# PSTAT 130 HW1

2022-04-01

## Question 1

Supervised learning is a type of machine learning approach that includes input variable and an output variable and treats the output variable as a supervisor.

Unsupervised learning is a type of machine learning method that does not employ a output variable and is used to find hidden patterns in a dataset.

## Question 2

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. Classification is the task of predicting a discrete class label. The response variable of classification is nominal such as genders, survived/died and other quality variable, but response variable of regression is numerical such as incomes, temperature and other quantity variable. Regression is the task of predicting a continuous quantity.

## Question 3

Metrics for regression ML problems:

Predict height by other features of body such as weight, length of feet and diameter of waist. We probably can use Polynomial Regression to estimate

Predict the incomes by demographic features such as age, gender, education level and others.We probably can use linear Regression to estimate

Metrics for classification ML problems:

Predict whether a person suffered from skin disease by their genes expression. We probably can use Logistic Regression to estimate

Predict whether a person would default by their economy state such as income, spending, jobs and other features. We probably can use K-Nearest Neighbours to estimate

## Question 4

- Descriptive models: To discribe the central trend, disperse trend or the whole distribution characterastics of the dataset.
- Inferential models: To evaluate the performance and confidence of prediction or estimations of models
- Predictive models: To predict the values of response variable exactly by predictors.

## Question 5

Mechanistic models: These are models that are built by machinism that describe how predictors determine the response. The influence of predictors on reaction can be understood. Empirically-driven models are built entirely by the data, with no machinism or causation involved. We don't know how predictors influence reaction. The two types of models, such as linear regression and tree models, are sometimes in the same format.

In general mechanistic models are easy-explained, due to the models are constructed based on the machinism of how predictors determine the response.

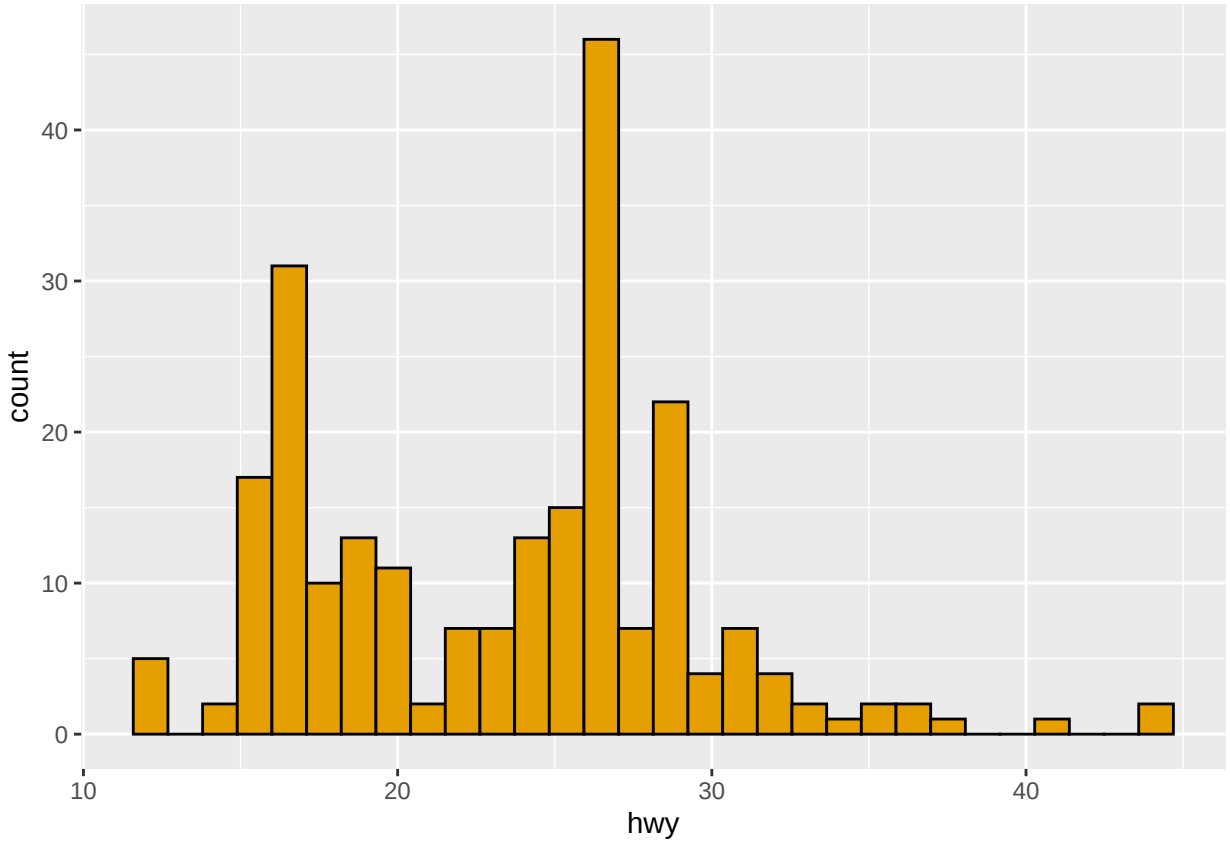
While we introduced elements that alter response from crucial to irrelevant in mechanistic models, the biases decreased but the variance increased. As a result, we remove elements that cause variance to increase more than biases decrease. In empirically-driven models, model complexity causes bias to decrease while variance increases, therefore we choose a reasonable complexity where variance increases while bias decreases.

