hw1

2022-09-30

Machine Learning Main Ideas

Question 1: Define supervised and unsupervised learning. What are the difference(s) between them? - Supervised learning is predicting or estimating the output based on the inputs added into the model created by the researcher (pg 1 of textbook*). Unsupervised learning on the other hand is having inputs but the researcher does not know the output, aka the response variable. The differences between both types of learning is that the response is known for the supervised learning but not unsupervised and unsupervised learning is more part of exploratory data analysis (pg. 498 of textbook). Question 2: Explain the diff between regression and classification models, specifically in context of machine learning. - Regression models often test for numerical values (aka quantitative values) and often have a continuous machine learning model (lecture 1). Meanwhile classification models test for categorical values (aka qualitative values) and therefore have a discrete machine learning model (lecture 1).

Question 3: Name 2 commonly used metrics for regression ML problems. Name 2 commonly used metrics for classification ML problems. - The two commonly used metrics for regression ML problems are testing and test mean square error (MSE) (lecture 1). The two commonly used metrics for classification ML problems are training error rate and test error rate (lecture 1).

Question 4: Provide a brief description of descriptive, inferential, and predictive models. - A descriptive model is choosing a visual model that demonstrates that there is a trend within the data (lecture 1). A inferential model is finding out the relationship between the outcome and predictor and trying to test out whether or not the predictor variable affects the response variable (lecture 1). The predictive model is trying to find the response values with “minimum reducible error” (lecture 1).

Question 5 (prt1): Define mechanistic and empirically-driven. How do they differ and are similar? - Mechanistic models hold a parametric form that would be more likely to predict the response values. Empirically-driven models are more observation-wise models that see the general trend between the response and predictor variables. They differ as mechanistic models assume a parametric form, while empirically-driven models have no assumption on whether its parametric or not. These models are similar as they can or are more flexible and can be or are overfitting.

Question 5 (prt2): In general, is mechanistic or empirically driven model easier to understand? explain - I think empirically driven models are easier to understand as you have the response and predictor values and are just seeing the trend between the two variables.

Question 5 (prt3): Describe how bias-variance tradeoff is related to use of mechanistic or empirically driven models. - Bias-variance tradeoff is related to the use of mechanistic or empirically drivenmodels as both mechanistic and empirically models are/can be flexible, which causes the model to have a low bias and high variance. If the mechanistic (empirically driven models are automatically flexible) model decides to be less flexible, the model would have a high bias, but low variance. It is very rare to have a low variance and bias and most of the time, models have to give on one of these factors due to the flexible mode. This phenomenon is called the bias-variance tradeoff.

Question 6 (prt 1): Given a voter’s profile/data, how likely is that they will vote in favor of the candidate? - This question is predictive as we don’t whether they voted for the candidate or not and we are just guessing based on the data.

