pstat131 hw1

Louis Lao

2022/3/31

R Markdown

Homework 1 PSTAT 131/231 Machine Learning Main Ideas Please answer the following questions. Be sure that your solutions are clearly marked and that your document is neatly formatted.

You don't have to rephrase everything in your own words, but if you quote directly, you should cite whatever materials you use (this can be as simple as "from the lecture/page # of book").

Question 1: Define supervised and unsupervised learning. What are the difference(s) between them?

Answer: The supervised one we known the outcome and the unsupervised learning we don't know the outcome.

Question 2: Explain the difference between a regression model and a classification model, specifically in the context of machine learning.

Answer: The Y in regression is quantitative, which means is numeric value but in the classification is qualitative, it's categorical.

Question 3: Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.

Answer: /

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 models: Aim is to predict Y with minimum reducible error, no focused on hypothesis test.(From Lecture Note)

Inferential models: Aim is to test theories, state relationship between outcome and predictor.(From Lecture Note)

Predictive models: Choose model to best visually emphasize a trend in data.(From Lecture Note)

Question 5: 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?

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

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

Answer: /

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?

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.

Exploratory Data Analysis This section will ask you to complete several exercises. For this homework assignment, we'll be working with the mpg data set that is loaded when you load the tidyverse. Make sure you load the tidyverse and any other packages you need.

Exploratory data analysis (or EDA) is not based on a specific set of rules or formulas. It is more of a state of curiosity about data. It's an iterative process of:

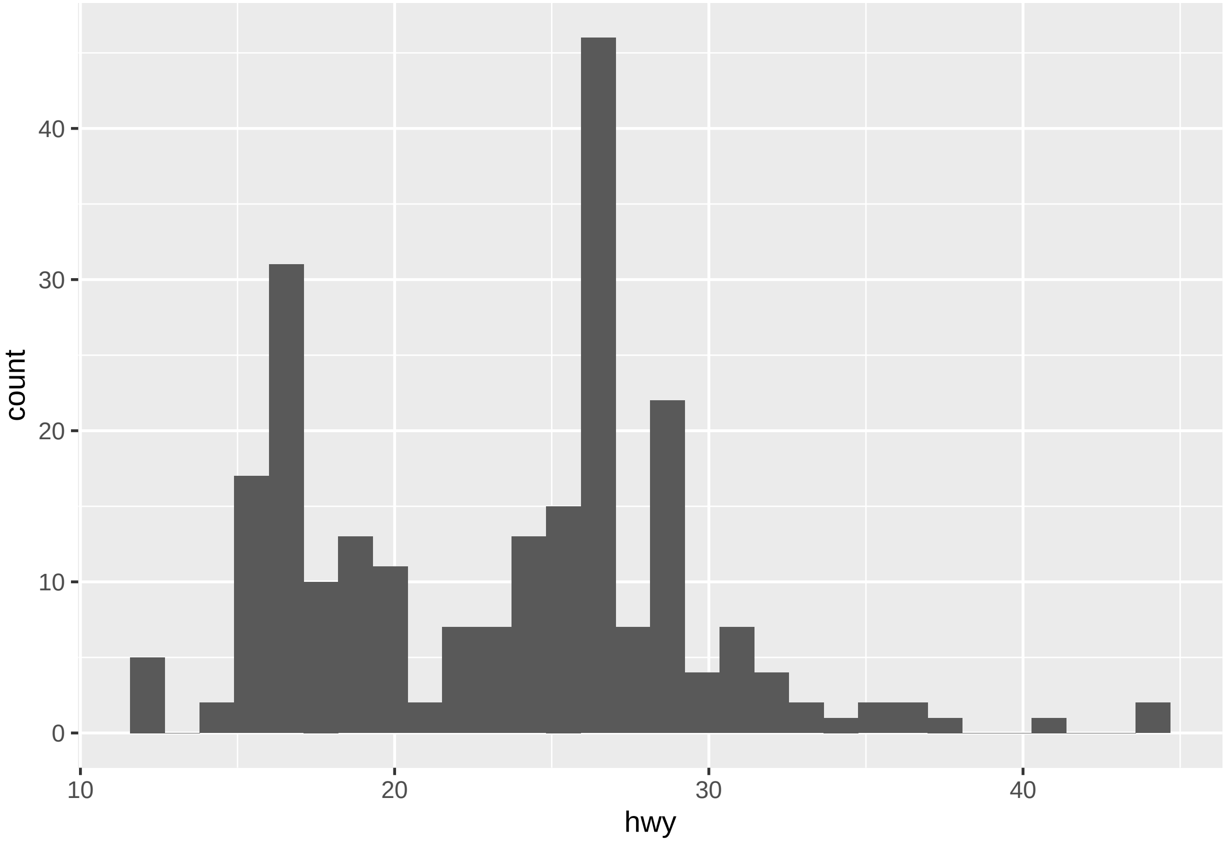
generating questions about data visualize and transform your data as necessary to get answers use what you learned to generate more questions A couple questions are always useful when you start out. These are "what variation occurs within the variables," and "what covariation occurs between the variables."

