Homework 1

Richard Zhao

2022-10-20

```
library(tidyverse)
library(ggplot2)
library(corrplot)
```

Question 1:

Supervised learning is when for each observation of the predictor measurements(s) xi, i = 1,. . . ,n there is an associated response measurement yi.

Unsupervised learning is when for every observation i = 1,. . . ,n, we observe a vector of measurements xi but no associated response yi.

The difference between the two is that we can fit a linear regression model for supervised learning but not for unsupervised learning. For supervised learning, the goal is to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference). For unsupervised learning, a linear regression model cannot be fitted because there is not a response variable to predict.

(from page #26 of An Introduction to Statistical Learning with Applications in R)

Question 2:

In the context of machine learning, regression models involve a continuous outcome with a quantitative response, while classification models involve a categorical outcome with a qualitative response.

Question 3:

Two commonly used metrics for regression ML problems are R-squared and MAE, or mean absolute error.

Two commonly used metrics for classification ML problems are accuracy and Bayes Classifier.

Question 4:

Descriptive models: Used to best visually emphasize a trend in data, such as using a line on a scatterplot.

Inferential models: Used to test theories and make causal claims regarding the relationship between the outcome and predictor(s).

Predictive models: Used to predict Y with minimum reducible error.

Question 5:

Mechanistic predictive models assume a parametric form for f. They won't match the true unknown f. Empirically-driven parametric models make no assumptions about f, the form of the relationship. They require a larger number of observations and are much more flexible by default. Both models have the same

slopes and intercepts. Mechanistic models have higher bias and lower variance, while empirically-driven models have higher variance and lower bias.

In general, a mechanistic model is easier to understand because it is easier to estimate f as mechanistic models generally fit simple parametric forms.

Mechanistic models have higher bias and lower variance because they are easier to understand. Empirically-driven models have higher variance and lower bias because they are harder to understand.

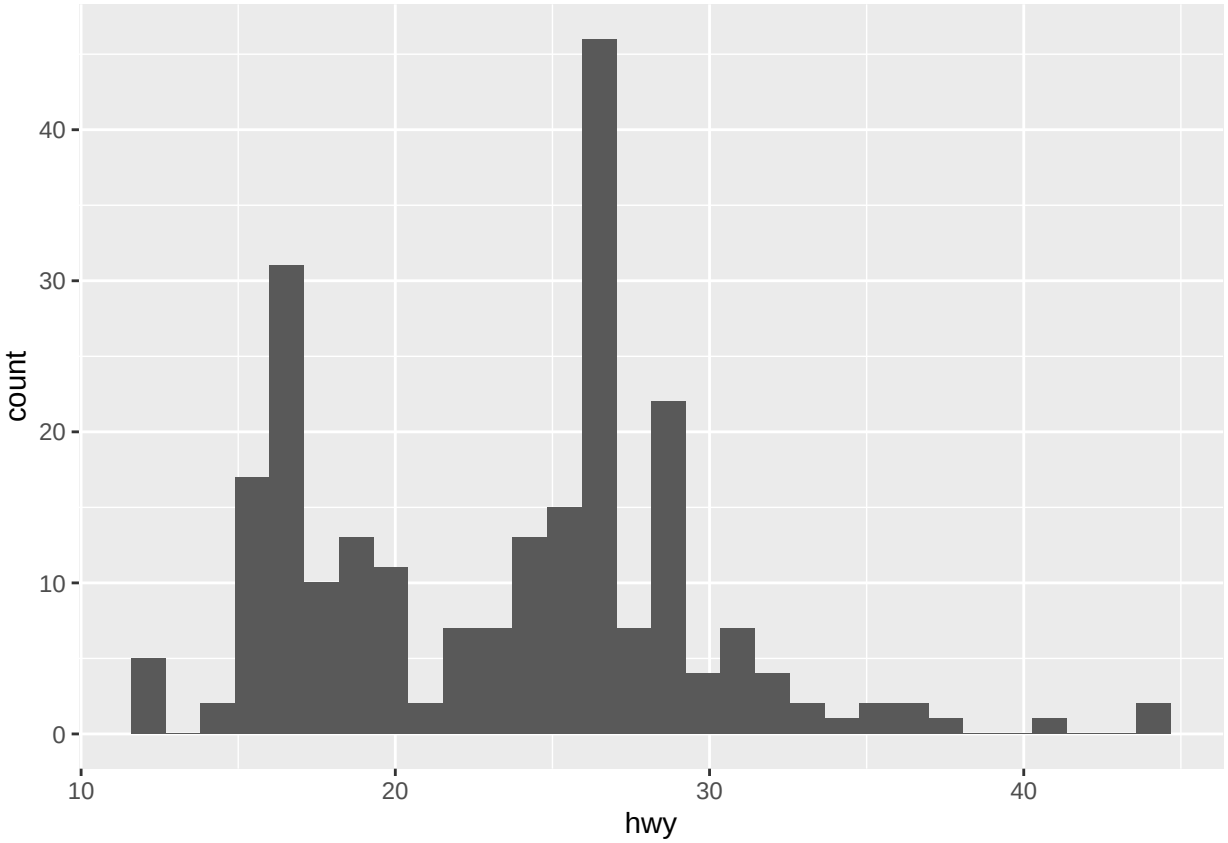
Question 6:

