hw1

Jiaxuan Sun

2022-10-09

#Question1

Supervised learning is to learn the model with observed output and input, with the actual data Y be the answer key for the learning model to make it accurate when making prediction, estimation, model selection, and inference. Unsupervised learning is to learn the model with only input variables, which is the predictors, while the response Y is unknown so that there is nothing to correct or examine the outcome. The difference between them is that whether there are supervisors or not, supervised learning learn model with known response to supervise the model, but unsupervised learning learn model by clustering data.

#Question2

The output variables of regression model are numeric values like price or temperature. Regression models input data and output continuous quantities to predict. And the output variables of classification model are categorical,qualitative outcomes. Classification models input data and classify them to discrete labels.

#Question3

two commonly used metrics for regression ML problems: R-Squared, Mean Suare Error(MSE) two commonly used metrics for classification ML problems: Accuracy, Precision

#Question4

Descriptive models: select appropriate model to illustrate a trend in data Inferential models: show relationship between input and output variables to test theories, whether something causes something else Predictive models: predict future response with minimum reducible error based on a combination of given set of input data

#Question5

-Mechanistic models make assumption in a parametric form for function f, the relationship between predictors and response. Empirically-driven models predict future outcomes based on observations, not theory. It is possible to add parameters to mechanistic models and the predicted values won't match true unknown f. The Empirically-driven models require a large number of observations. They are both predictive models and they might overfit data.

-A mechanistic model is easier to understand. Because we define the function ourselves and by adding or deleting parameters, we could see the relationship between different predictors and responses directly and then we could understand the function f.

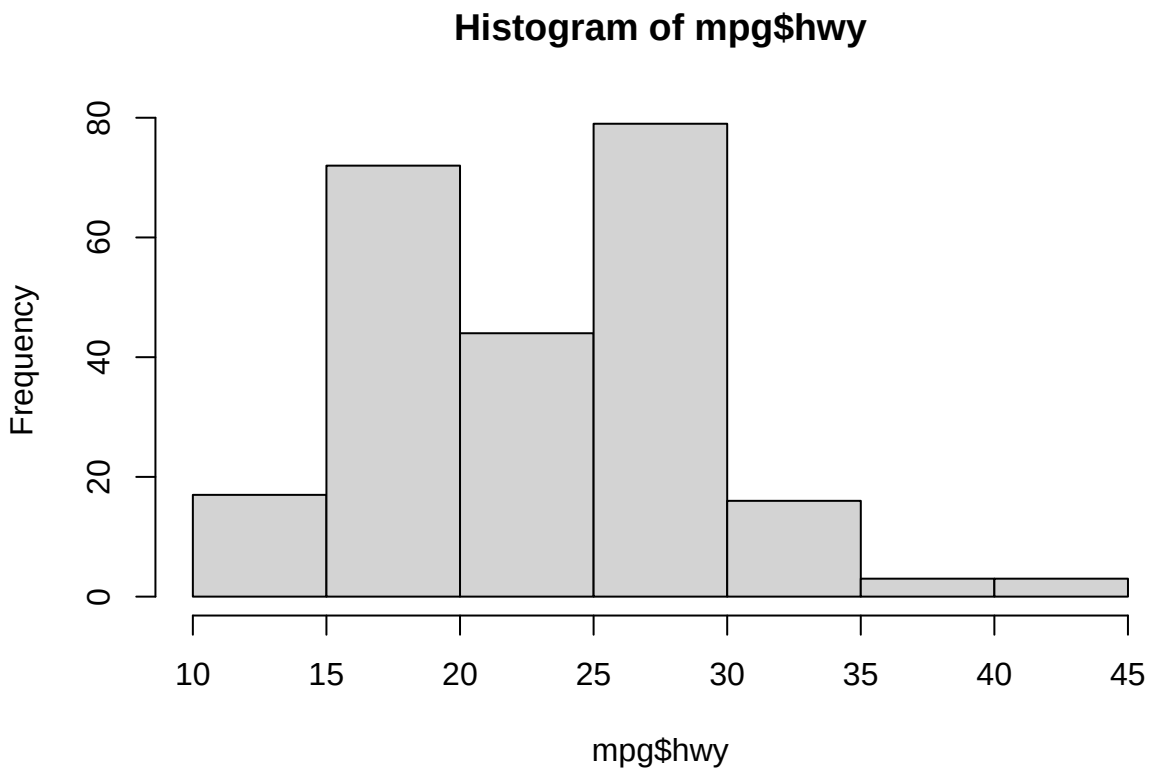
-Mechanistic model has high bias and low variance because of undetermined flexibility, and empirically-driven model has high flexibility, with low bias and high variance since it match existing observations. In both cases, higher flexibility means overfitting to the data points.