You should use the tidyverse and ggplot2 for these exercises.

Exercise 1: We are interested in highway miles per gallon, or the hwy variable. Create a histogram of this variable. Describe what you see/learn.

```
plot1 <- ggplot(mpg, aes(x=hwy)) + geom_histogram()
print(plot1)
```

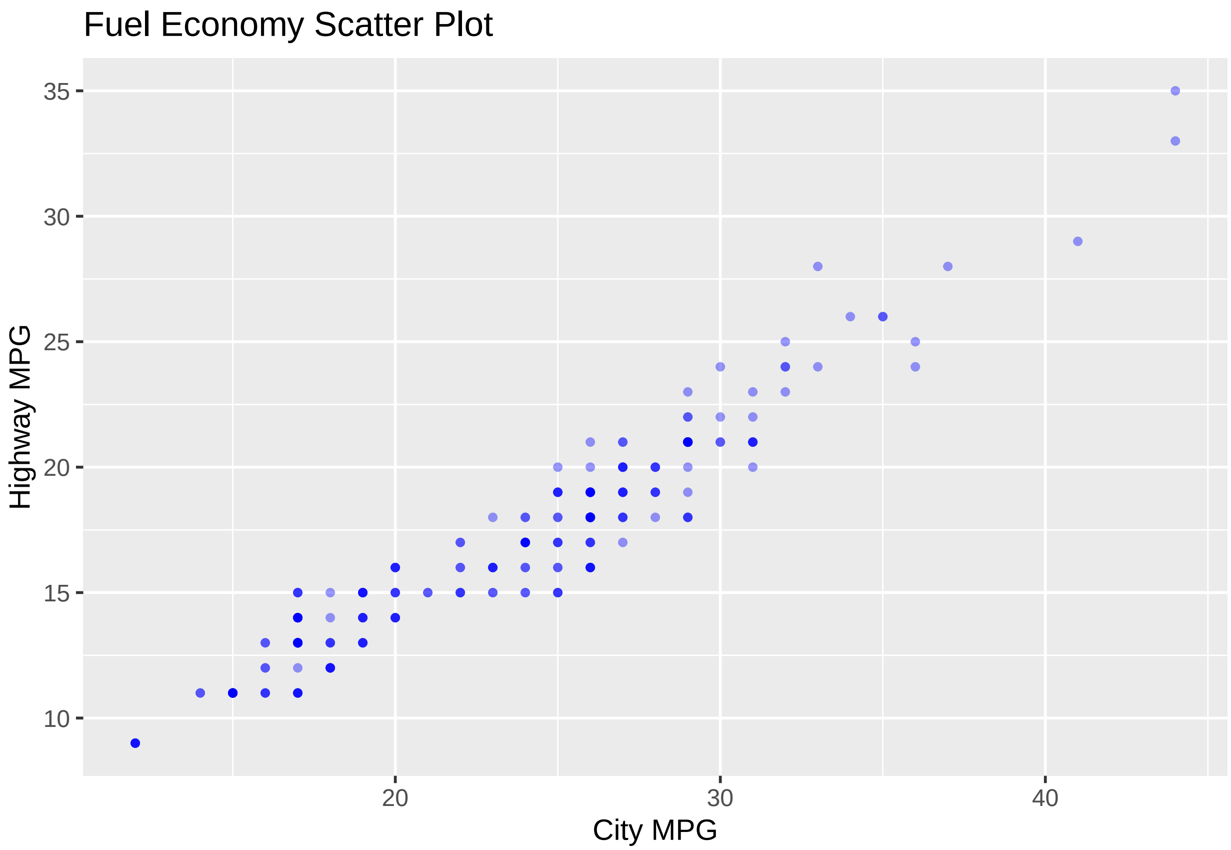
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



