

PSTAT131 HW1

Henry Lin

4/3/2022

```
install.packages("usethis") library(usethis) use_git_config(user.name = "Henry Lin", user.email = "henrylin@ucsb.edu")
```

Question 1-6

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

Supervised and unsupervised learning refers to the type of data set machine learning uses to analyze. For supervised learning, the data sets are known so the machine learning could proceed to use techniques such as “classification and regression models” for the data set. On the other hand, unsupervised learning refers to data sets (such as response variable) that are unknown. Thus, techniques such as clustering and PCA could be conducted on these types of data sets.

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

In the context of machine learning, the differences between regression model and a classification model is that regression model consists of quantitative data while classification model is qualitative.

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

The two commonly used metrics for regression ML problems are mechanistic and parametric. The two commonly used metrics for classification ML problems are empirically-driven and non-parametric.

4. Descriptive models: According to lecture, descriptive model consists of models to best visually emphasize a trend in data.

Inferential models: According to lecture, inferential models consists of models to test theories and causal claims to state the relationship between outcome and predictors.

Predictive models: According to lecture, predictive model consists of models to predict specific variables with minimum reducible error. (“Not focused on hypothesis tests”)

5a. Define mechanistic. Define empirically-driven. How do these model types differ? How are they similar?

Mechanistic technique uses theoretical approach to predict a data-set. On the other hand, empirically-driven technique uses data-sets (generally in larger # of observations) to create a theory on the data set. These model types are similar in the sense that both are used in Machine learning and also fitting the best match for specific data sets.

5b. In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice.

In general, I believe that empirically-driven model would be easier to understand because since the data sets are known and the model is based off of pure data, it becomes straight forward for me to understand the relationship between the fitted model and the data set.

5c. Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models.

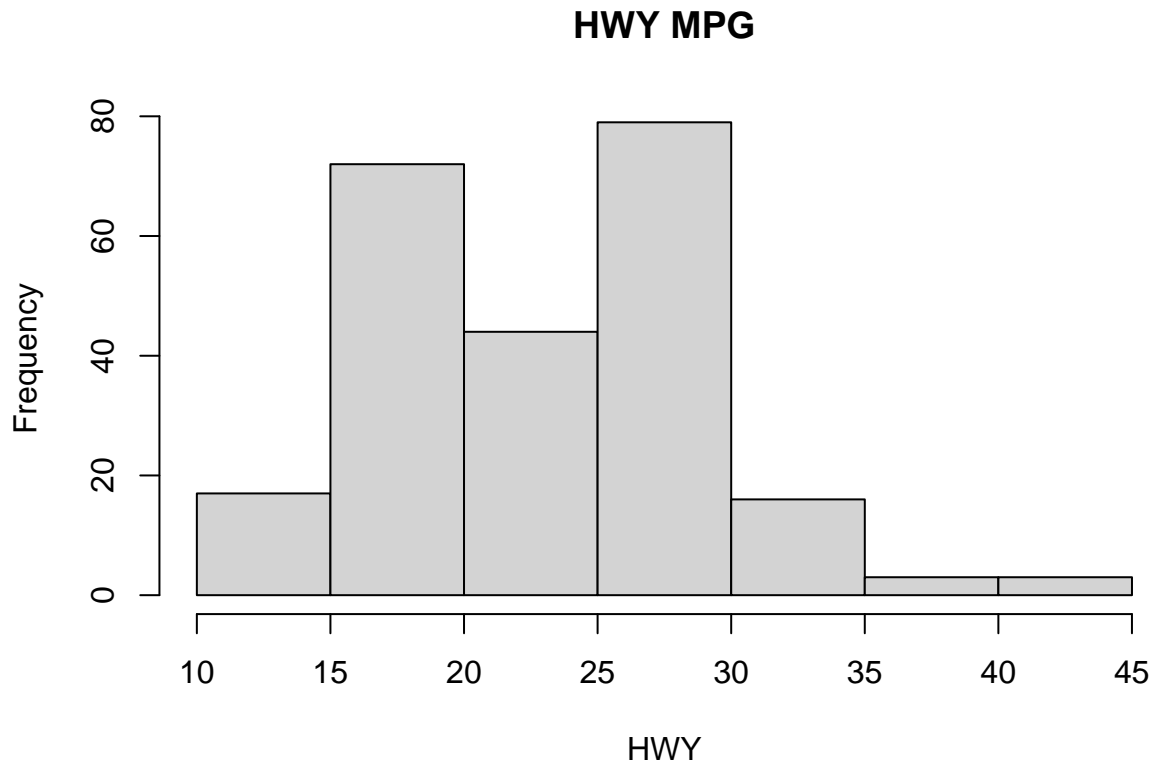
The bias-variance tradeoff is related to the use of mechanistic or empirically-driven models in the sense that both models could reduce the variance of the “estimated parameters by increasing the bias in the estimated parameters.” 6. For the first part of this question, I believe that it is a predictive approach because the voter’s profile/data was used to predict the likelihood of him/her voting in favor of the candidate. On the other hand, I believe that it is a inferential approach because predicting a person’s vote based off of their interaction with the candidate is not based off of solid data.