The first question is predictive because it is aiming to determine the probability of a voter's vote in favor of the candidate based on their profile/data.

The second question is inferential because it is aiming to determine the relationship between the outcome, which is a voter's likelihood of support for the candidate, and the predictor, which is personal contact with the candidate.

Exercise 1:

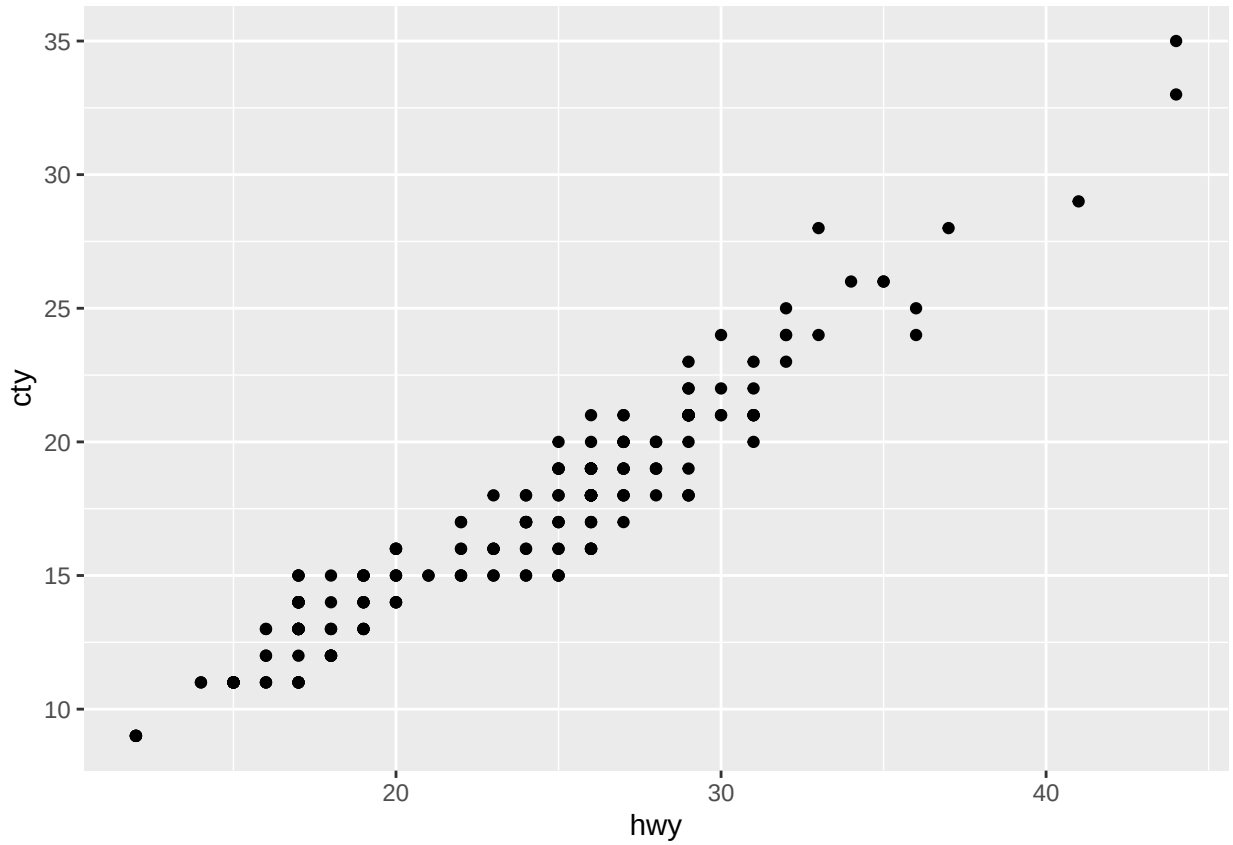
```
ggplot(mpg, aes(x = hwy)) + geom_histogram()
```