So, mostly the miles car could use per gallon are under 30.

Exercise 2: Create a scatterplot. Put hwy on the x-axis and cty on the y-axis. Describe what you notice. Is there a relationship between hwy and cty? What does this mean?

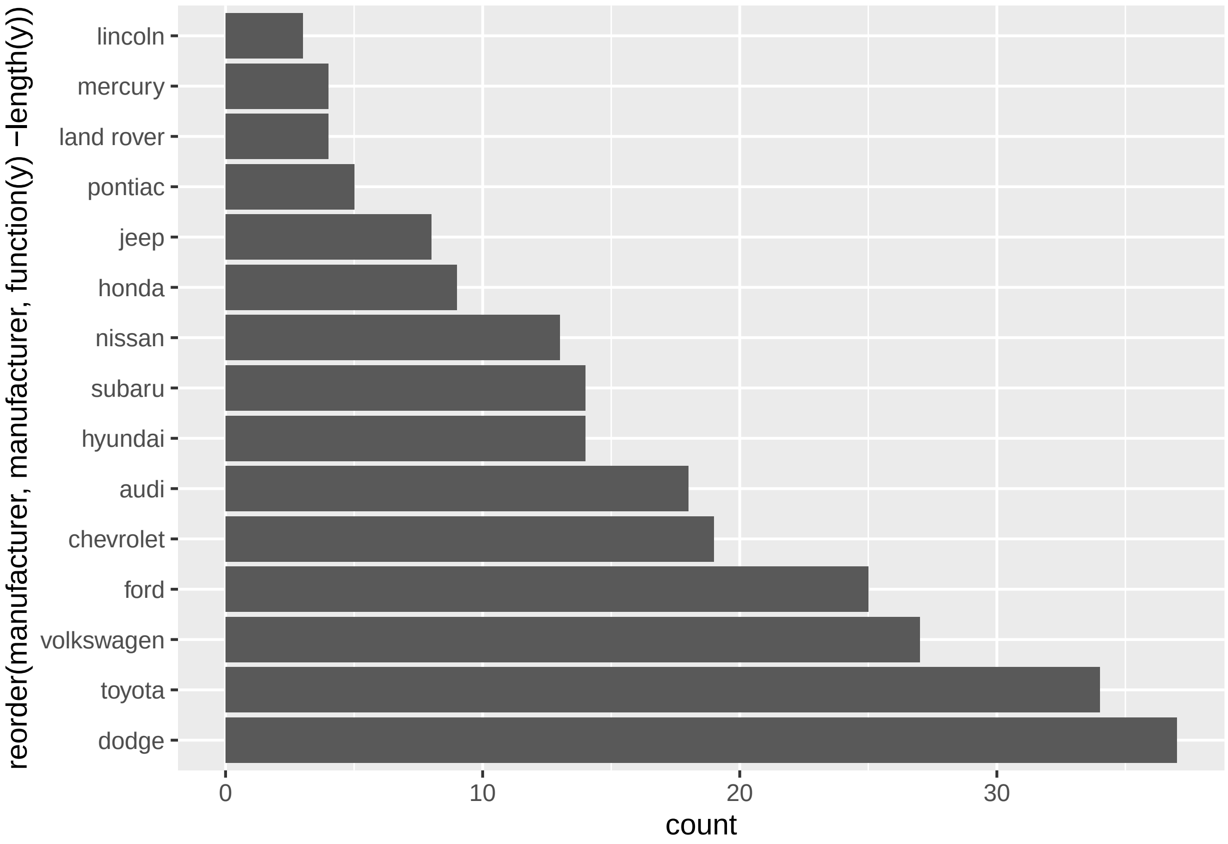
```
plot2 <- ggplot(mpg, aes(hwy, cty)) +
  geom_point(color='blue', size= 1, alpha=0.4) +
  labs(x="City MPG", y="Highway MPG", title="Fuel Economy Scatter Plot")
print(plot2)
```



