PSTAT131HW01

Yifei Zhang

2022-04-02

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

Supervised learning is the type of machine learning where we give the model observed output and input, given the actual data Y being the supervisor. Supervised learning can accurately predict future response given predictors, understand how predictors affect response, find the best model for response given predictors, and assess the quality of our predictions and (or) estimation

Unsupervised learning, just as the name suggests it is referring to the learning that does not have a supervisor, no answer keys are given.

The main difference between the two is that under supervised learning response is known, and under unsupervised learning response is unknown. Supervised learning is more involved in Linear regression, Logistic regression, k-nearest neighbors, Decision trees, Random forests, Support vector machine(s). Unsupervised learning is more involved in Principal Component Analysis (PCA), k-means clustering, Hierarchical clustering. They can both work with Neural networks.

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

The main difference between a regression model and a classification model is that in the regression model the Y is quantitative(numerical values) and in a classification model the Y is qualitative(categorical values).

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

Two of the most commonly used metrics for regression ML problems are mechanistic, and parametric. And two of the most commonly used metrics for classification ML problems are empirically-driven and non-parametric.

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, Inferential models, Predictive models).

Descriptive models: best used to visually emphasize the trends in give data, graphing charts, scatter plot, tendency lines.

Inferential models: best used to show what features are significant, test theories, claims, and state relationship between outcome and predictor(s).

Predictive models:best used to find fit, the combo of features that fits best, predict Y with minimum reducible error, but it does not really focus on hypothesis tests.

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?

```
Mechanistic is when we use a theory to predict the outcome in our model. Empirically-driven is when we u
The Mechanistic one assumes a parametric form for f, but wont match the true f which is unknown. And the
They are both very flexiable, but one requires us to add more parameters to be more flexible, and the ot
```

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

```
In general, I think the empirically-driven model, we are taking a large number of observations, and we c
```

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

```
The bias-variance tradeoff is related to the use of mechanistic or empirically-driven models in the sens
```

Question #6: A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions: 1. Given a voter's profile/data, how likely is it that they will vote in favor of the candidate? 2. 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.

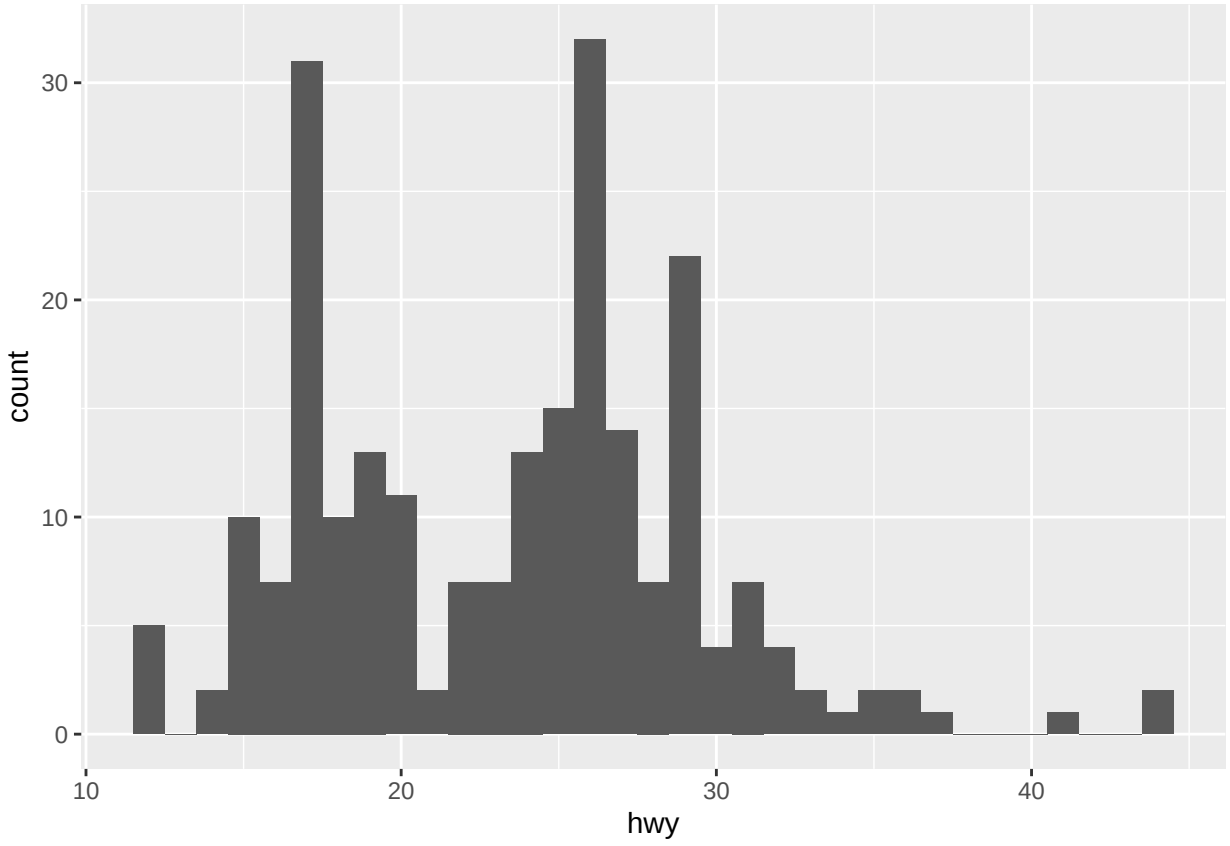
The first question should be classified as predictive, because we are trying to predict the vote of the voter based on their profile/data, we are aiming to predict Y in this case in favor of the candidate or not.

