hw1-pstat131

Ziyi Fang

2022-10-02

Question 1

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

Supervised learning include a response column(Y) and unsupervised learning does not contain the respond column(Y). With supervised learning, you can measure the model's performance. With unsupervised learning, you can cluster the observations together based on liked observations, but you can't measure the model performance.

Question 2

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

For Regression model, the respond column (Y) is quantitative so it only contains numerical values. For classification model, the respond column (Y) is qualitative, so it only contains categorical values.

Question 3

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

2 commonly used error metrics for regression based problems are Mean Squared Error and Root Mean Squared Error.

2 Commonly used metrics for classification based problems are accuracy and AUC (area under the curve).

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: Choose model to best visually emphasize a trend in data. For example, using a line on a scatterplot.

Predictive: The goal is to predict respond column with minimum reducible error. Not focused on hypothesis tests.

Inferential: The goal of inferential models is to determine which predictors are most associated with a response and to make some type of causal claim.

Question5

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. Define empirically-driven. How do these model types differ? How are they similar? Mechanistic makes an assumption of the function that fits the data, but we won't know the true function. We can add as many parameters as needed, but risk overfitting the model.

Empirical models do not assume a particular function fits the data. These models can be very flexible and require a lot of data, but can also overfit the data.

They are similar because they both can overfit the data. They are different because mechanistic assumes a function should fit the data, whereas empirical models do not assume this.

In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice. Mechanistic is easier to understand because we know what the estimated function fit to the data is. and the parameters included in the function.

Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models. Bias-variance tradeoff is related to the various models because this is something we need to consider when fitting any model to the data. We want to reduce how wrong we are (bias) and when our models sees different or new datasets, and we want to minimize the variance in the error metric we see each time.

Question 6

A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions:

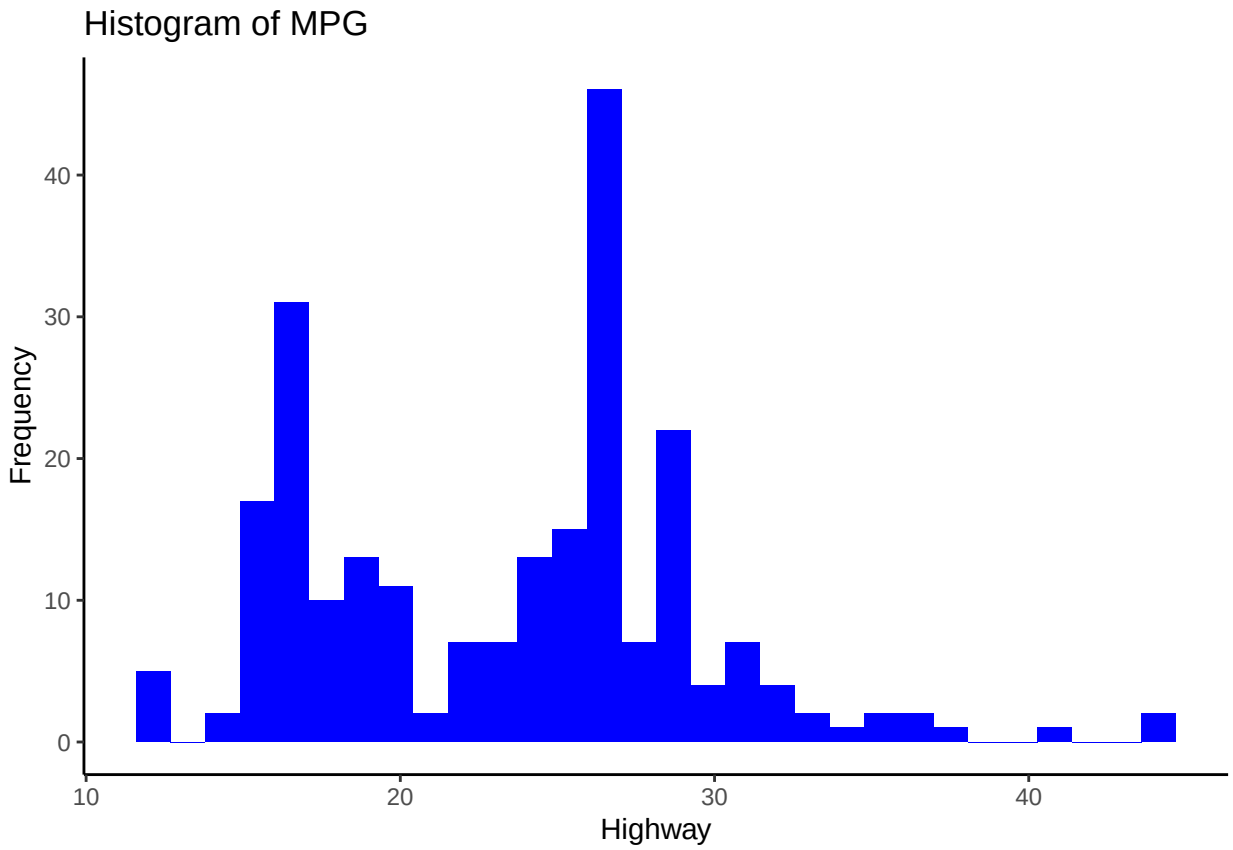
Given a voter's profile/data, how likely is it that they will vote in favor of the candidate? This is predictive because we are trying the predict a probability or likelihood we vote for a particular candidate.

