# PSTAT131 HW1

2022-04-03


## Liangchen Xia


## Question 1

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

Supervised learning is a kind of machine learning methods have the Response variable, such output, target, dependent variable. In supervised learning, the Response variable is the supervisor. But the Unsupervised learning is a kind of machine learning methods do not have the Response variable. It learn without a supervisor. We use it to discover the potential patterns hidden behind the dataset.

## Question 2

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

Both regression models and classification models are supervised learning methods with response variable. The main difference between the two kinds of models is type of response variable. The response variable of Regression is Numerical values such as price, blood pressure, population. And the response variable of classification is Categorical values such as genders, survived or died, recoverable or unrecyclable.

## Question 3

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

Two metrics for regression ML problems:

- Predict age by lift style such as living expenses, shopping habits, or networks of friends
- Predict the incomes by personal features such as Education, residential address, consumption level

Two metrics for classification ML problems:

- Predict whether a person would crime or not by their behavior such as family, spending, jobs and other features
- Predict the fit partner or not by personal characteristics such as Personality, family income, interests and hobbies

## 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.

- Descriptive models: Choose this model to best visually emphasize a trend, like discribe the central trend, disperse trend or the whole distribution characterastics in data.
- Inferential models: Choose this model to test theories, like evaluate the performance and confidence of prediction or estimations of models.
- Predictive models: Choose this model to predict the values of response variable exactly by predictors which mean predict the value with minimum reducible error.

## Question 5

Q1

- The Mechanistic models uses a theory to predict what will happen in the real world which mean effects of predictors on response. This kind of models are constructed by the machinism which describe how predictors determine the response
- Empirically-driven models are constructed by the data itself without any machinism or causality such mean empirical modeling, studies real-world events to develop a theory. In this kind of models, we we cannot explained how predictors affect response
- Both Mechanistic model and Empirically-driven model are overfitting. They are in the same format sometimes such as linear regression or tree models

Q2

- I think the Mechanistic models are easier to understand. Becausethe models are uses a theory to predict what will happen in the real world which mean effects of predictors on response. It's constructed based on the machinism

Q3

- The Mechanistic models add factors affect response from critical to irrevelant, when the biases decreasing but variance increasing. So we remove factors will make variance increasing more than biases decreasing
- For the Empirically-driven models, the complexity of model would make biased decreasing but variance increasing. So we need select a suitable complexity when variance increasing and biased decreasing balanced

