HW1

Hanying Feng

9/27/2022

Question 1: In supervised learning, there is a supervisor when the machine learns, which means there is an answer key for each observation. However, there is no supervisor in unsupervised learning. The model is trying to find rules by itself.

Question 2: In the regression model, the outcomes (response variables) are numerical values. In the classification model, the outcomes are categorical values.

Question 3: Regression model: Mean squared error (MSE), Root mean squared error (RMSE), Mean absolute error (MAE) Classification model: Accuracy, Confusion matrix, F1 score

Question 4: Descriptive models aim to best visualize the trend in data. Inferential models aim to find the significant features and test theories. They also aim to find the relationship between predictors and response variables. Predictive models aim to predict the response variable with minimum reducible error. They do not focus on hypothesis test.

Question 5: A mechanistic model has an assumption of functional form and it is trained based on the f, while empirically-driven model does not have a certain f. They can both be used for prediction. Mechanistic model may become more flexible when we add parameters. Empirically-driven model is always more flexible. Mechanistic model always assumes a parametric form for f, but empirically-driven model doesn't assume f. Empirically-driven model also requires larger data sets.

I think mechanistic model is easier to understand, because we always know what f is and we just need to choose parameters. It might be harder to think what an empirically-driven model should be like.

The mechanistic model may have higher bias but lower variance, while empirically-driven model may have lower bias and higher variance.

Question 6: The first one is predictive, because the model will use information to predict if the voter will vote certain candidates. The second one is inferential, because the model will try to analyze the significance of the feature "if voter has personal contact with the candidate".

```
library(tidyverse)
```

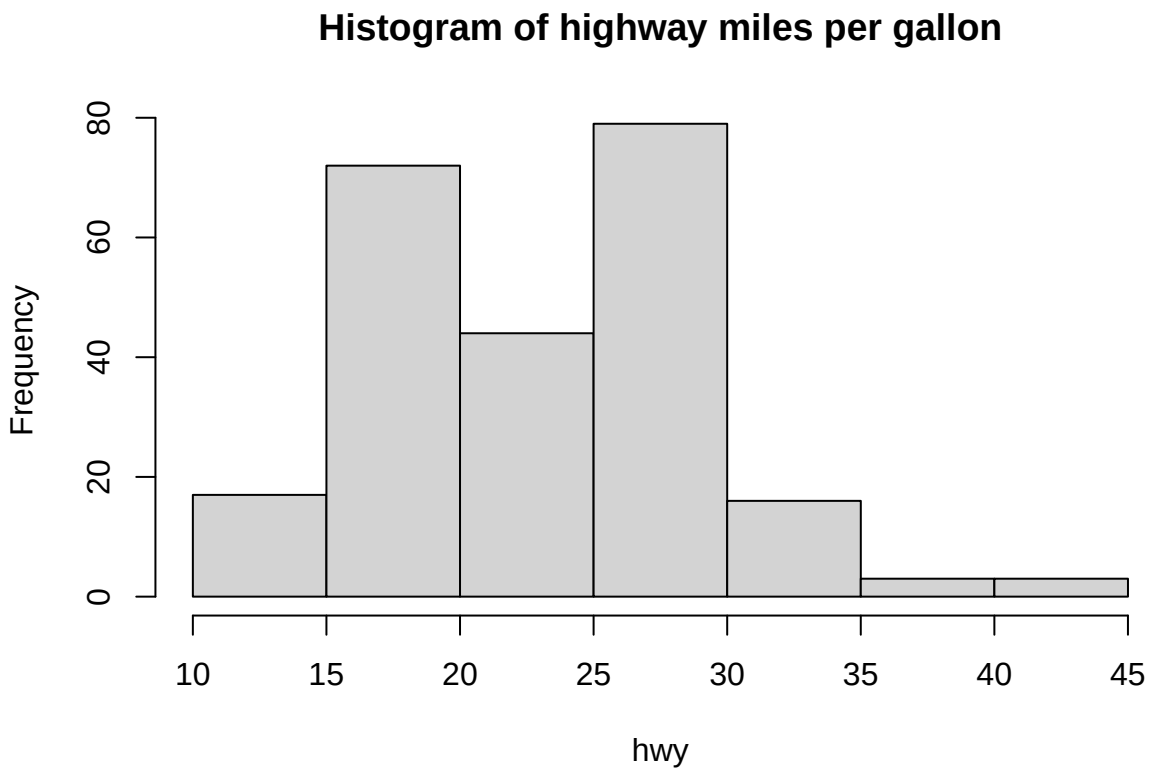
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

```
## corrplot 0.92 loaded
```

Exercise 1:

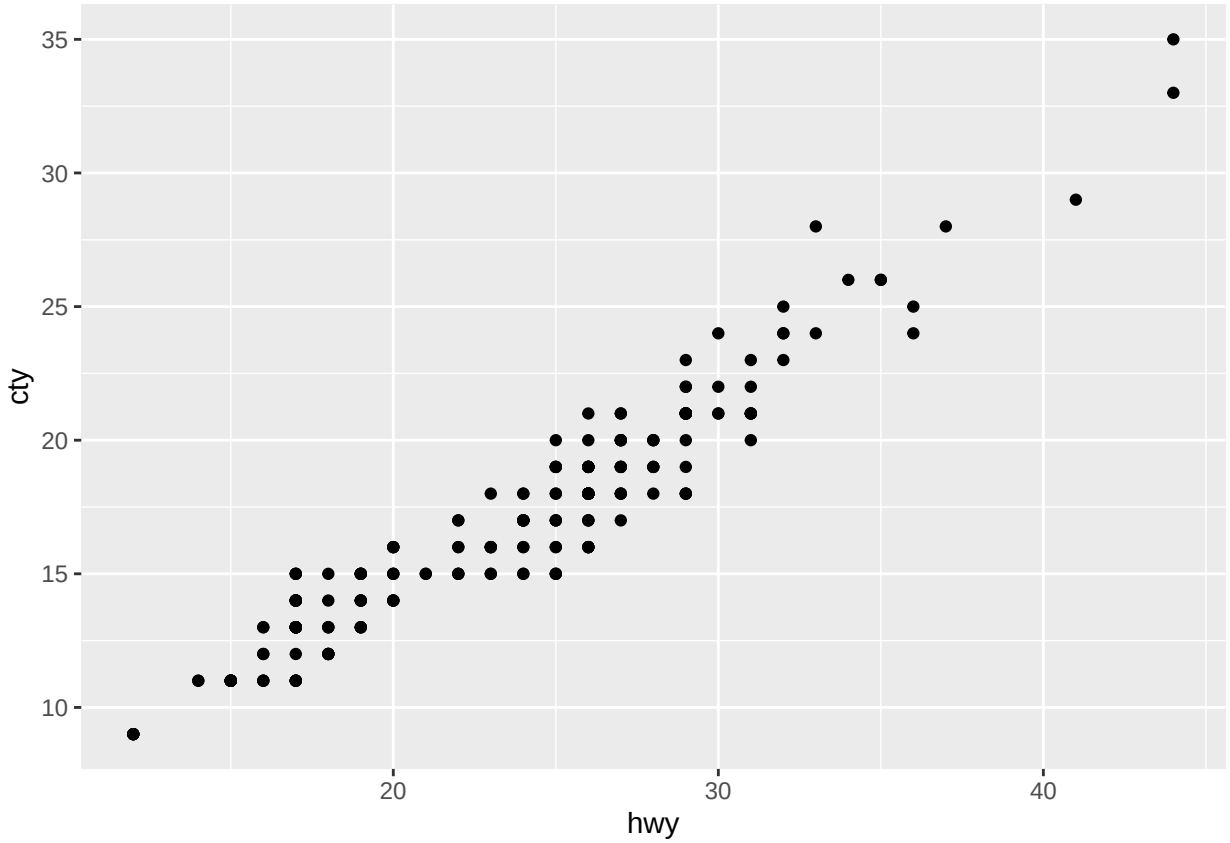
```
hist(mpg$hwy, xlab = "hwy", main = "Histogram of highway miles per gallon")
```



We can see that most of highway miles per gallon are in the range of [15, 30]. The range of all highway miles per gallon is [10, 45].

Exercise 2:

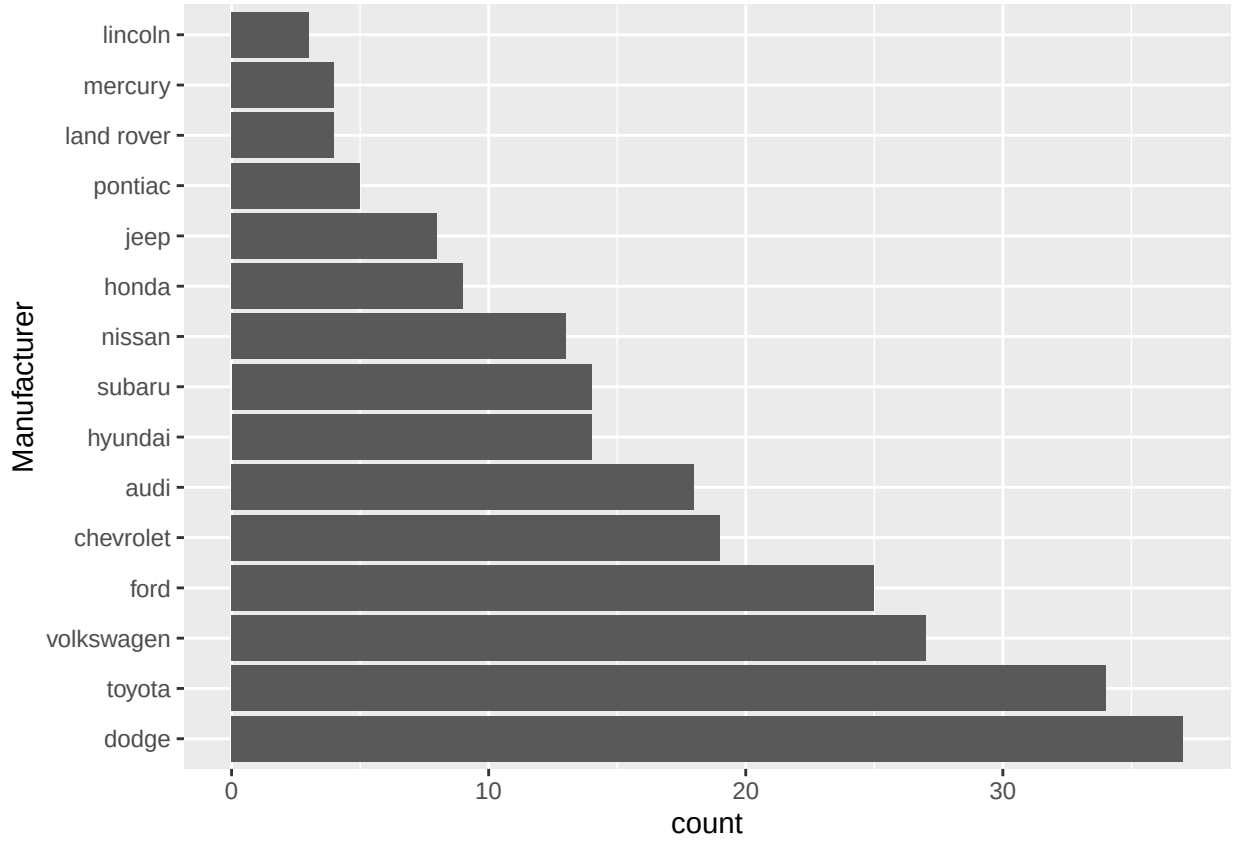
```
ggplot(mpg, aes(x=hwy, y=cty))+geom_point()+labs(xlab = 'hwy', ylab = 'cty')
```



Yes, there is a positive relationship between hwy and cty. cty increases as hwy increases. Based on the plot, it seems like there is a linear trend between cty and hwy.

Exercise 3:

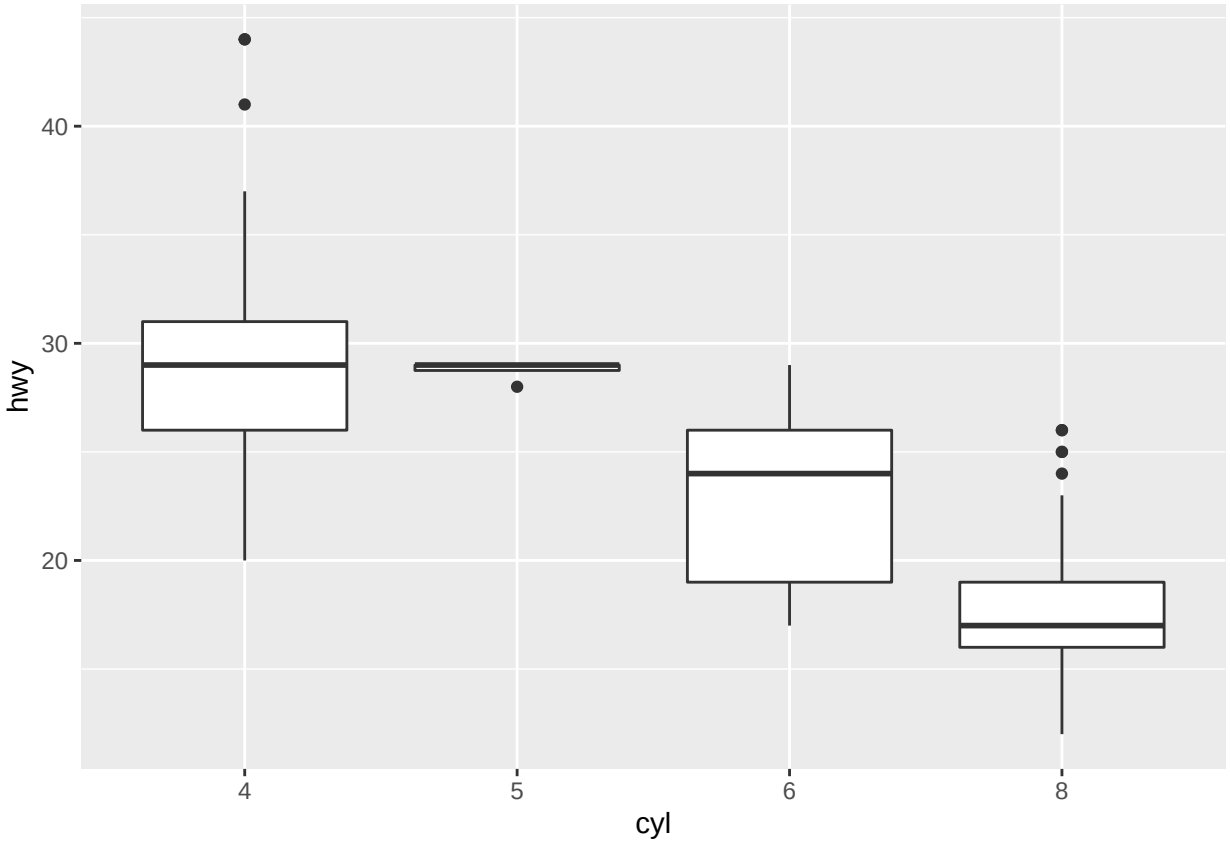
```
mpg %>%
  ggplot(aes(x=fct_infreq(manufacturer)))+geom_bar()+labs(x="Manufacturer")+coord_flip()
```



