Homework 1

Adam Alcala

2022-10-02

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Homework 1

Adam Alcala

10-01-2022 *Question prompts taken from homework-1.html

Question 1

Define supervised and unsupervised learning. What are the difference(s) between them? Supervised and Unsupervised learning are the two ways of how machines learn through models. Supervised learning means that the response variable, Y, is known, or 'supervised'. We can directly see what the machine model predicted and compare it to the known response variable. This lets us see how effective the model is at predicting the response variable with the predictor variables. Unlike supervised learning, unsupervised learning does not have any response variable known. Since the response is unknown, what is shown from an unsupervised model is what was shown from the predictors.

Question 2:

Explain the difference between a regression model and a classification model, specifically in the context of machine learning. For Machine learning, the difference between a regression and classification model is concerned with the characteristic of the response variable. If the response is quantitative, it is a regression model. If it is qualitative, it is a classification model.

Question 3:

Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems. Two commonly used metrics for regression: Age and Height

Two commonly used metrics for classification: Quality and Medical Conditions

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: These models are concerned with visually representing relationships or trends in data.

Inferential models: This model wants to test questions asked about the data and the outcome. It wants to test any theories involving the data and find out what is significant.

Predictive models: For this model, the main goal is to 'predict' the response with the least amount of reducible error. It wants to find the combination of features that reduce the most amount of error.

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 models assume that the function is in a parametric form. That means that the function is defined by various constants (for example, $\beta_0 + \beta_1 + ...$). For an empirically-driven model, there is no assumption of a parametric form. Empirically-driven models also require a much larger number of observations. Between the two models, an empirically- driven model is much more flexible to start, but mechanistic models can become more flexible with the addition of more parameters. In a similar manner, both models can become too flexible and cause over-fitting.

In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice. A mechanistic model is simpler to understand. Assuming the model function to be parametric clearly defines what the model is taking in and how those may affect the model's outcome.

Describe how the bias-variance trade off is related to the use of mechanistic or empirically-driven models. As a model gets more flexible, it gains less bias and more variance. In the context of mechanistic or empirically-driven models, the former is initially more biased with less variance, but the latter is less biased with more variance.

Question 4:

Classify each question as either predictive or inferential. Explain your reasoning for each.

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? Predictive, this question wants to find out how likely a voter will choose a certain candidate. It would be most beneficial to answer this question with trying to have the least amount of error in predicting the choice of a voter.

How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate? Inferential, this question is concerned with testing a theory about the voters. Seeing if a voter's support for a candidate will change with personal contact is the main theory that want to be tested with this question. It will test if there is a significant relationship between the predictor and the outcome.