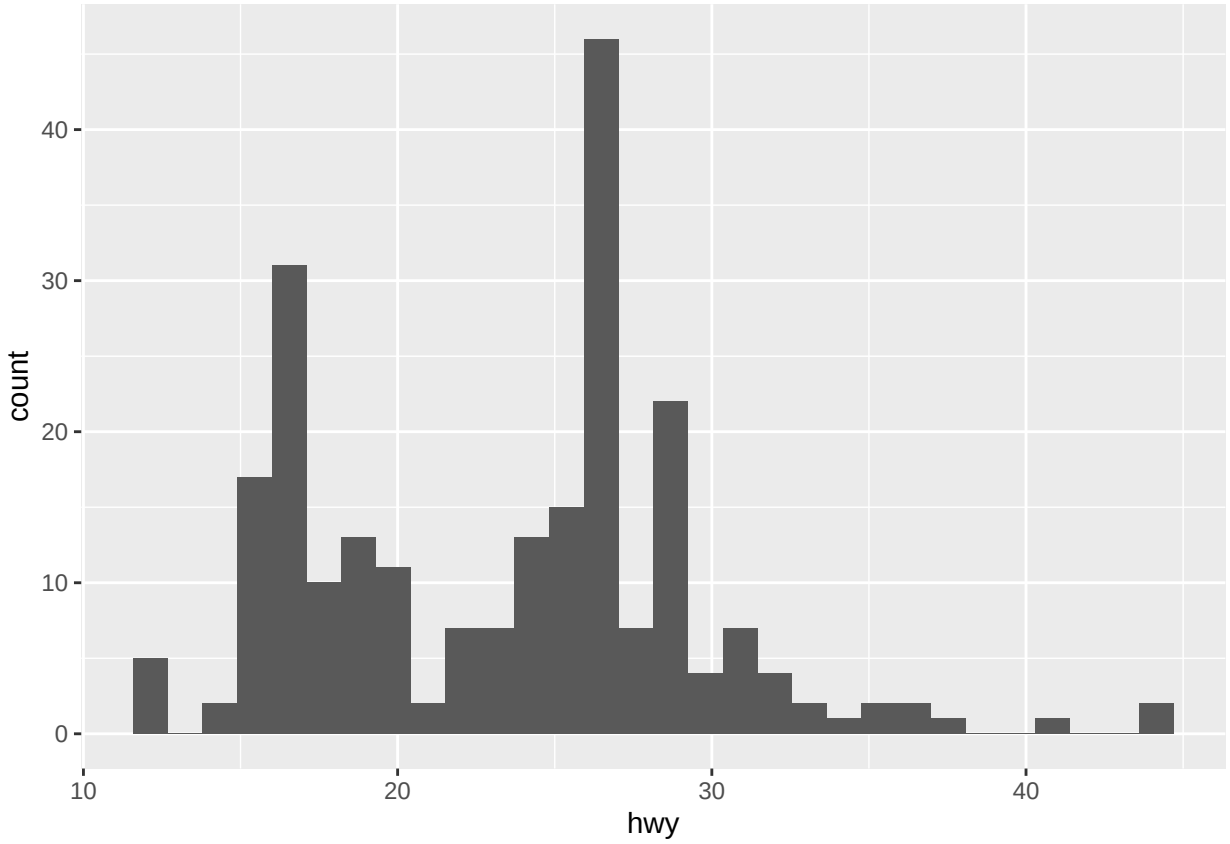
## Question 6

For both of question, we are interested in the probability of voting for candidate. The first question is focused on the quality of our prediction on voting for candidate. We need know how likely the vote in favor of the candidat. And for the second question is focused on the quakity change on voting for ccandidate by our prediction. The predictive model only consider which candidate is voted by these voters. So, I think both the two question are inferential model.

