HW1

Roy Zhang

2022-10-02

MLMI:

1. A supervised model has a response variable that the explanatory variables can refer to, while an unsupervised model does not.

   Source: textbook p.26

2. A regression model has its outcome on an array of numerical values, whereas a classification model only has a fixed number of outcomes in different categories. Therefore, the response of a regression model is quantitative and the response of a classification model is qualitative.

   Source: textbook p.28.

3. - Regression: mean squared error, residual sum squared.
   - Classification: error rate, Bayes classifier.

   Source: lecture slides, textbook p.37.

4. - Descriptive models: serve to visually display a trend in data.
   - Inferential models: serve to find the variables associated with the outcome and the kind of association between them.
   - Predictive models: serve to accurately predict the outcome with minimum reducible error.

   Source: lecture slides, textbook pp.17-20.

5. - A mechanistic model assumes a particular model fits well into the data, and uses the data to estimate the parameters of the model. An empirically-driven model does not make such an assumption and they try to directly find an estimate of the model, which will fit well into the data. The most significant difference between the two kinds of model is the assumption process, which makes an empirically-driven model more flexible but more demanding in terms of the quantity of the data. They are similar that both are vulnerable to overfitting, as the mechanistic model might contain too many parameters and the empirically-driven model might generate an estimate too specific to the training data.
   - In general, a mechanistic model is easier to understand because it is built upon an assumed model chosen by the constructor. Therefore, the constructor is more likely to have a clearer knowledge of the structure of the model and the function of each parameter.
   - The bias-variance tradeoff serve to evaluate the complexity of the mechanistic or empirically-driven model to ensure that they are still rather accurate in fitting into the data but do not suffer too much from overfitting due to having too many parameters.

   Source: lecture slides, textbook pp.21-24, 33-36

6. - This question is predictive because we are trying to predict the outcome (the likelihood of voting in favor of the candidate) using a series of data variables (a voter's profile/data).

- This question is inferential because we are trying to whether there exists an association between a particular variable (personal contact with the candidate) and the outcome and the kind of association in place if it exists.

EDA:

```
library(tidyverse)
```

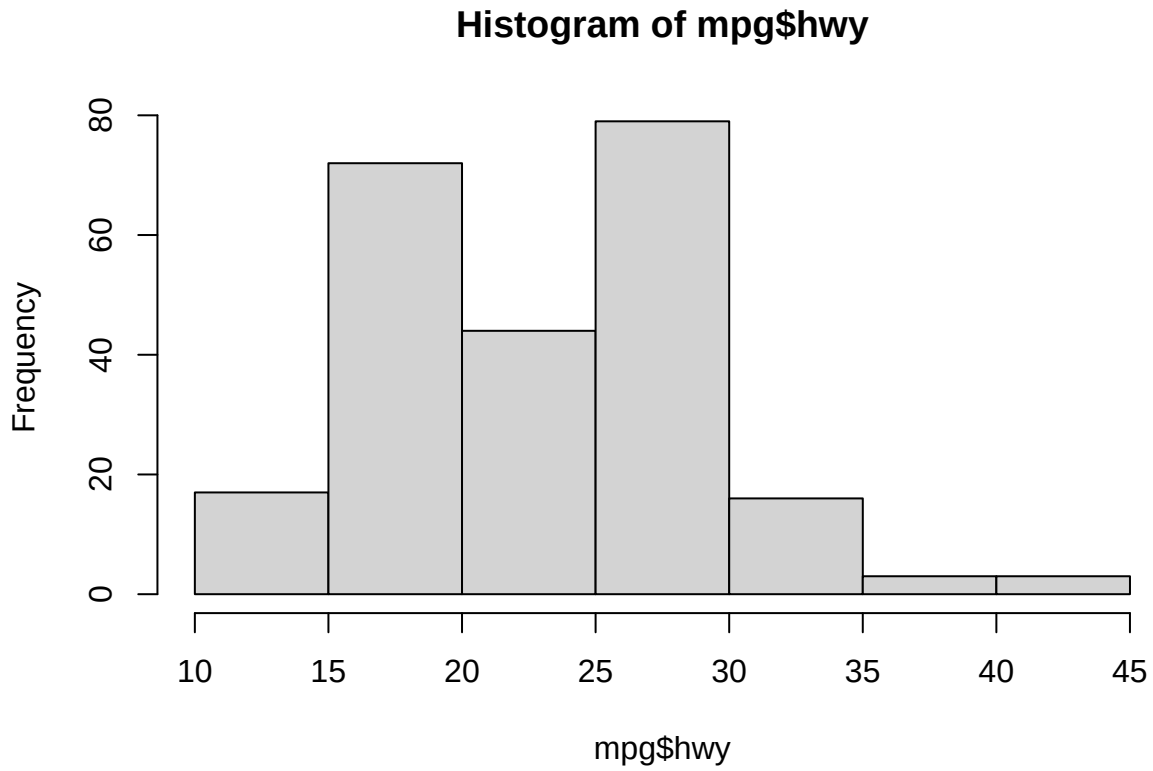
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------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

