PSTAT131HW1

Tao Wang

2022-04-02

Machine Learning Main Ideas

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

Answer:

1. Defination of supervised learning:

- According to IBM.com, "supervised learning is a machine learning approach that's defined by its use of labeled datasets. These datasets are designed to train or 'supervise' algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time." (from https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning)
- To be more specific, according to the textbook, for supervised learning, "each observation of the predictor measurement(s) x_i, i =1,. . . ,n there is an associated response measurement y_i. We wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference)." (from page #26 of our textbook)

2. Defination of Unsupervised learning:

- According to IBM.com, "unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention (hence, they are 'unsupervised')." (from https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning)
- To be more specific, according to the textbook, "unsupervised learning describes the somewhat more chal- lenging situation in which for every observation i =1,. . . ,n, we observe a vector of measurements x_i but no associated response y_i." (from page #26 of our textbook) Also, we should know that, " in this setting, we are in some sense working blind; the situation is referred to as unsupervised because we lack a response variable that can supervise our analysis." (from page #26 of our textbook)

3. The differences between supervised and unsupervised learning

- There are so many differences between supervised and unsupervised learning. The main difference is that the supervised learning has associated response y_i for each observation, while the unsupervised learning doesn't. (The supervised learning requires input and output, while the unsupervised learning only requires input)

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

Answer:

- The difference between a regression model and a classification model is:
- The response variable Y of regression model is quantitative, which take on numerical values. For example, Y can be price, blood pressure and so on.
- The response variable Y of classification model is qualitative, which take on categorical values. According to the textbook, "examples of qualitative variables include a person's marital status (married or not), the brand of prod- uct purchased (brand A, B, or C), whether a person defaults on a debt (yes or no), or a cancer diagnosis (Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia)." (from page # 28 from the book)

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

Answer: (In office hour, professor said that we can skip this question since we haven't learnt yet.)

Question 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.

Answer:

Descriptive models: According to what we learned in the lecture, we use descriptive models in order to "best visually emphasize a trend in data." (from the lecture note)

Inferential models: According to what we learned in the lecture, the aim of inferential models is to test theories, causal claims, and "state the relationship between outcome and predictor".(from the lecture note)

Predictive models: According to the textbook, the aim of predictive models is to accurately predict "the response for future observations." (the page #26 from the textbook) To be more specific, we want to "predict Y with minimum reducible error." (from the lecture note)

Question 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.

Answer:

a) Define empirically-driven. How do these model types differ? How are they similar?

- 1. The definition of mechanistic
- Mechanistic models, which are parametric models,"are based on assumptions about the distribution of population from which the sample was taken". (from the website https://www.ibm.com/docs/en/db 2woc?topic=procedures-statistics-parametric-nonparametric) We can add parameters to make the mechanistic model more flexibility. Have a problem of overfitting if we add too many parameters to it. (from the lecture)
- 2. The definition of empirically-driven
- Empirically-driven models, which are nonparametric models, "are not based on assumptions, that is, the data can be collected from a sample that does not follow a specific distribution." (from the website https://www.ibm.com/docs/en/db2woc?topic=procedures-statistics-parametric-nonparametric) The

empirically-driven model requires a larger number of observations. By default, It is much more flexible. Have a problem of overfitting.(from the lecture)

- 3. The difference between these models:
- Mechanistic models "are based on assumptions about the distribution of population from which the sample was taken," while empirically-driven models are not.(from the website https://www.ibm.com/docs/en/db2woc?topic=procedures-statistics-parametric-nonparametric)
- 4. The similarity between these models:
- The Empirically-driven model has a problem of overfitting, and the Mechanistic model also has a problem of overfitting if we add too many parameters to it.

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

- By the definition of the mechanistic model and the definition of the empirically-driven model, we know that empirically-driven model are much more flexible than the mechanistic model. According to what we learned in the lecture, we know that more flexible means less interpretative and it also means bigger error. Therefore, I think mechanistic model is much easier to understand since it is more interpretative.

c) Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models.