## Exercise 1

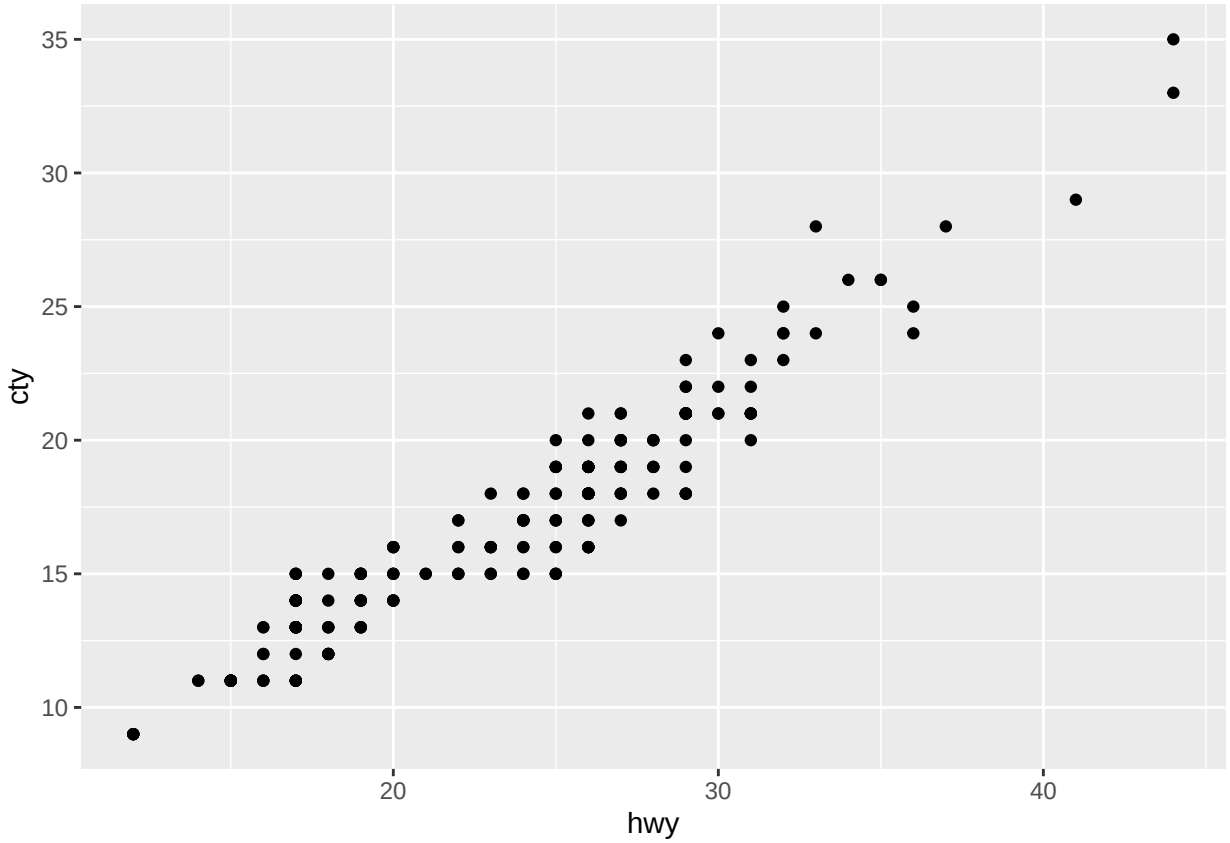
```
data(mpg)
mpg%>%ggplot(aes(hwy))+geom_histogram()
```

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




## Exercise 2

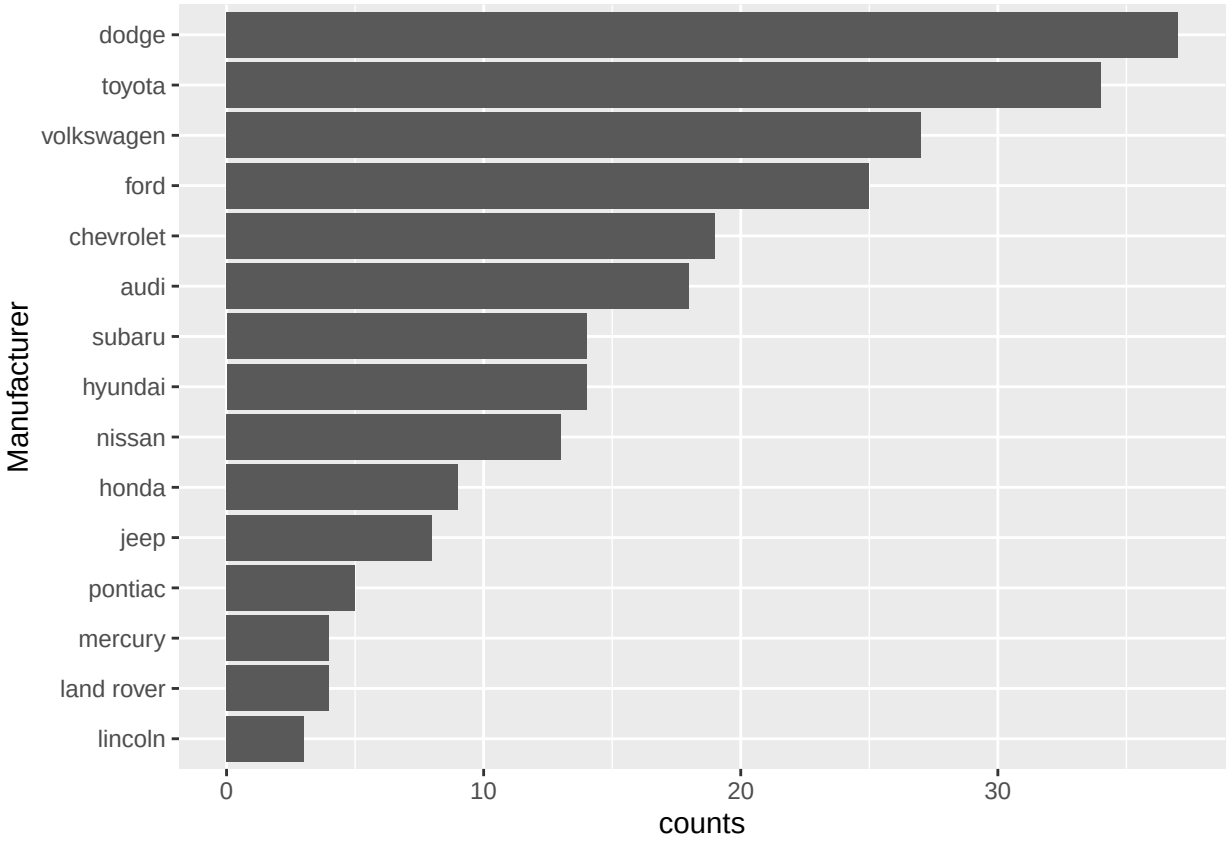
```
mpg%>%ggplot(aes(x=hwy,y=cty))+geom_point()
```




