131hw1_Yiting

Yiting Zhang

2022-04-08

Machine Learning Main Ideas

Question 1

- Supervised learning: we know the outcome just like we know the "answer key". We can supervised the accuracy of prediction and modeling. The response variable is the supervisor.
- Unsupervised learning: we cannot supervise its learning because we don't know the outcome.It is like we never know the answer key as there is no response thus learning without a supervisor.

Question 2

- Regression model: the outcome is continuous or say the response variable Y is quantitative.
- Classification model: the outcome is categorical or say the response variable Y is qualitative.

Question 3

- Two commonly used metrics for regression ML problems: Mean Squared Error (MSE) and Mean Absolute Error (MAE)
- Two commonly used metrics for classification ML problems: Accuracy and F1-score;

Question 4

- Descriptive models: choose model to best visually emphasize a trend in data
- Inferential models: what features are significant; to test theories; state relationship between outcome & predictors
- Predictive models: What combo of features fits best; to predict Y with minimum reducible error; not focused on hypothesis tests
  (Cited from Lecture 2 Page 7)

Question 5

- A mechanistic model uses a theory to predict what will happen in the real world. A empirical-driven model studies real-world events to develop a theory. Empirical model does not make any assumptions about f while mechanistic models does; it requires a larger number of observations than mechanistic model; we can add parameters to improve flexibility on mechanistic models while empirical-driven models are mroe flexible by default. (Cited from Lecture 2 Page 6)

- Empirically-driven model. Because real observations are used to estimate the empirically-driven model, it is easy to understand.

- Because a mechanistically-driven model is more flexible, it has a lower bias and a higher variance. Simplier models, on the other hand, will have larger bias and lower variance. As a result, the bias-variance trade-off would influence our choice of a mechanistic or empirical model.

Question 6

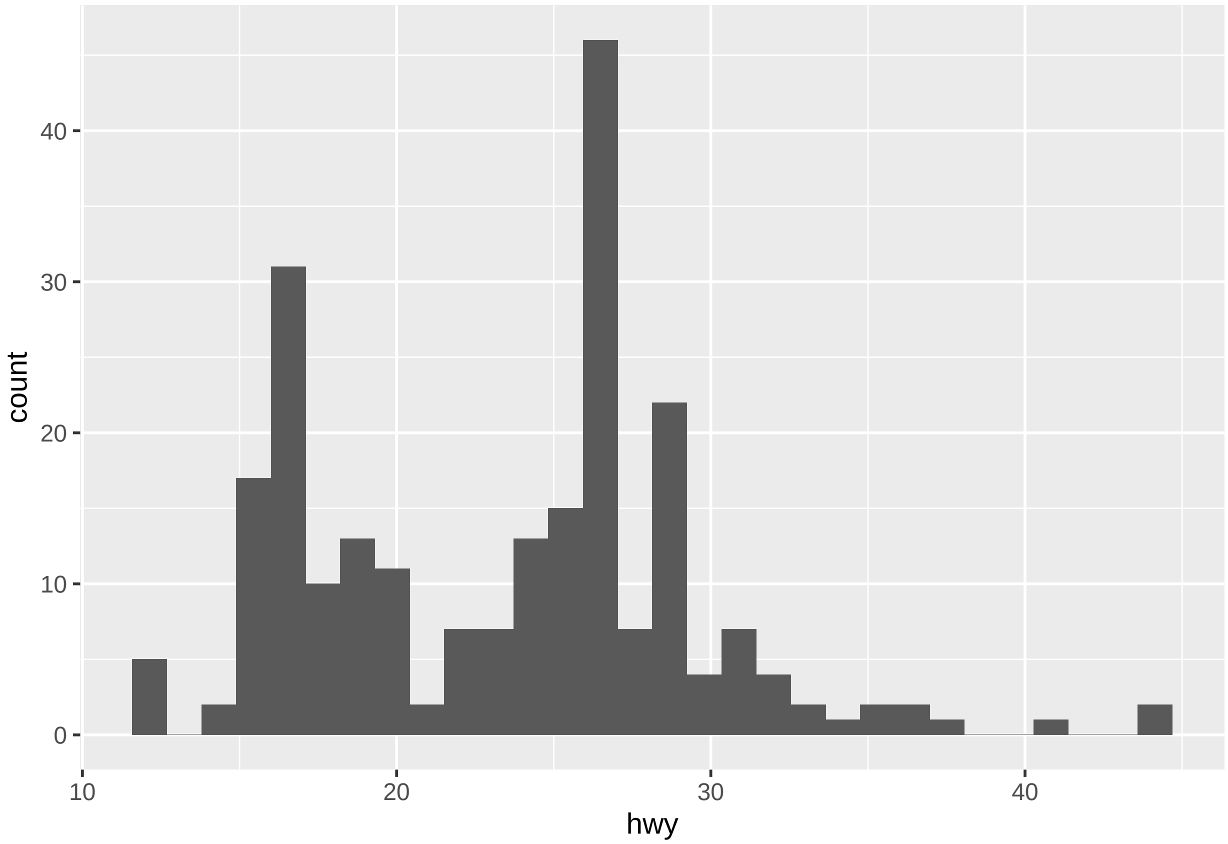
- Predictive, because the profile data about the voter is given in order to predict the possible choice that he/she will make in voting for the candidate.
- Inferential, because the goal is to understand the relationship between whether the vote has personal contact with the candidate and the voter's likelihood of support for the candidate.

