PSTAT131_HW1_DWH

2022-10-02

Machine Learning Main Ideas

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

Answer: The notable difference between supervised and unsupervised learning is the corresponding datasets. Supervised learning is only suitable for datasets with predictor variables x and response variables y, while unsupervised learning fits data that only have predictor variables. Tasks such as prediction and estimation are standard; for example, prediction's main goal is to generate a solution that predicts y using x most efficiently, and since unsupervised learning data doesn't have response variables, the major goal is to learn and analyze the distribution of the datasets. Information about the dataset's distribution in a multi-dimension data space allows for tasks such as clustering.

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

Answer: The main difference between a regression model and a classification model is their corresponding type of response variables. Regression models predict quantitative response variables, which are numerical; however, the classification model deals with categorical data. Even if the response variables are illustrated by numbers, the values don't have any purposes and are only labels that encode different categories.

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

Answer: Regression metrics: Mean absolute error and mean square error Classification metrics: Confusion matrix(recall and precision), F1 Score, and the AUC-ROC 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 models, Inferential models, Predictive models:

Answer: A descriptive model aims to record and visualize a pattern or trend in data, and the inferential model aims to find the relationship between the predictors and the responses. And the predictive model aims to predict the response variables most efficiently (optimizing the evaluation metric chosen for the model). For the example tasks could be such as finding the significant variables or interpreting the model into arguments.

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: A mechanistic model is a mathematical equation (a fixed theory) describing the relationship between the predictors and the responses. On the other hand, an empirically-driven model is based on empirical observations rather than a fixed theory. Mechanistic models assume a parametric form for the

relationship between the predictors and the responses, whereas the empirically-driven model doesn't, which could say empirically-driven models are more flexible. Lastly, for similarities, they both could suffer issues of over fitting, and they both use predictors to predict responses.

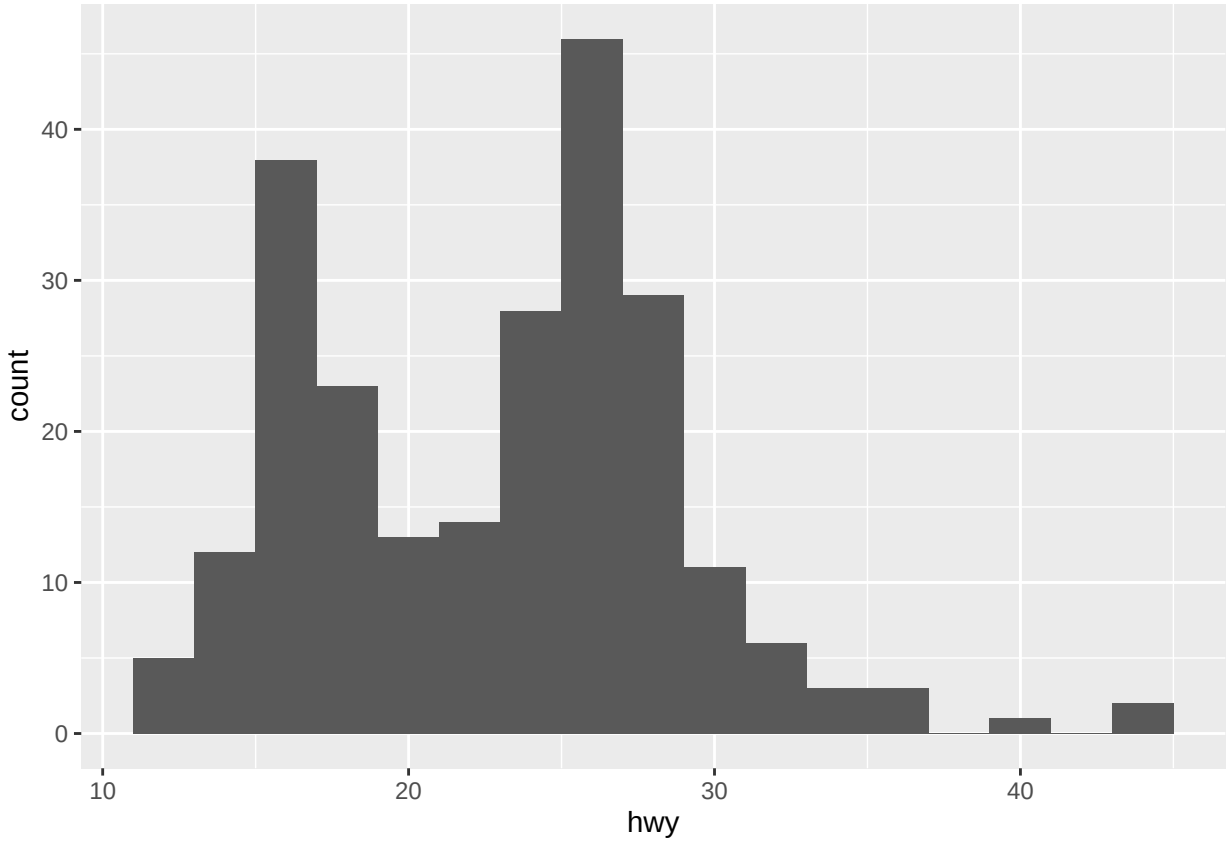
Question 6: A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions: - Given a voter's profile/data, how likely is it that they will vote in favor of the candidate? - How would a voter's likelihood of support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