How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate? This is inferential because we are trying to establish a case/effect relationship of personal contact with a candidate and the likelihood of voting for that candidate.

Exploratory Data Analysis

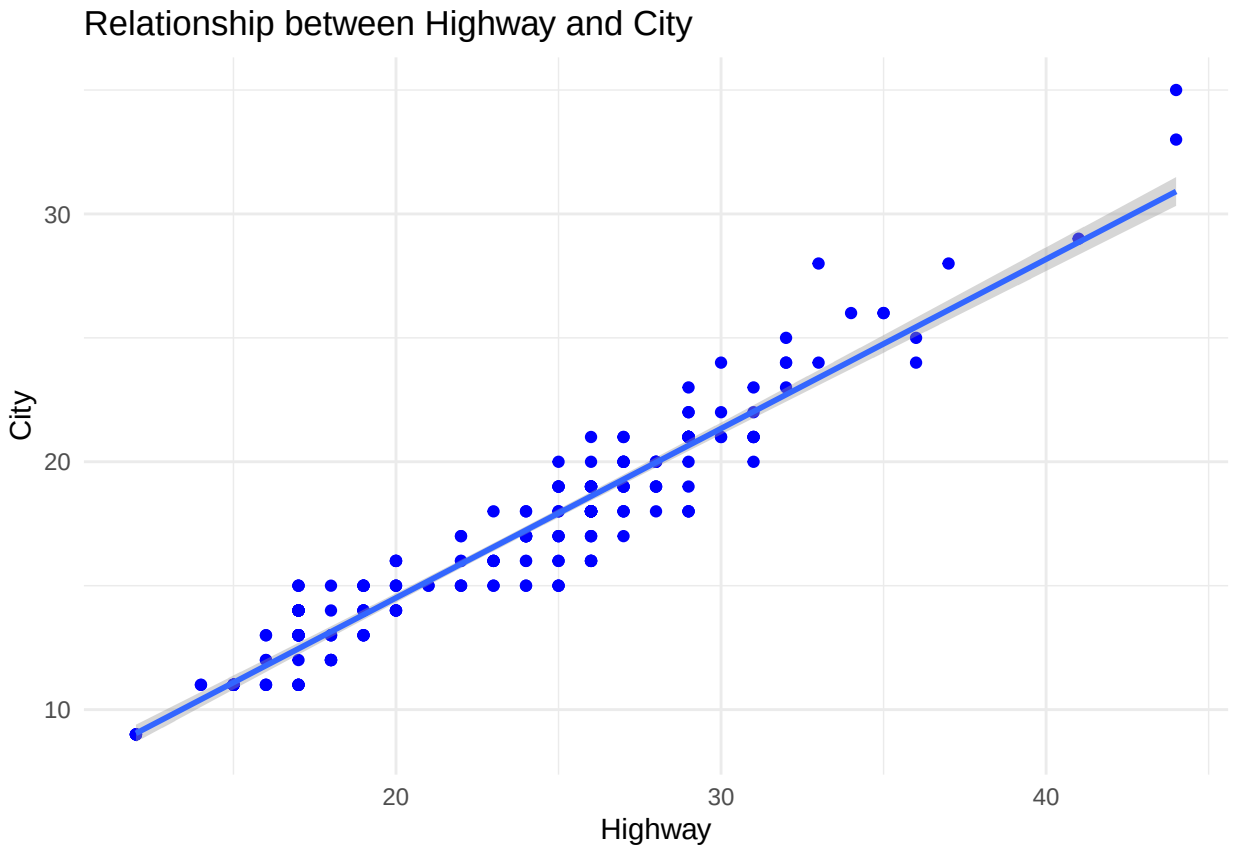
```
#load mpg data
data("mpg")

# Exercise 1
ggplot(data = mpg, aes(x = hwy)) + geom_histogram(fill = "blue") +
  labs(x = "Highway", y ="Frequency", title = "Histogram of MPG") +
  theme_classic()
```

