Hw1_JiangShu

JIANG SHU

2022-10-02

##1. Solution to homework 1 Q1. Define supervised and unsupervised learning. What are the difference(s) between them?

- Supervised learning is learning the model for data by using input variables(predictors) and output variables (Y). According to lecture, we can do prediction, estimation, model selection and inference by applying supervised learning. Generally, we do regression problem and classification problem with supervised learning.
- Unsupervised learning only have input variables and no output variables. One main task to do is clustering. Therefore, unsupervised learning explores the structure of raw data and there's no correct output value to examine the model.
- The major difference between supervised learning and unsupervised learning is the answer key. In supervised learning, we intervene with the model by giving training examples with output values, while we only give unknown data in unsupervised learning. In short, we teach computer using data in supervised learning while computer learns data by itself in unsupervised learning.

##2. Solution to homework 1 Q2. Explain the difference between a regression model and a classification model, specifically in the context of machine learning.

- A regression model maps input variables to output variables so that in the future any new input variables can be used to predict output variables. They are generally continuous function, A classification model takes input variables and classify them according to certain criteria. The output is generally discrete. For instance 0 (don't pass the exam), 1 (pass the exam).

##3. Solution to homework 1 Q3. Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.

- Regression ML problem: Mean Squared Error (MSE), accuracy
- Classification ML problem: Precision, F1 score

##4. Solution to homework 1 Q4. 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: Present data by using appropriate models to show trend.

- Inferential models: Learn the parameter that affects result. Test hypothesis relationship.

- Predictive models: Predict future data points by using historical data. Not on hypothesis tests. (According to lectures)

##5. Solution to homework 1 Q5.

Define mechanistic. Define empirically-driven. How do these model types differ? How are they similar?

- Mechanistic model proposes a function f to predict the event we are studying while empirically-driven model uses observations to derive a function/theory about the relationship between input and output. We can add parameters in mechanistic model and they won't match the unknown true function. Empirically-driven model requires a large amount of examples to derive the relation. They are both predictive models and they both might subject to the problem of over fitting.

In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice.

- I would say a mechanistic model is easier to understand. We can add parameters in the process of deciding relationship. Also, we determine the formula at the beginning so we have a better idea of what's going on. However, empirically-driven model looks more accurate when we have a large amount of observations, it's just we are harder to understand the function it derives.

Describe how the bias-variance trade off is related to the use of mechanistic or empirically-driven models.

- A mechanistic model is more general, which also mean it has higher bias and low variance. An empirically-driven model is more specific to the dataset, which means it has lower bias and higher variance. When we are predicting the new outputs, the model with higher complexity(empirically-driven) might perform bad due to overfitting problem. While mechanistic model might perform better on new data, it might have the problem of underfitting.

##6. Solution to homework 1 Q6.