Answer: The first one is predictive because the purpose is to estimate the likelihood of voting for a candidate that the result rather than the relationship is focused. For the second one is inferential since it tries to determine the influence of the change of one variable on another. In this case, the relationship is focused much more than the result, which makes the question inferential.

Exploratory Data Analysis

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.

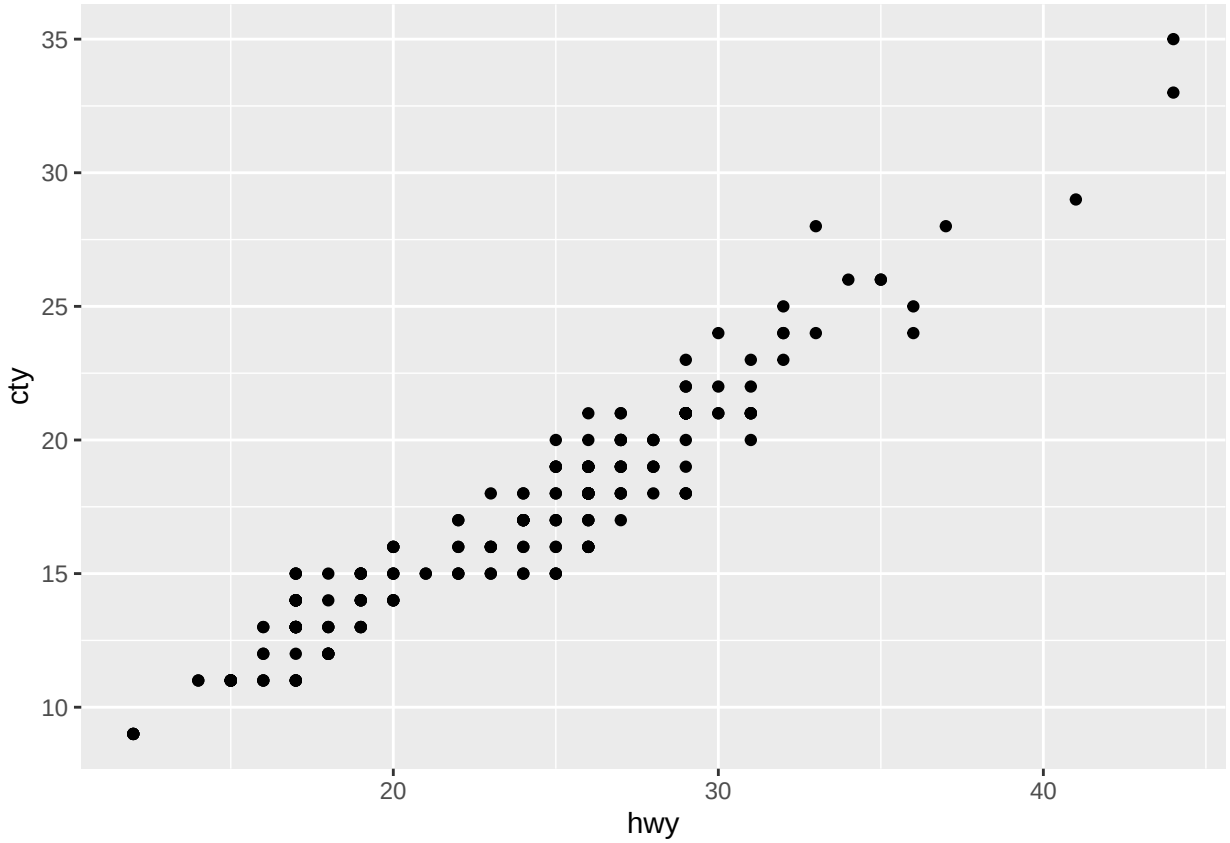
```
data('mpg')
ggplot(mpg) + geom_histogram(binwidth = 2, aes(hwy))
```



