PSTAT131 HW#1

Cynthia Cao

2022-09-29

Machine Learning Main Ideas

Q1: Supervised learning: give the model observed output and input, with the actual data *Y* (supervisor), accurately predict future response with given predictors. Unsupervised learning: learn without a supervisor, no responses, find the pattern of data without a specific goal. Most of unsupervised learning are clustering. Differences: Supervised learning has supervisor while unsupervised learning has not.

Q2: Differences: Y is quantitative in regression model, data are numerical values, but Y is qualitative in classification model and data are categorical values.

Q3: Regression model: R Square, Mean Square Error(MSE) Classification model: Accuracy, Confusion matrix

Q4: Descriptive models: used to choose model to best visually emphasize a trend in data; Inferential models: used to detect significant features, test theories, (possibly) causal claims, and state relationship between outcome & predictor(s); Predictive models: used to find what combo of features fits best, predict Y with minimum reducible error, and not focused on hypothesis tests.

Q5: Mechanistic: models are fitted with parameters, can add parameters to has more flexibility; Empirically-driven: models are not fitted with parameters, require a larger number of observations, much more flexible by default; Both may overfitting.

Mechanistic will be easier to understand because it assume a parametric form for f, then the relationship between x and y will be easier to interpret and understand.

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. Mechanistic models can delete parameters to make the model less flexible and hence may increase the bias and decrease variance. The Empirically-driven model can lead to overfitting.

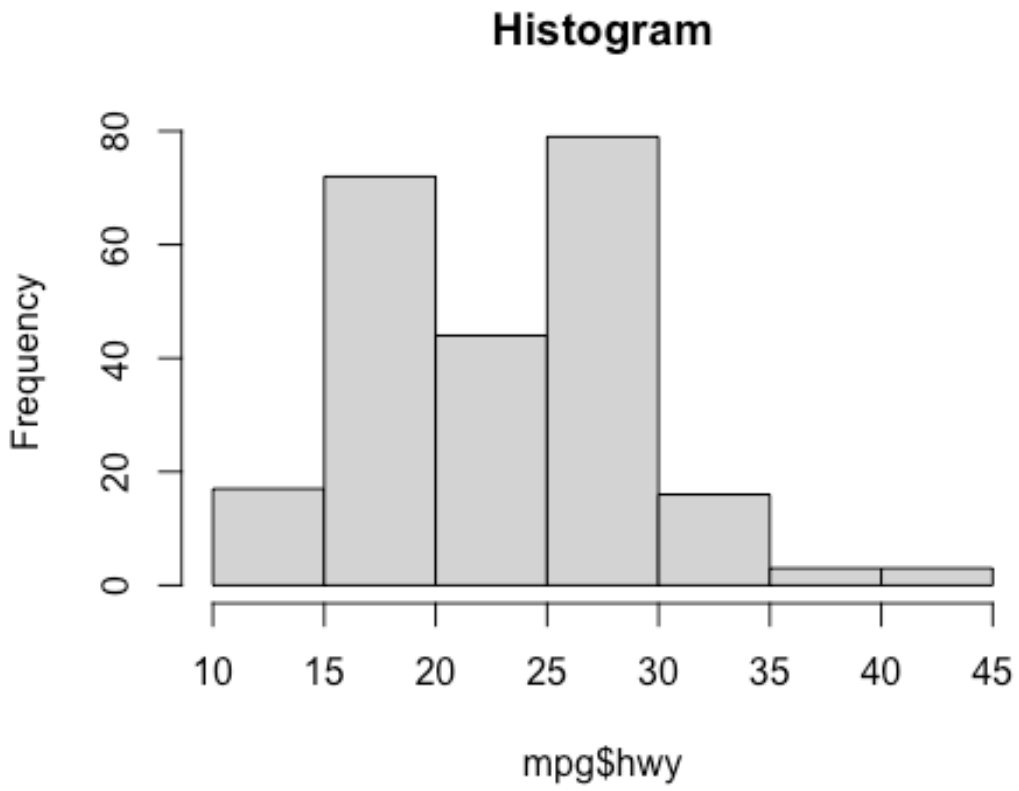
Q6: (1) Predictive, because it asks to predict a probability of a voter that they will vote in favor of the candidate; (2) Inferential, because it asks for the likelihood of support for the condidate and relationship between personal contact with the candidate.