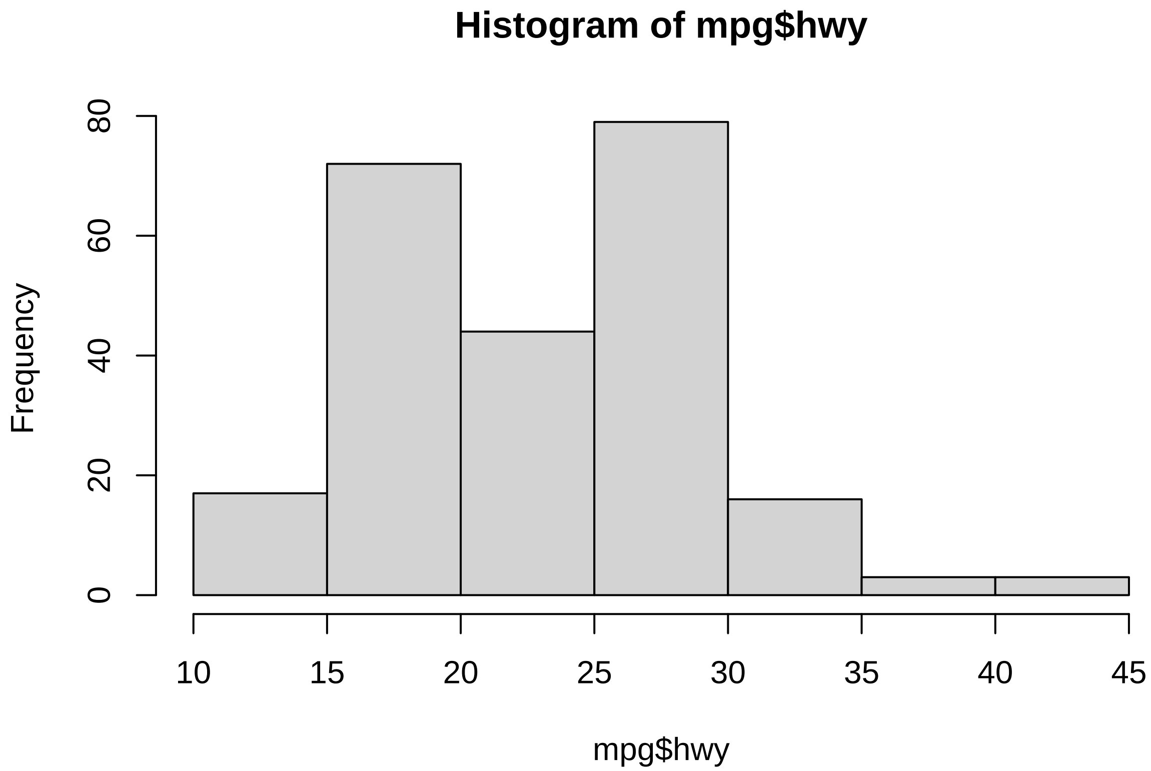
A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions: Given a voter's profile/data, how likely is it that they will vote in favor of the candidate? How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate? Classify each question as either predictive or inferential. Explain your reasoning for each.

- The first question is predictive as it tries to predict the decision of a voter by analyzing the voter's profile. The voter hasn't voted, and they are using historical data to predict his/her action, so this is predictive.The second question is inferential as it tries to explore the relationship between predictors and response variable. This is essentially testing relationship so it's inferential.