## Question 6

We are interested in the probability of voting for candidate in both questions because they are inferential models. The first question concerns the quality of our prediction on candidate voting, while the second concerns the change in the quality of our prediction on candidate voting. Only which candidate is voted for by these voters is taken into account by the predicted algorithm.

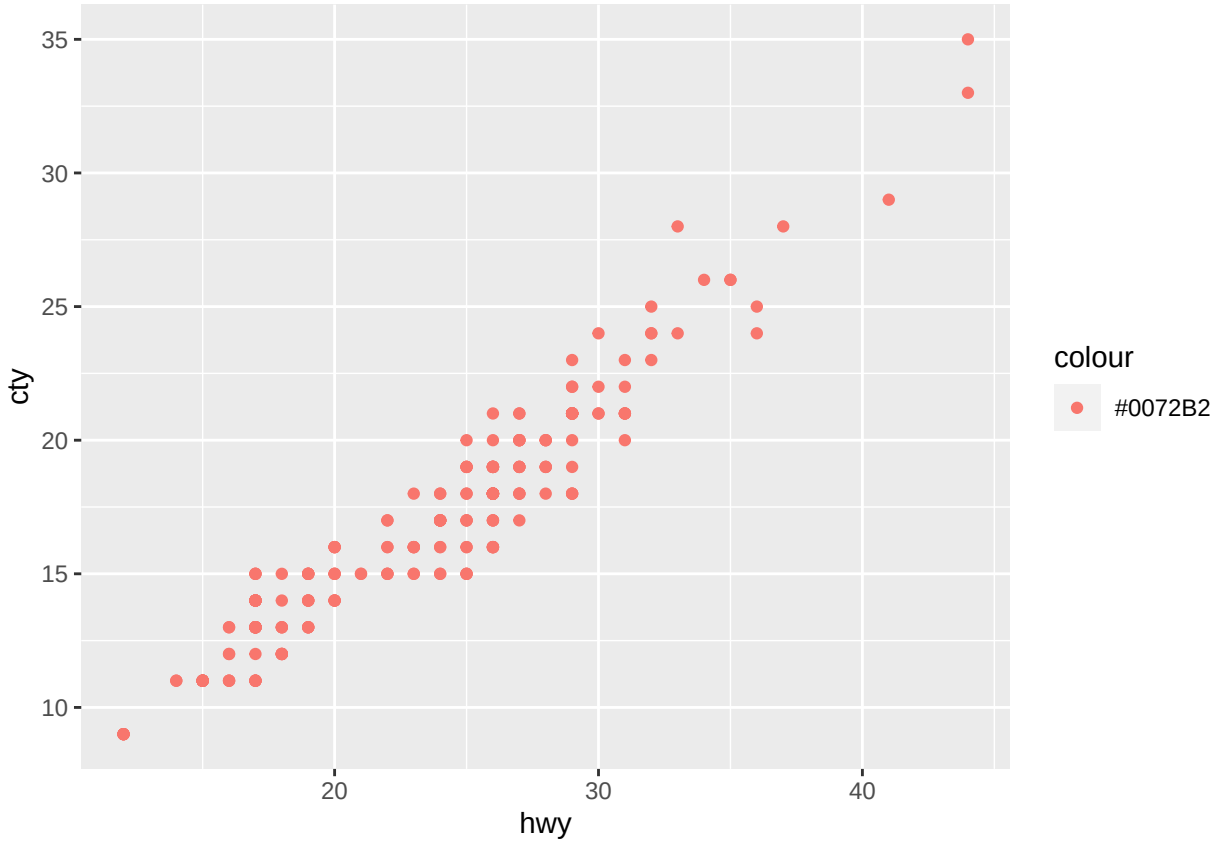
## Exercise 1

```
data(mpg)
mpg%>%ggplot(aes(hwy))+geom_histogram( colour="black",fill= "#E69F00")
```