The scatters of `hwy` and `cty` are on a straight line, which mean that the two variables are linearly positively related. So as the increasing `hwy` would lead to `cty` also increasing linearly.

## Exercise 3

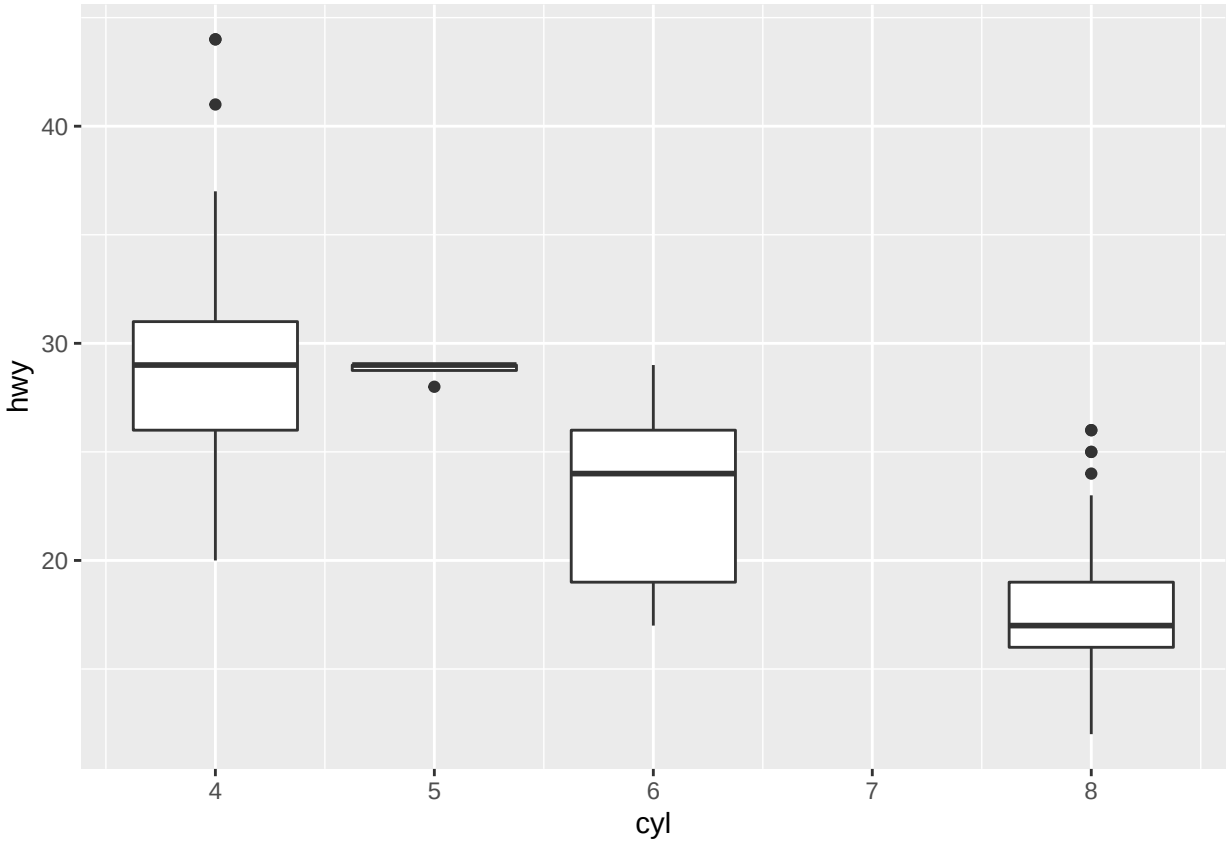
```
mpg%>%group_by(manufacturer)%>%summarise(counts=n())%>%
  ggplot(aes(x=reorder(manufacturer,counts),y=counts))+
  geom_bar(stat='identity')+coord_flip()+labs(x='Manufacturer')
```