The graph has two peaks at hwy = 17 and 27 which means that most cars have highway miles per gallon around 17 or 27 for some reason.

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?

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



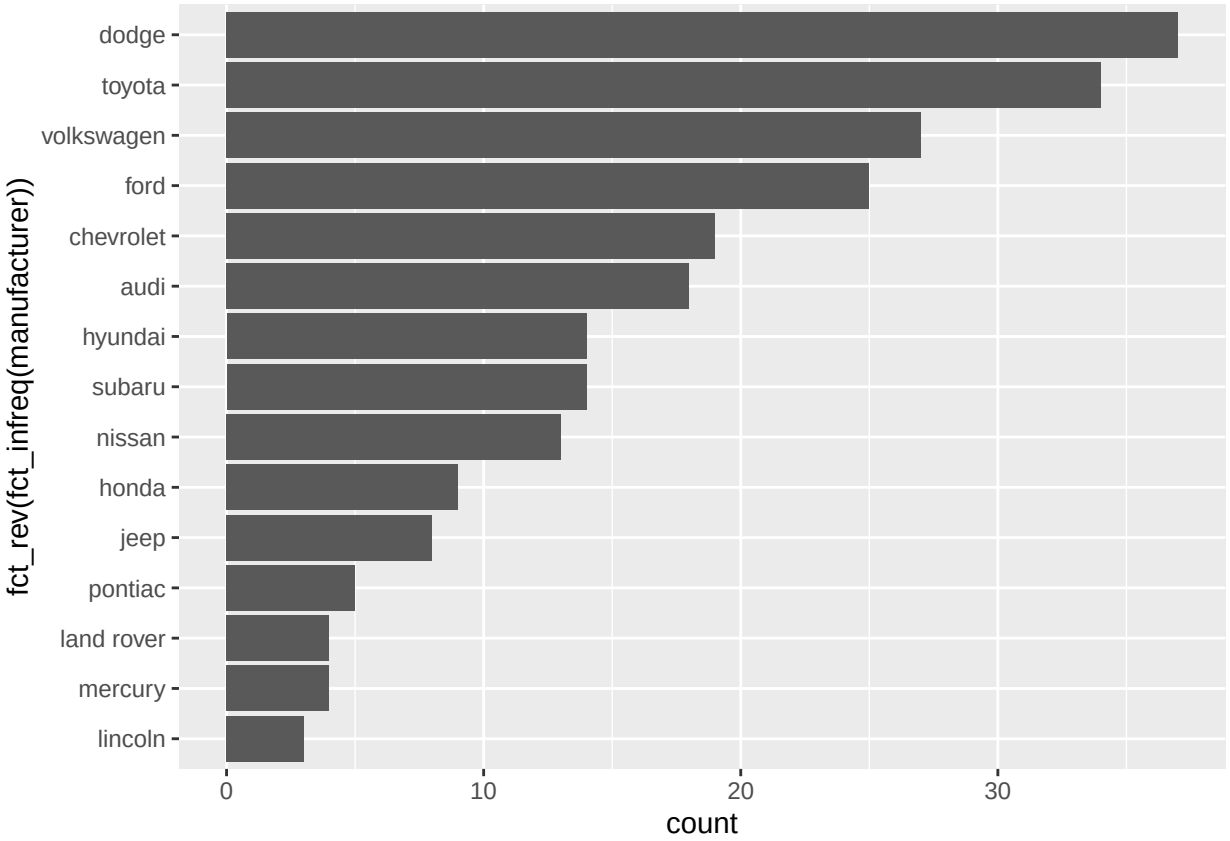
From the graph, it is fair to say that there is a positive linear relationship between hwy and cty because the more highway miles per gallon, the more city miles per gallon.