Exploratory Data Analysis

```
library(tidyverse)
library(ggplot2)
# head(mpg)
```

Question 1

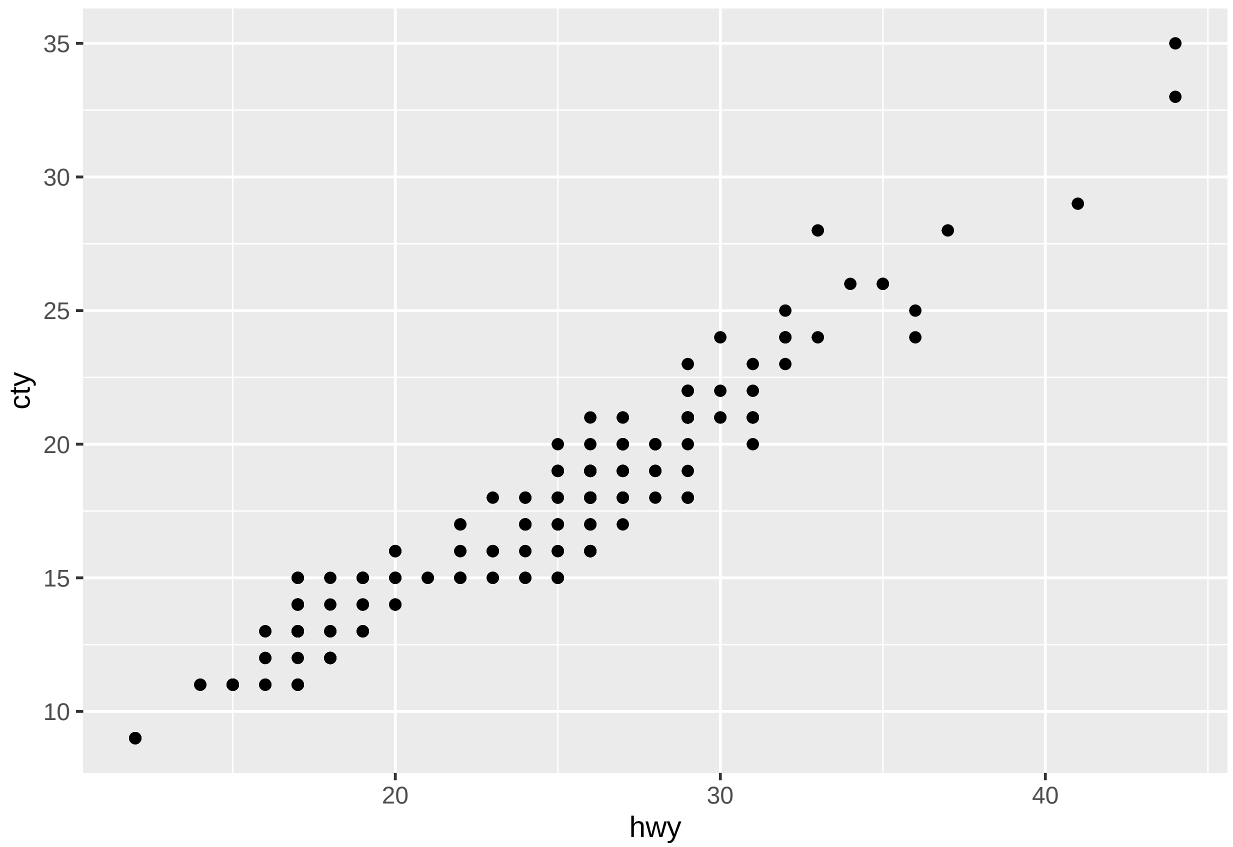
```
mpg%>%ggplot(aes(hwy)) + geom_histogram()
```



From what I observed, there are two peaks on the graph: one appears around 16 the other appears around 26.