The Highway and City is correlated. The more fuel spend on highway will spend the most on city, vice versa.

Exercise 3: Make a bar plot of manufacturer. Flip it so that the manufacturers are on the y-axis. Order the bars by height. Which manufacturer produced the most cars? Which produced the least?

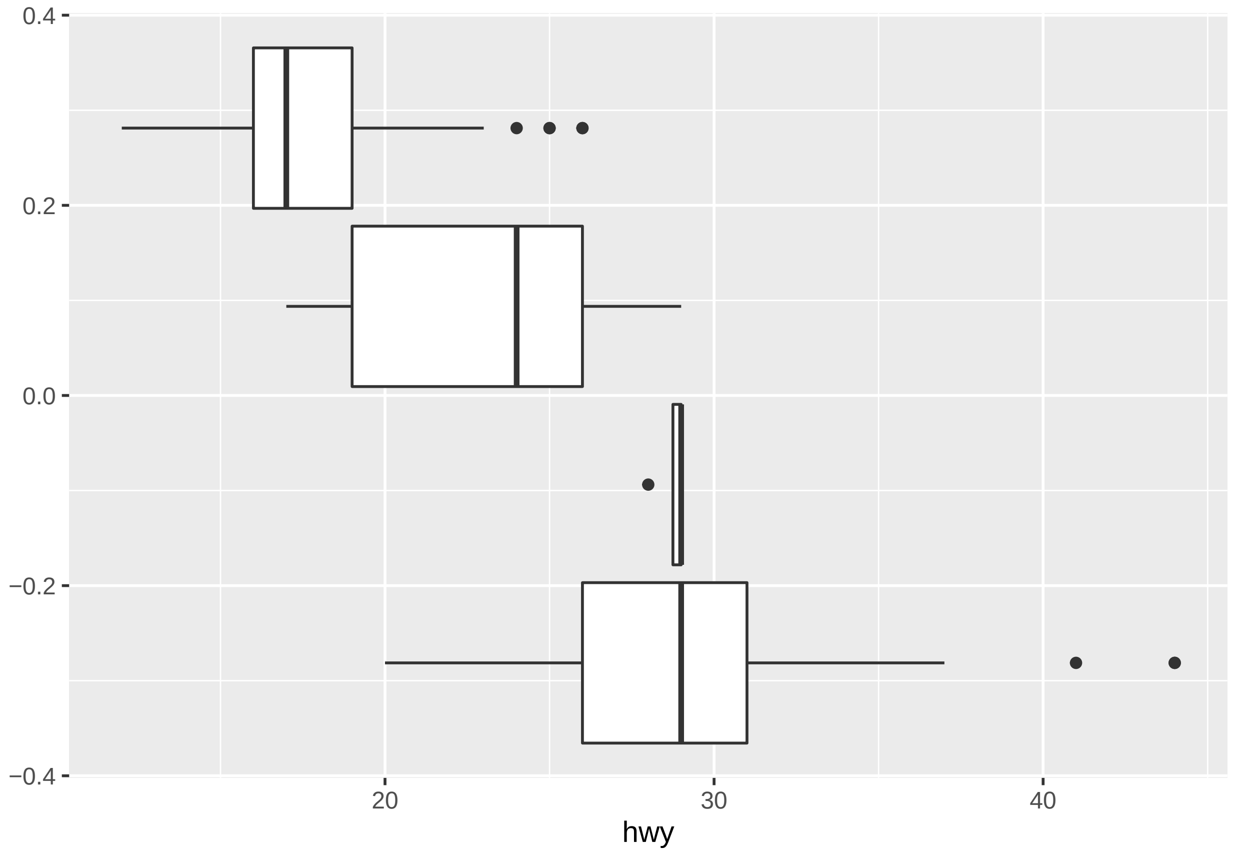
```
plot3<-ggplot(mpg, aes(y=reorder(manufacturer, manufacturer,function(y)-length(y)))) + geom_bar()
plot3
```



