Homewoek 1

Machine Learning

Question 1

Solution: From the lecture, in supervised learning, the actual data Y is the supervisor, we need to give the model observed output and input. In unsupervised learning, we learn without a supervisor. The difference between them is whether they have supervisor.

Question 2

Solution: When the supervisor Y is quantitative, it is a regression model. When the supervisor Y is qualitative, it is a classification model.

Question 3

Solution: For regression ML problems, the two commonly used metrics are test MSE and test mean absilute error (MAE). For classification ML problems, the two commonly used metrics are test error rate and accuracy.

Question 4

Solution: From the lecture,

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

Inferential models: Aim is to test theories, (Possibly) causal claims. State relationship between outcome & predictor(s).

Predictive models: Aim is to predict Y with minimum reducible error, which is not focused on hypothesis tests.

Question 5

Solution: From the lecture,

Mechanistic assumes a parametric form for f, empirically-driven has no assumptions about f. The difference between them is whether they have specific form for f. They both can lead to overfit.

In general, mechanistic model is easier to understand because it assumes a parametric form for f.

For mechanistic model, with the increasing of parameters, the bias decreases and the variance increases. For empirically-driven model, with the increasing of the number of observations, the bias decreases and the variance increases.

Question 6

Solution:

The first question is predictive, because we are interesed in the result of the voter.

The second question is inferential, becaues we are interesed in the relationship between the result of the voter and whether they had personal contact with the candidate.

Exploratory Data Analysis

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

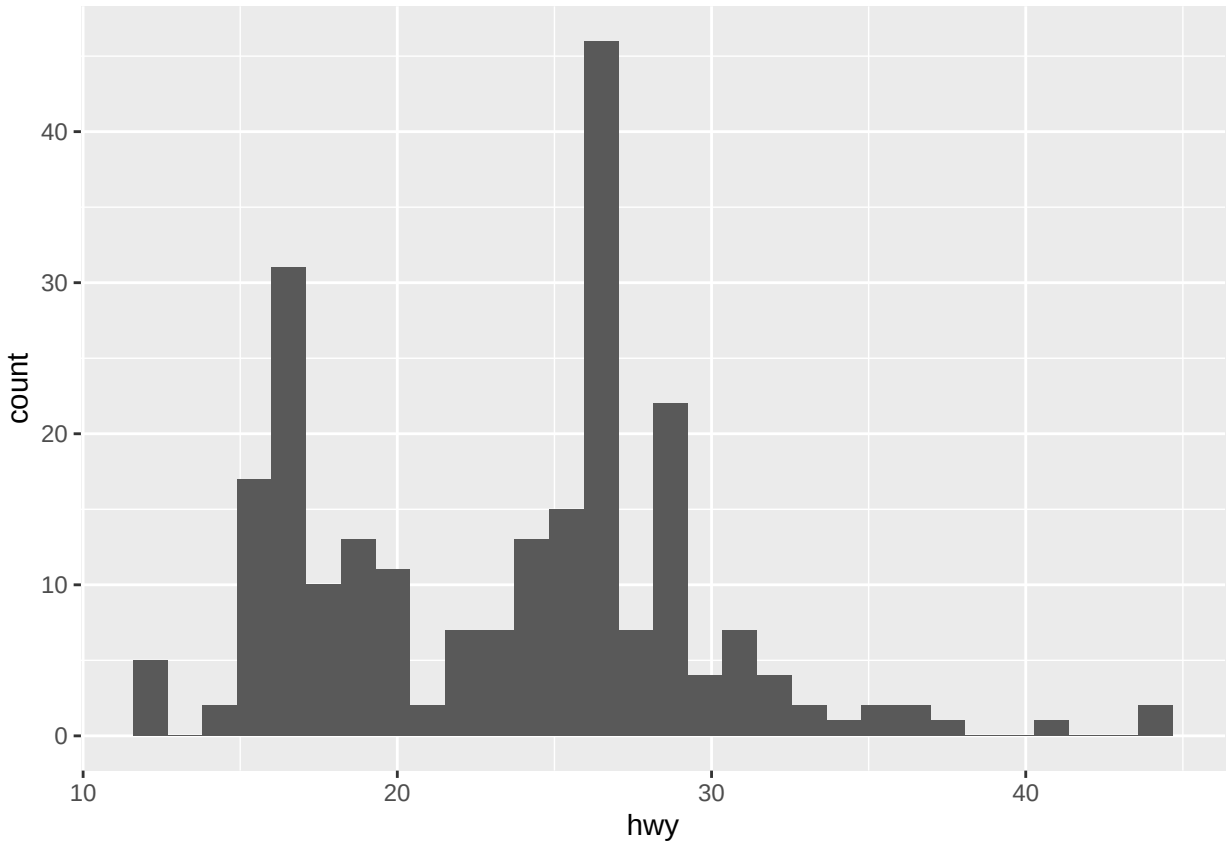
```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Exercise 1