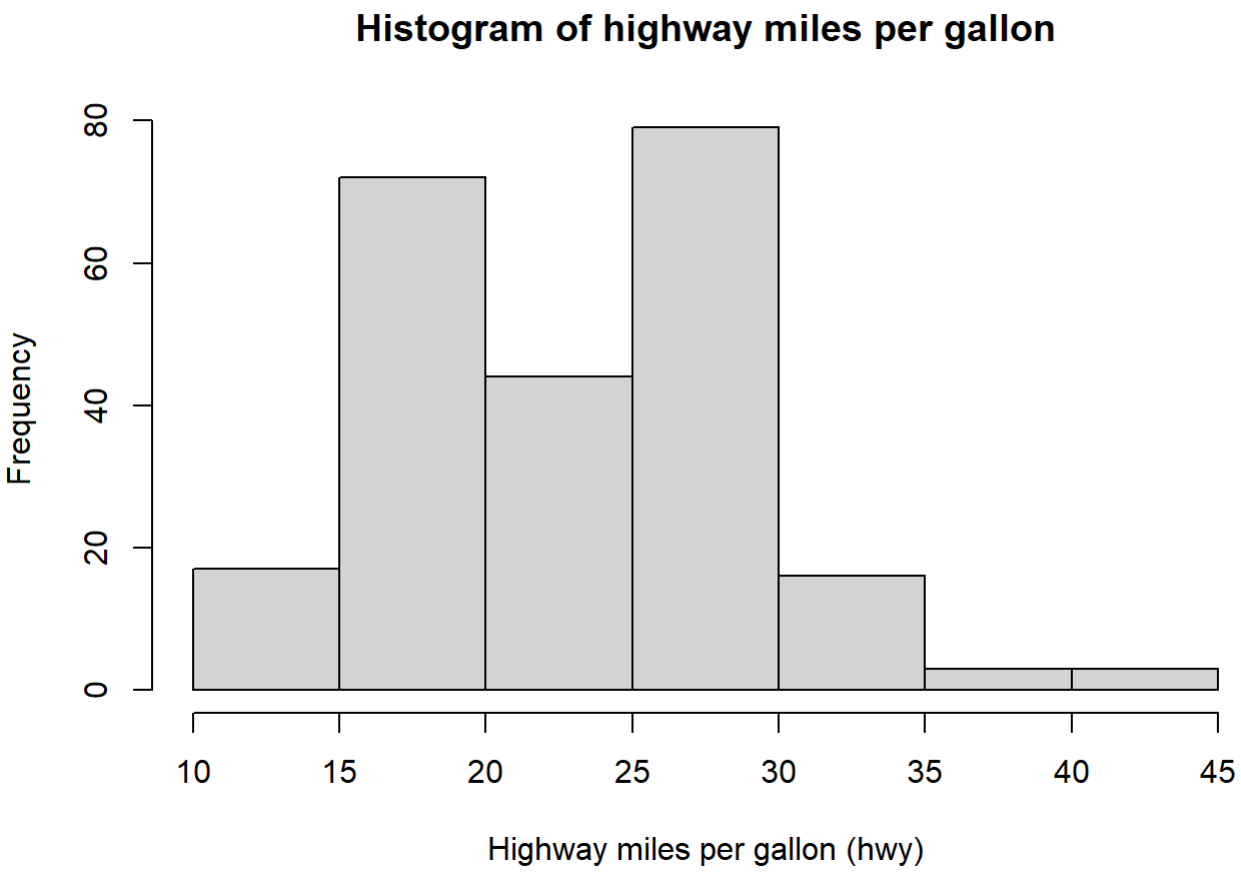
Question 6 (prt 2): How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate? - This question is inferential because we know if they had contact with the candidate or not and know whether they voted for the candidate or not. Therefore, we are analyzing the affect of the predictor variable (meeting the candidate) on the response variable (voting for him), making this a inferential model.

Exploratory Data Analysis

Setting up for exercises:

Exercise 1:

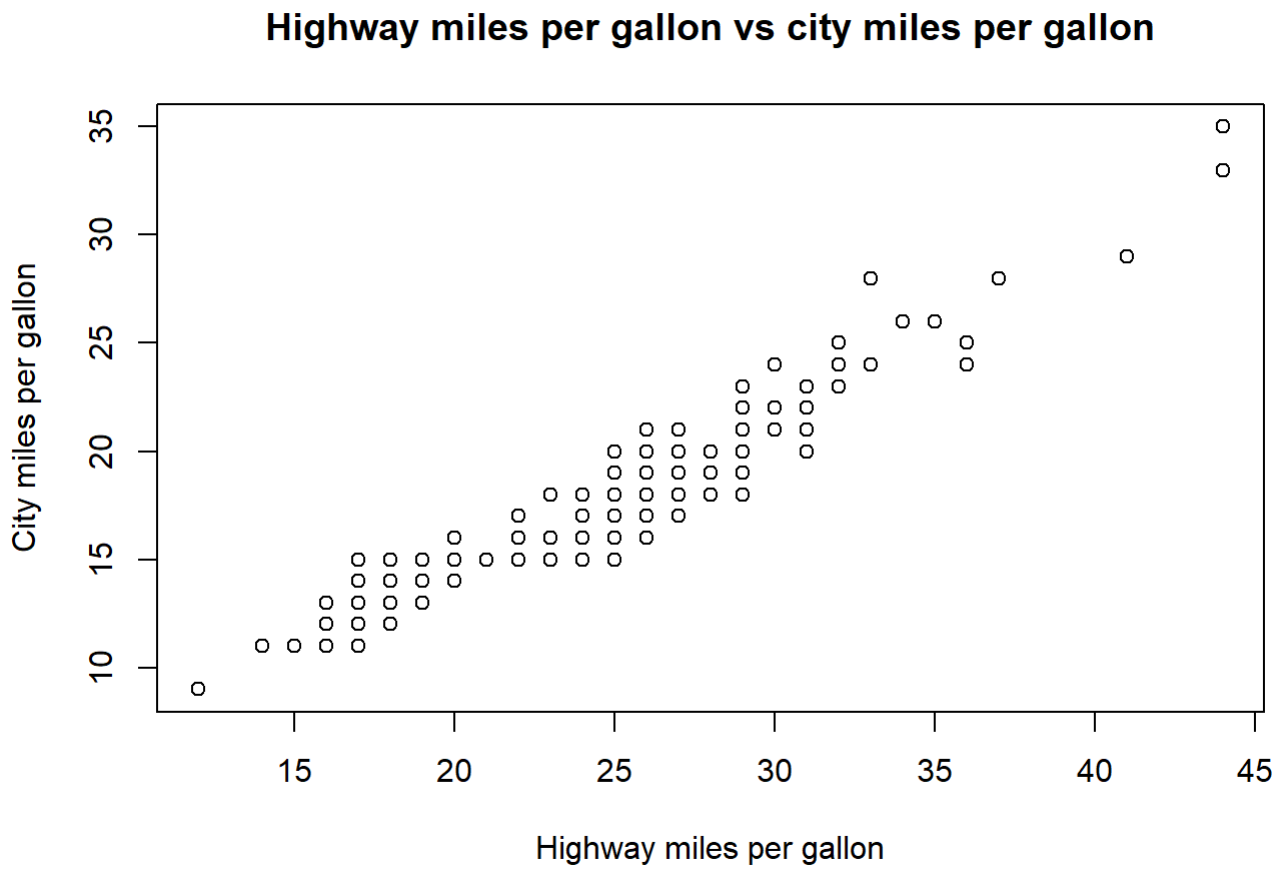
```
data(mpg)
hist(mpg$hwy, main= "Histogram of highway miles per gallon", xlab = "Highway miles per gallon (h
wy)")
```



We see that this histogram has a bimodal distribution around the 15-20 highway miles per gallon and the 25-30 highway miles per gallon. Most recorded highway miles per gallon are centered around 10-35.

Exercise 2:

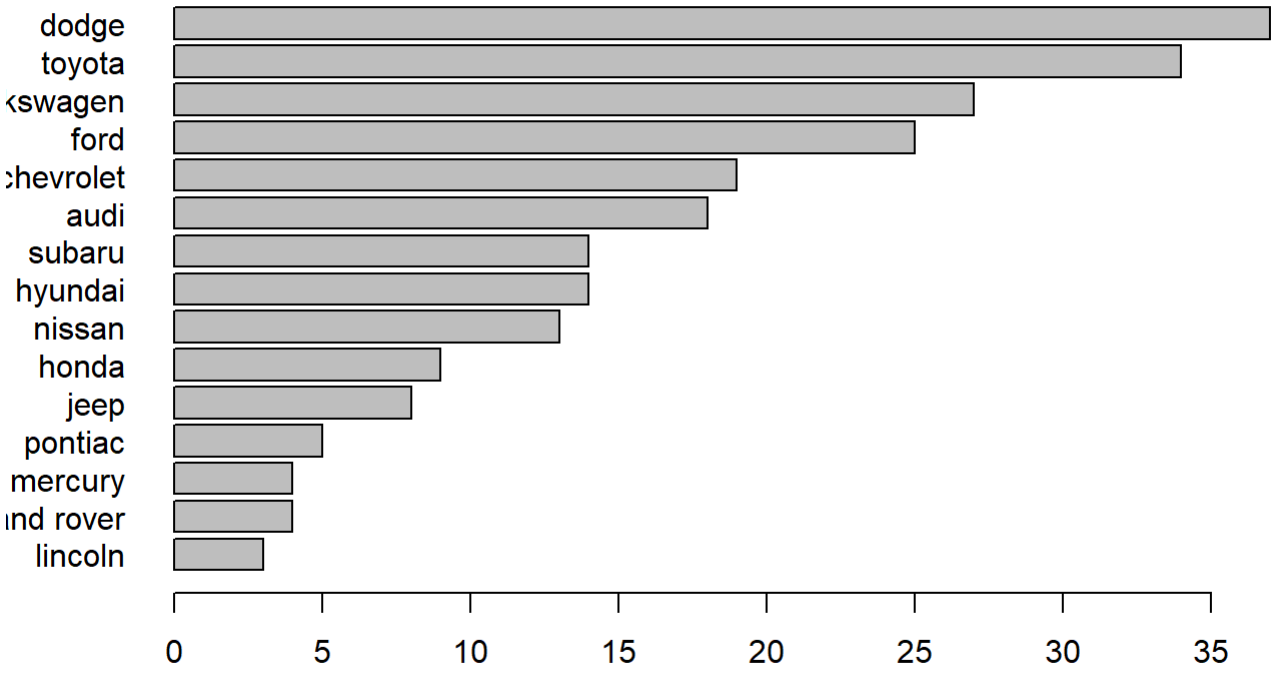
```
plot(mpg$hwy, mpg$cty, main = "Highway miles per gallon vs city miles per gallon", xlab = "Highw
ay miles per gallon", ylab = "City miles per gallon")
```