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

```
## corrplot 0.92 loaded
```

1.

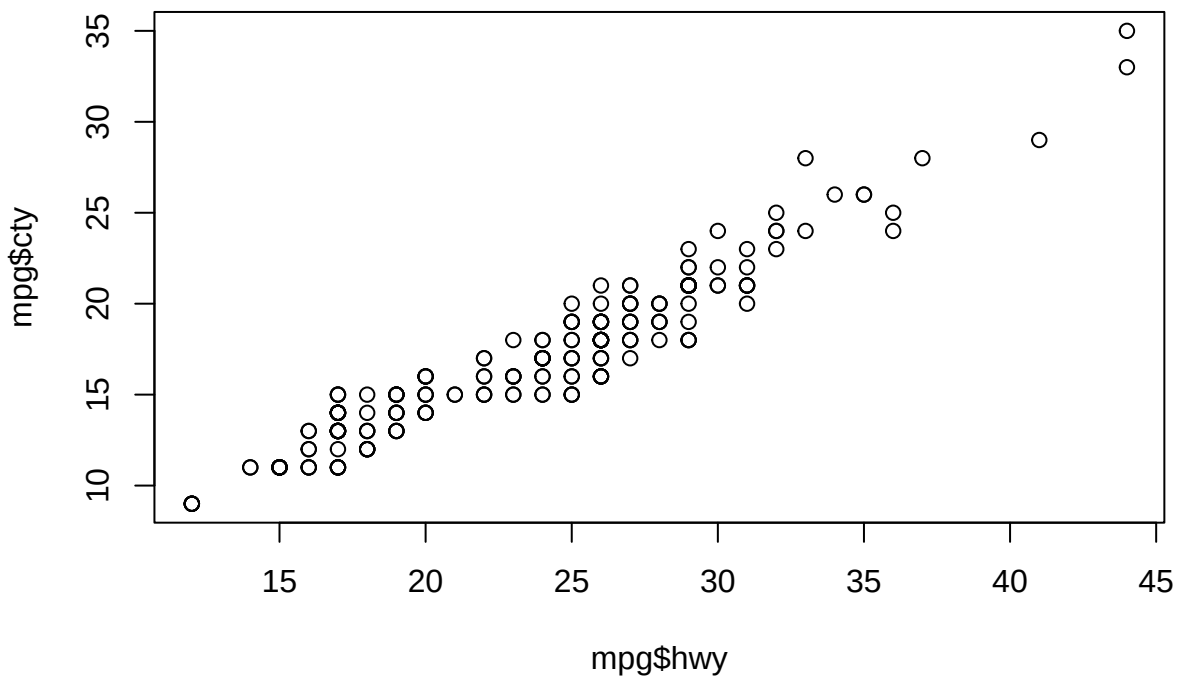
```
hist(mpg$hwy)
```



Based on the historgram, the highway miles per gallon is skewed to the right. Most of the mpg frequency land at 15-20 and 25-30 mpg.

2.

