PSTAT131HW#1

Praveen Manimaran

10/2/2022

Question 1: Define supervised and unsupervised learning. What are the difference(s) between them? Supervised learning uses predictor(s) and an associated response to try to predict the response of future observations or try to find/understand the relationship between the predictors and the response. Unsupervised learning describes the situation of having predictor(s) but there is no associated response for the predictors. The main difference between them is that supervised learning uses an associated response for a predictor while unsupervised learning does not.

Question 2: Explain the difference between a regression model and a classification model, specifically in the context of machine learning. Regression model uses y values that are quantitative while classification uses y values that are categorical.

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

2 metrics for regression ML problems are mean square error(MSE) and root mean square error(RMSE) while the 2 metrics for classification ML problems are accuracy and F1-score.

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: Choosing the model to best visually emphasize a trend in the data Inferential models: Try to these theories and to identify significant features and state the relationship between outcome and predictor(s) Predictive models: The aim is to predict Y with minimum reducible error and see which combination of features works best.

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.

Mechanistic assumes a parametric form for f and can add parameters which leads to more flexibility while empirically-driven makes no assumptions about f and requires a larger number of observations and is much more flexible by default. The similarities between the two is they both can be prone to overfitting.

Mechanistic is easier because the models are more likely to be simpler such as linear regression while empirically-driven models can be more complex which is harder to describe such as decision trees that contain several nodes. Mechanistic models usually have high bias and low variance while empirically driven models usually have high variance and low bias. This means that fitting the model for new data for mechanistic models would make the model inaccurate and that fitting the model to new data for empirically-driven models would not really lower the accuracy.

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.

1. The question is predictive because we are trying to use all variables to predict the outcome
2. The question is inferential because we are only changing one variable to simply test how a voter's likelihood of support for the candidate would change.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.6     ✓ dplyr   1.0.8
## ✓ tidyr   1.2.1     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ISLR)
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