## Exercise 2

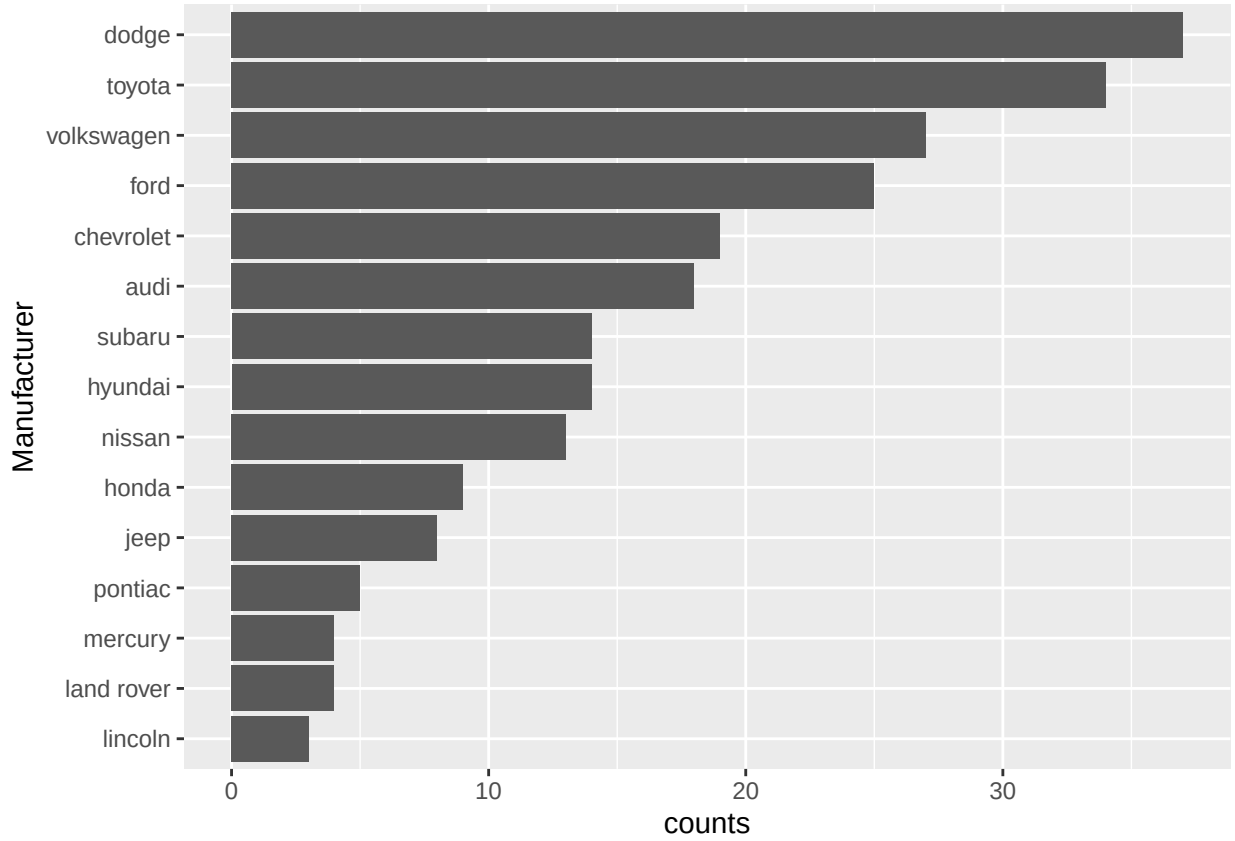
```
mpg%>%ggplot(aes(x=hwy,y=cty, colour="#0072B2"))+geom_point()
```