- (In office hour, professor said that we can skip this question since we haven't learnt yet.)

Question 6: A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions:1. Given a voter's profile/data, how likely is it that they will vote in favor of the candidate? 2.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.

Answer:

1. Given a voter's profile/data, how likely is it that they will vote in favor of the candidate?

- For question 1, this question should be predictive.
- From the lecture, we know that the aim of predictive models is to accurately predict "the response for future observations, and the aim of inferential models is to test theories, causal claims, and"state the relationship between outcome and predictor". (the page #26 from the textbook) Clearly, for question 1, we can know that the voter's profile/data is the predictor, and since we want to predict whether they will vote in favor of the candidate, we can know that the probability of they vote in favor of the candidate is our Y. Then, we can know that this question 1 is predictive.

2.How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?

- For question 2, this question should be inferential.
- In this question, we focus on how a voter's likelihood of support for the candidate change if they had personal contact with the candidate, which means that we want to find the relationship between the voter's likelihood of support for the candidate and whether they had personal contact with the candidate. So, clearly, we can know that this question is inferential.

Exploratory Data Analysis

This section will ask you to complete several exercises. For this homework assignment, we'll be working with the mpg data set that is loaded when you load the tidyverse. Make sure you load the tidyverse and any other packages you need.

Exploratory data analysis (or EDA) is not based on a specific set of rules or formulas. It is more of a state of curiosity about data. It's an iterative process of:

generating questions about data visualize and transform your data as necessary to get answers use what you learned to generate more questions A couple questions are always useful when you start out. These are "what variation occurs within the variables," and "what covariation occurs between the variables."

You should use the tidyverse and ggplot2 for these exercises.

Exercise 1:

- We are interested in highway miles per gallon, or the hwy variable. Create a histogram of this variable. Describe what you see/learn.

```
# install packages
# install.packages("tidyverse")
# install.packages("tidymodels")
# install.packages("ISLR")

# load the packages we need
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   2.1.2     v forcats 0.5.1
```

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

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 0.2.0 --
```

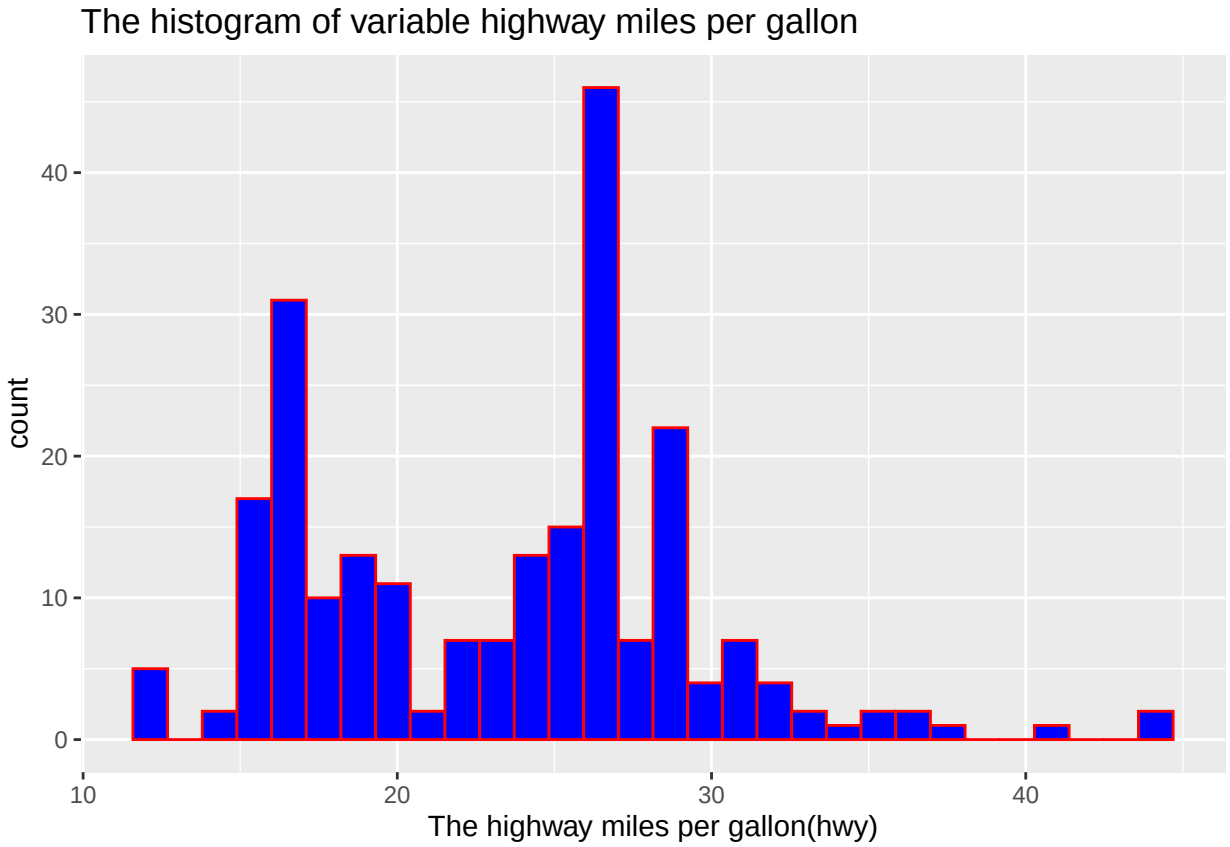
```
## v broom        0.7.12     v rsample      0.1.1
## v dials        0.1.0      v tune         0.2.0
## v infer        1.0.0      v workflows    0.2.6
## v modeldata    0.1.1      v workflowsets 0.2.1
## v parsnip      0.2.1      v yardstick    0.0.9
## v recipes      0.2.0
```