Question 2

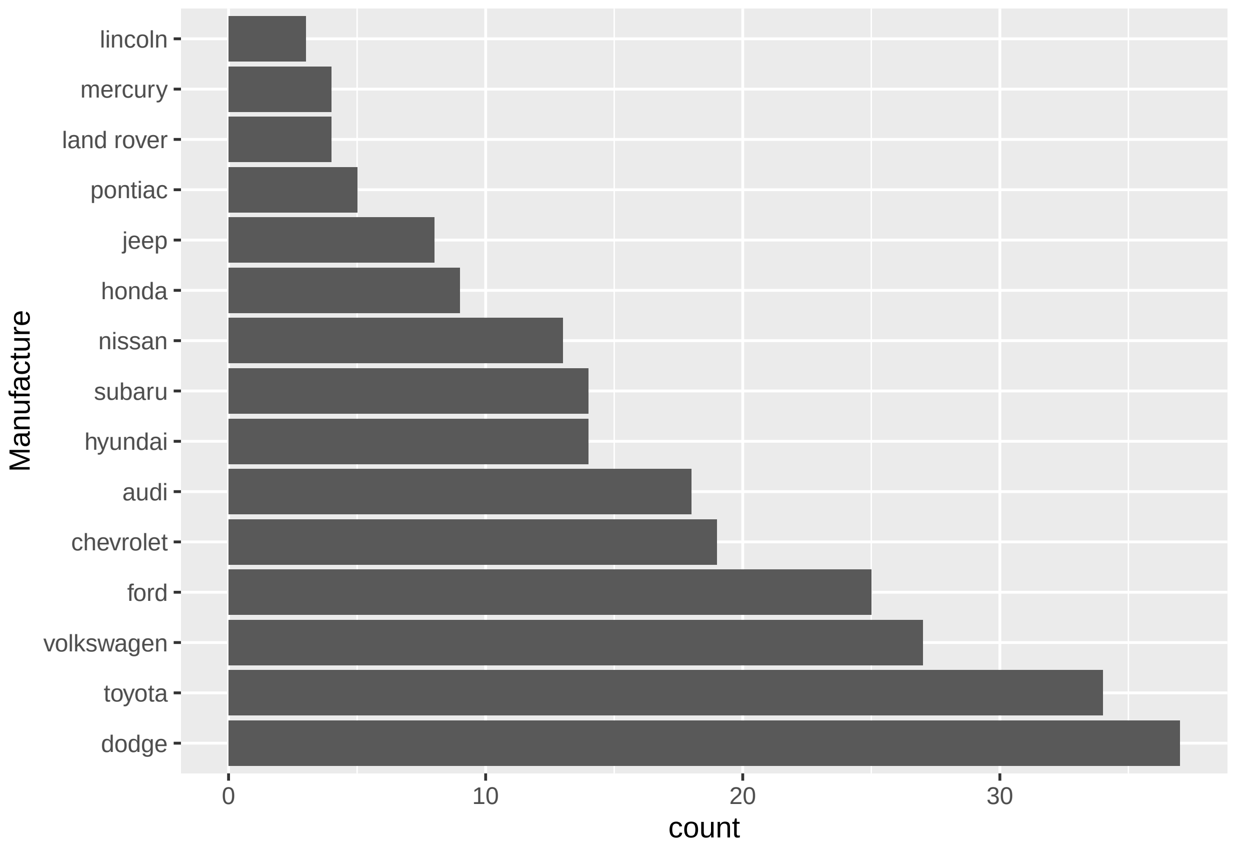
```
ggplot(mpg, aes(hwy, cty)) + geom_point()
```



I notice that there is a positive linear relationship between hwy and cty. As hwy increases, cty also increases.