Exploratory Data Analysis

Exercise 1

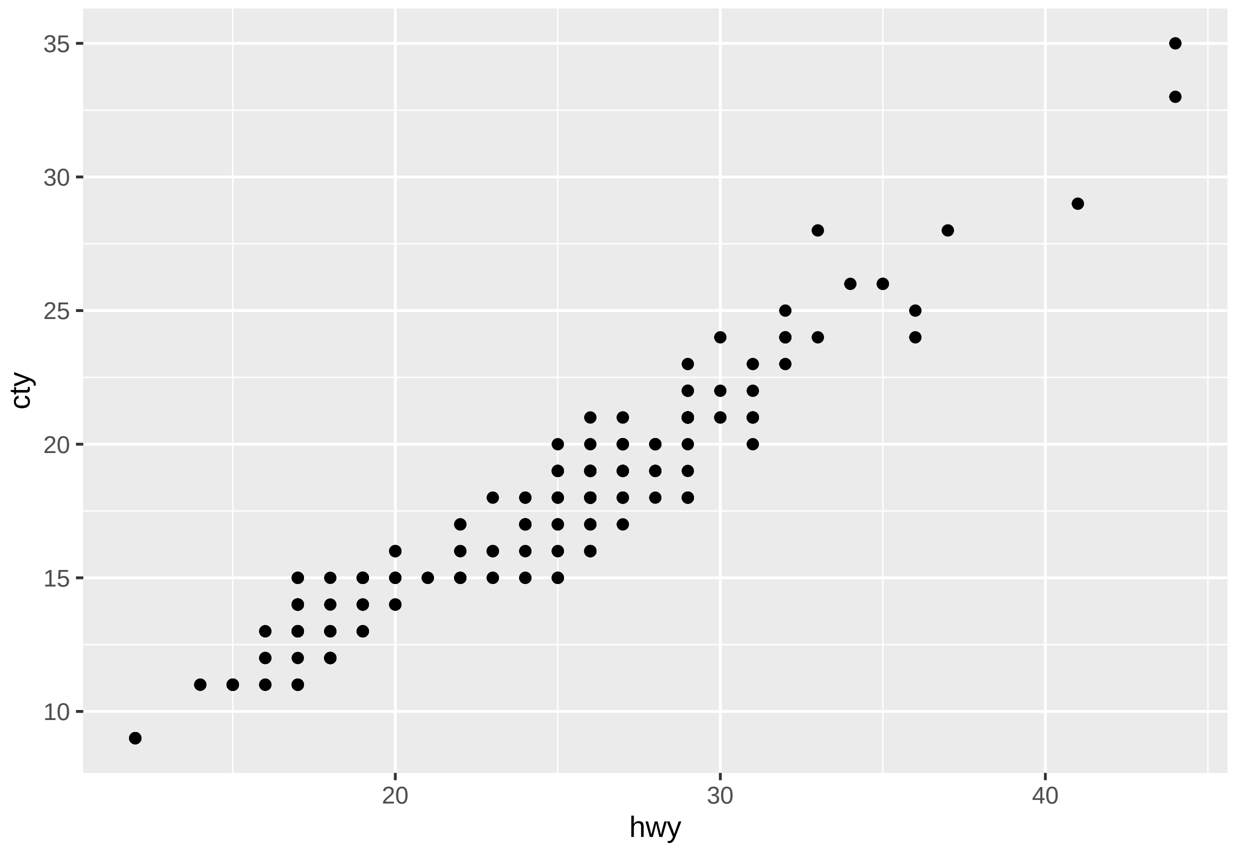
```
hist(mpg$hwy)
```



We can see most vehicles' mile per gallon on highway is in the range of 25-30. While only a little has mpg greater than 35 on highway. Also, vehicles with mpg lower than 15 is also rare.

Exercise 2

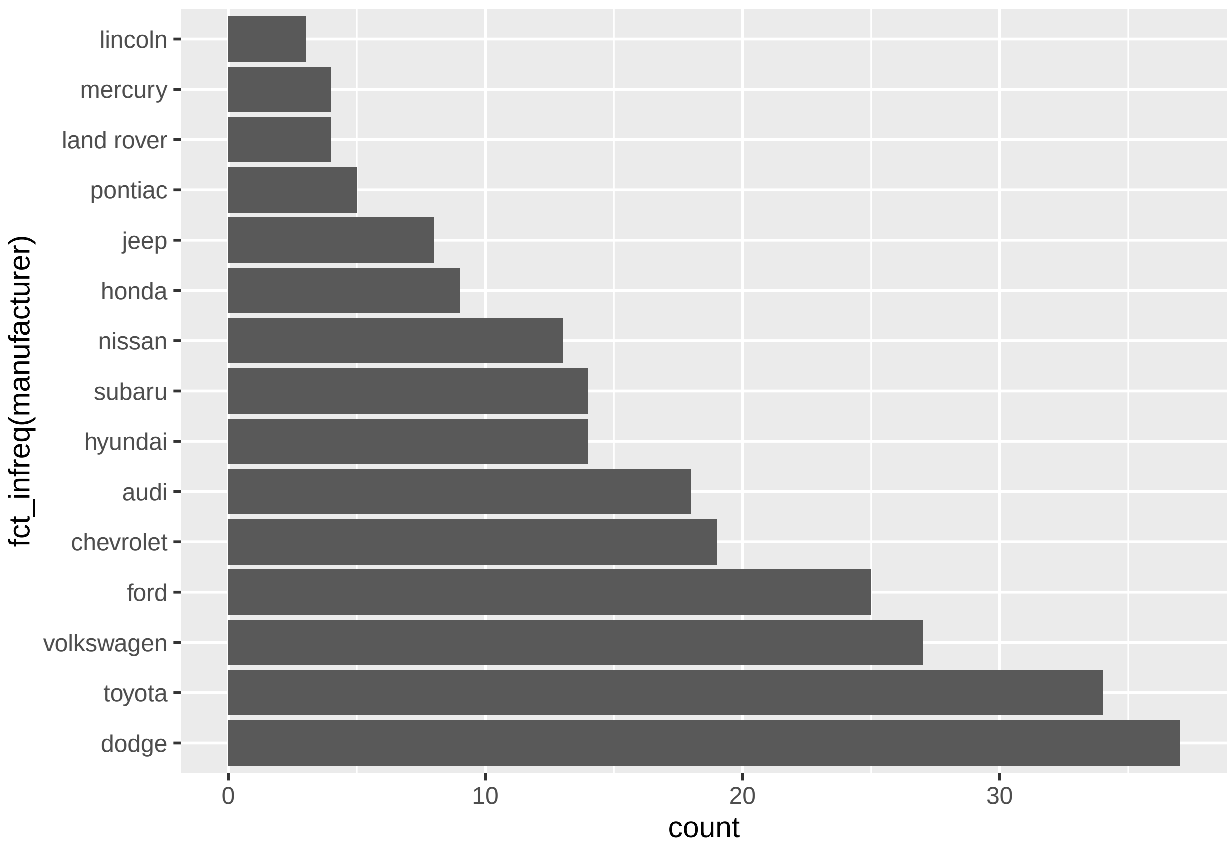
```
scatterplot1 <- ggplot(mpg, aes(x=hwy, y=cty)) +
  geom_point()

print(scatterplot1)
```