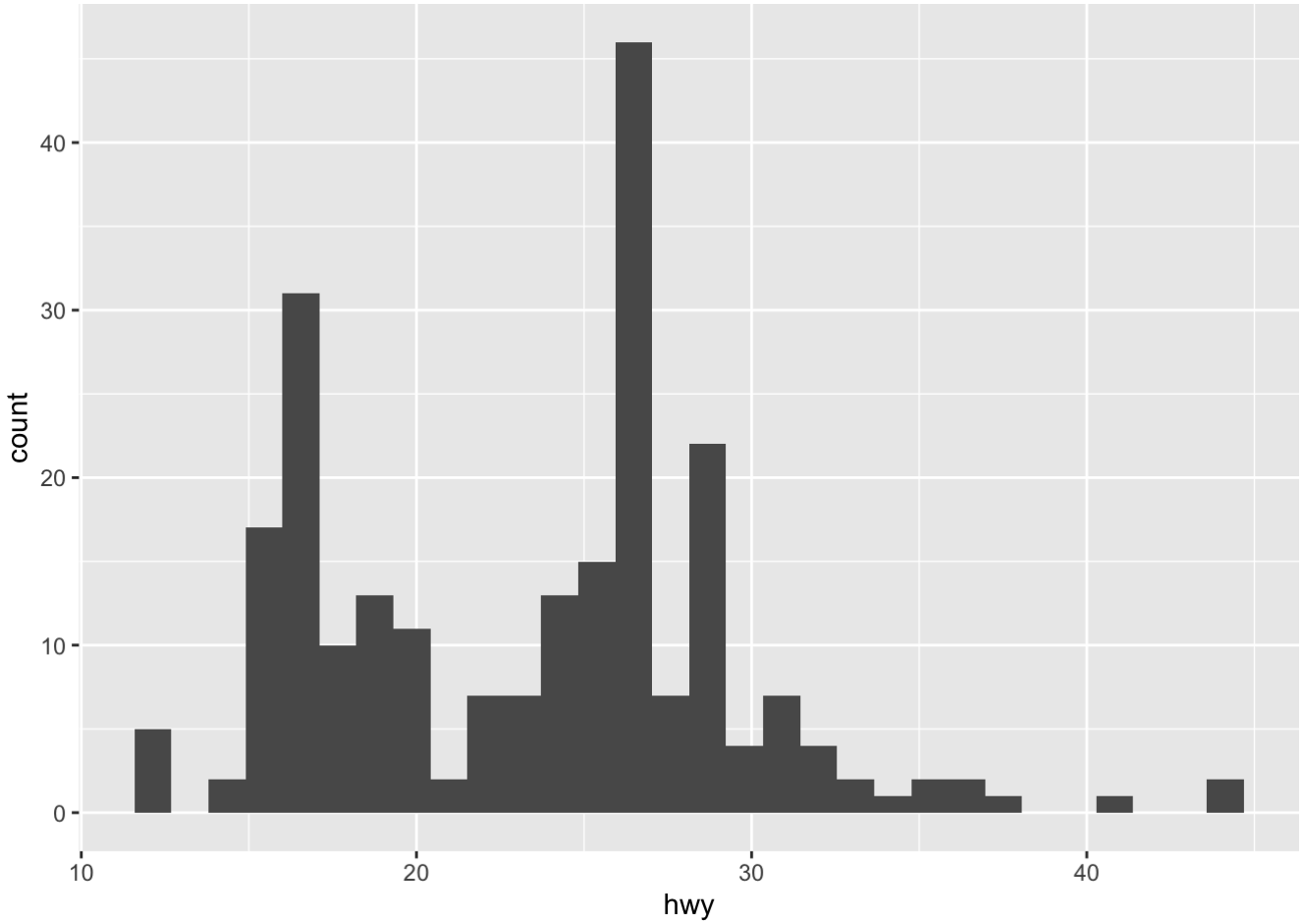
```
data(mpg)
```

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.

```
hwy_hist <- ggplot(mpg, aes(hwy)) + geom_histogram()
hwy_hist
```