The second question should be classified as inferential, because we are trying to see if having personal contact with the candidate can be a significant feature, see if that predictor affect the outcome.

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.

From the figure down below we can see there are a few outliers on the extremely low(~12) and extremely high ends(~44), but more on the low ends relatively. There are two big clumps one centers around roughly 17 and the other one centers around roughly 26, and there are two big spikes at these two value points, and another spike near 29. The highway miles per gallon variable is not evenly distributed.

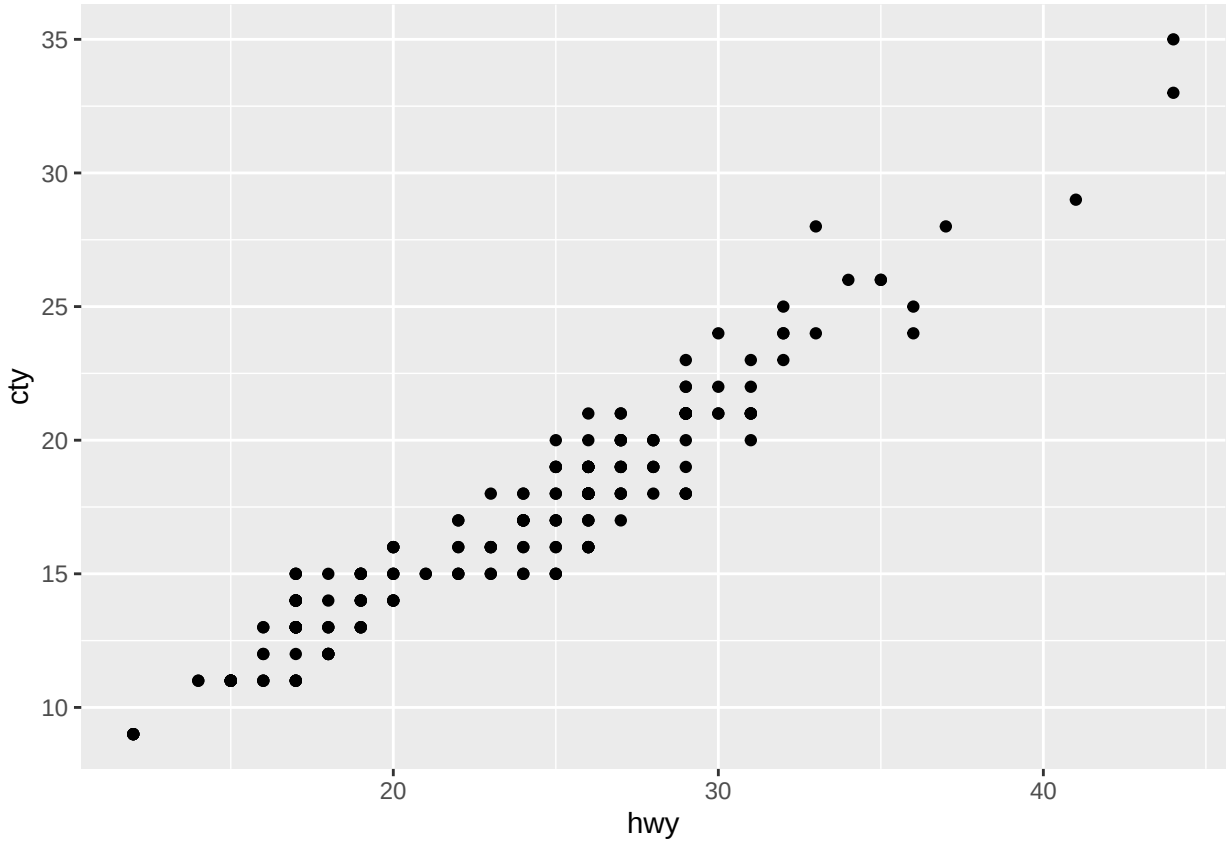
```
ggplot(data = mpg, aes(x = hwy)) +
  geom_histogram(binwidth = 1)
```