```
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Search for functions across packages at https://www.tidymodels.org/find/
```

```
library(ISLR)
```

```
# Create a histogram of this variable of hwy
ggplot(mpg, aes(x = hwy)) + geom_histogram(fill = 'blue', color = 'red') +
  labs(title = "The histogram of variable highway miles per gallon",
       x = 'The highway miles per gallon(hwy)')
```

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



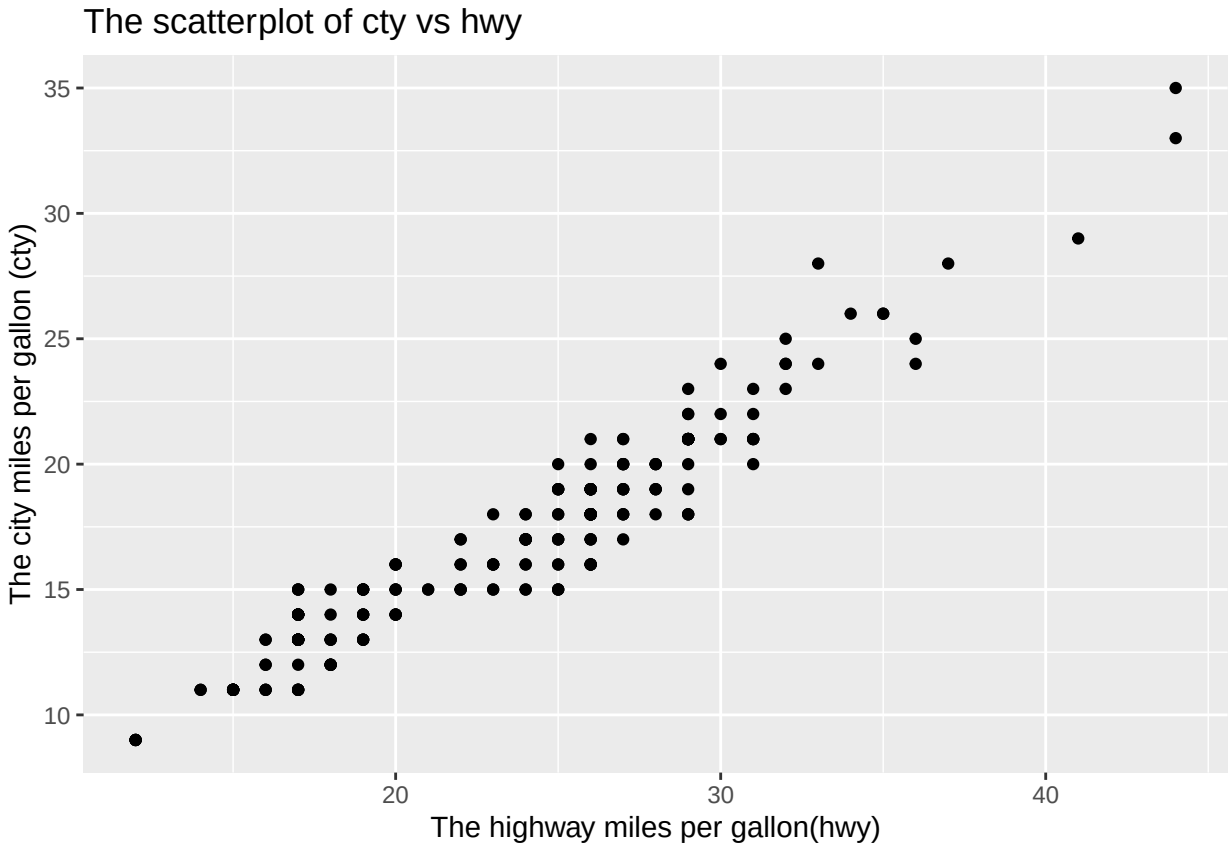
Answer:

- For the histogram of this hwy, we can see that the maximum count occurs when highway miles per gallon is close to 26, and there are two peaks in our graph (the first one is close to 17, and the second one is close to 26). We can also see that this data is left-skewed. (Most of hwy data is less than 30)

Exercise 2:

- Create a scatterplot. Put hwy on the x-axis and cty on the y-axis. Describe what you notice. Is there a relationship between hwy and cty? What does this mean?