The scatters of `hwy` and `cty` are locted on a straight line, which indicates that the two variables are linearly positively related. This means that increasing `hwy` would lead to `cty` also increasing linearly.

## Exercise 3

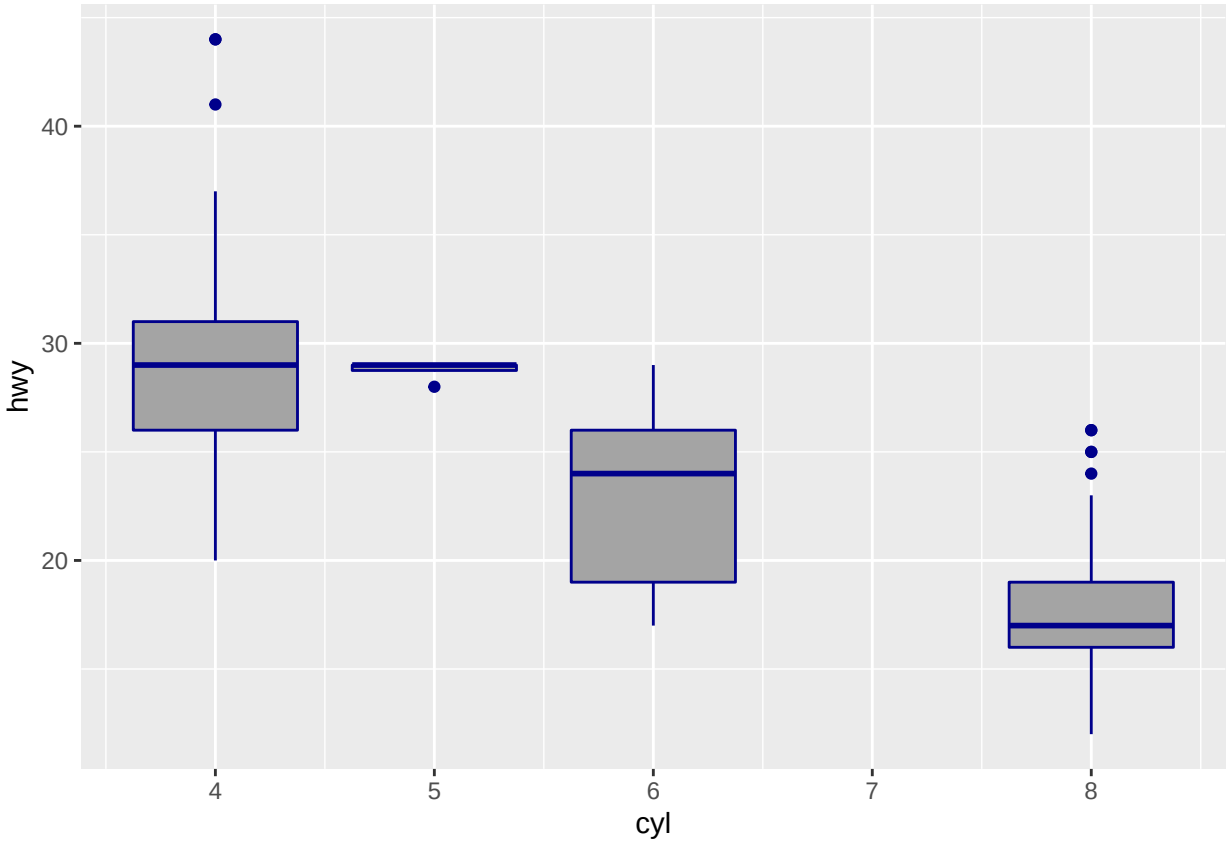
```
mpg%>%group_by(manufacturer)%>%summarise(counts=n())%>%
  ggplot(aes(x=reorder(manufacturer,counts),y=counts))+
  geom_bar(stat='identity')+coord_flip()+labs(x='Manufacturer')
```