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?

From the scatter plot down below we can see there is tendency hinting at a positive linear relationship between hwy and cty. As hwy increases, cty increases. We should be able to conclude there is a relationship between hwy and cty. This means they are correlated, and it is worth noting when we make our model parameters. hwy can be a reliable predictor for cty.

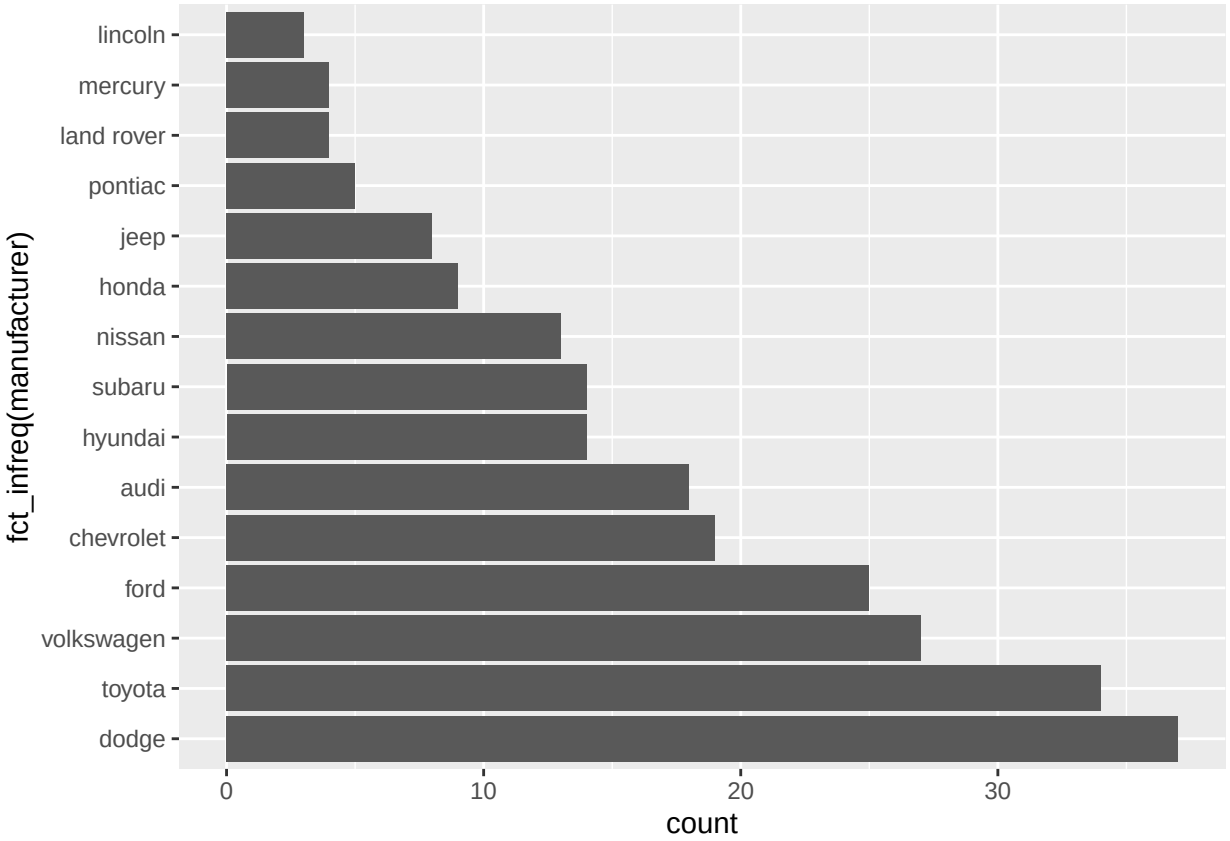
```
ggplot(data = mpg, aes(x = hwy, y = cty)) +
  geom_point()
```