```
mpg %>%
  ggplot(aes(x = hwy)) +
  geom_histogram()
```

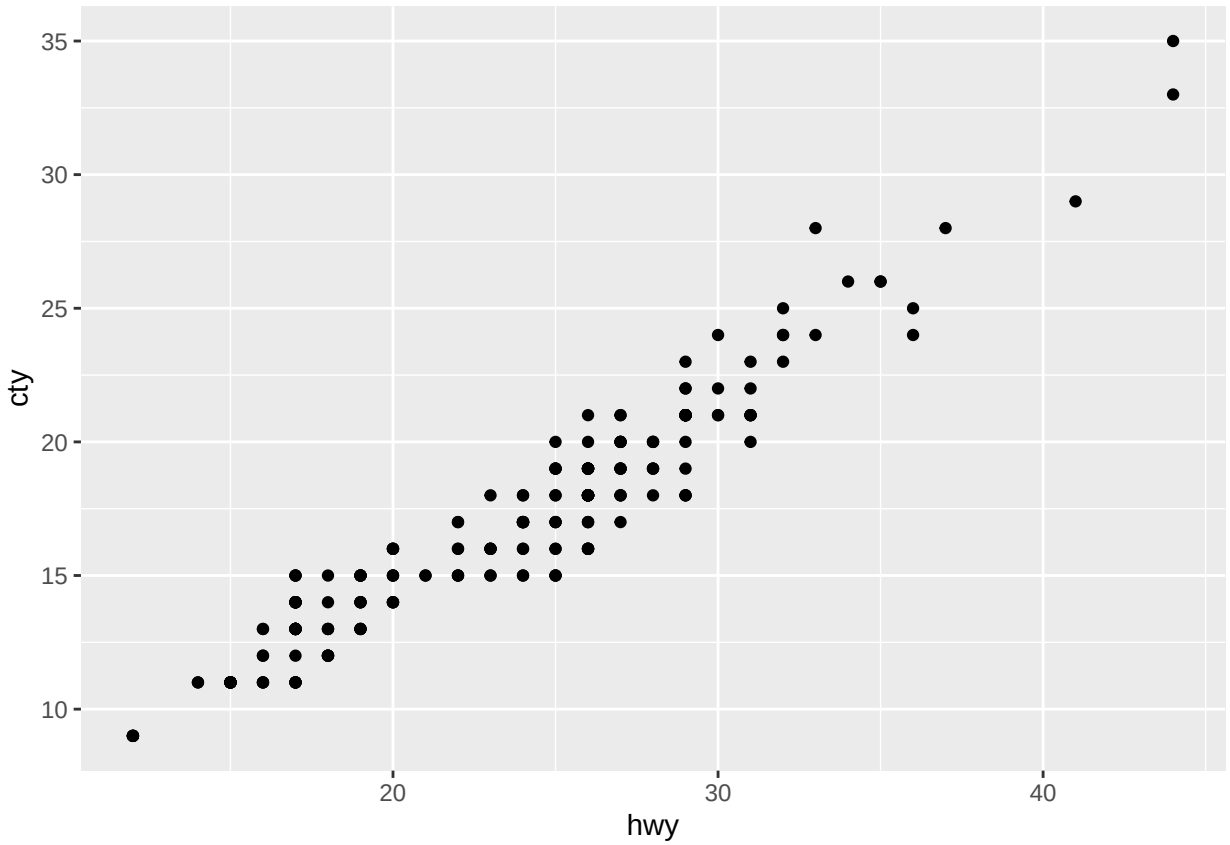
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Solution: There are two peaks from the histogram.

Exercise 2

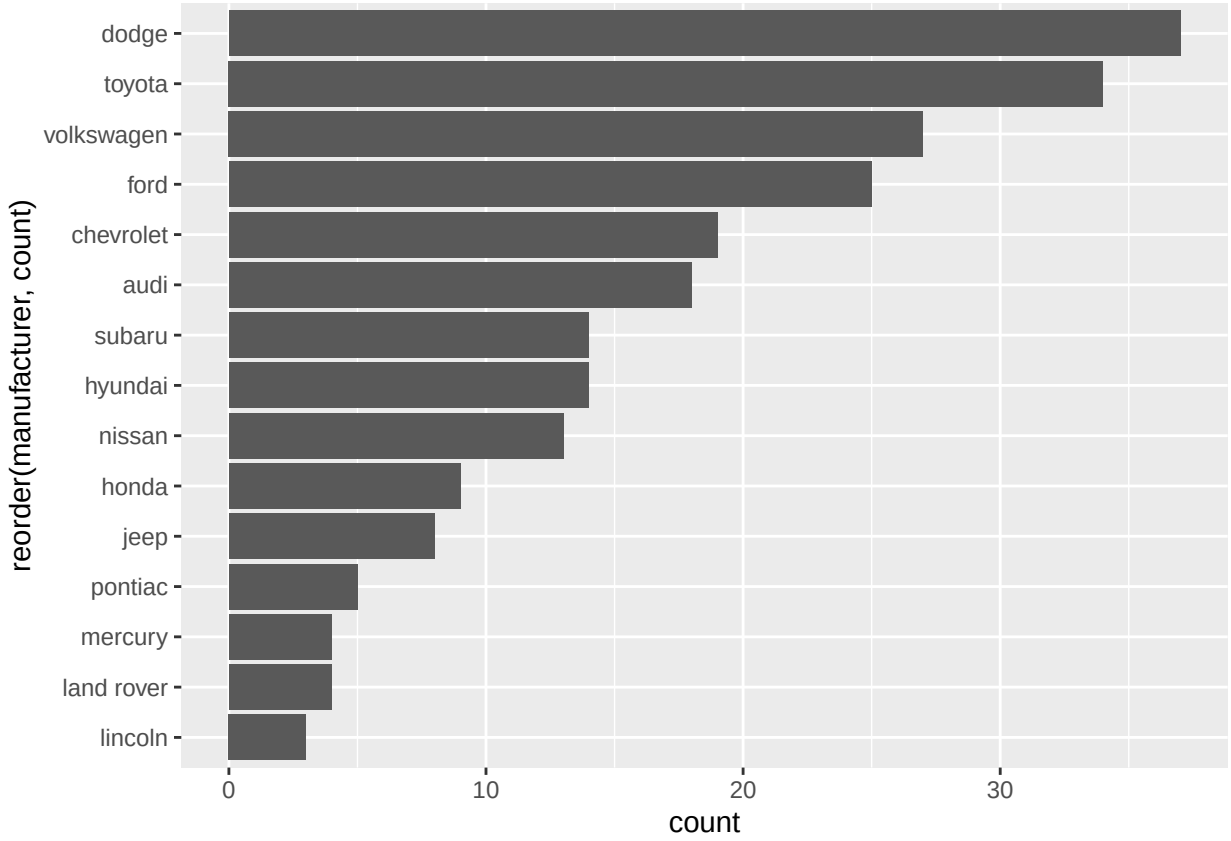
```
mpg %>%
  ggplot(aes(x = hwy, y = cty)) +
  geom_point()
```



Solution: The points almost lie on a straight line. There are linear relationship between