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

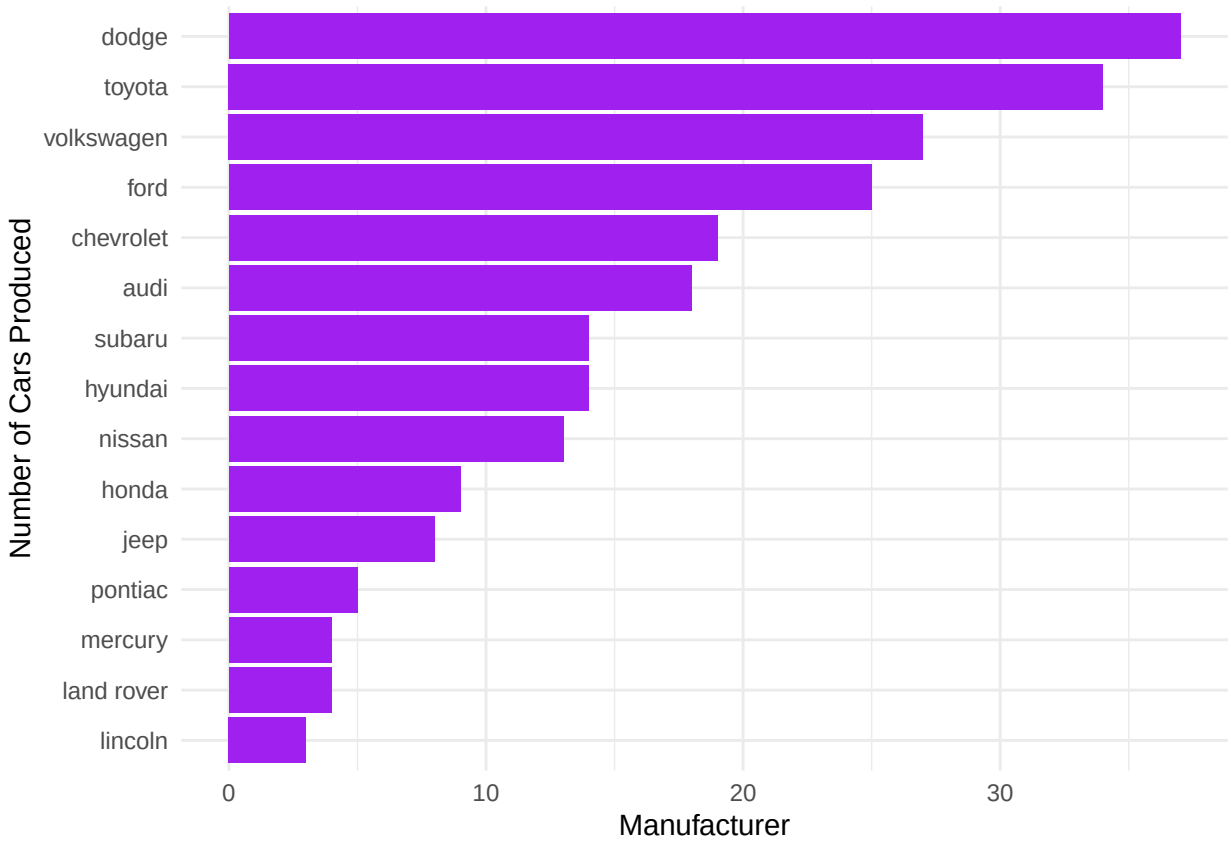



```
#A lot of cars are clustered in the 15-20 mpg range and there's another cluster in the 25-25 range.

# Exercise 2
ggplot(data = mpg, aes(x = hwy, y = cty)) + geom_point(color = "blue") +
  geom_smooth(method = "lm") +
  labs(x = "Highway", y ="City", title = "Relationship between Highway and City") +
  theme_minimal()
```