The highway mpg is skewed right. The distribution seems to be bimodal, with a peak around 17 and another peak around 26. The highway mpg drops significantly after 30.

Exercise 2:

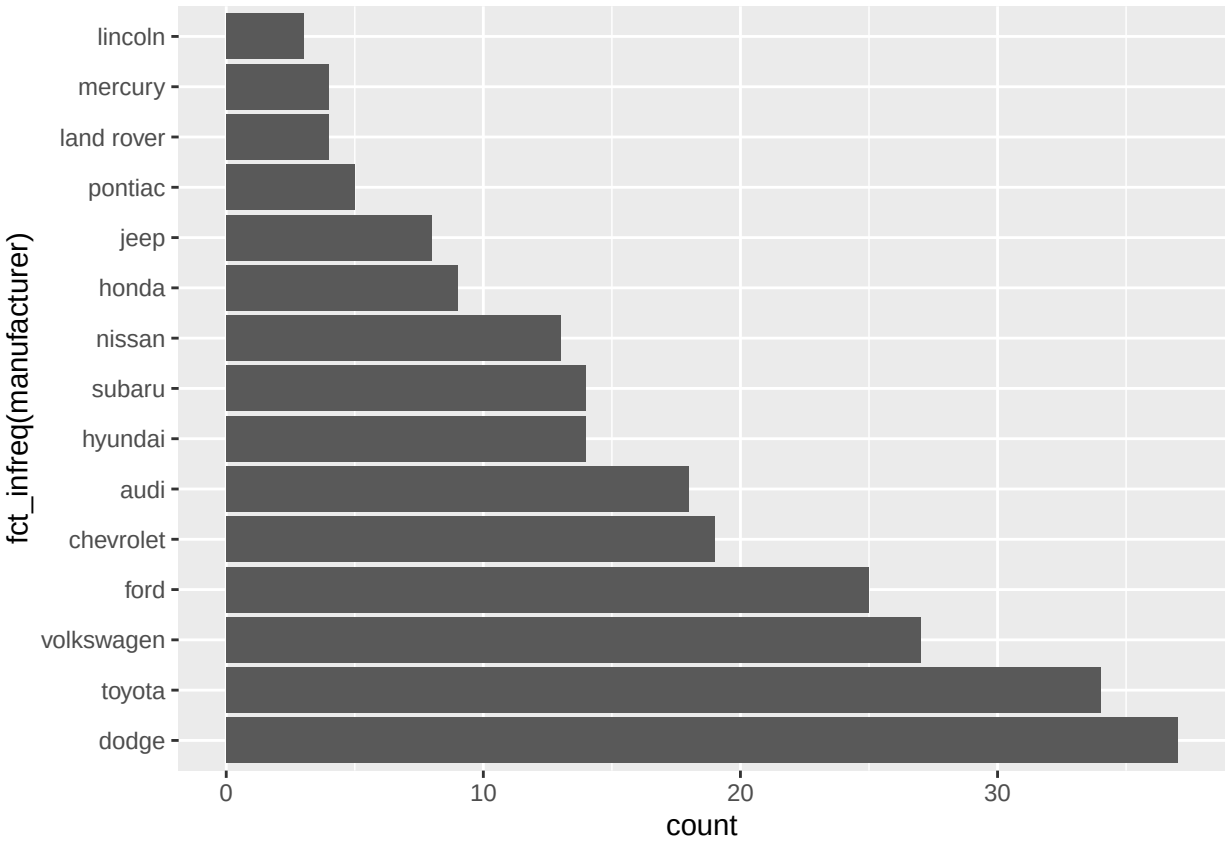
```
ggplot(mpg, aes(x = hwy, y = cty)) + geom_point()
```