I see that as highway miles per gallon (hwy) increase then city miles per gallon also increase (cty), so there is a relationship between hwy and cty. This means that as a car drives a lot on the highway, then the cars also drive a lot on the city. This makes sense as cars who are driven a lot could travel on highway and city.

Exercise 3:

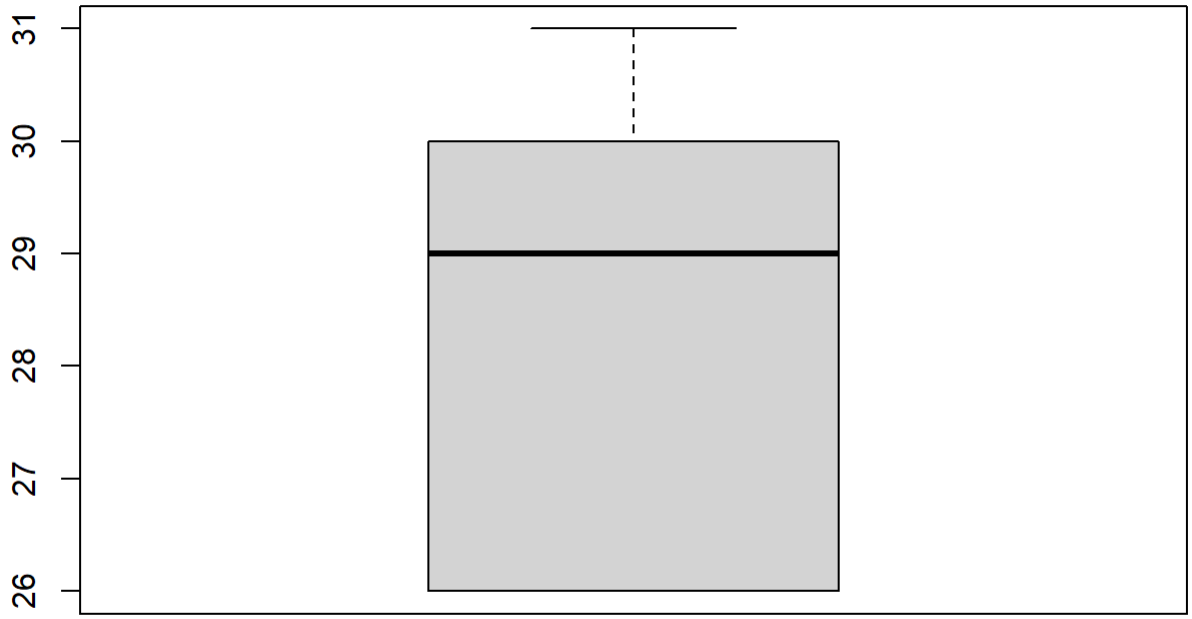
```
small <- mpg %>%
  group_by(manufacturer) %>%
  summarise(count = n()) %>%
  arrange(count)
#from section notes
barplot(small$count,names.arg = small$manufacturer, horiz = TRUE, las=1)
```



Dodge produced the most cars, while Lincoln produced the least cars.