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

Exercise 4:

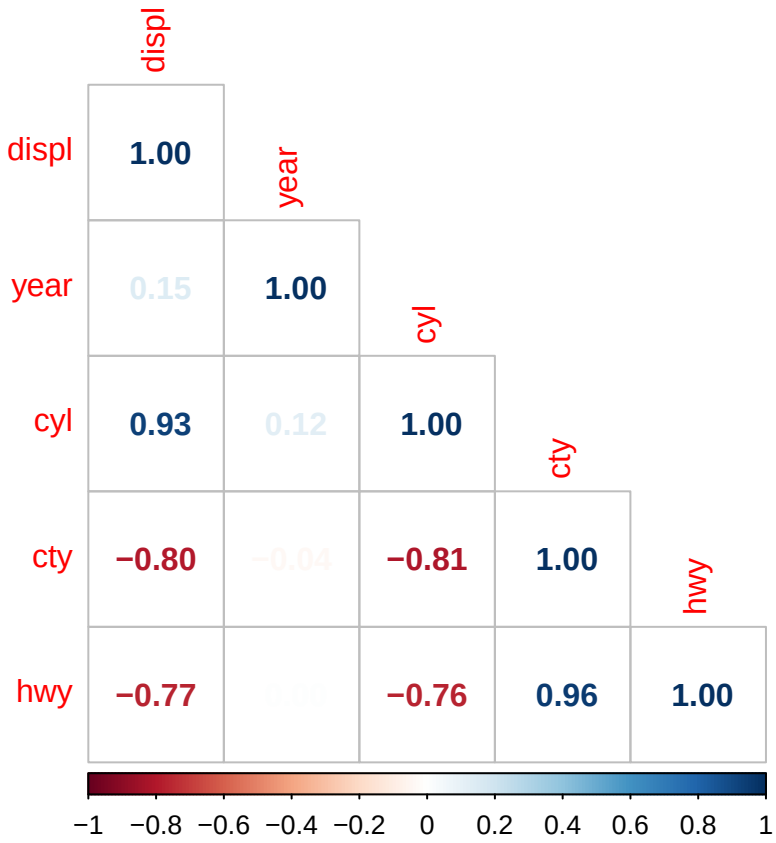
```
mpg %>%
  ggplot(aes(x=hwy, y=as.character(cyl)))+geom_boxplot()+labs(y="cyl")+coord_flip()
```



The pattern is that hwy decreases as cyl increases.

Exercise 5:

```
data = cor(mpg[, c(3, 4, 5, 8, 9)])
corrplot(data, method="number", type='lower')
```



Positively correlated: year and displ, cyl and displ, cyl and year, hwy and cty Negatively correlated: cty and displ, cty and year, cty and cyl, hwy and displ, hwy and year, hwy and cyl

cyl anf displ have a strong positive relationship. This makes sense to me because when the number of cylinders increases, the engine displacement should also increase. hwy and cty also have a strong positive relationship because higher highway mileage implies higher city mileage.

cyl and hwy have negative relationships with displ and cyl respectively. This is because for cars with higher engine displacement or higher number of cylinders can drive less distance as they consume more oil.