#Question6

The first one is predictive model because we want to predict future decisions of voters,response, based on their profiles, existing data. The second one is inferential since we want to determine the relationship between personal contact and likelihood of support, the relationship between predictors and responses.

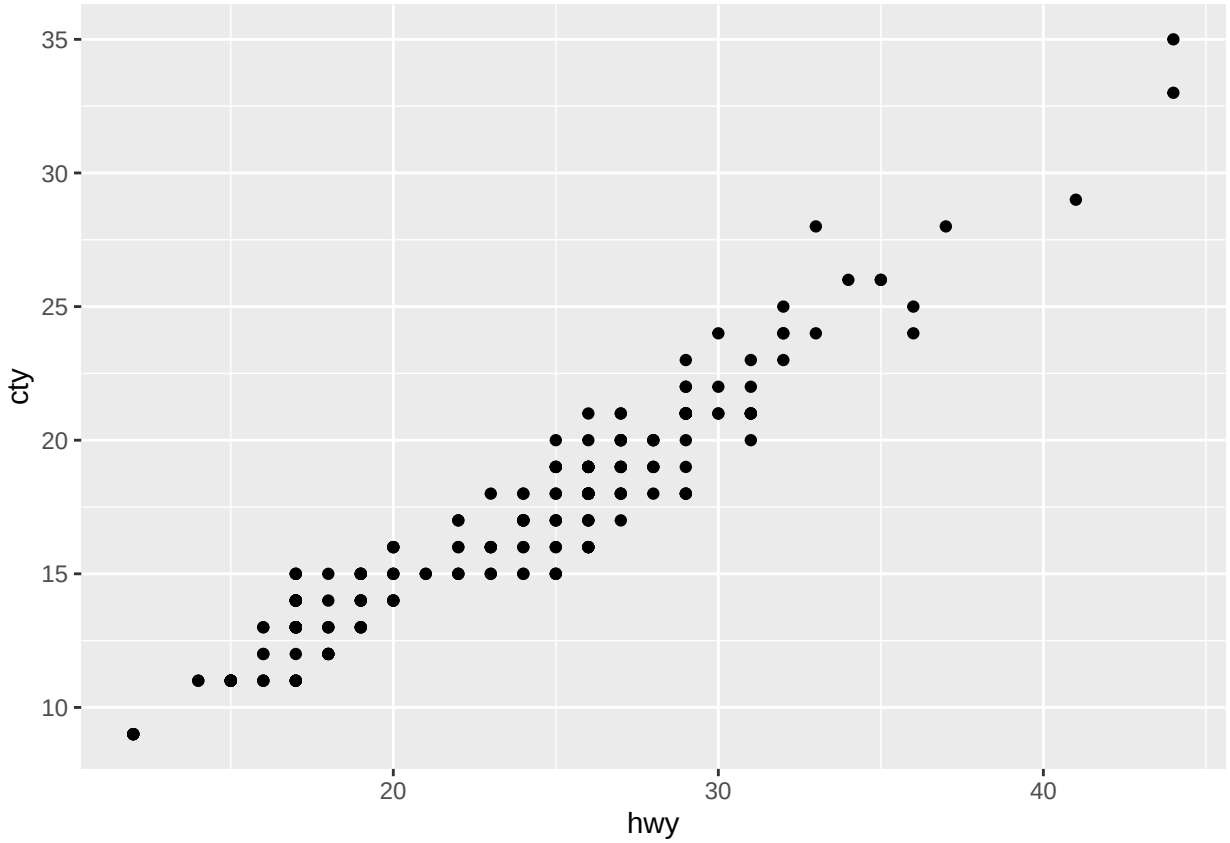
```
library(tidyverse)
library(tidymodels)
library(ISLR)
```

```
##Exercise1
hist(mpg$hwy)
```