Exploratory Data Analysis

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
mpg
```

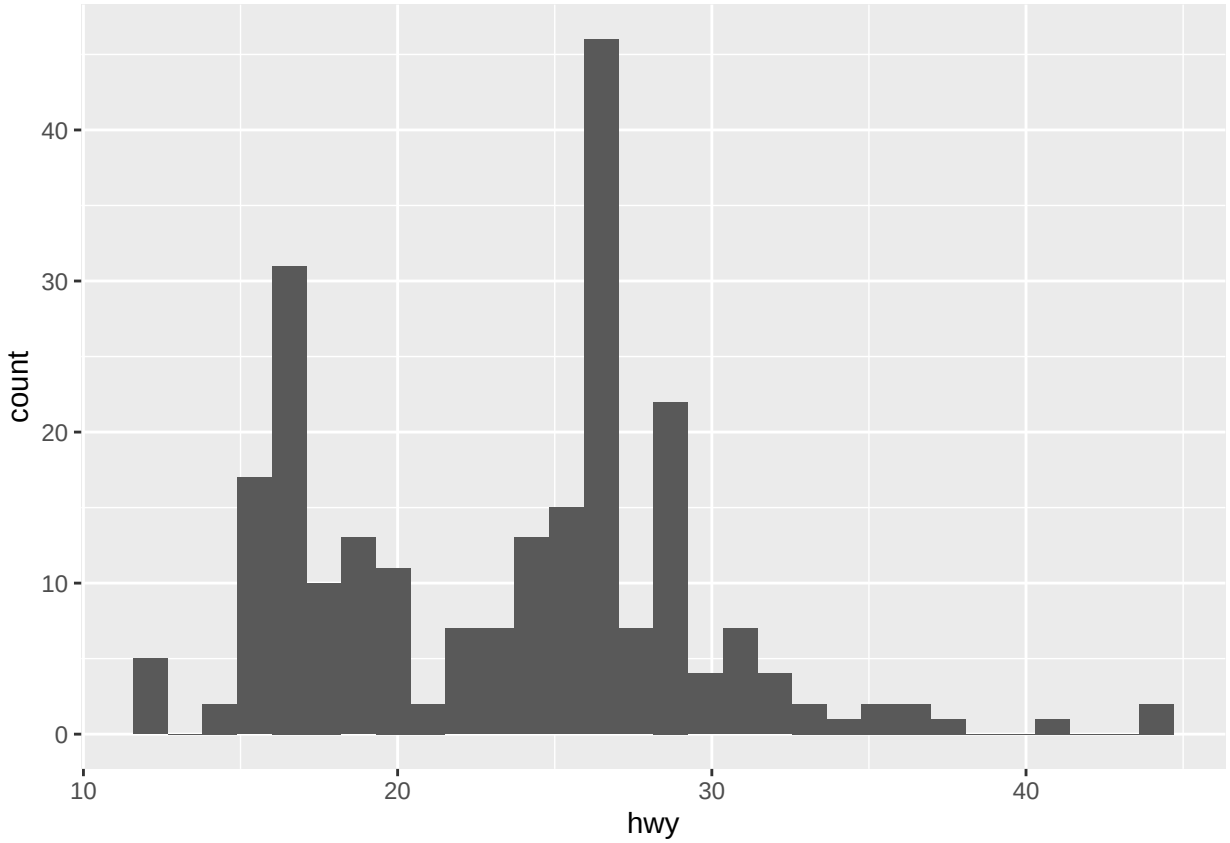
```
## # A tibble: 234 x 11
##    manufacturer model      displ  year   cyl trans drv     cty   hwy fl    class
##    <chr>        <chr>      <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
##  1 audi         a4           1.8  1999     4 auto~ f        18    29 p     comp~
##  2 audi         a4           1.8  1999     4 manu~ f        21    29 p     comp~
##  3 audi         a4           2    2008     4 manu~ f        20    31 p     comp~
##  4 audi         a4           2    2008     4 auto~ f        21    30 p     comp~
##  5 audi         a4           2.8  1999     6 auto~ f        16    26 p     comp~
##  6 audi         a4           2.8  1999     6 manu~ f        18    26 p     comp~
##  7 audi         a4           3.1  2008     6 auto~ f        18    27 p     comp~
##  8 audi         a4 quattro   1.8  1999     4 manu~ 4        18    26 p     comp~
##  9 audi         a4 quattro   1.8  1999     4 auto~ 4        16    25 p     comp~
## 10 audi         a4 quattro   2    2008     4 manu~ 4        20    28 p     comp~
## # ... with 224 more rows
```

Exercise 1:

Histogram of the 'hwy' variable