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

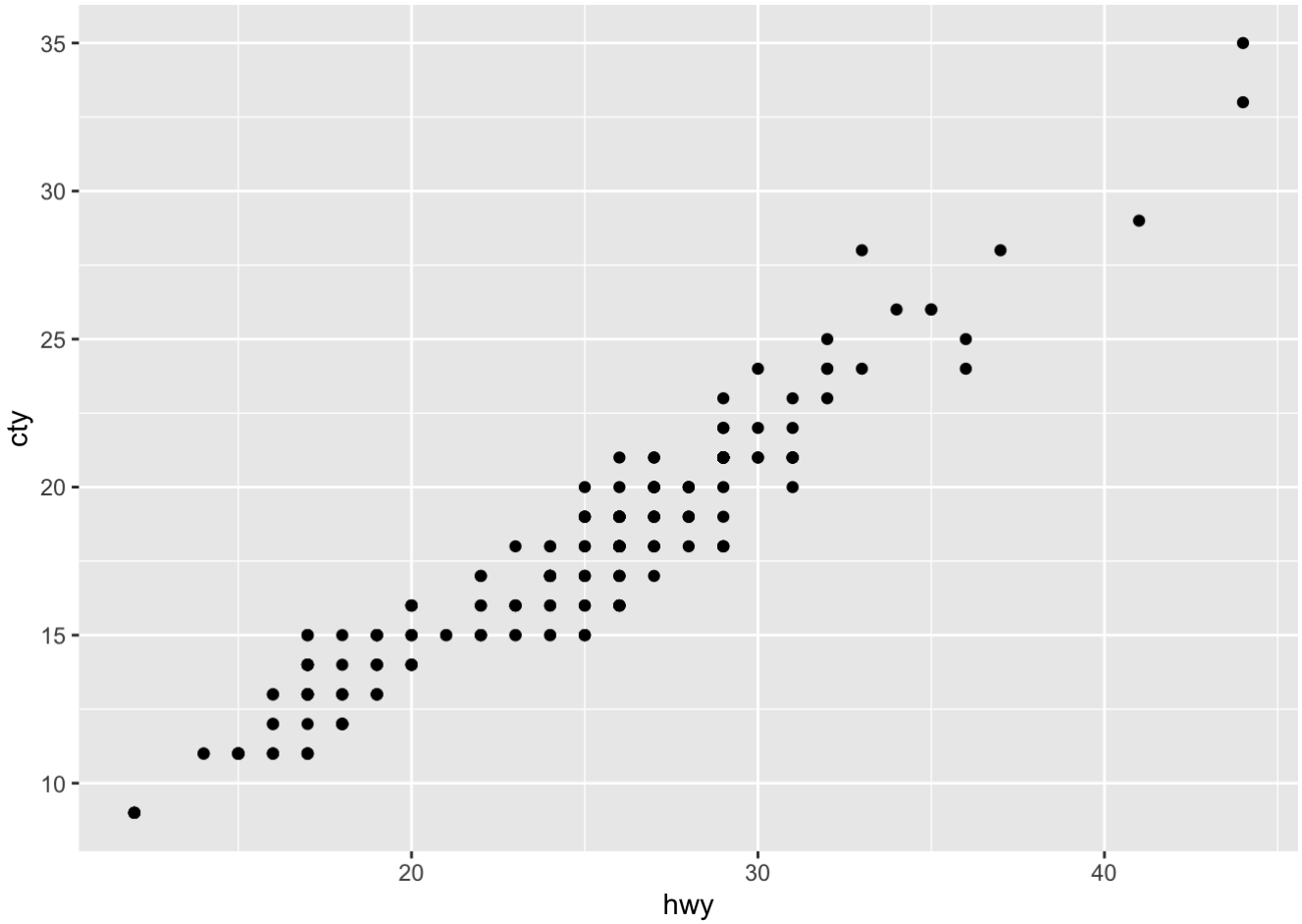


The histogram appears to be skewed to the right and has 2 peaks at mpg=17 and mpg=27 with the x-axis starting at 10 and ending at approximately 46.