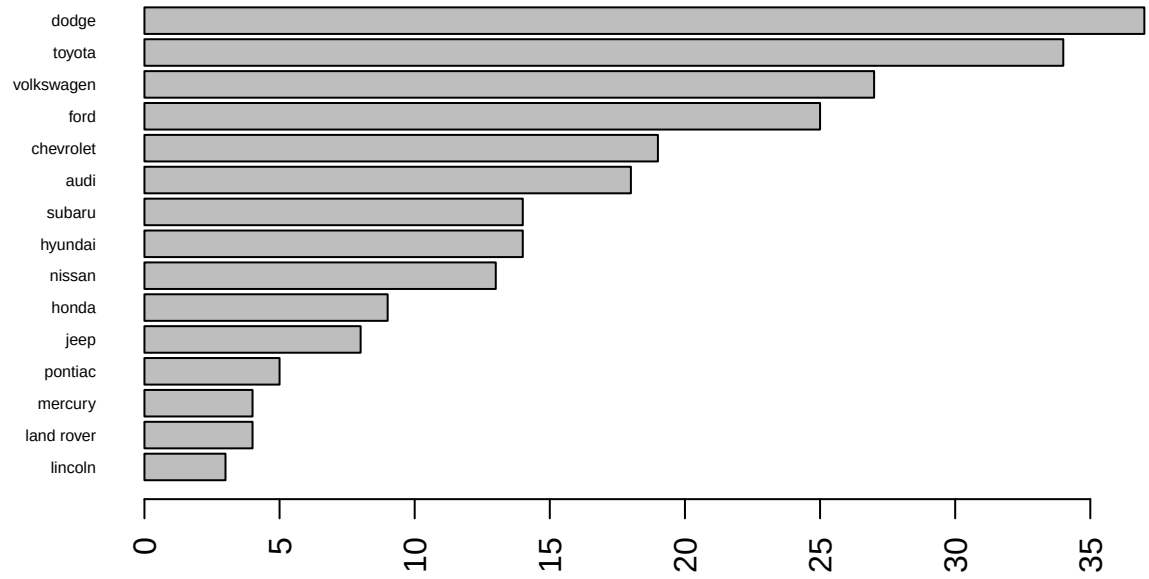
```
plot(mpg$hwy, mpg$cty)
```



According to the scatterplot, `hwy` and `cty` show a upward sloped relationship. This means highway mpg and city mpg are likely to be positively correlated.

3.

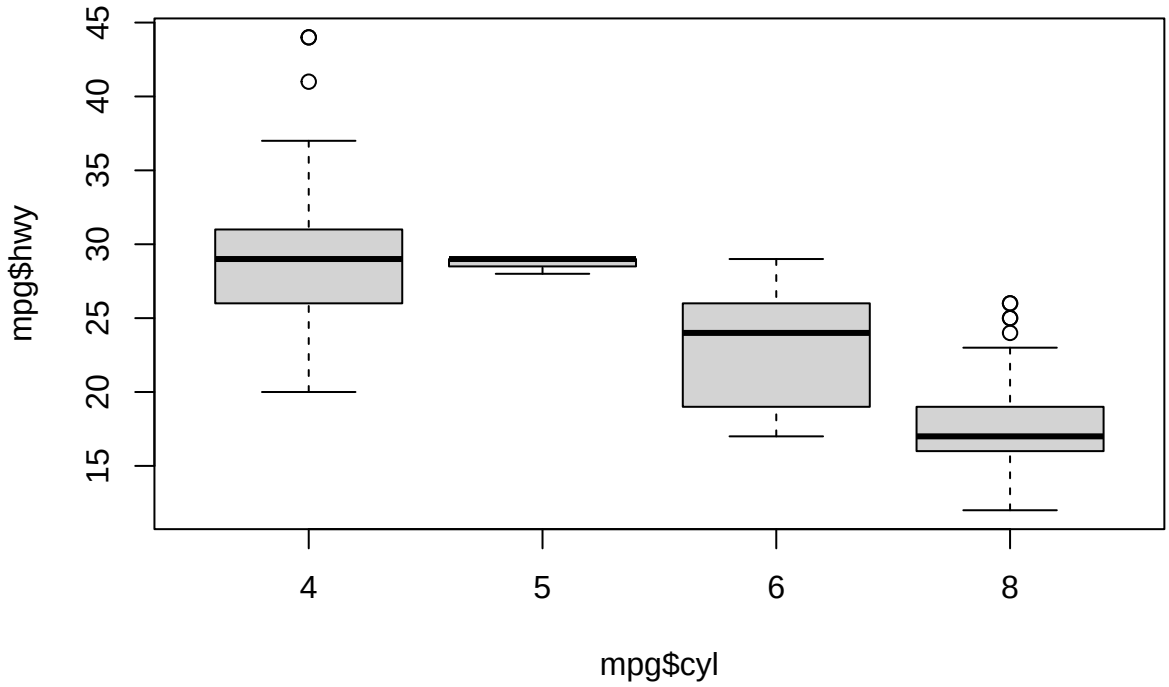
```
barplot(sort(table(mpg$manufacturer)), horiz = TRUE, las = 2, cex.names = 0.5)
```