There is a positive linear correlation between hwy and cty. City mileage seems to increase as highway mileage increases.

Exercise 3:

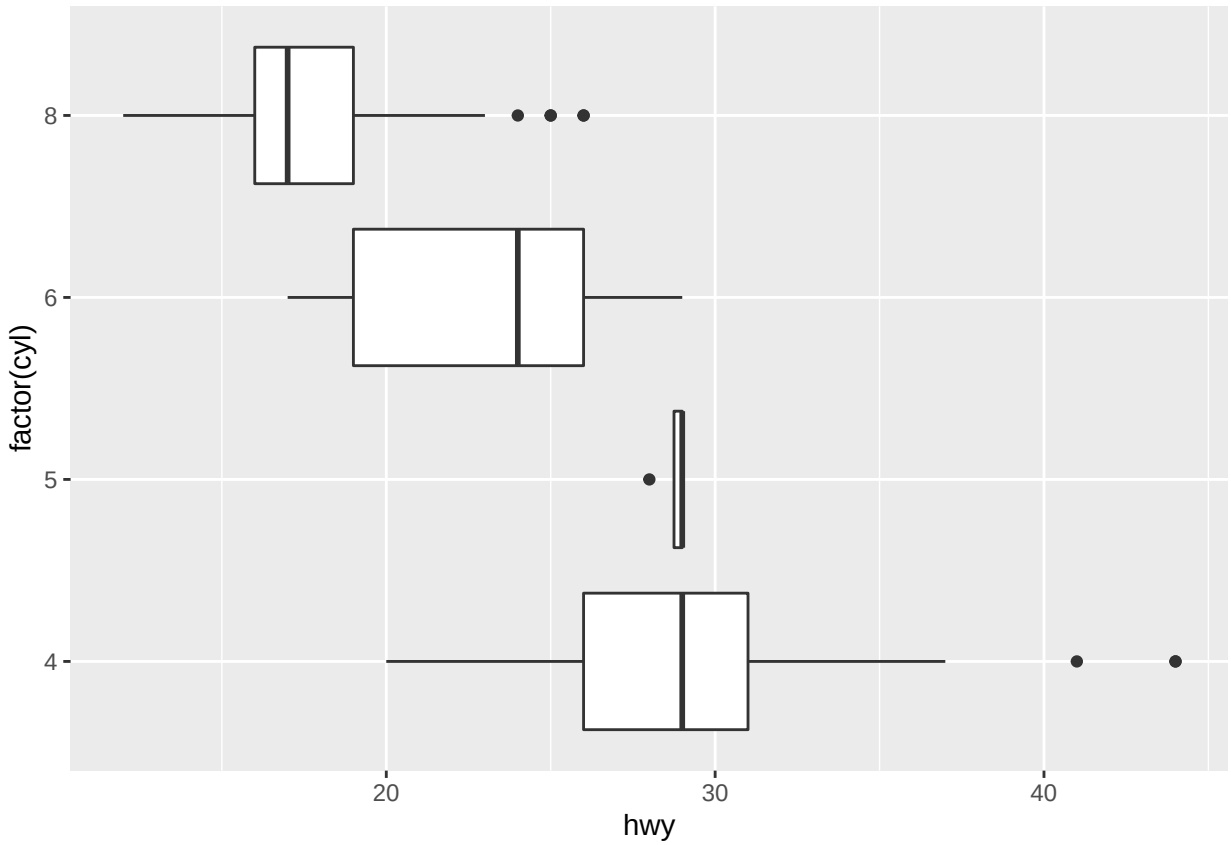
```
ggplot(mpg, aes(x = fct_infreq(manufacturer))) + geom_bar() + coord_flip()
```