Dodge produced the most and Lincoln produced the least.

Exercise 4: Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?

```
plot4<- ggplot(mpg,aes(group = cyl, x=hwy))+geom_boxplot()
print(plot4)
```



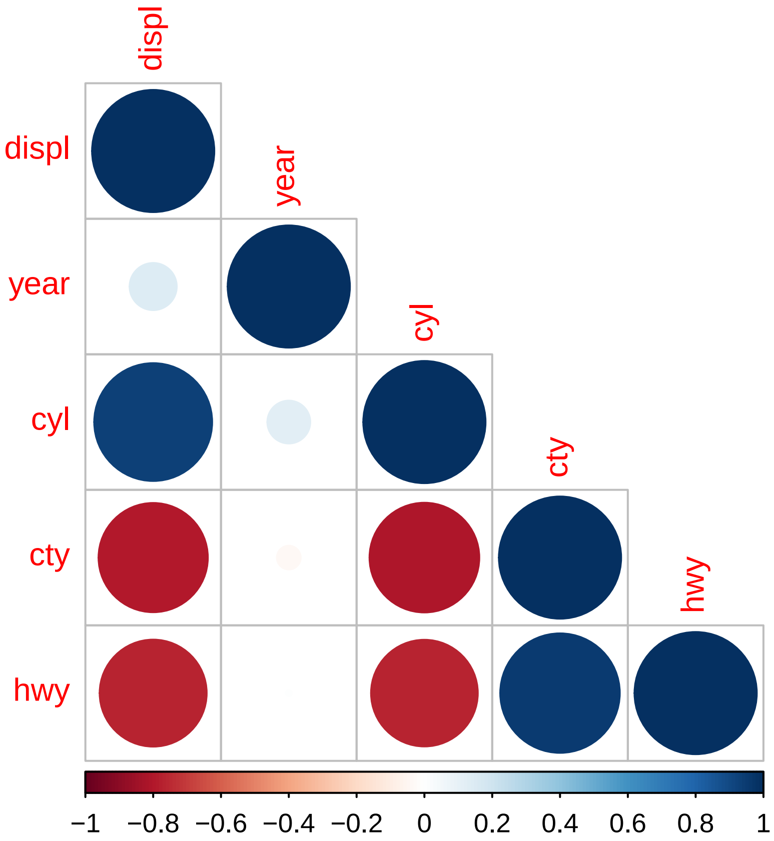
Exercise 5: Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset. (Hint: You can find information on the package here.)

```
newmpg = subset(mpg, select = -c(manufacturer, model, trans,drv, class, fl))
head(newmpg)
```

```
## # A tibble: 6 x 5
##   displ  year   cyl   cty   hwy
##   <dbl> <int> <int> <int> <int>
## 1   1.8  1999     4    18    29
## 2   1.8  1999     4    21    29
## 3   2     2008     4    20    31
## 4   2     2008     4    21    30
## 5   2.8  1999     6    16    26
## 6   2.8  1999     6    18    26
```

```
M = cor(newmpg)

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



Which variables are positively or negatively correlated with which others? Do these relationships make sense to you? Are there any that surprise you?

Negative correlated: City miles per gallon with engine displacement, highway miles per gallon with engine displacement. Number of cylinders with city miles per gallon and number of cylinders with highway miles per gallon. Positive correlated: Engine displacement with number of cylinders. City miles per gallon with highway miles per gallon.

I think the correlated that cty and hwy makes sense since the oil consuming between city and highway won't have big difference.