Confusion matrix

- Question 4

Descriptive models: Choose the model to best visually emphasize a trend in data

Inferential models: Use the model to state relationship between outcome and predictors.

Predictive models: Use the model to predict the outcomes with minimum reducible error.

*from lecture slides page #39

- Question 5

a. Mechanical models use scientific theories to predict what will happen in the real world. Empirical models develop theories based on a large number of facts or observations about real-world events.

b. mechanistic model is easier to understand, because mechanistic model's predictions are based on scientific theory, not on potentially inaccurate observations. The empirically-driven model requires requires developing a theory and applying it, but this theory is only observed and inferred in your model.

c. Bias and variance are inversely connected. So there are trade-offs to be made in the process of modifying the algorithm. In many cases, Mechanistic models require multiple parameters to make the model more flexible, but this may also increase variance. In other words, the output prediction results are not concentrated enough. Conversely, simplistic models can lead to overfitting. Empirically driven model discovery should be less empirically accurate because it requires a lot of empirical observation. However, the Empirical models are not able to predict beyond a particular operating range.

*from lecture slides page #38

- Question 6

The first one is predictive, because campaigns predict future behavior based on past information about voters.

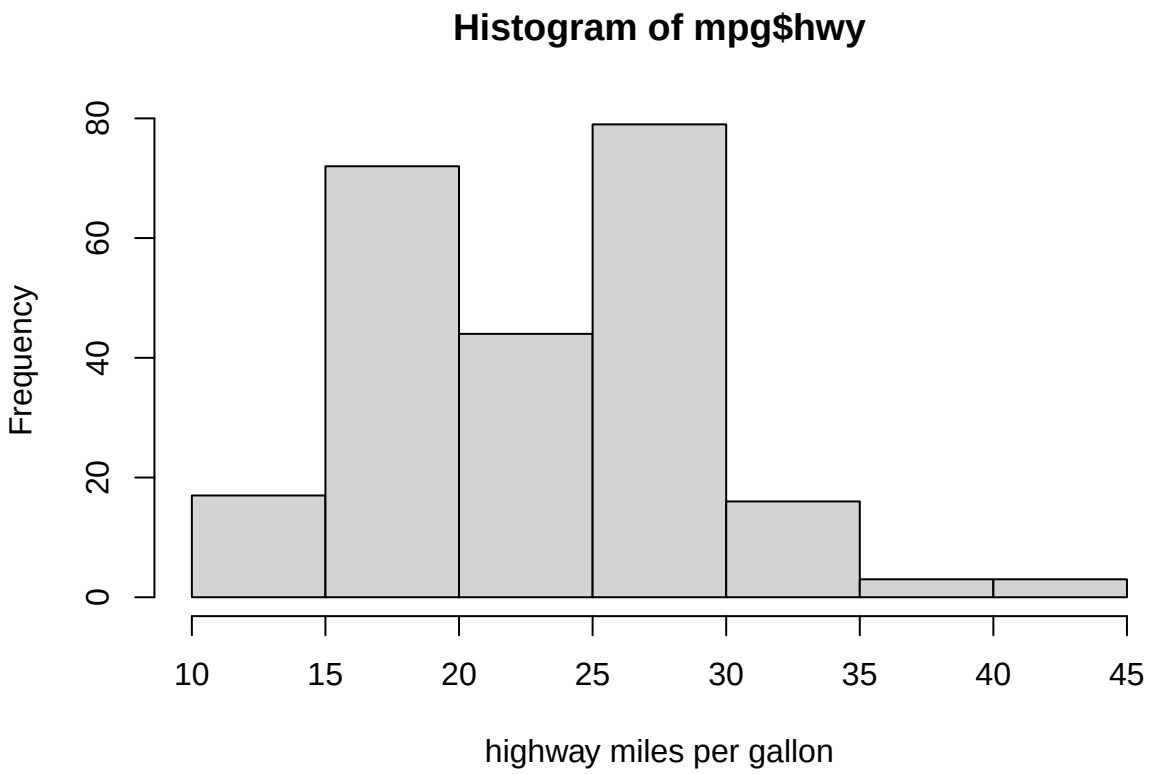
The second one is inferential, because the campaign wants to discover the relationship that exists between politicians and voters. This question focuses on the hypothesis that the probability that a voter will vote increases after meeting a candidate.