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?

```
mpg[order()]
```

```
## # A tibble: 234 x 0
```

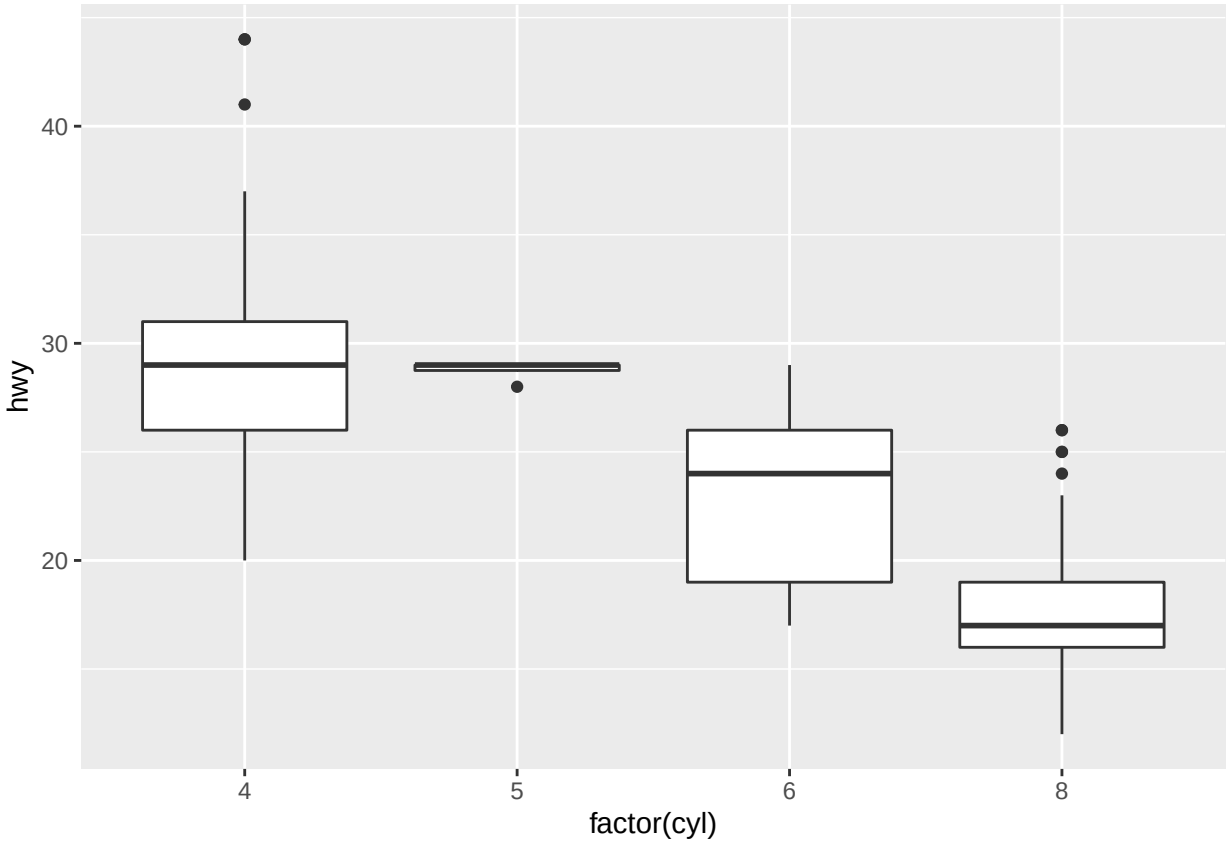
```
ggplot(mpg) + geom_bar(aes(fct_rev(fct_infreq(manufacturer)))) + coord_flip()
```