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

## Exercise 4

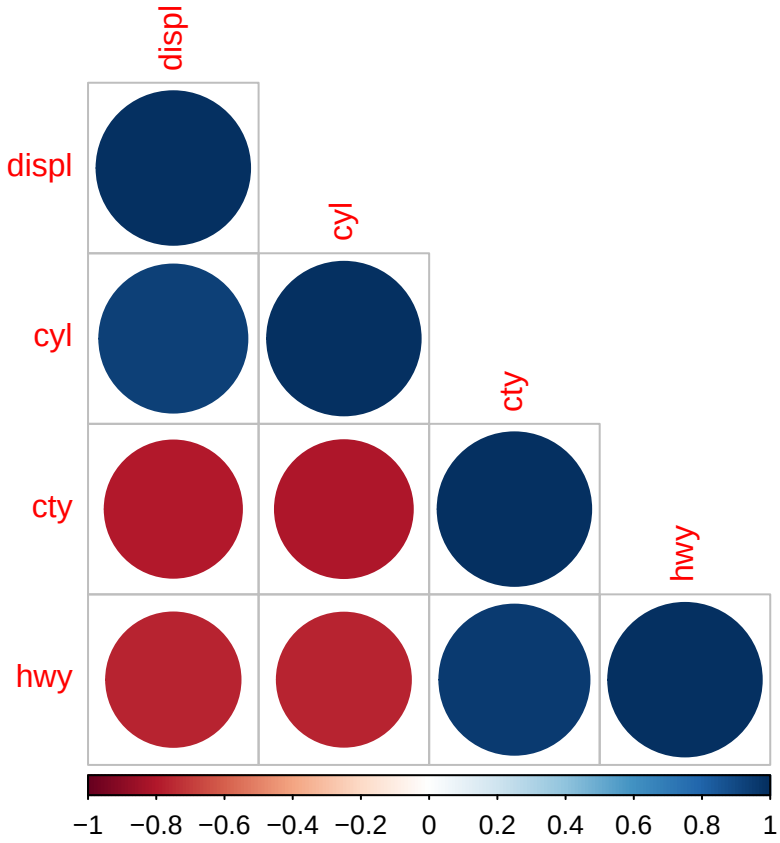
```
mpg%>%ggplot(aes(x=cyl,y=hwy,group=cyl))+geom_boxplot(fill='#A4A4A4', color='darkblue')
```




It shows that as cyl increasing, hwy would decrease.

## Exercise 5

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




Categorical variables and predictor `year` are removed firstly. As a result, `hwy` is negatively correlated with `displ` and `cyl` but positively correlated with `cty` and `hwy`. `cty` is negatively correlated with `displ` and 'cyl' but positively correlated with `cty`. `cyl` is positively correlated with `displ` and `cyl`. The correlations between these variables are strong, thus it make sense to me.