- Exercise 1:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## 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()
```

```
hist(x = mpg$hwy, xlab = "highway miles per gallon")
```

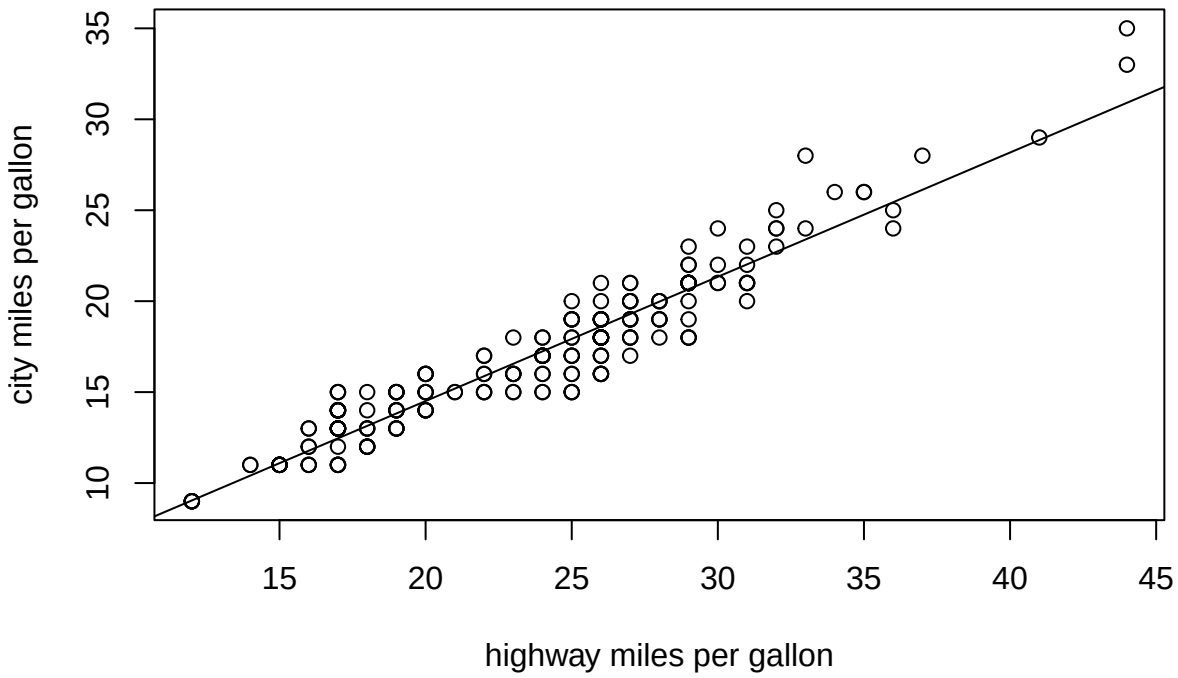