Dodge produces the most cars, and Lincoln produces the least cars.

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

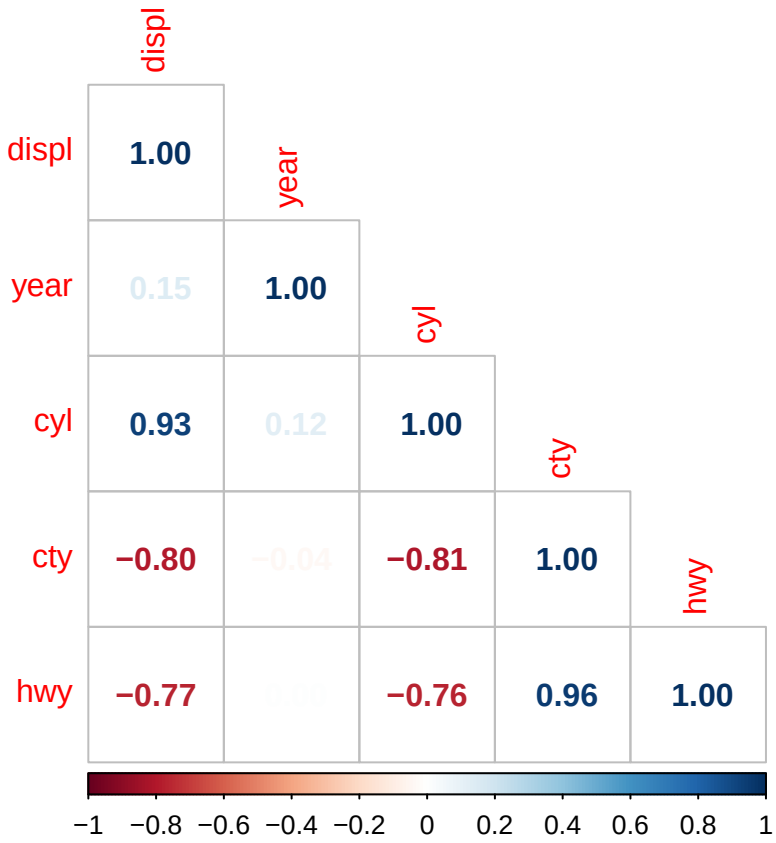
```
ggplot(mpg, aes(x = factor(cyl), y = hwy)) + geom_boxplot()
```



There seems to be a negative linear relationship between hwy and cyl because the more cylinders a car model has, the fewer (less) highway miles per gallon it is likely to have.

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?

```
M = cor(mpg[,sapply(mpg, is.numeric)])
corrplot(M, method = 'number', type = 'lower')
```



First of all, displ is positively related to cyl and negatively related to cty and hwy. Secondly, cty and hwy both have a negative relationship with cyl, while there is a positive relationship between them. Lastly, year has little correlation with any other variables, telling that the year of manufacture doesn't matter. And yes, it surprised me, and it has corresponding relationships, so it doesn't make sense to me.