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?

According to the chart down below, dodge produced the most and lincoln produced the least.

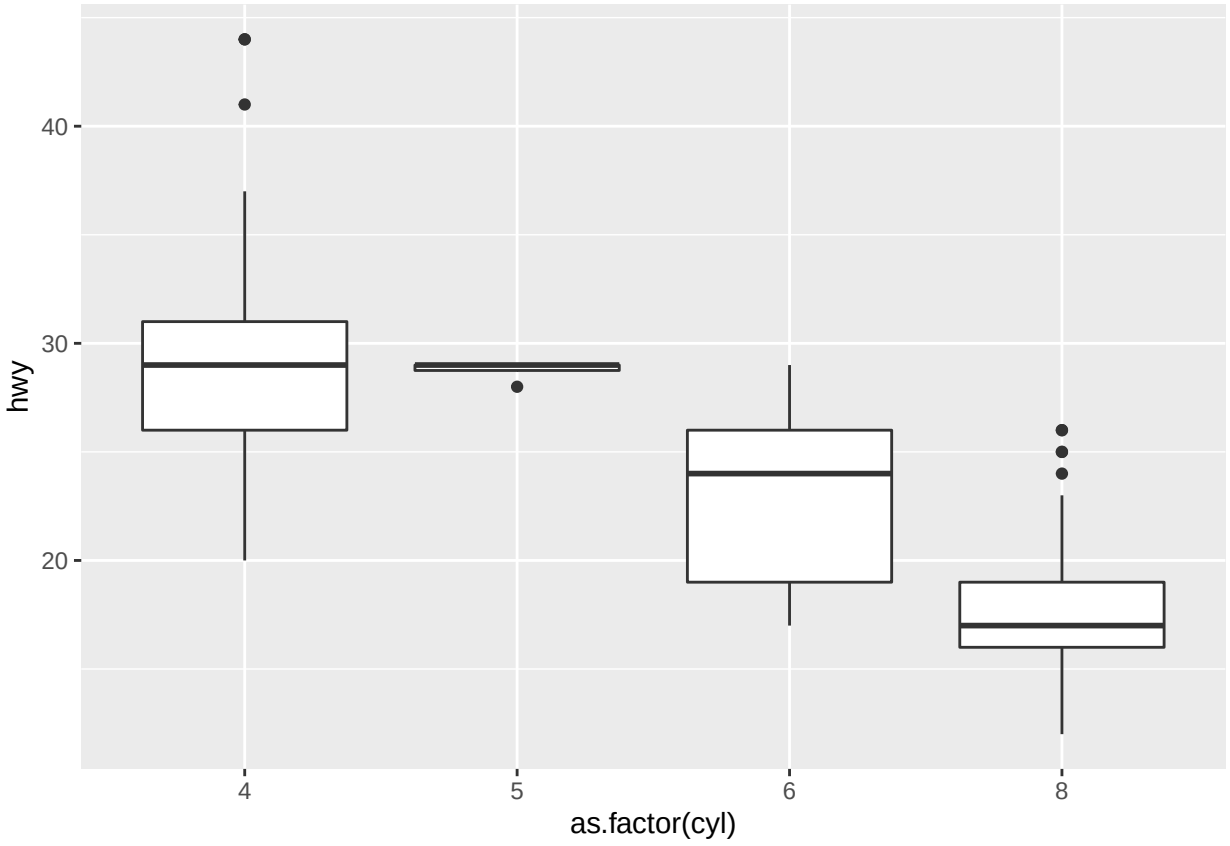
```
ggplot(data = mpg, aes(y = fct_infreq(manufacturer))) +
  geom_bar(stat = 'count')
```