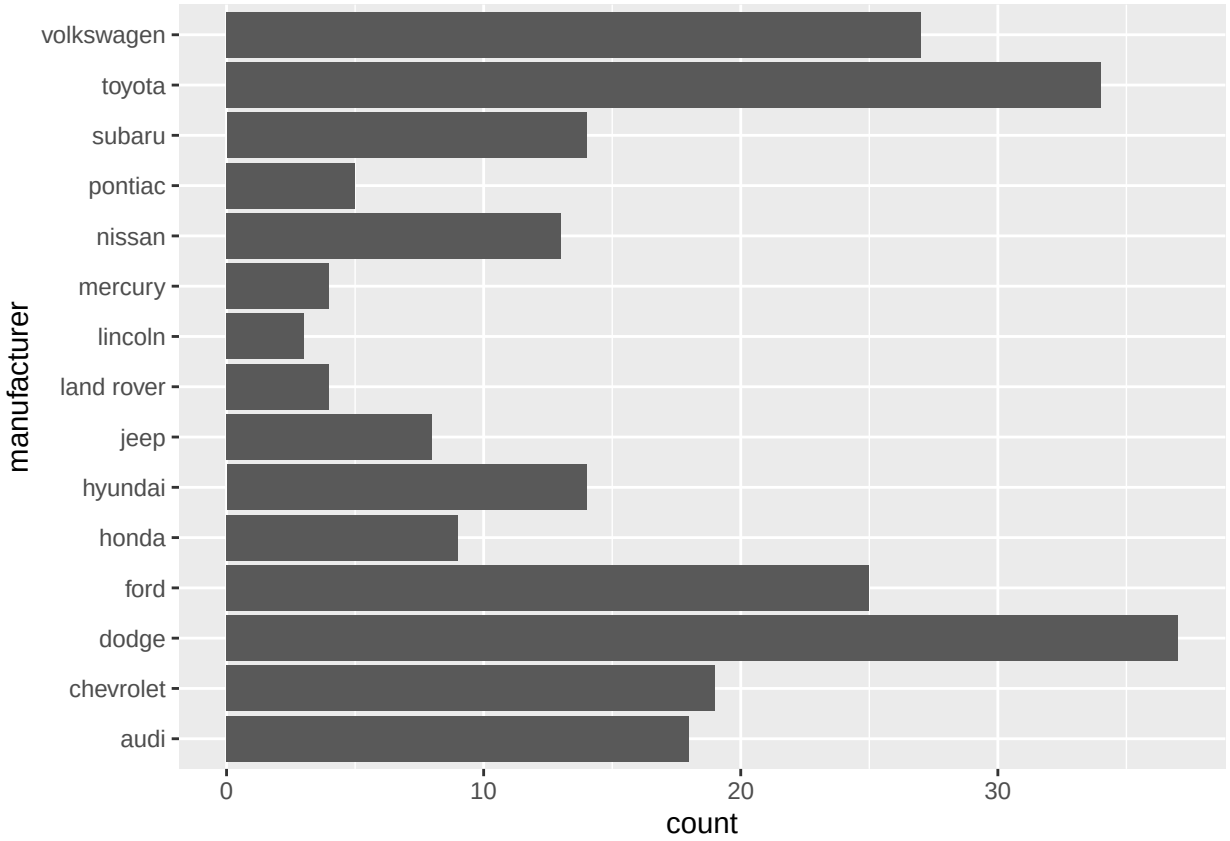
Most vehicles' highway miles per gallon values are in the range of 15-30, and a few vehicles have mpg value greater than 30 or less than 15.

```
##Exercise2
scatterplot <- ggplot(mpg,aes(x=hwy,y=cty)) + geom_point()
print(scatterplot)
```