hwy and cty, which means we can use linear regression to analysis them.

Exercise 3

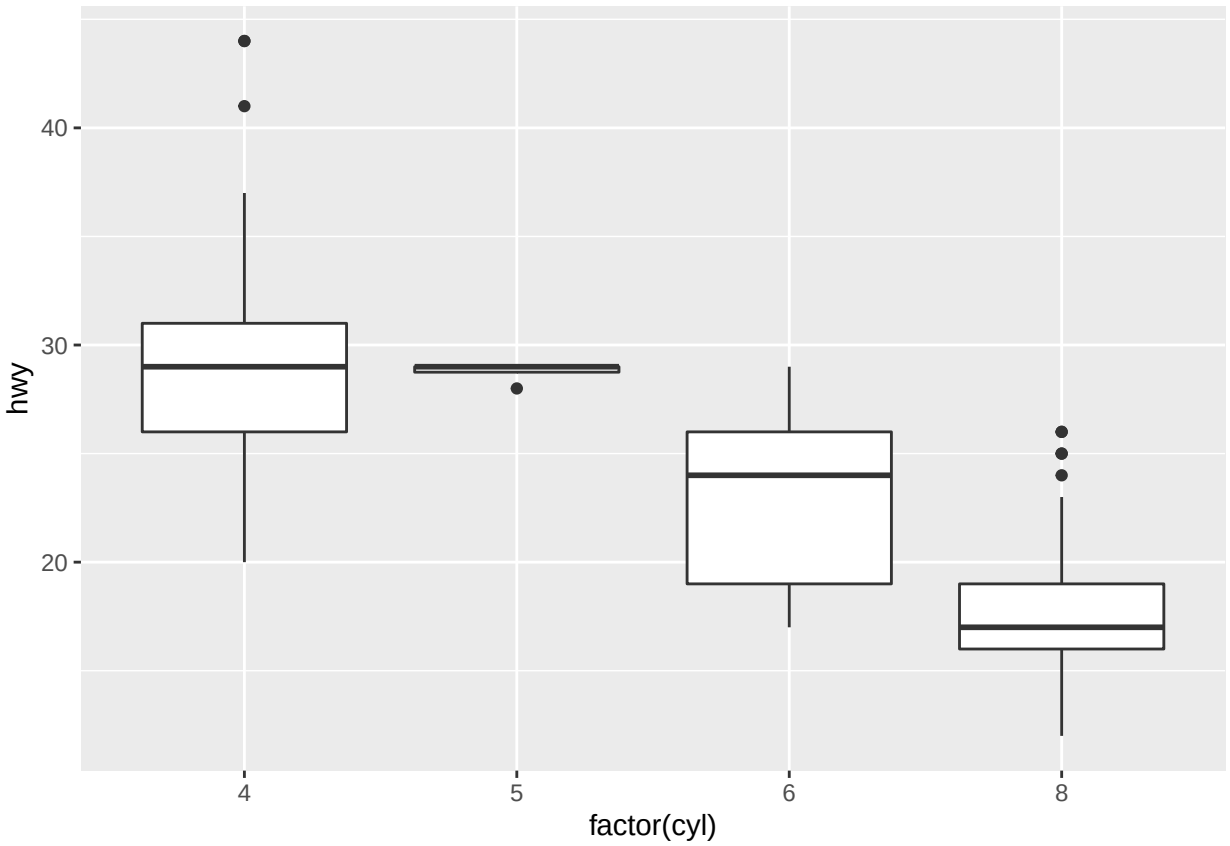
```
mpg %>%
  group_by(manufacturer) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  ggplot(aes(x = reorder(manufacturer, count), y = count)) +
  geom_bar(stat = "identity") +
  coord_flip()
```



Solution: Dodge produced the most cars and lincoln produced the least.

Exercise 4