Question 3

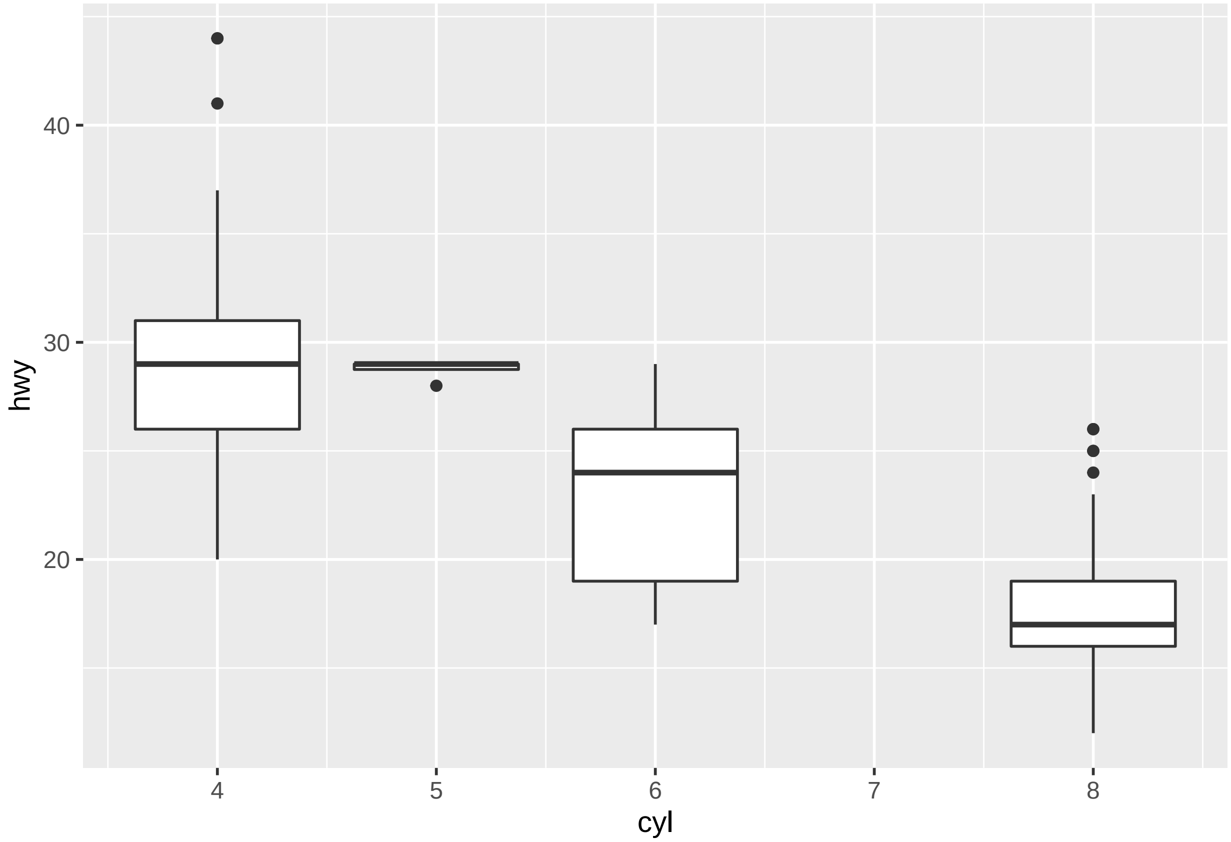
```
mpg %>%
  group_by(manufacturer)%>%
  summarise(count = n()) %>%
  ggplot(aes(x = count, y = reorder(manufacturer, -count))) +
  geom_bar(stat='identity')+
  labs(y='Manufacture')
```



Dodge produced the most cars, and Lincoln produced the least car.

Question 4

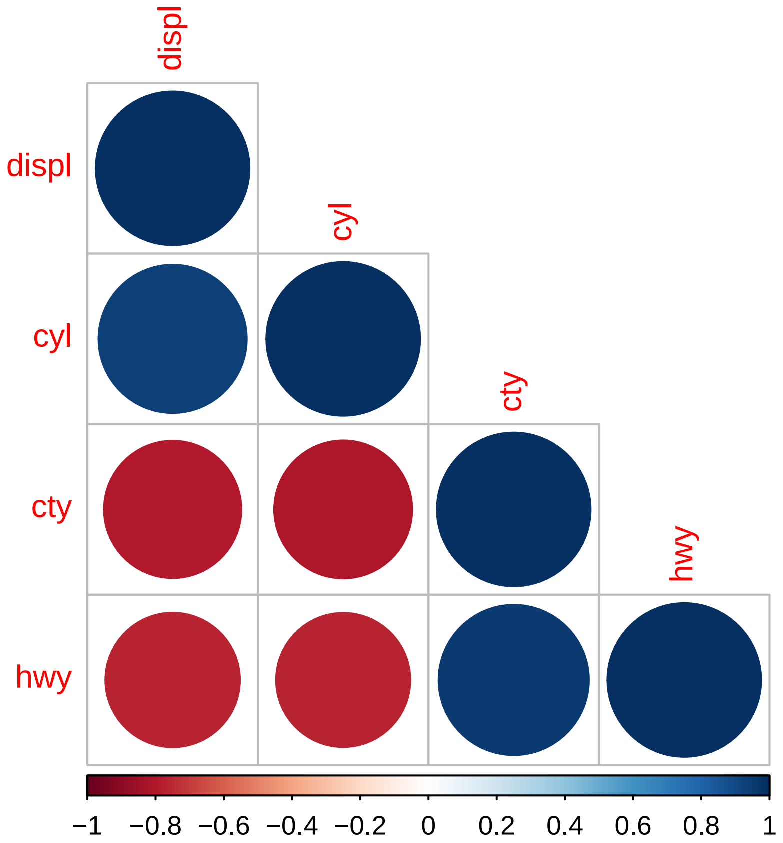
```
ggplot(data=mpg, aes(x=cyl, y=hwy, group=cyl)) + geom_boxplot()
```



From the graph, I can observe a negative relationship between cyl and hwy. As cyl increases, hwy decreases.As the car cylinders increase, it has lower miles per gallon left.

Question 5

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



As it is shown on the graph above:

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

The correlations make sense to me and do not surprise me.