```
mpg_hwy <- mpg %>% select(hwy)
hist <- ggplot(data = mpg_hwy) + geom_histogram(mapping = aes(x=hwy))
hist
```

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

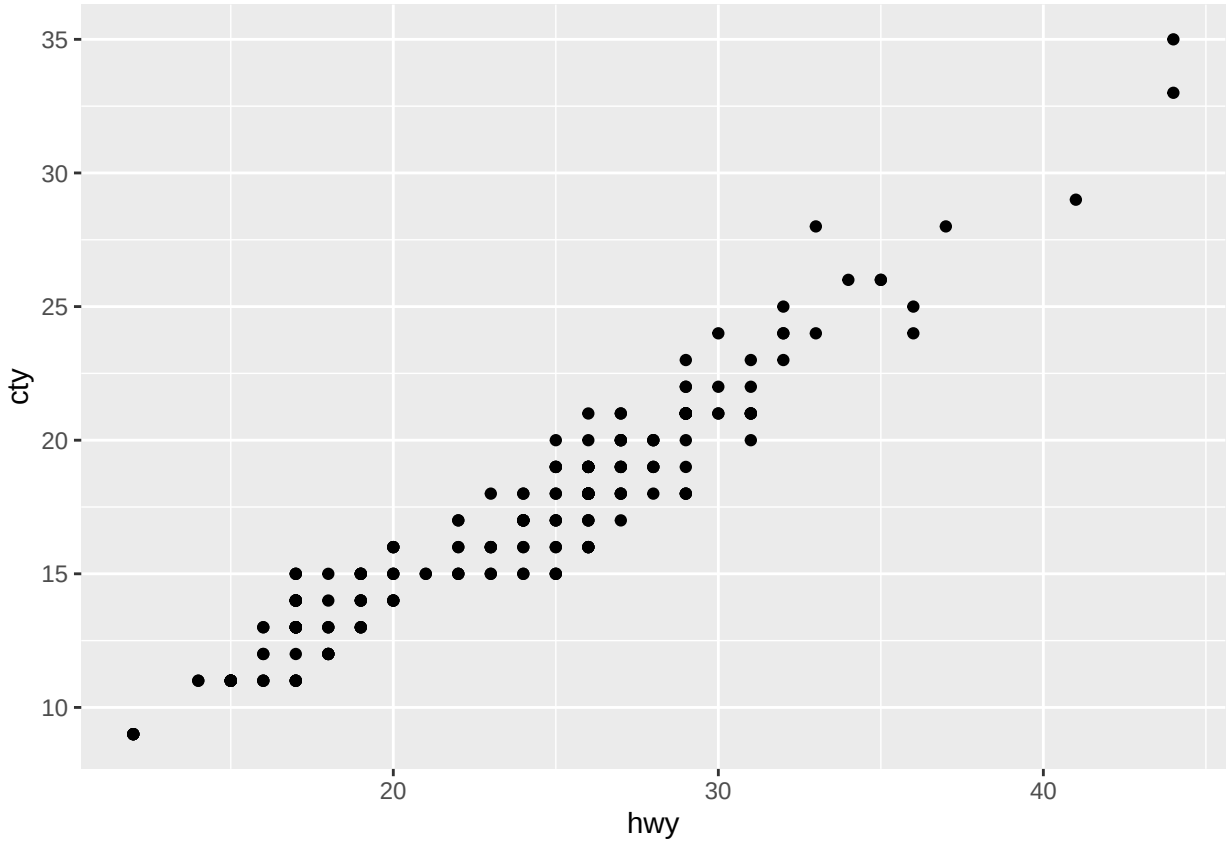


I notice that the majority of the observations are between 15 and 30. There are also two peaks in the observations, one at ~16 and another at ~26.

Exercise 2:

Scatter plot of hwy and cty variables

```
mpg_scatter <-mpg %>% select(hwy,cty)
splot <- ggplot(data = mpg_scatter) + geom_point(mapping = aes(x=hwy, y = cty))
splot
```

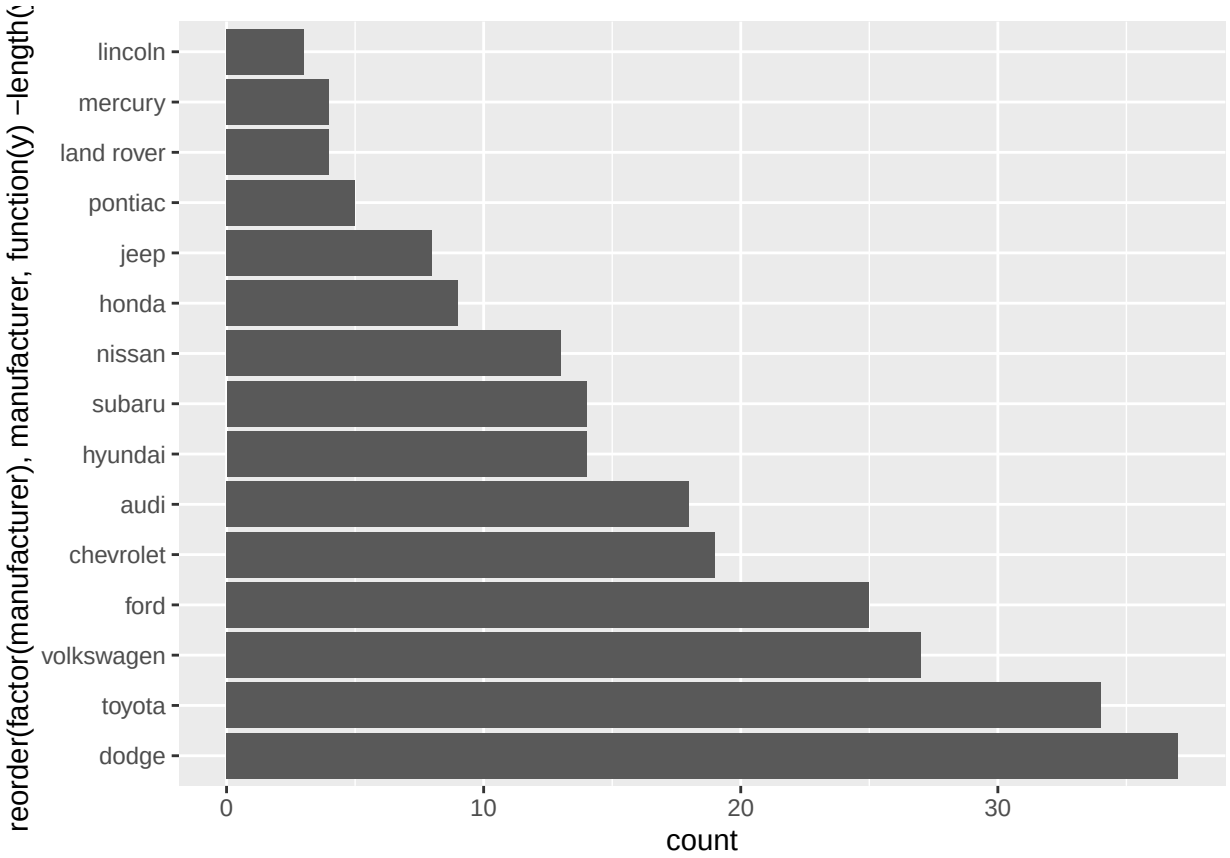


From the scatter plot, there is a direct positive relationship between the hwy and cty variables. As hwy increases, so does cty.

Exercise 3:

Bar plot of manufacturer variable.