Exploratory Data Analysis Q1:

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✓ ggplot2 3.3.6      ✓ purrr   0.3.4
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.2      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

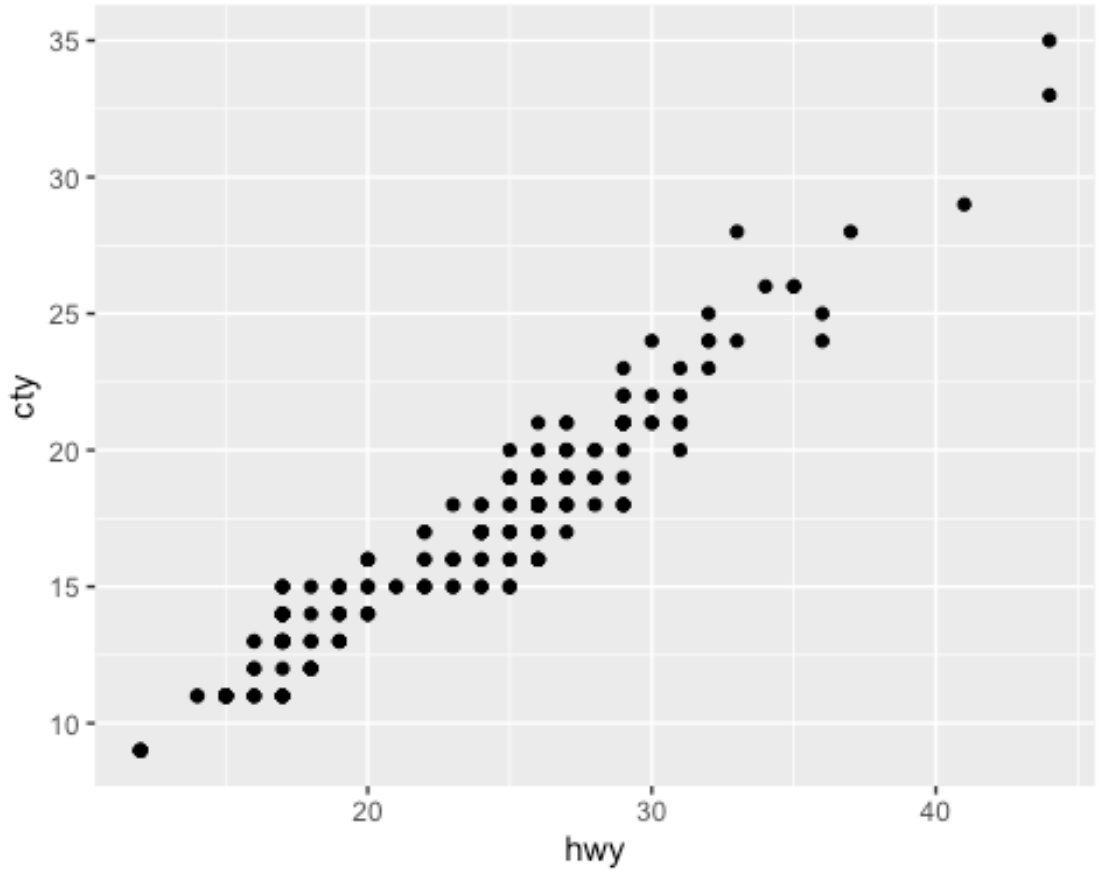
```
hist(mpg$hwy,main = "Histogram")
```



The highway miles per gallon between 15-30 have high frequency while the other mpg are with low frequency.

Q2:

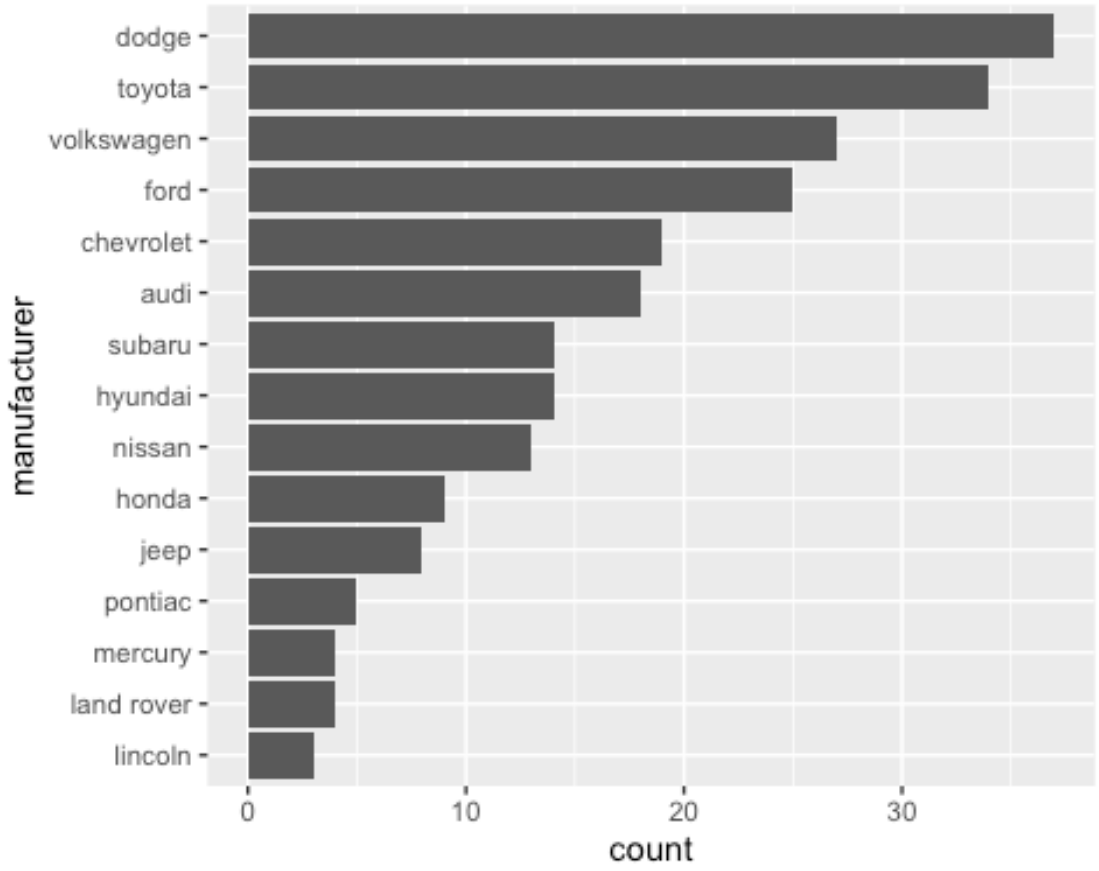
```
library(ggplot2)
ggplot(mpg, aes(hwy, cty)) + geom_point()
```



The hwy and cty have strong positive linear relationship by the plot, which means that the cty may increases with hwy increases with a certain rate, they may have linear relationship between each other.

Q3:

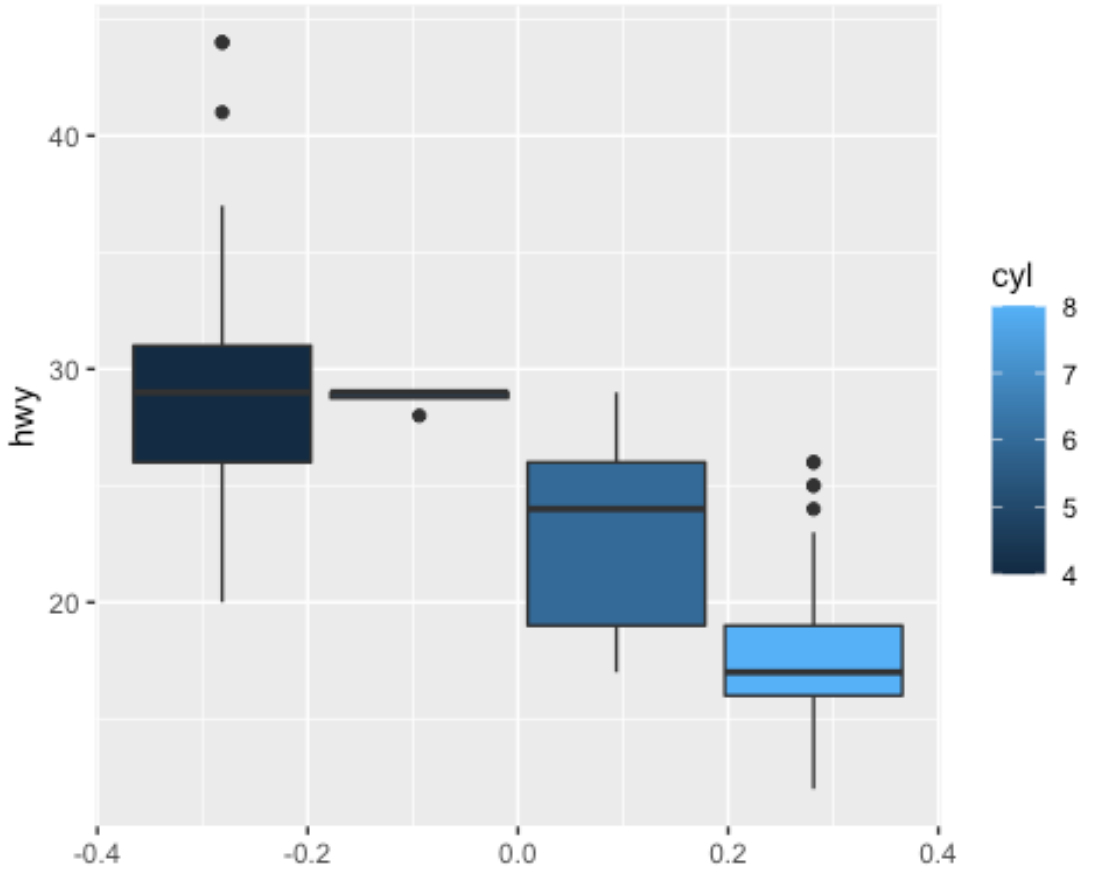
```
manufacturer <- mpg$manufacturer
counts <- table(manufacturer)
A <- factor(manufacturer,levels = names(sort(counts)))
mpg1 <- within(mpg,manufacturer <- A)

ggplot(data = mpg1, aes(x = manufacturer)) + geom_bar(stat="count") +
coord_flip()
```