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

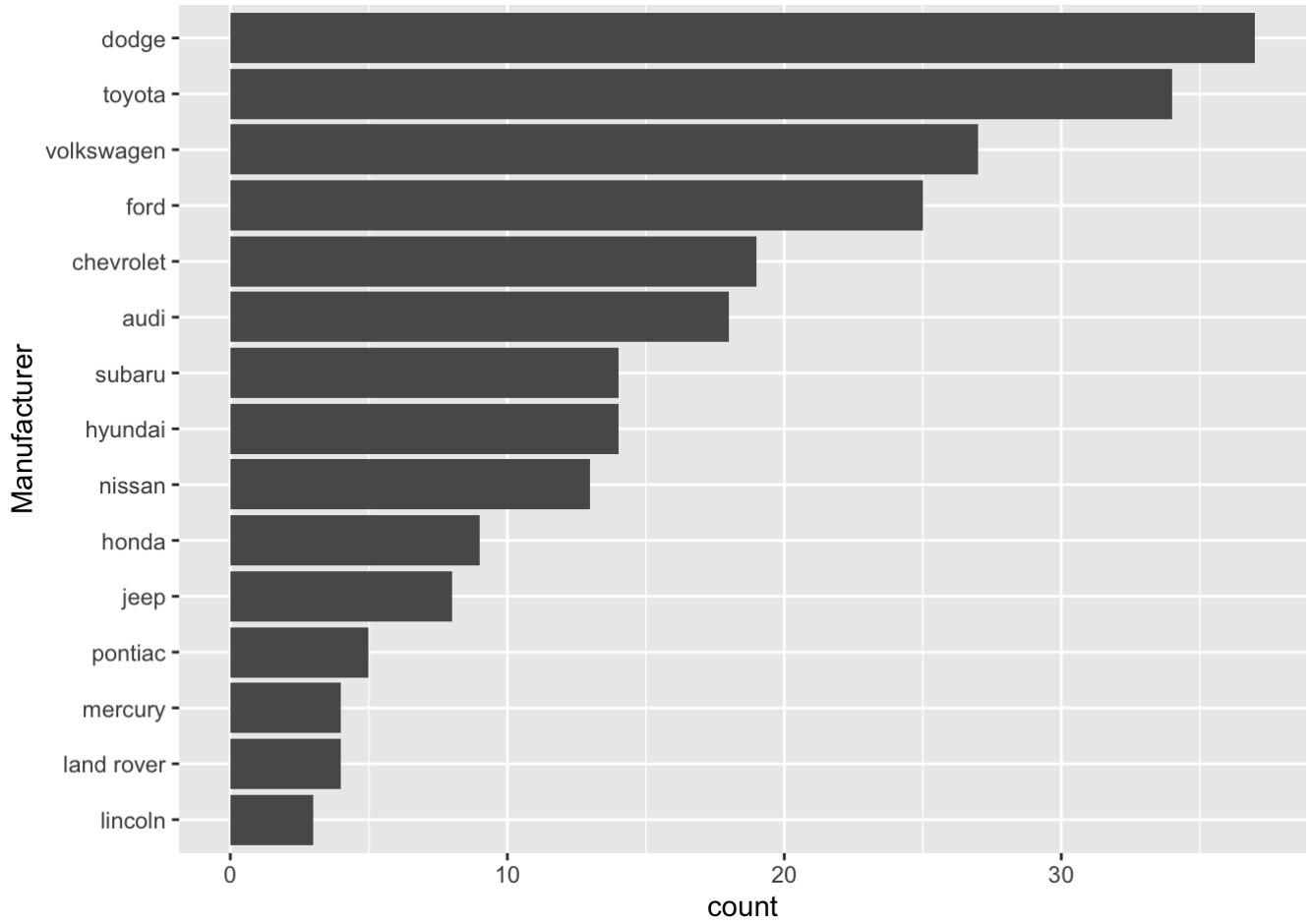


There appears to be a linear relationship between hwy and cty. As cty increases so does hwy which indicates a strong positive linear relationship.

Excercise 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?

```
ggplot(mpg,aes( y = reorder(manufacturer, manufacturer, length) ) ) + geom_bar() + labs
(y="Manufacturer")
```