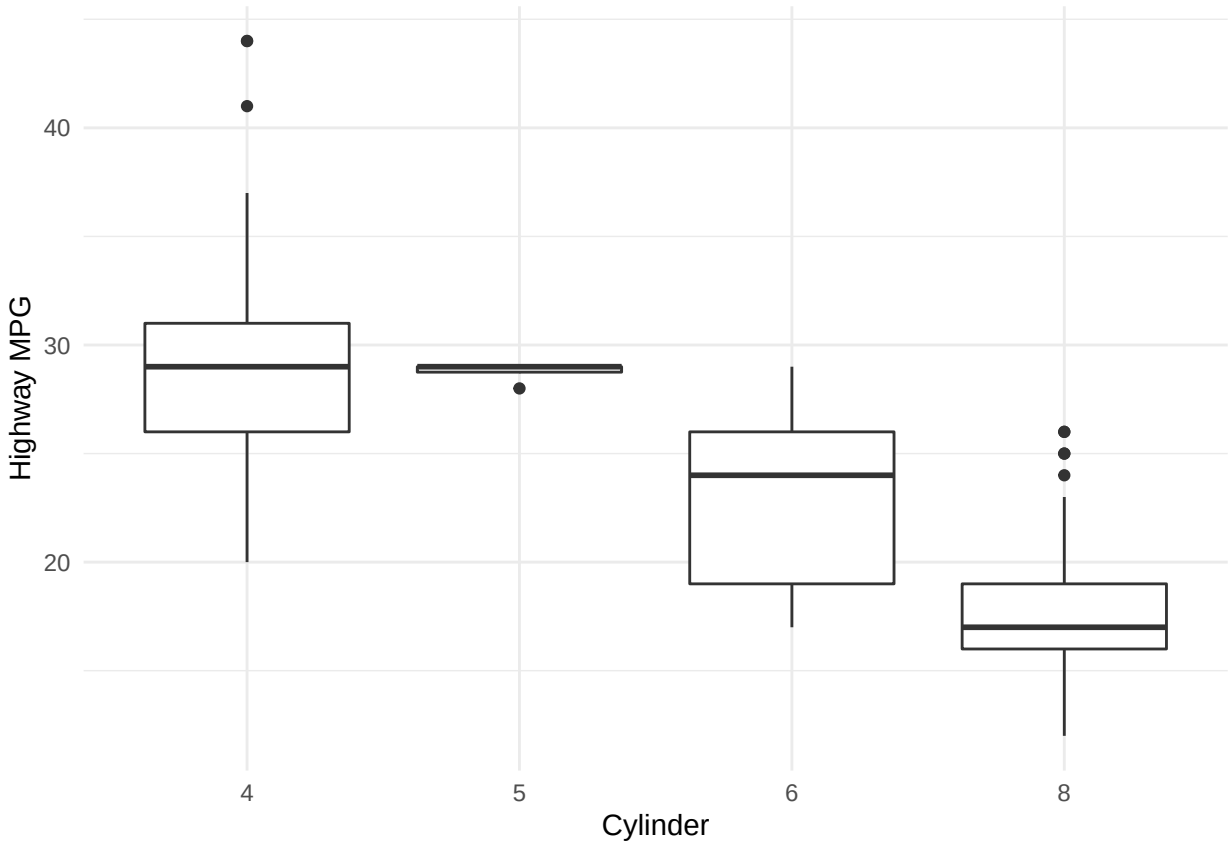
```
## `geom_smooth()` using formula 'y ~ x'
```