Exploratory Data Analysis

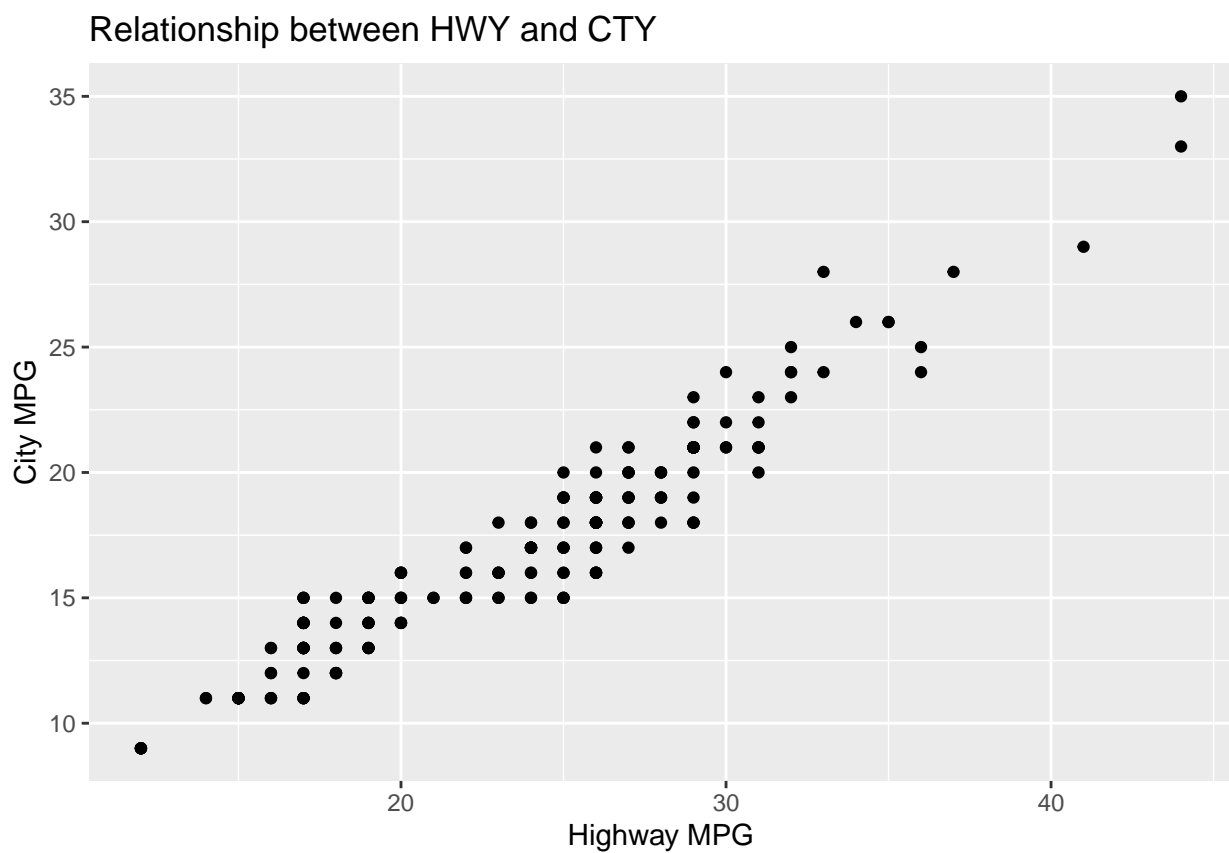
```
library(tidyverse) library(ggplot2) library(tidymodels) library(ISLR)
```