Dodge produced the most cars, Lincoln produced the least.

Q4:

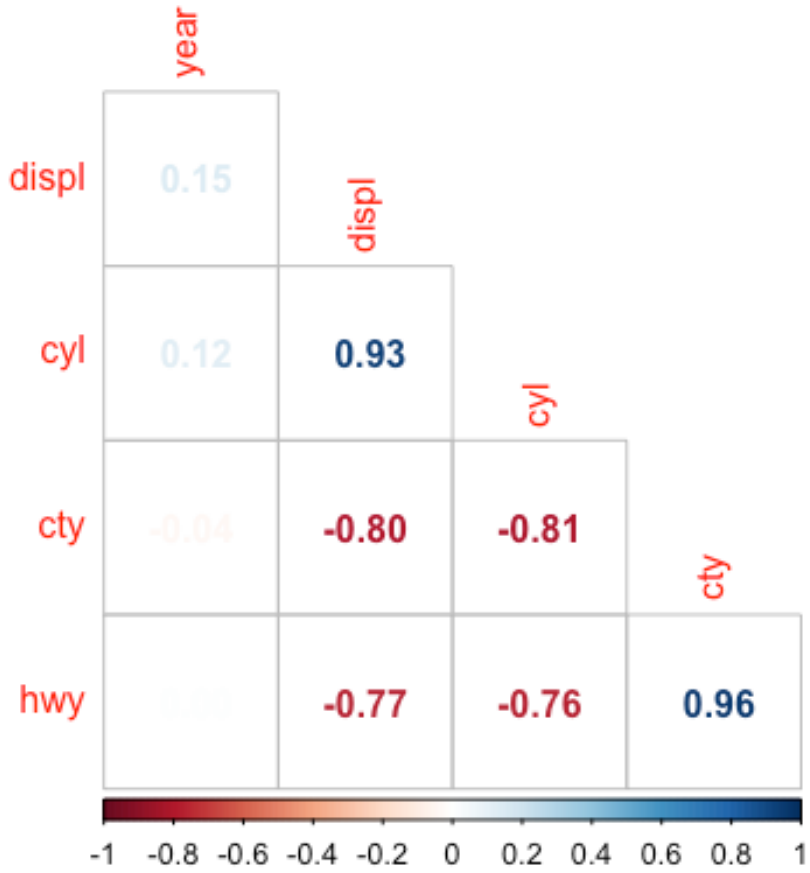
```
ggplot(mpg, aes(y = hwy, group = cyl,fill=cyl)) + geom_boxplot()
```



The highway mpg decreases with the cylinder increases.

Q5:

```
library(corrplot)

## corrplot 0.92 loaded

mpg2 <- select_if(mpg, is.numeric)
M <- cor(mpg2)
corrplot(M,method = 'number',type = 'lower',order = 'AOE',diag = F)
```



Positively correlated: displ & year(about zero), cyl & year(about zero), cyl & displ, hwy & cty; negatively correlated: cty & year(about zero), cty & displ, hwy & displ, cty & cyl, hwy & cyl; The relationships are make sense to me, it's reasonable to say that larger clinder would have larger engine displacement, but less city and highway mpg.