Homework 1

Yixin Wang

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

   Supervised learning: For each observation of the predictor measurement(s) $x_i, i = 1, ..., n$ there is an associated response measurement y_i.

   Unsupervised learning: For every observation $i = 1, ..., n$, we observe a vector of measurements x_i but no associated response y_i.

   The difference is that unsupervised learning lacks a response variable that can supervise the analysis.

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

   In regression model, we tend to refer to problems with a quantitative response. In classification model, it involves a qualitative response.

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

   Regression: Training Mean Squared Error(MSE) and test MSE

   Classification: test error rate and training error rate

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.
   - Predictive models: It aims to predict Y with minimum reducible error.
   - Inferential models: It aims to test theories or causal claims or tate relationship between outcome and predictor(s).

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 model uses a theory to predict what will happen in the real world. The empirical modeling studies real-world events to develop a theory.

     They are different in the way the model made. One has the assumption, the other does not have.

     The similarity is that they may have overfitting and they may be more flexible.

   - In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice.

     The mechanistic is easier to understand because there is an explicit assumption of the function. Since non-parametric models do not have explicit assumption, it may need a lot of observations for the function.

   - Describe how the bias-variance trade-off is related to the use of mechanistic or empirically-driven models.

     Higher model flexibility may represent the lower training MSE and lower test MSE. They may have low bias and high variance.

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.

   The first question is predictive, since it is a probability between 0 and 1. The second question is inferential, since it has the relationship between the predictors and results.

Exploratory Data Analysis

```
library(tidyverse)
```

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

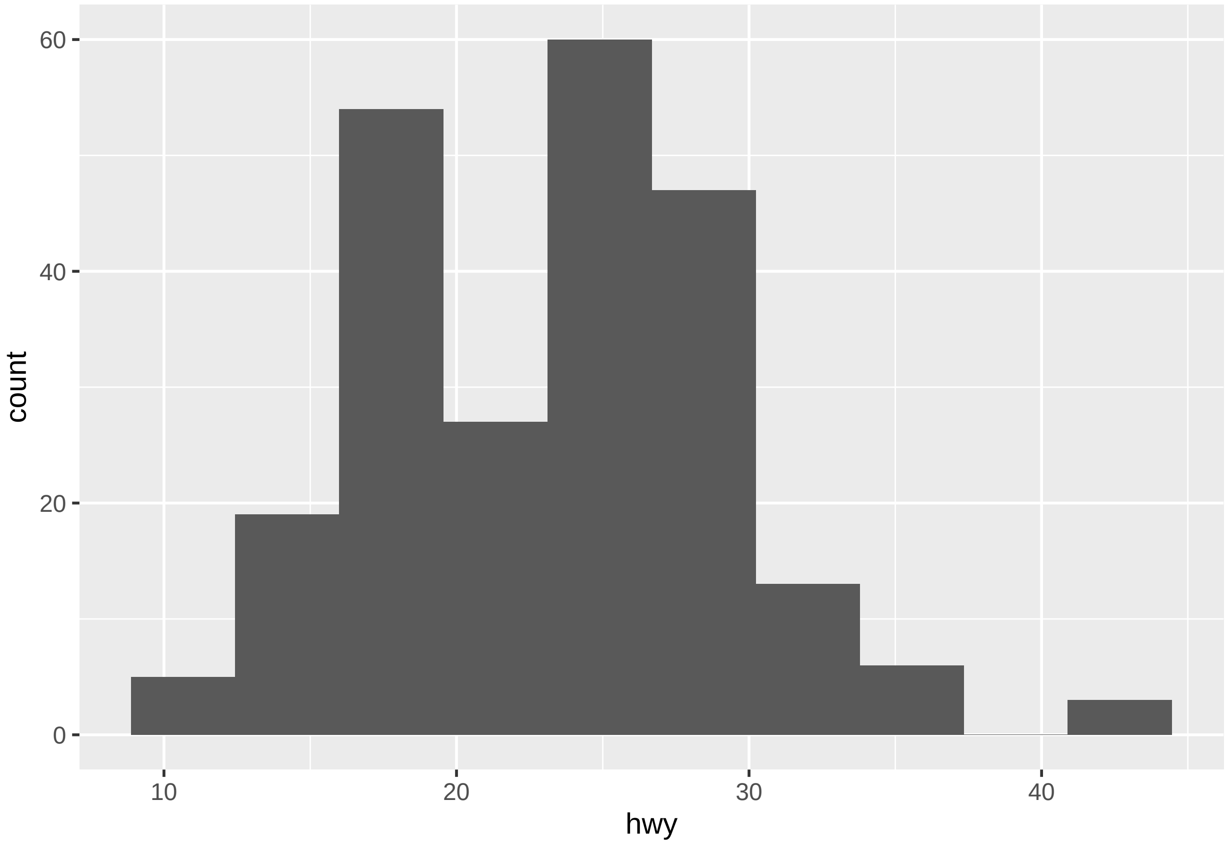
```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

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