According to the barplot, Dodge produced the most cars, and Lincoln produced the least.

4.

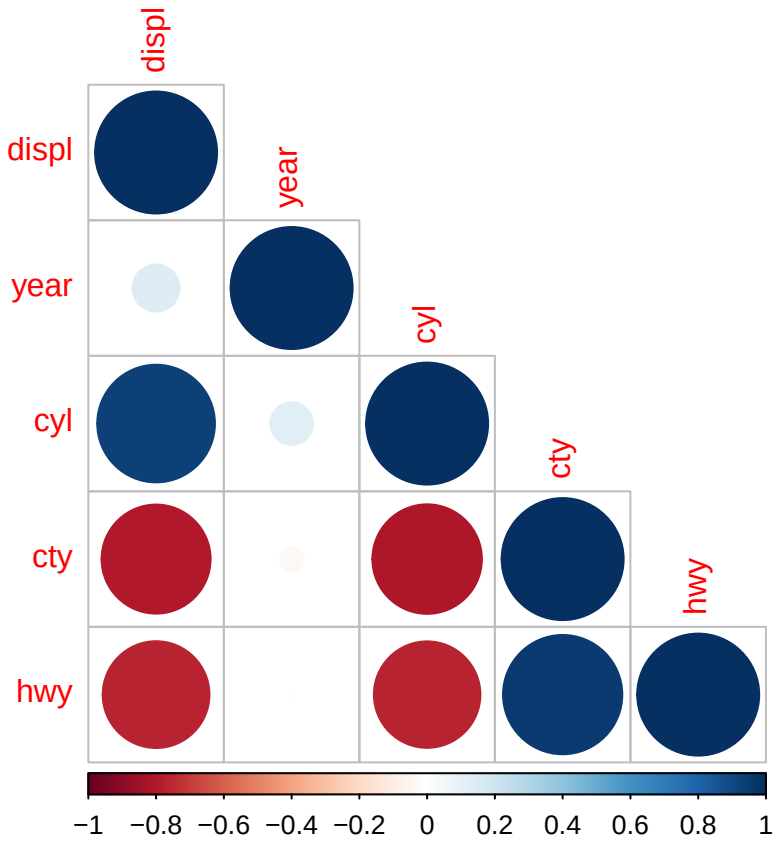
```
boxplot(mpg$hwy ~ mpg$cyl)
```



The box plots show an overall downward trend between `cyl` and `hwy`. There is little difference between cars with 4 and 5 cylinders, and for cars with 6 cylinders there starts to be a certain degree of drop. Cars with 8 cylinders have a significantly smaller amount of highway mpg compared to others.

5.

```
corrplot(cor(select(mpg, where(is.numeric))), type = "lower")
```



`displ` is positively correlated to `cyl`, `cty` is positively correlated to `hwy`, and each of the two pairs is negatively correlated to both variables of the other pair. `year` is uncorrelated to any variables. These correlations do not surprise me as engine displacement should go up with the number of cylinders, and it will result in lower mpg. However, I am somewhat surprised by the lack of correlation to `year`, as I thought 10 years would have brought some kind of improvement to the cars' performances.