```
help(mpg)
# Create a scatterplot. Put hwy on the x-axis and cty on the y-axis.
ggplot(mpg, aes(x = hwy, y = cty)) + geom_point() + labs(title = "The scatterplot of cty vs hwy",
       x = 'The highway miles per gallon(hwy)', y = 'The city miles per gallon (cty)')
```

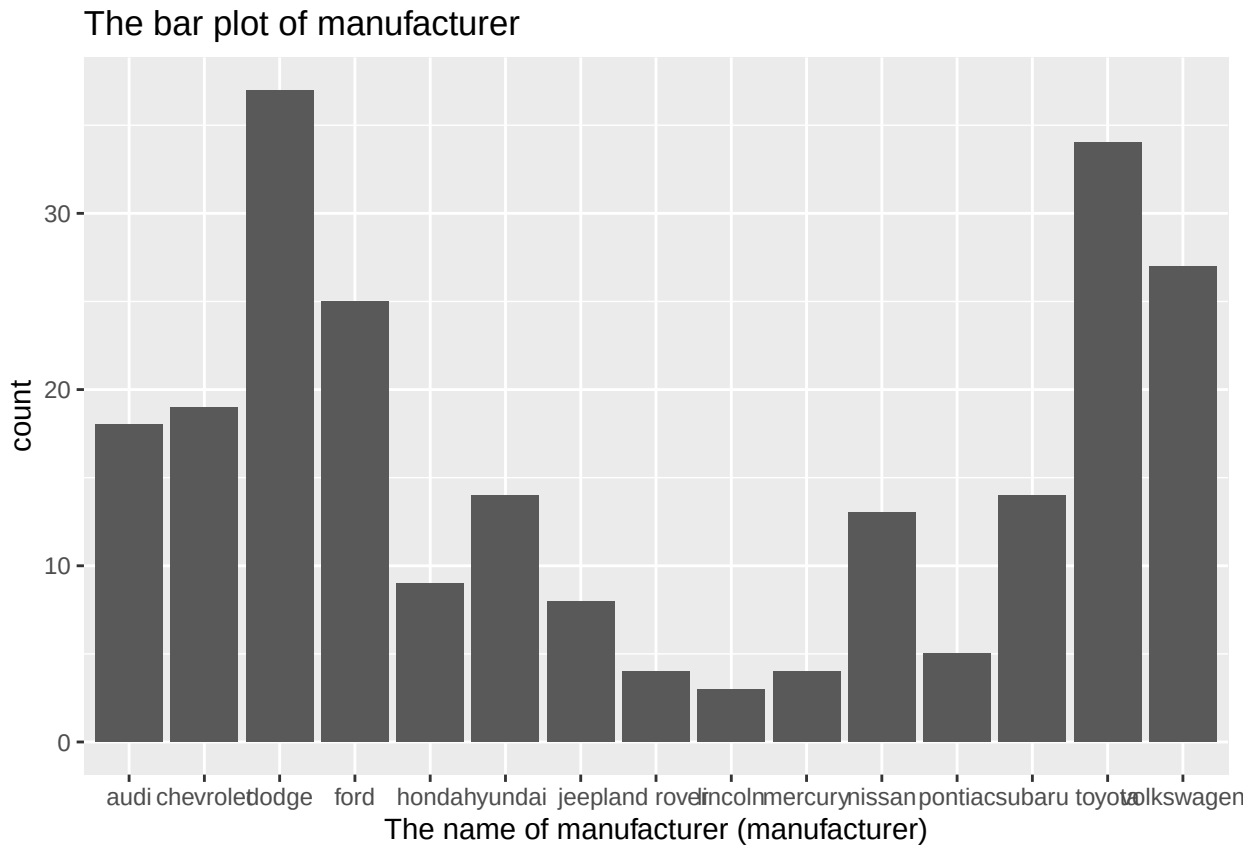


Answer: From the scatterplot, we can see that there is a linear relationship between hwy and cty. According to the scatterplot, we can know that the linear relationship between hwy and cty is