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

Yes there seems to be a pattern. In group setting the number of cyl increases the hwy tend to decrease with a few outliers.

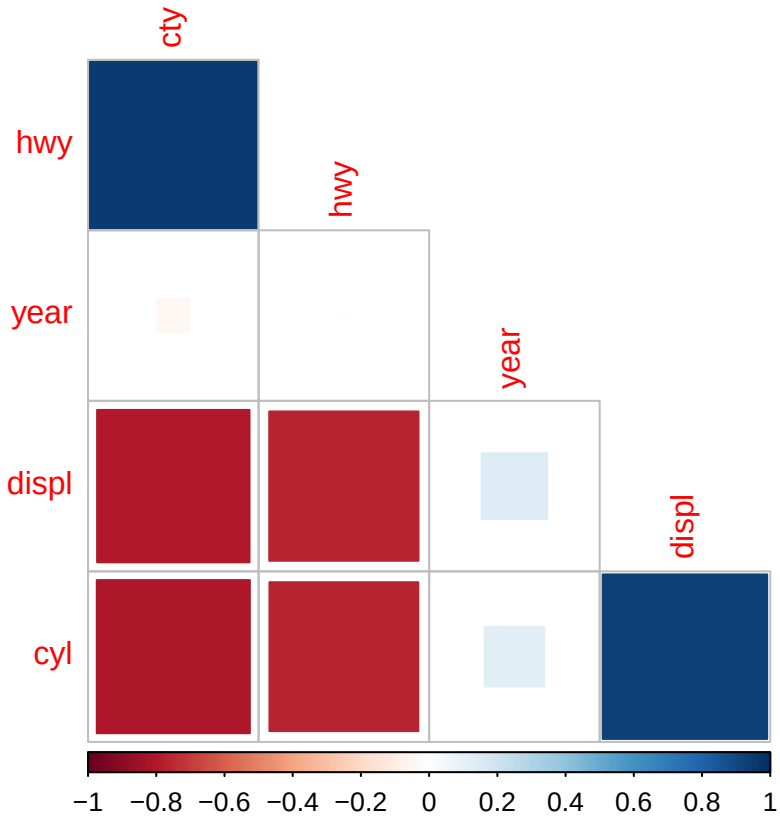
```
ggplot(data = mpg, aes(x = as.factor(cyl), y = hwy)) +
  geom_boxplot()
```



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

I only selected numeric variables, and the result for the numeric variables are not surprising, I dont know if any of them is not supposed to be like that since I dont know much about cars.Hwy and cty are positively correlated, year and cty is slightly negatively correlated, displ is negatively correlated with both cty and hwy, but is positively correlated with year, cyl is negatively correlated to both cty and hwy but is positively correlated with year and more so with displ.

```
num_mpg <- mpg %>%
  select(displ, year, cyl, cty, hwy)
M = cor(num_mpg)
corrplot(M, method = 'square', order = 'FPC', type = 'lower', diag = FALSE)
```


cty
hwy
year
displ
cyl
hwy
year
displ
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1