It looks like there's a linear relationship between hwy and cty. Generally, as hwy increases, cty also increases. This makes sense because mpg of a car would be high in city when its mpg is high on highway. Regardless where the vehicle is, its general fuel consuming ability won't change much relative to itself.

Exercise 3

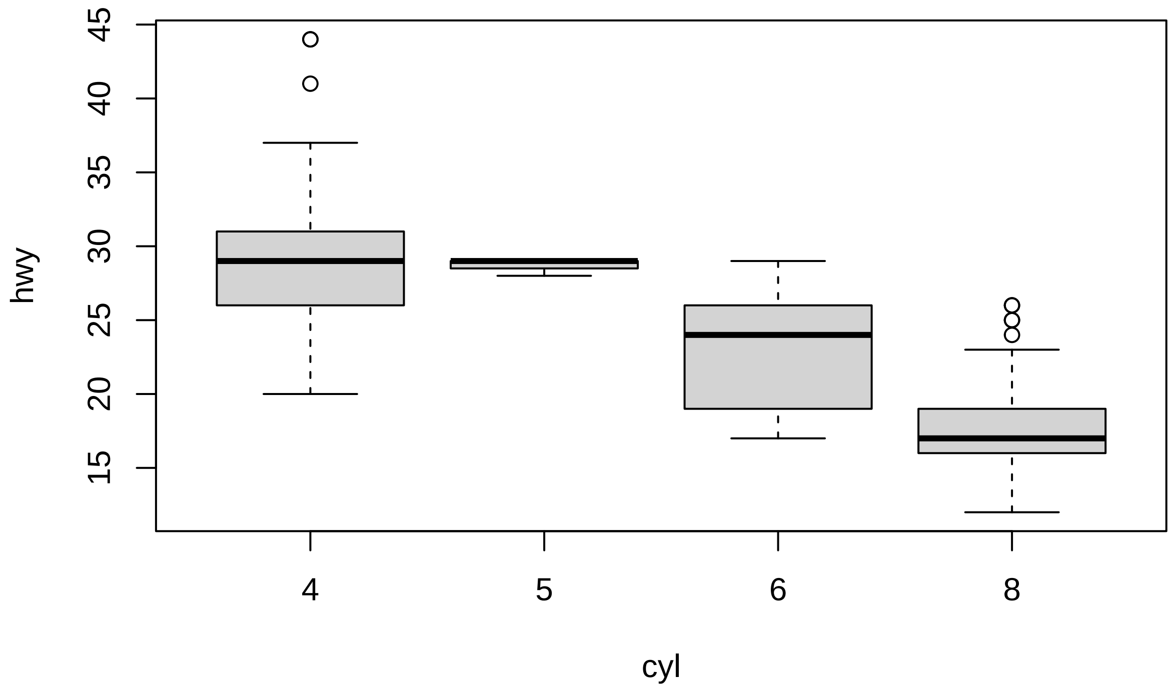
```r
barplot1 <- ggplot(mpg, aes(x=fct_infreq(manufacturer))) +
  geom_bar() +
  coord_flip()

print(barplot1)
```



Dodge produced most cars while Lincoln produced least cars.