positive (if we draw a approximate line, we can find the slope of the line is positive), then there is a positive relationship between hwy and cty, which means that as the amount of highway miles per gallon increases, the amount of city miles per gallon will also increase.

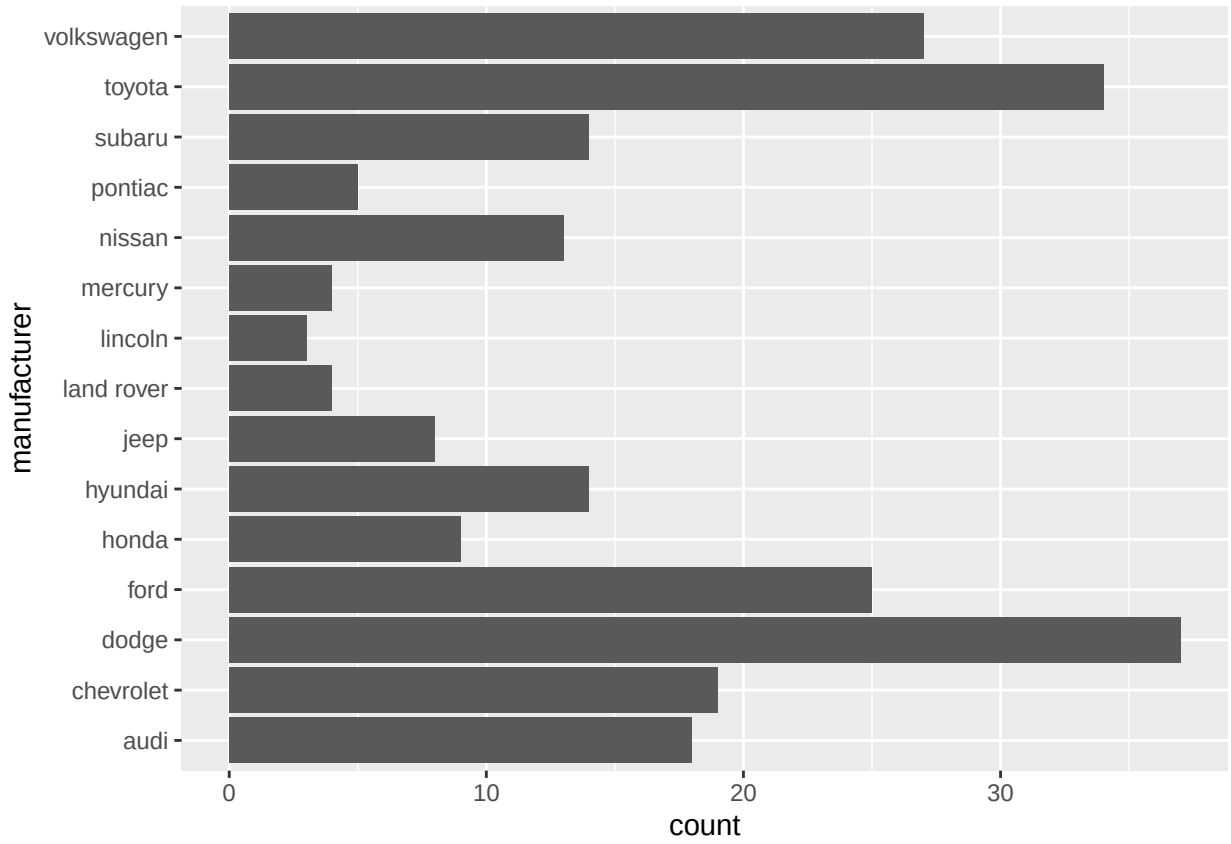
Exercise 3:

- Make a bar plot of manufacturer. Flip it so that the manufacturers are on the y-axis. Order the bars by height. Which manufacturer produced the most cars? Which produced the least?