```
knitr::opts_chunk$set(echo = TRUE)
```

Most of these popular cars drive 15 to 30 miles per gallon on the highway, because 15-30 have high frequency.

- Exercise 2:

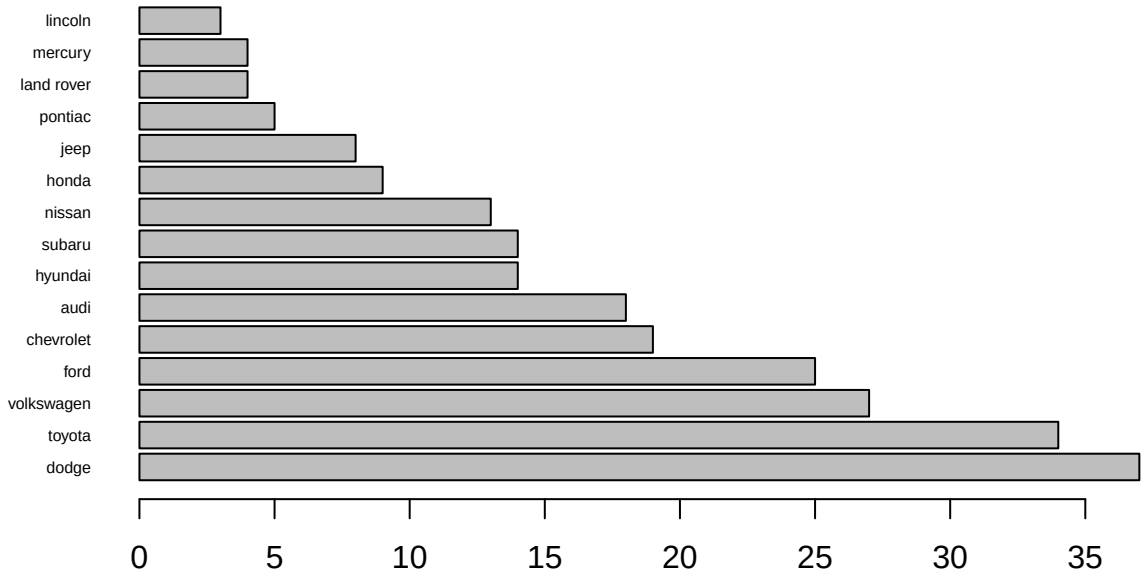
```
x = mpg$hwy
y = mpg$cty
plot(x = mpg$hwy, y = mpg$cty, type="p", , xlab = "highway miles per gallon",
     ylab ="city miles per gallon")
abline(lm(y~x))
```



From the graph, a positive correlation is found between Highway miles per gallon and City miles per gallon.

- Exercise 3:

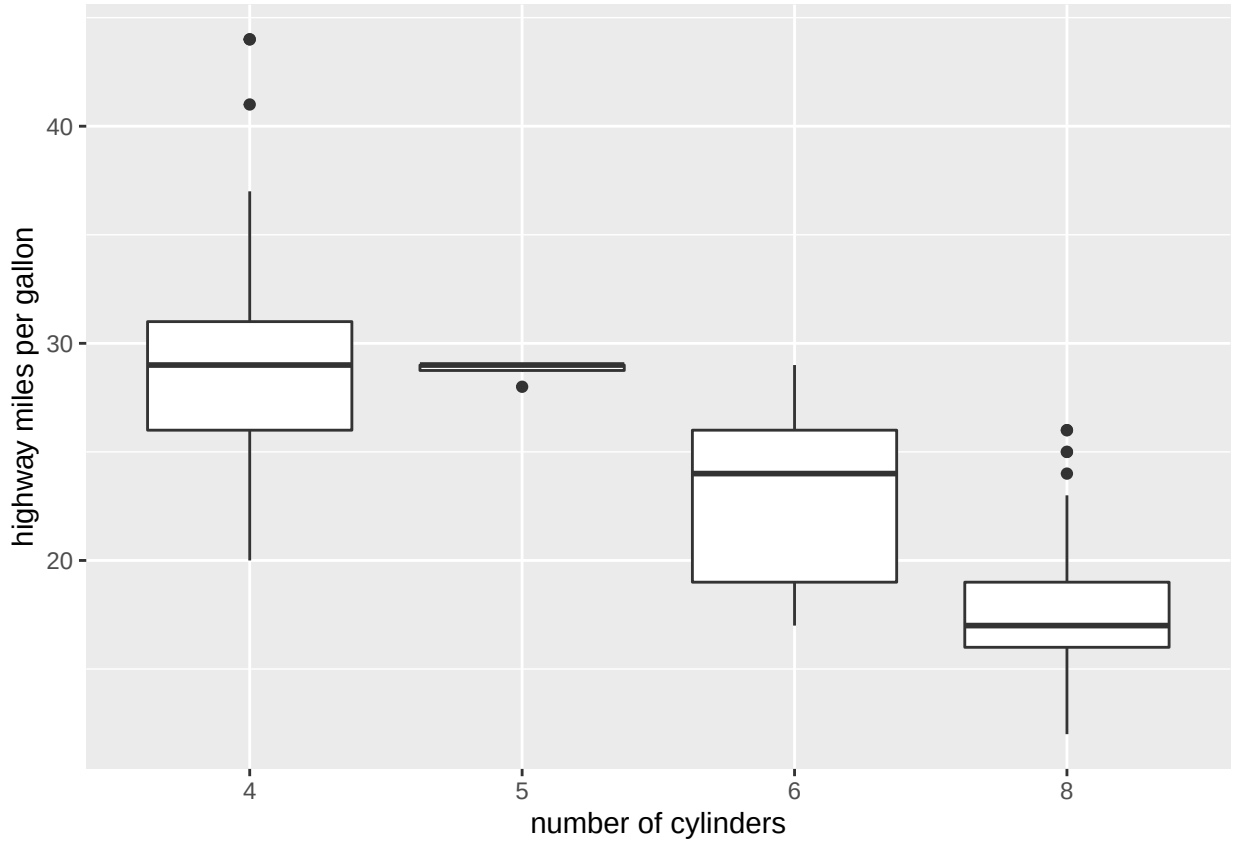
```
x = sort(table(mpg$manufacturer), decreasing = TRUE)
y = data.frame(x)
barplot(x, names.arg = y$Var1, horiz = TRUE, cex.names = 0.5, las = 1)
```



Dodge produced the most cars and Lincoln produced the least number of cars.

Exercise 4:

```
library(ggplot2)
ggplot(mpg, aes(x = factor(cyl), y = hwy)) +  geom_boxplot() +
  xlab("number of cylinders") + ylab("highway miles per gallon")
```



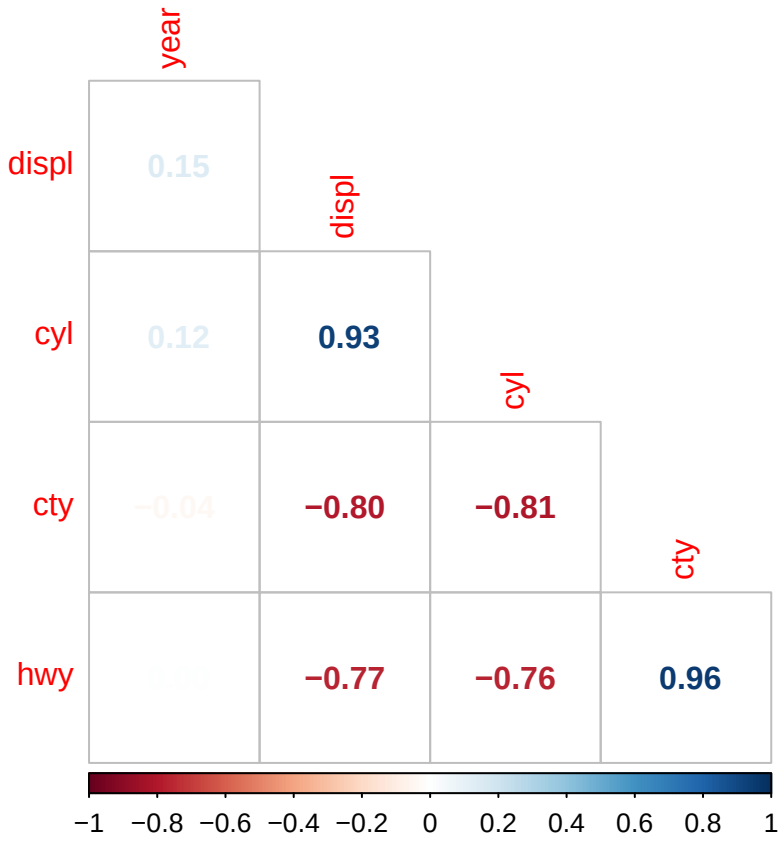
The higher the number of cylinders, the lower the highway miles per gallon.

Exercise 5:

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
M <- cor(select_if(mpg, is.numeric))
corrplot(M, method = 'number', order = 'AOE', type = 'lower', diag = FALSE, )
```



Engine displacement and number of cylinders are positively correlated; city miles per gallon and highway miles per gallon are positively correlated.

City/highway miles per gallon is negatively correlated with both engine displacement and number of cylinders.

Lastly, none of the correlations between year and other variables were significant.