Exercise 1:

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

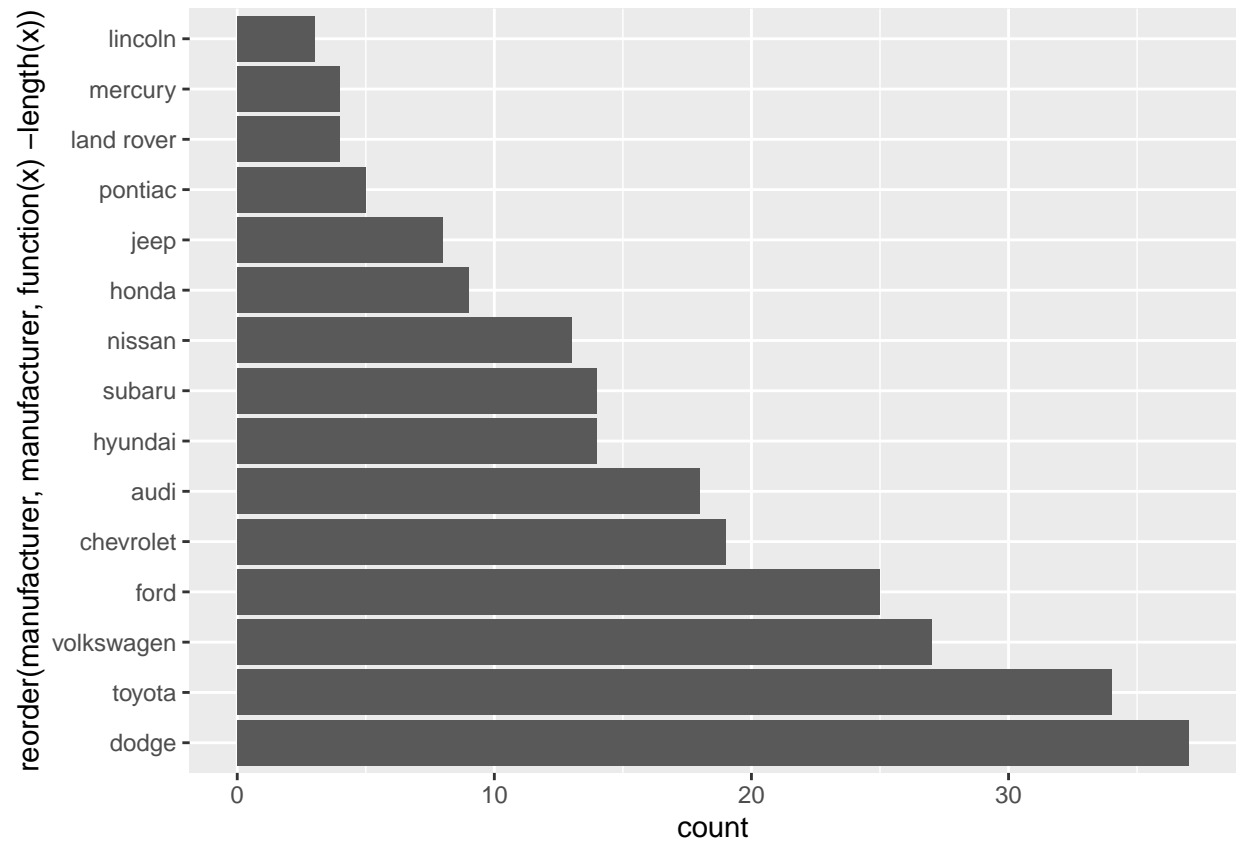


From the histogram, I see that the highest two frequencies occurs in the 15-20 MPG and 25-30 MPG range. On the other hand, the lowest frequencies occurs in the 35-45 MPG range.