```
# 1. Make a bar plot of manufacturer
ggplot(mpg, aes(x=manufacturer)) + geom_bar()+
  labs(title = "The bar plot of manufacturer", x = 'The name of manufacturer (manufacturer)')
```




```
# 2.Flip it so that the manufacturers are on the y-axis.
ggplot(mpg, aes(y= manufacturer)) + geom_bar()
```



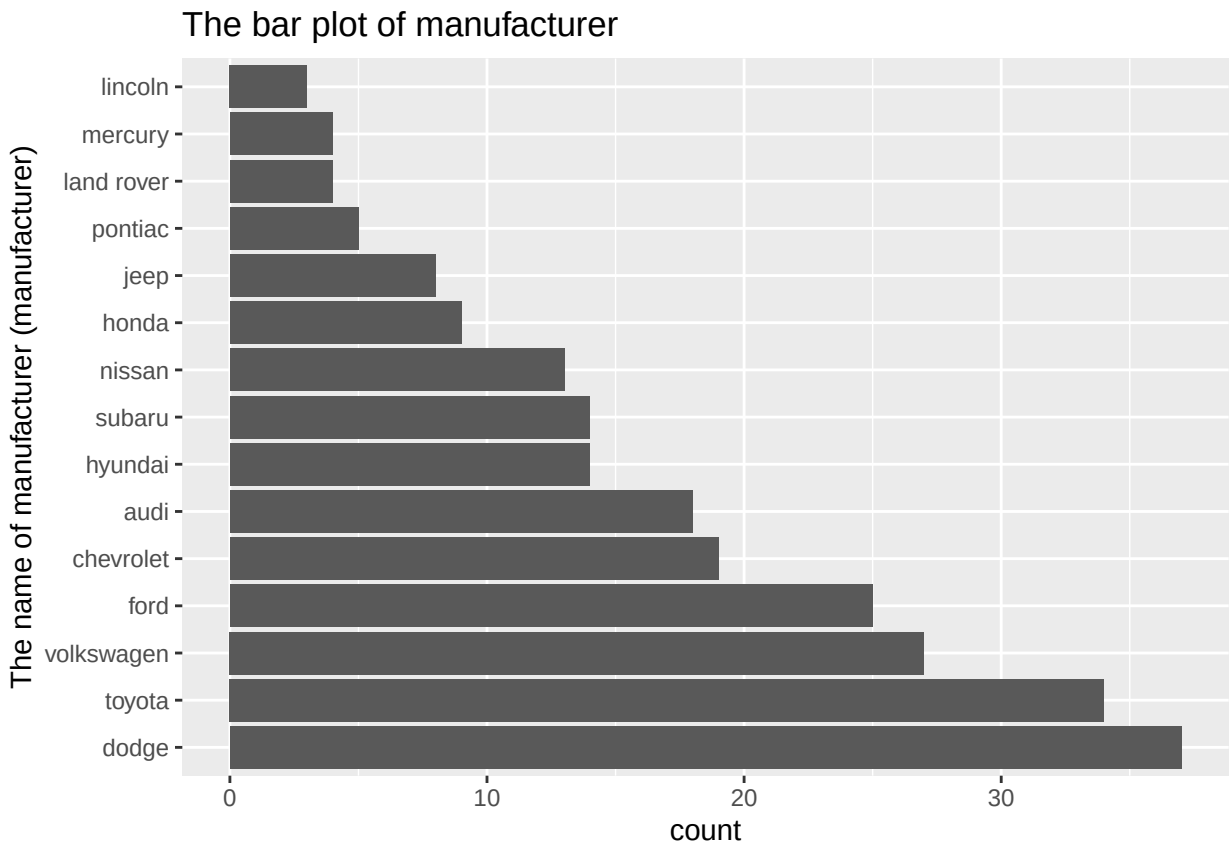

```
# 3. Order the bars by height.
newmanufacturer<- mpg %>%
  group_by(manufacturer) %>%
  summarise(count = n()) %>%
  arrange(count)
newmanufacturer
```

```
## # A tibble: 15 x 2
##    manufacturer count
##    <chr>        <int>
##  1 lincoln          3
##  2 land rover       4
##  3 mercury          4
##  4 pontiac          5
##  5 jeep             8
##  6 honda            9
##  7 nissan          13
##  8 hyundai         14
##  9 subaru          14
## 10 audi            18
## 11 chevrolet       19
## 12 ford            25
## 13 volkswagen      27
## 14 toyota          34
## 15 dodge           37
```

```
is.data.frame(newmanufacturer)
```

```
## [1] TRUE
```

```
# make a new bar plot of manufacturer so that the manufacturers are on the y-axis and order the bars by
ggplot(newmanufacturer,aes(y= reorder(manufacturer, -count),  x=count)) + geom_bar(stat='identity') +
  labs(title = "The bar plot of manufacturer", y = 'The name of manufacturer (manufacturer)')
```



Answer: From the bar plot we made, we can know that Dodge produced the most cars, and Lincoln produced the least cars.

Exercise 4:

- Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?

```
# Recall what we learn in PSTAT 10
# A model formula is given by y~x where  y is a numeric vector which is grouped according to the
# value of x
# In this question hwy is grouped by cyl.
# So, x = cyl, y = hwy
ggplot(mpg, aes(x= as.factor(cyl), y = hwy))+ geom_boxplot() +
  labs(title = "The box plot of hwy, grouped by cyl", y = 'The highway miles per gallon(hwy)',
       x= 'The city miles per gallon (cyl)')
```

Answer: Clearly, we can find that there is pattern in this box plot, which shows us that as the amount of cyl increases, the amount of hwy will decrease. It means that, there is a negative relationship between cyl and hwy.

Exercise 5:

- Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset. (Hint: You can find information on the package here.)
- Which variables are positively or negatively correlated with which others? Do these relationships make sense to you? Are there any that surprise you?