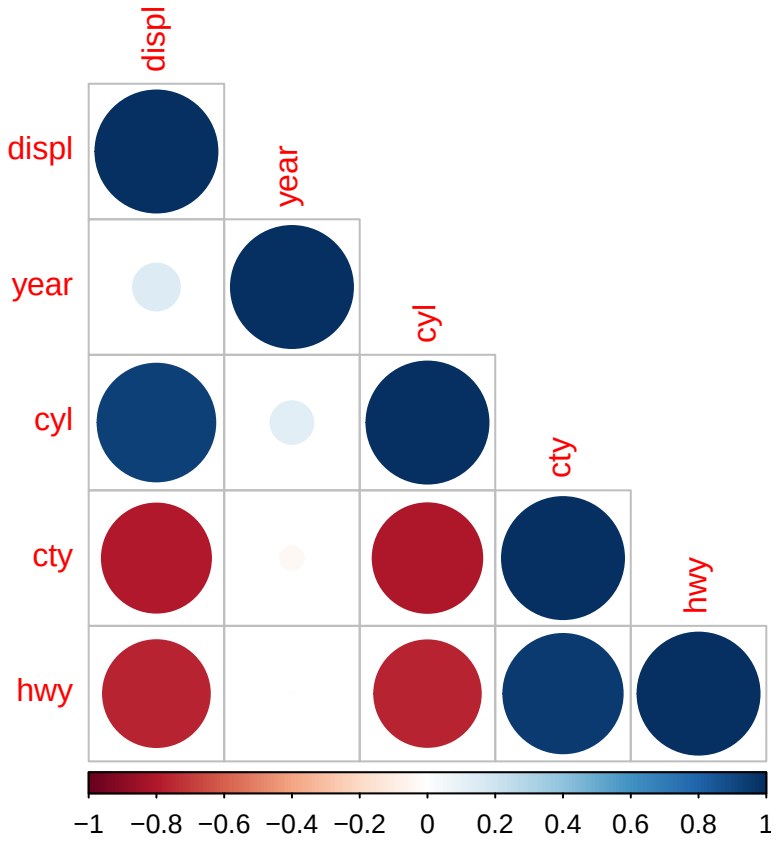

```
# There is a positive and linear relationship between highway and city. Which means more cars in a city

# Exercise 3
# reorganize data to get the manufacturer and the number of cars they produced.
df<- mpg %>% dplyr::group_by(manufacturer) %>% dplyr::summarise(count = n())
ggplot(data = df, aes(x = reorder(manufacturer, count), y = count)) + geom_bar(stat = "identity", fill =
  coord_flip() +
  labs(x = "Number of Cars Produced", y = "Manufacturer") +
  theme_minimal()
```




```
# Dodge produced the most cars and Lincoln produced the least cars.

# Exercise 4
ggplot(data = mpg, aes(x = factor(cyl), y = hwy)) + geom_boxplot() +
  labs(x = "Cylinder", y  = "Highway MPG") +
  theme_minimal()
```




```
# The pattern is that higher cylinder cars get less MPG.
# Cylinder presents iteself as integer, but it's really categorical (discrete) -- so we need to change

# Exercise 5

#Create correlation matrix of only numeric columns
df_cor<- mpg %>% select_if(is.numeric) %>% cor()

corrplot(df_cor, type = "lower" )
```




```
# Highly positively correlated variables are cylinder & dispacement. hwy & cty.
# Highly negatively correlated variables are cty & displ, hwy & displ, cty & cyl, hwy & cyl.
# These relationships make sense to me.
```