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

Exercise 4:

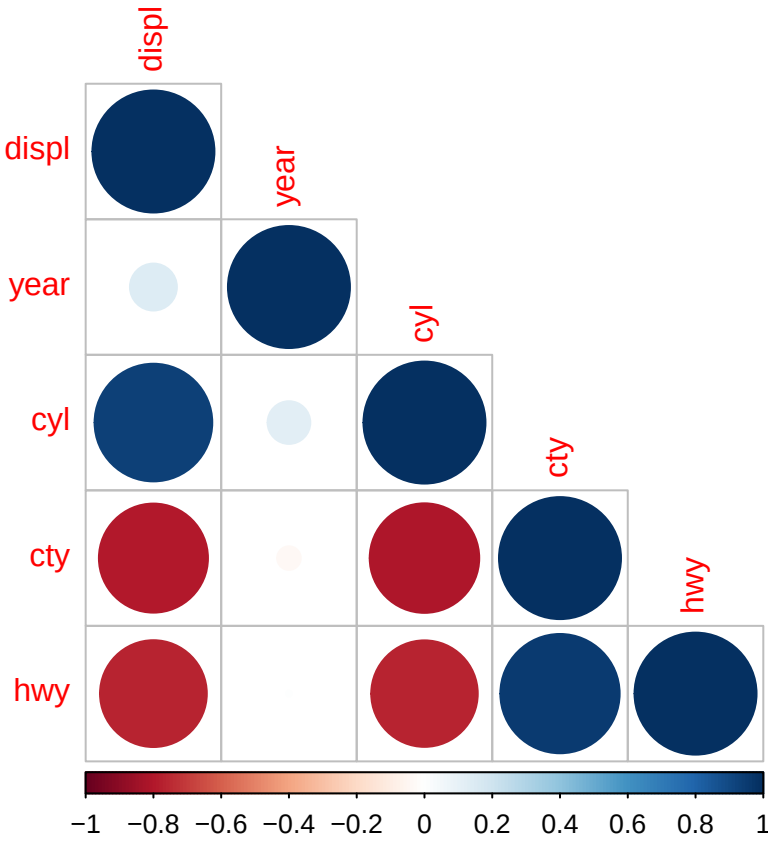
```
ggplot(mpg, aes(x = hwy, y = factor(cyl))) + geom_boxplot()
```



As the number of cylinders increases, highway mpg decreases.

Exercise 5:

```
mpg %>%
  select(is.numeric) %>%
  cor() %>%
  corrplot(type = "lower")
```



There is a positive correlation between city mileage and highway mileage. There is also a positive correlation between displacement and the number of cylinders. However, there is a negative correlation between displacement and city mileage, displacement and highway mileage, the number of cylinders and city mileage, and the number of cylinders and highway mileage. These relationships make sense to me.