```
# install.packages('corrplot')
# load the package
library(corrplot)
```

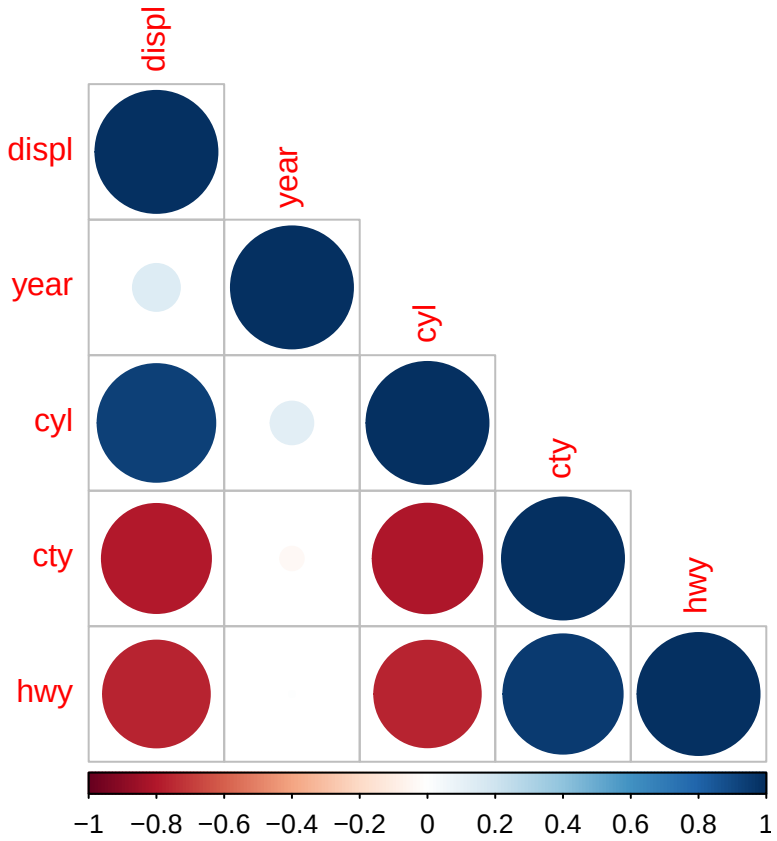
```
## corrplot 0.92 loaded
```

```
# let see the first 10 rows of mpg dataset
head(mpg)
```

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
```

```
## 6 audi         a4      2.8  1999     6 manual(m5) f        18     26 p      compa~
```

```r
# since we want to find the correlation matrix of the mpg dataset,
# we should use only numerical variable
newmpg <- mpg %>% select(-c(manufacturer,model,trans,drv,fl,class))
# make a lower triangle correlation matrix of the mpg dataset
corrplot(cor(newmpg), type = 'lower')
```



Answer:

According to the corrplot, we can know that:

- 1. hwy (highway miles per gallon) has negatively correlated with displ (engine displacement, in litres)
- 2. hwy (highway miles per gallon) has negatively correlated with cyl (number of cylinders)
- 3. hwy (highway miles per gallon) has positively correlated with cty (city miles per gallon)
- 4. cty (city miles per gallon) has has negatively correlated with displ (engine displacement, in litres)
- 5. cty (city miles per gallon) has has (little) negatively correlated with year (year of manufacture)
- 6. cty (city miles per gallon) has has negatively correlated with cyl (number of cylinders)
- 7. cyl (number of cylinders) has positively correlated with displ (engine displacement, in litres)
- 8. cyl (number of cylinders) has (little) positively correlated with year (year of manufacture).
- 9. Year (year of manufacture) has (little) positively correlated with displ (engine displacement, in litres)

These relationships make sense to me. Explanation:

We can understand these relationships through our everyday experience. For example, Year (year of manufacture) has positively correlated with displ (engine displacement, in litres) because when people use a car for a long time, the probability of needing to replace the engine is getting higher and higher. So, it is very clear for us to understand that the Year has positively correlated with displ.

No, there is no surprise.