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

## Exercise 4

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



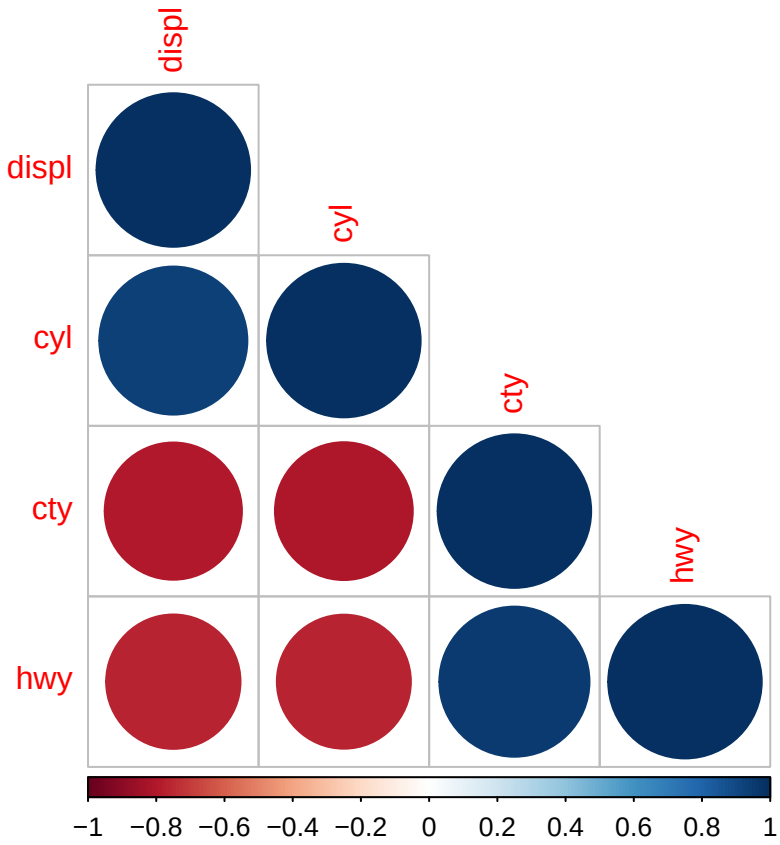

It shows as `cyl` increasing, `hwy` would decrease.

## Exercise 5

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(mpg%>%select(displ,cyl,cty,hwy)),type='lower')
```




In this question, we can see the

- `cyl` is positively correlated with `displ` and `cyl`.
- `cty` is negatively correlated with the `displ` and `cyl` and positively correlated with `cty`.
- `hwy` is negatively correlated with the `displ` and `cyl` and positively correlated with `cty` and `hwy`.

We realize the correlations between these variables are strong, so I believe it is make sense.