```
mpg %>%
  ggplot(aes(x = factor(cyl), y = hwy)) +
  geom_boxplot()
```



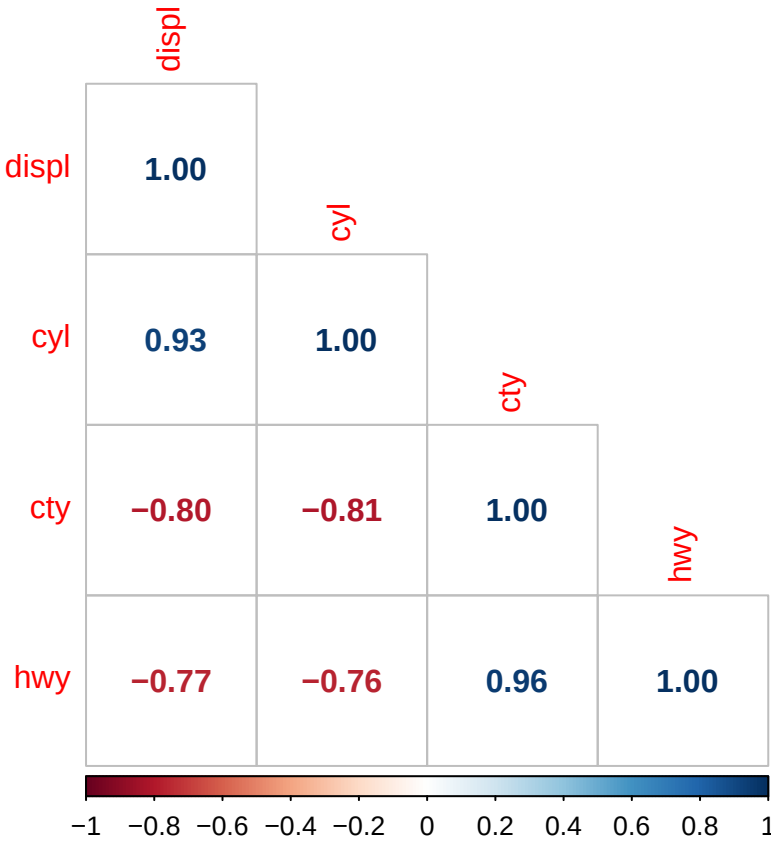
Solution: With the increasing of cyl, the hwy decreases.

Exercise 5

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
M = cor(mpg[, c(3,5,8,9)])
corrplot(M, method = "number", type = "lower")
```



Solution: Displ is positively correlated with cyl, and it is negatively correlated with cty and hwy. Cyl is negatively correlated with cty and hwy. Cty is positively correlated with hwy. The relationship between hwy and cyl makes sense to me. No.