Exercise 4

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

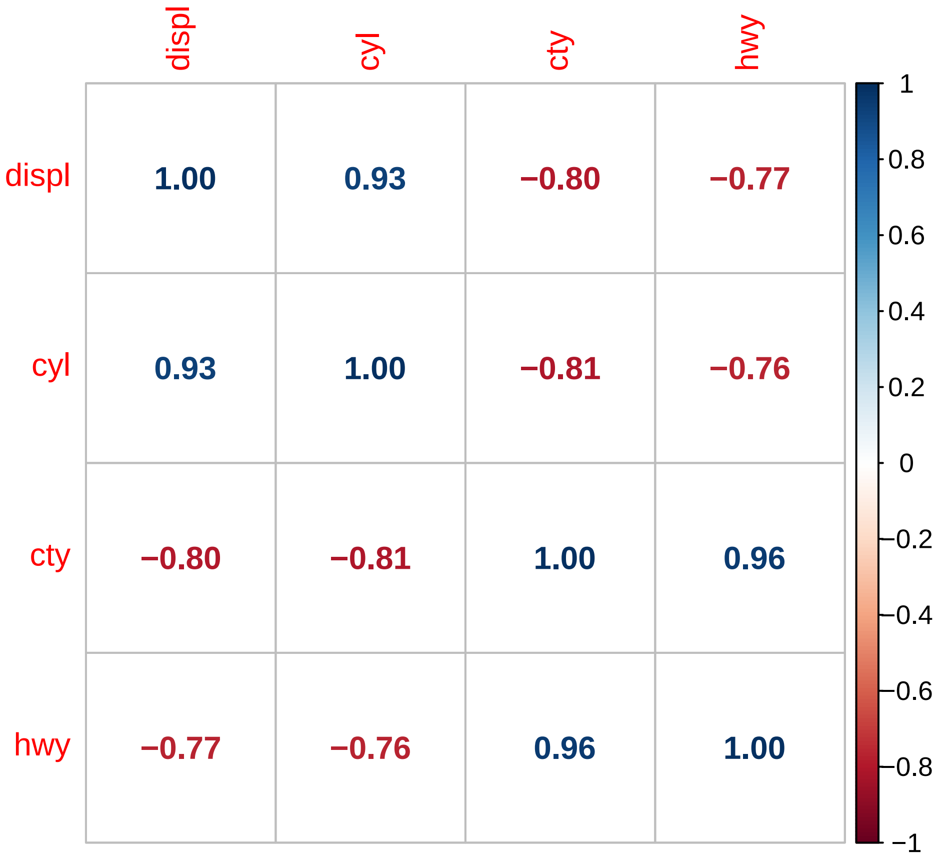


With more cylinders, the mpg of cars on highway is lower. That's true because more cylinders consume more fuels, thus resulting in lower miles per gallon.

Exercise 5

Since we are interested in correlation of cars' fuel economy, let's get rid of not useful columns.

```
mpg_revised <- mpg %>%
  select(displ,cyl,cty,hwy)
M <- cor(mpg_revised)
corrplot(M, method = 'number')
```



	displ	cyl	cty	hwy
displ	1.00	0.93	−0.80	−0.77
cyl	0.93	1.00	−0.81	−0.76
cty	−0.80	−0.81	1.00	0.96
hwy	−0.77	−0.76	0.96	1.00

As we can see, each variable is perfectly correlates with itself. displ is positively related to cyl and negatively related to cty & hwy. This makes sense because high-displacement engine would take more fuel/air mixture per revolution and more fuel is consumed, thus lower cty and hwy. Also, more cylinders makes displ higher because we have more total volume to burn the fuels. cyl is also negatively related to cty and hwy since more cylinders would simply consume more fuels. Thus mile per gallon in city and highway would both decrease. cty is positively correlated to hwy since a higher hwy would also means a higher cty. One thing that surprises me is the correlation between cty and hwy seems really linear.