```
mpg_bar <- mpg %>% select(manufacturer)
bplot <- ggplot(data=mpg_bar, aes(y=reorder(factor(manufacturer),manufacturer,function(y)-length(y)))) +
bplot
```

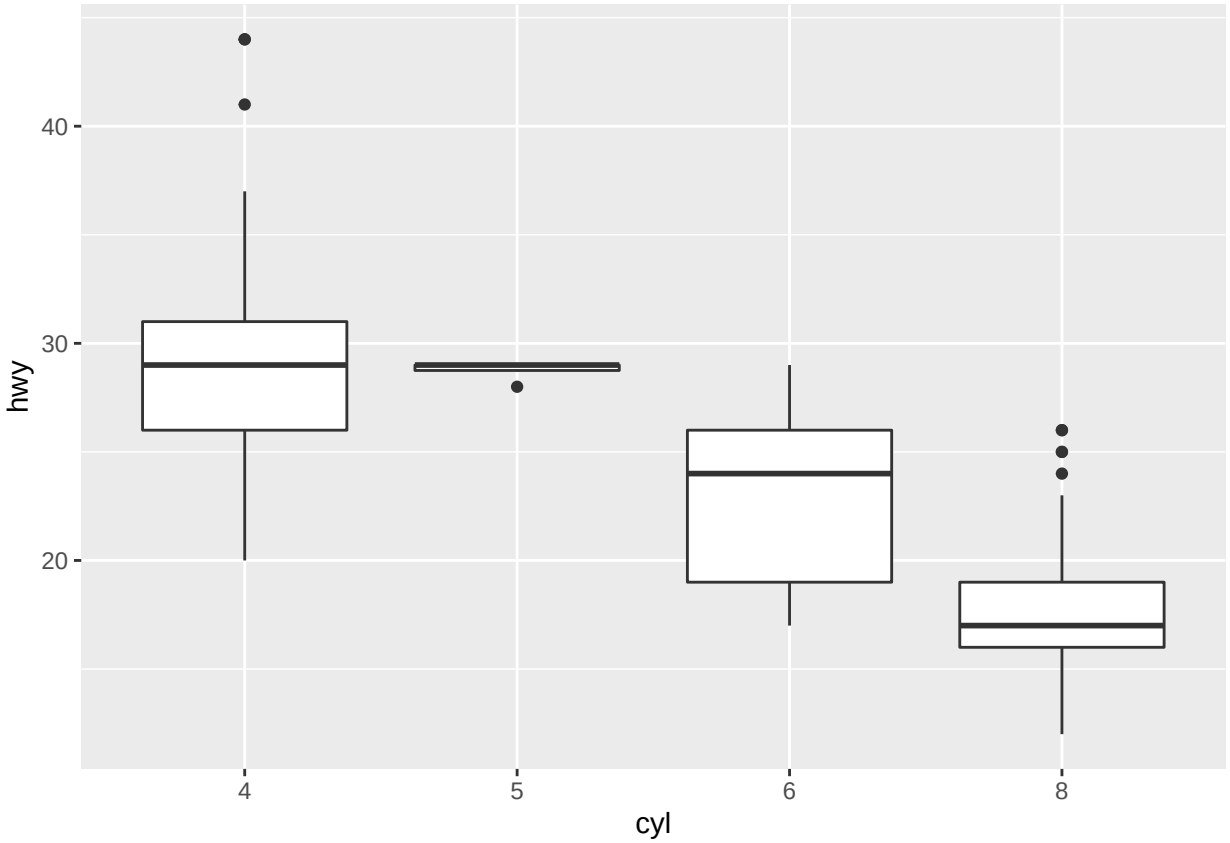


The manufacturer that produced the most cars was Dodge, and the one that produced the least was Lincoln.

Exercise 4:

Box Plot of Hwy grouped by cyl