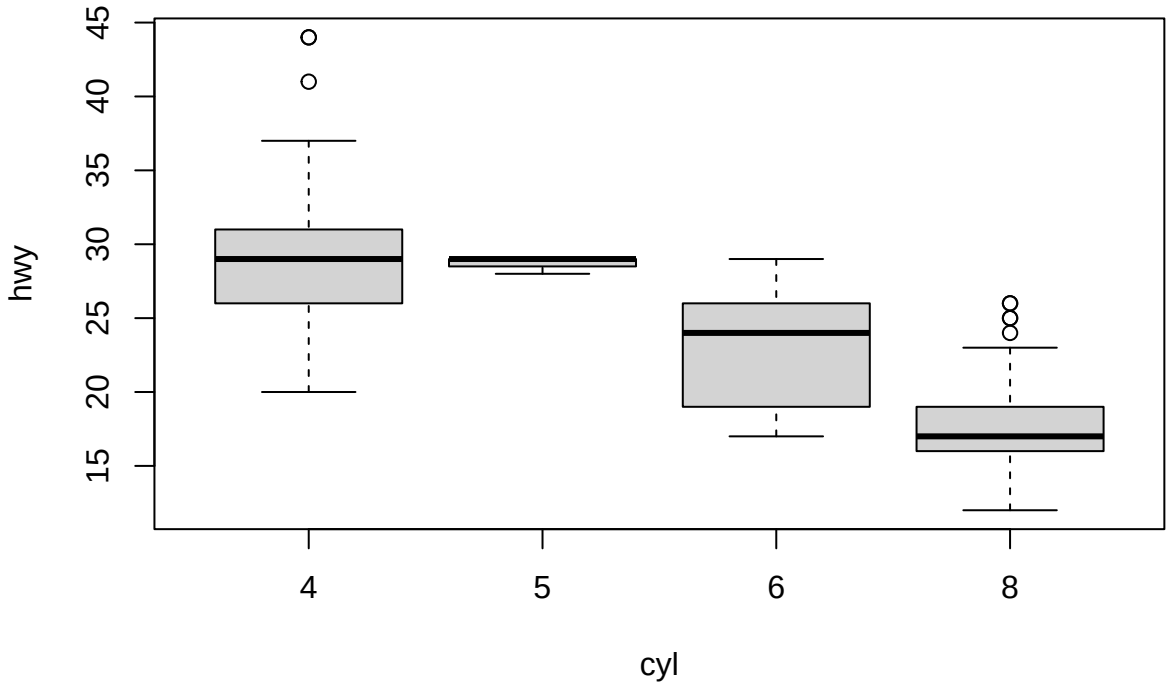
There is a roughly positive linear relationship between these two variables, this means that generally, the vehicles having high value of city miles per gallon would have high mpg in highway.

```
##Exercise3
barplot <- ggplot(mpg,aes(x=manufacturer)) + geom_bar() +coord_flip()
print(barplot)
```



Dodge produces the most cars and Lincoln produces the least.

```
##Exercise4
boxplot(hwy~cyl,data=mpg)
```



There is a pattern that as the number of cylinders increases, value of miles per gallon decreases. More cylinders consume more fuel so that the vehicle has lower value of mpg.

```
##Exercise5
library(corrplot)
mpg1 <- mpg%>%
  select(displ,cyl,hwy,cty)
corr <- cor(mpg1)
corrplot(corr,method='number')
```

'displ' is positively related to 'cyl' and negatively related to 'hwy' and 'cty'. These relationship make sense because more cylinders causes more engine displacement, consuming more fuel so that result in low mpg. 'cyl' is also negatively related to 'hwy' and 'cty' due to previous explanation. According to exercise2, 'cty' is positively related to 'hwy', one thing that surprises me is that the correlation between these two variables are close to 1.