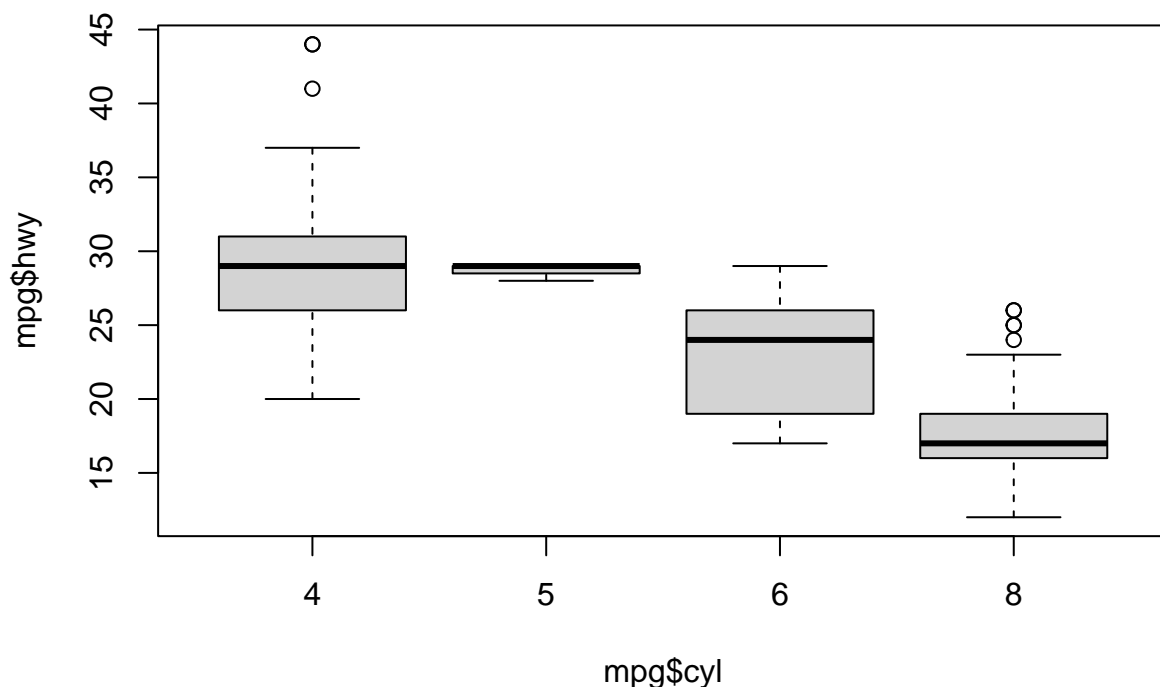
Exercise 2:

From this scatterplot, I noticed a clear linear relationship between hwy and cty. This means that an increase of City MPG would also mean an increase to the Highway MPG.



Exercise 3:

From this bar plot, Lincoln produced the least amount of cars while Dodge produced the most.



Exercise 4:

From this box plot, the plots shows a negative relationship between HWY MPG and CYL MPG. From the plots, we may assume that an increase of CYL MPG will suggests an decrease of HWY MPG. Thus, vice versa.

Exercise 5:

```
## Warning: package 'corrplot' was built under R version 4.1.3
```

```
## corrplot 0.92 loaded
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
## [1,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2,] TRUE  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [3,] TRUE  TRUE  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [4,] TRUE  TRUE  TRUE  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [5,] TRUE  TRUE  TRUE  TRUE  FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [6,] TRUE  TRUE  TRUE  TRUE  TRUE  FALSE FALSE FALSE FALSE FALSE FALSE
## [7,] TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  FALSE FALSE FALSE FALSE FALSE
## [8,] TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  FALSE FALSE FALSE FALSE
## [9,] TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  FALSE FALSE FALSE
## [10,] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE FALSE
## [11,] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE
## [12,] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [13,] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [14,] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15,] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [16,] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

[illegible]

[illegible]

[illegible]

[illegible]

```
## [233,]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [234,]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

From this data, majority of the elements in this data set seems to be correlated. I am not quite sure whether the relationship makes sense to me or not because I'm not sure if I am doing this problem right (Will contact TA). Therefore, I do not have anything that surprises me from this exercise.