```
mpg_box <- mpg %>% select(hwy,cyl)
mpg_box$cyl = as.character(mpg_box$cyl)
bxplot <-ggplot(mpg_box,aes(x= cyl,y = hwy)) + geom_boxplot()
bxplot
```



I notice that as the number of cylinders on a car increases, the highway miles per gallon begin to decrease.

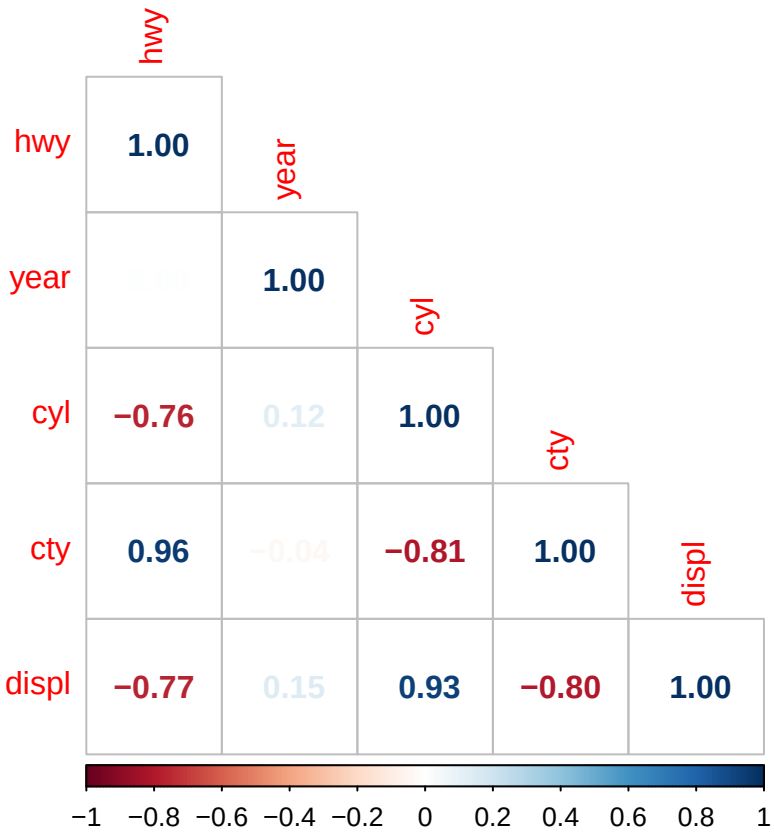
Exercise 5:

Use corrplot to make a lower triangle correlation matrix of the mpg data set

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
mpg_numbers <- mpg %>% select(hwy,year,cyl,cty,displ)
cplot <- cor(mpg_numbers)
corrplot(cplot, method = 'number',type = 'lower')
```



In order to correlate variables in the data set, only numerical variables have been chosen. These are the positive relationships:
'hwy and cty' 'year with cyl' 'year with displ' 'cyl with displ'

For hwy and cty, both have to do with miles per gallon, so a car that would be better at having more miles per gallon would be effective in both the city and highway.

With year with cyl and displ, it makes sense that newer models will improve and find ways to increase both the amount of cylinders a vehicle model has as well as more potential power with engine displacement.

As for cyl and displ, it makes sense that more cylinders in a car engine generating more power would create more engine displacement.

When it comes to negative relationships, There are these combinations:

'hwy and cyl' - More cylinders means less miles per gallon of gas used. 'hwy and disp' - More engine displacement means less miles per gallon of gas used. 'year and cty' - Very small negative correlation, not sure about reason, maybe there are more cars in future years, so there is more time spent in the city? 'cyl and cty' - More cylinders using gas means less miles being used per gallon. 'cty and displ' - More engine displacement means less city miles per gallon.

One relationship that surprises me is the relationship between city miles per gallon and the year of manufacture. There is very little difference in the city miles per gallon used, and I cannot think of a distinct reason as to why.