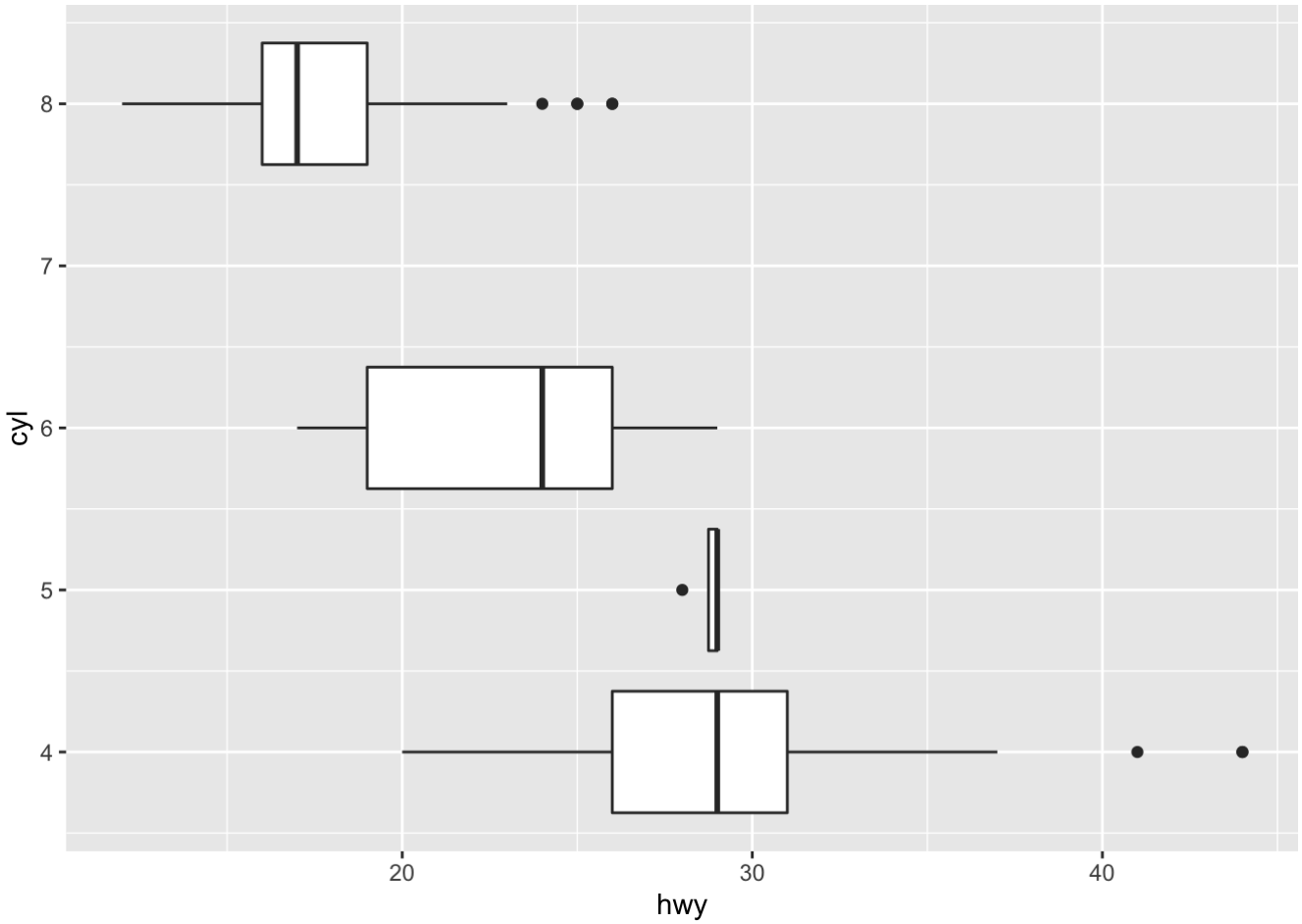


Dodge produced the most cars while Lincoln produced the least cars.

Excercise 4

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

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

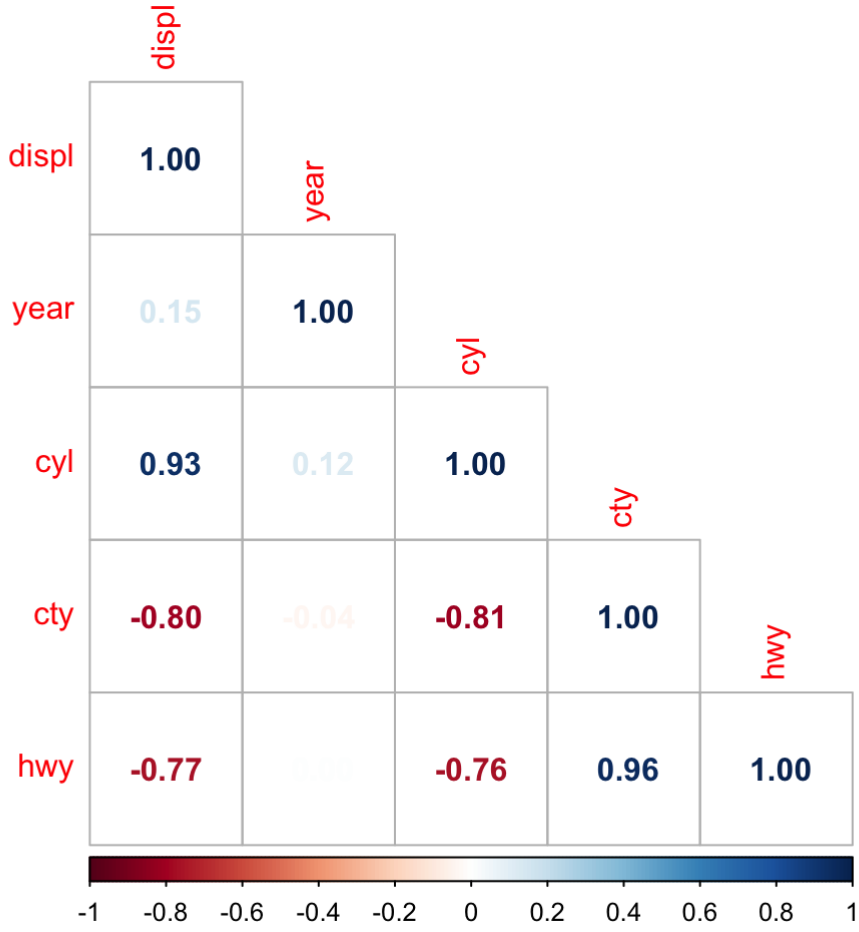


As cyl increases, hwy tends to decrease.

Excercise 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?

```
mpg2 <- mpg[, !names(mpg) %in% c("manufacturer", "model", "trans", "drv", "fl", "class"
)]
corrplot(cor(mpg2), method="number", type="lower")
```



Year appears to have no correlation with the other variables. Cty and hwy have a correlation and this makes sense because a car gets more mpg in a city will also get more on the highway as well. There is a positive correlation between displ and cyl and this makes sense because displ is how much engine space is taken by cyl. There is a negative correlation with cyl and cty, cyl and hwy because more cylinders will cause the car to weigh more which it makes it less efficient in terms oof mpg for both cty and hwy. There is a negative correlation with displ and hwy, displ and cty because displ is the amount of space taken by cylinders so the greater the displ means the greater the cyl which would translate to a lower hwy & cty.