```
library(ggplot2)
```

Exercise 1

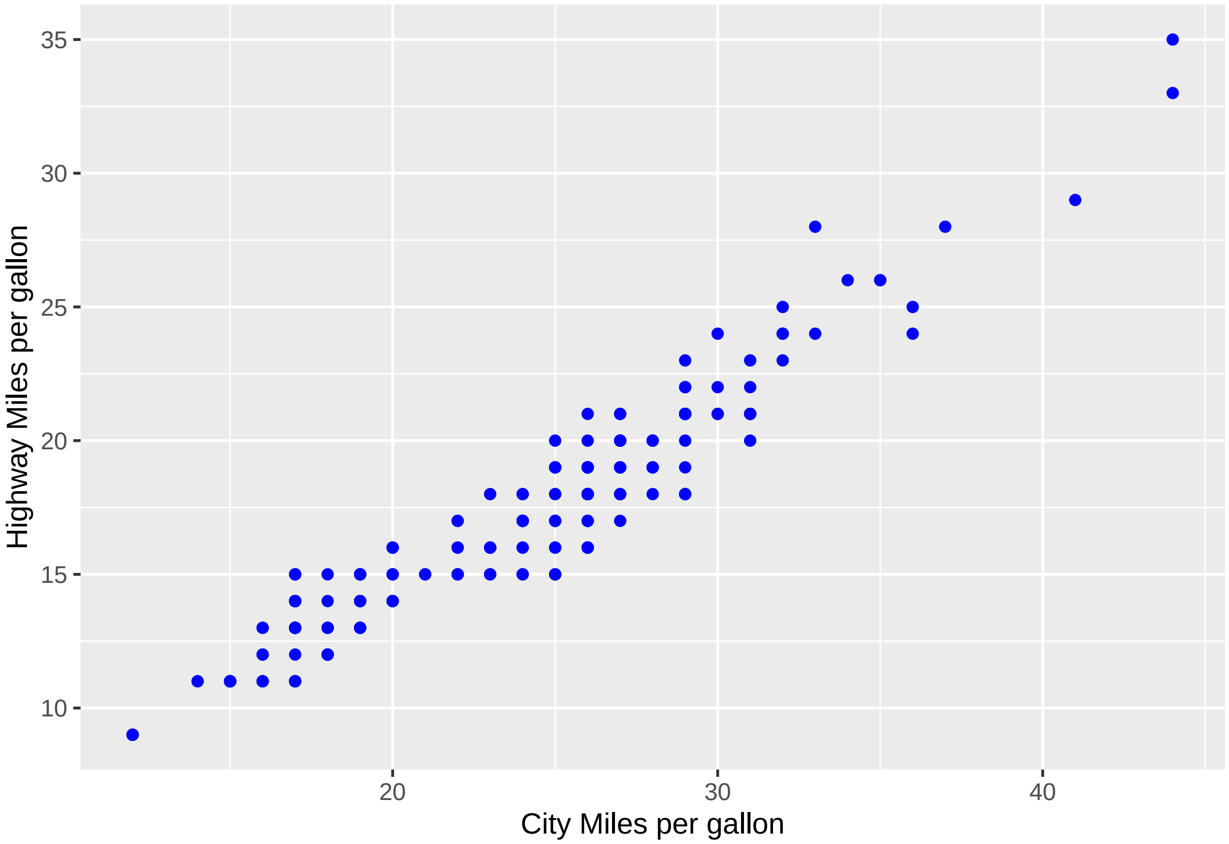
```
hist <- ggplot(mpg, aes(x=hwy)) + geom_histogram(bins = 10)
hist
```



In the histogram, the highest frequency of highway miles per gallon is happening between 20 and 30 gallons.

Exercise 2

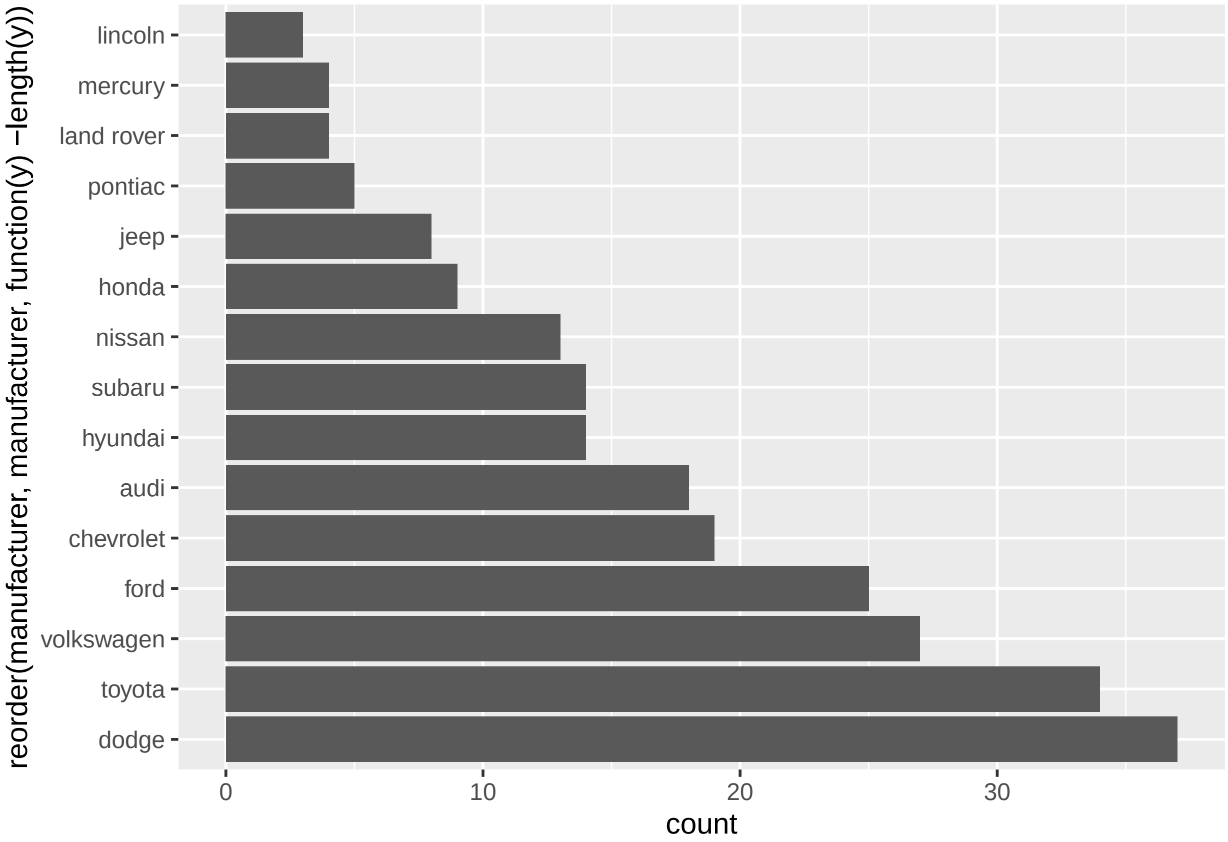
```
scatter <- ggplot(mpg, aes(hwy, cty)) + geom_point(color = 'blue') + labs(x = "City Miles per gallon", y
scatter
```