Exercise 4:

```
#hwy grouped by cyl. what's pattern
boxy <- mpg %>%
  group_by(cyl) %>%
  select(hwy, cyl) %>%
  head()
#from section notes
boxplot(boxy$hwy)
```



It seems that most values are around 26 to 30 highway miles per gallon (with the exception to 31).

Exercise 5:

```
#install.packages("corrplot")
#install.packages("caret")
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(caret)
```

```
## Loading required package: lattice
```

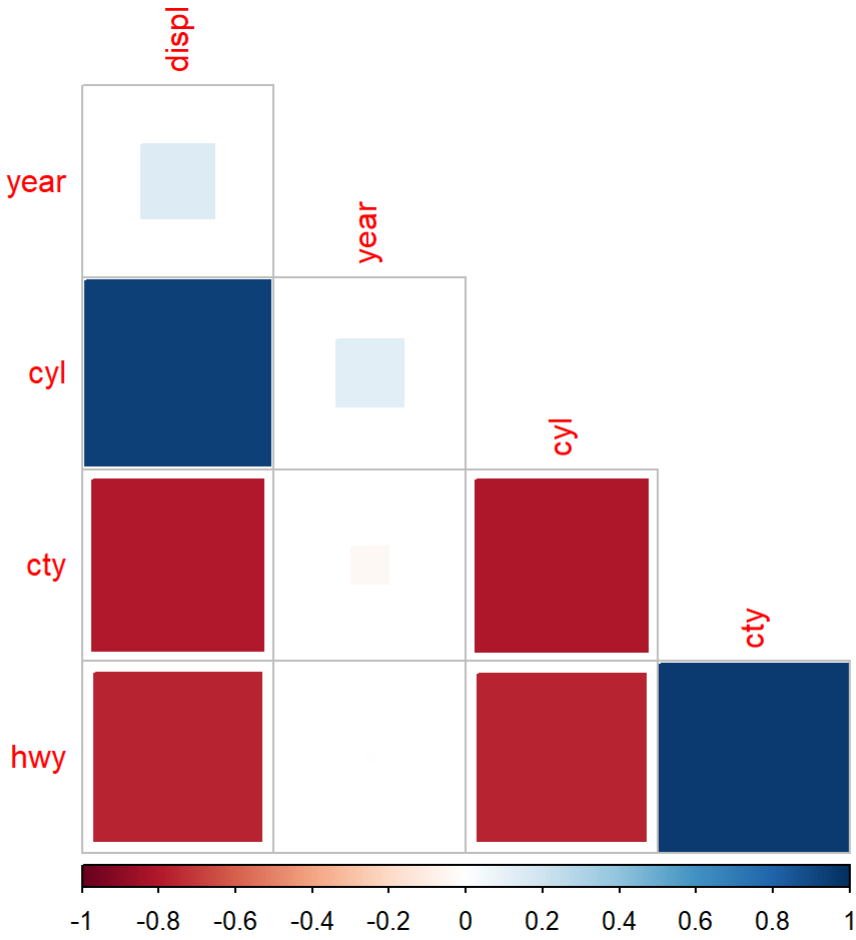
```
##
## Attaching package: 'caret'
```

```
## The following objects are masked from 'package:yardstick':
##
##     precision, recall, sensitivity, specificity
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
num_mpg <- mpg %>%
  select_if(is.numeric)

m = cor(num_mpg)
corrplot(m, method = 'square', type = 'lower', diag = FALSE)
```




```
#Link given by professor about corrplot
```

The variables cty (city miles per gallon) vs displ (engine displacement), hwy (highway miles per gallon) vs displ, cty vs cyl (number of cylinders), and hwy vs cyl are strongly negatively correlated. While variables like cyl vs dipl and hwy vs cty are strongly positively correlated. These relationships do make sense to me. For example, it makes sense that driving a lot (cty or hwy) can have a negative impact on engine displacement. I was kind of surprised that year had almost zero correlation between the other variables.

*textbook refers to “An Introduction to Statistical Learning with Applications in R” by G. James, D. Witten, T. Hastie, R. Tibshirani