There is a positive correlation between city miles per gallon and highway miles per gallon. It represents that more city miles per gallon may lead to more highway miles per gallon.

Exercise 3

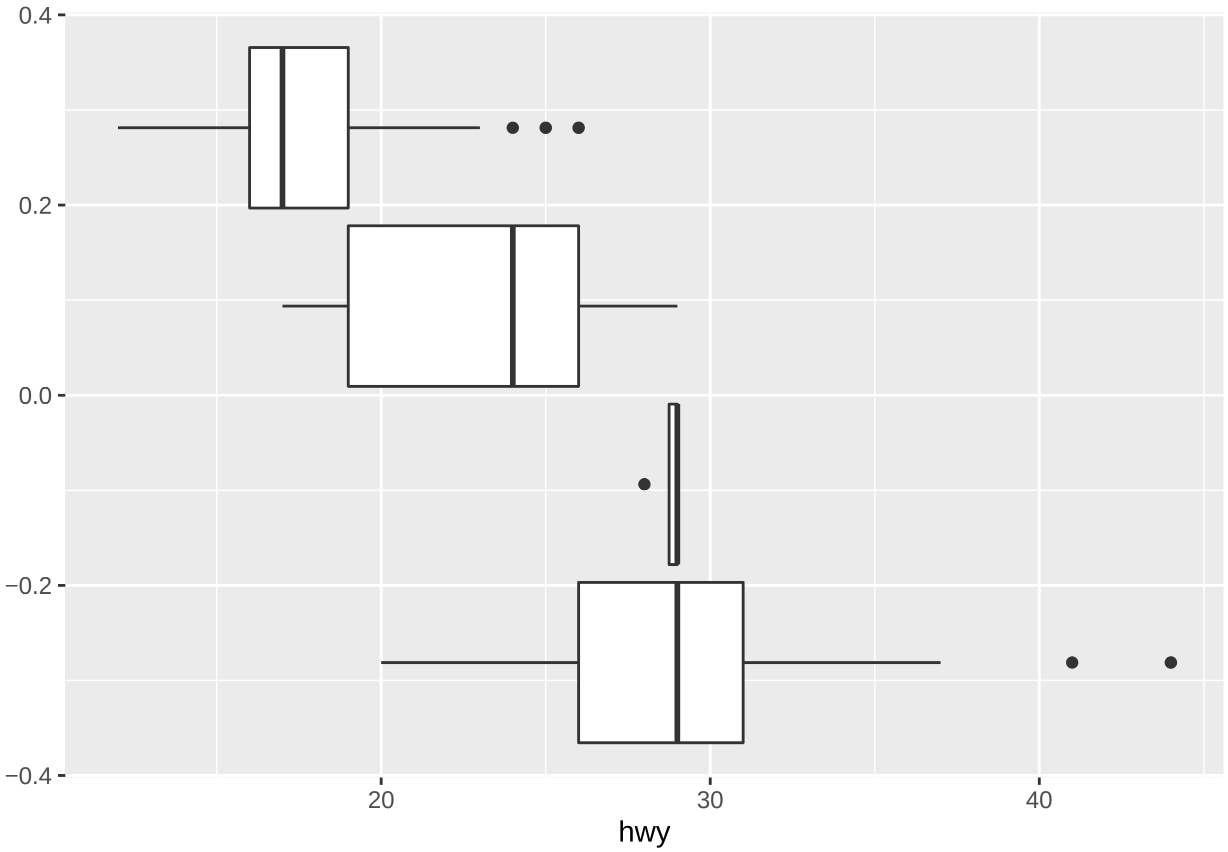
```
manufac <- ggplot(mpg, aes(y = reorder(manufacturer, manufacturer, function(y)-length(y)))) + geom_bar(
manufac
```



Based on the graph, Dodge produced the most cars and Lincoln produced the least cars.

Exercise 4

```
bar_plot <- ggplot(mpg, aes(group = cyl, x = hwy)) + geom_boxplot()
bar_plot
```



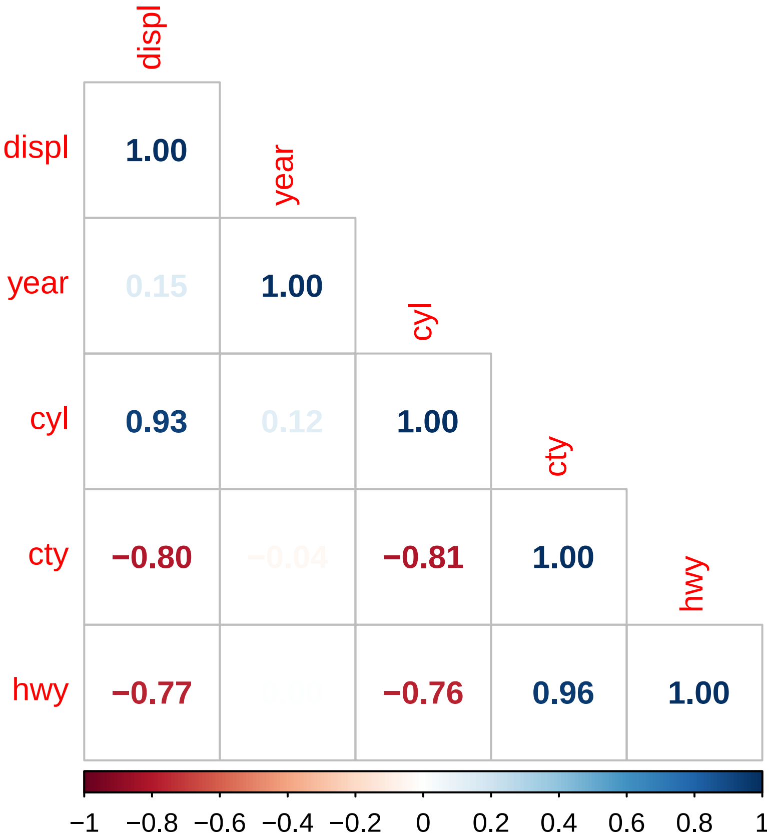
It seems that there is a negative correlation between cyl and highway miles per gallon.

Exercise 5

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
M = cor(mpg %>% dplyr::select(where(is.numeric)))
corrplot(M, method = 'number', type = 'lower')
```



The positive correlations are cyl~displ, hwy~cty. The negative correlations are cty~displ, hwy~displ, cty~cyl, hwy~cyl. The more city miles per gallon will lead to more highway miles per gallon. The displacement of engine is also positively correlated to the cylinders the car have.

It is surprised to see that the engine displacement is negatively correlated to city and highway miles per gallon.