Pstat131 Hw1

erasmo-rivas

2022-09-30

Question 1:

Supervised learning is the application of statistical lerning methods to data, in order to extract useful patterns within the data. Unsupervised learning is the application of statistical learning methods to data, in order to identify clusters in the data. The difference, is that in supervised learning we have some response variable to help train our models, in unsupervised learning we do not have this supervisor.

Question 2:

A regression model would be used when our response is continuous and classification model would be used when our response is discrete (categorical).

Question 3:

Two common metric for regression machine learning problems are training mean squared error (training MSE) and test mean squared error (test MSE). Two common metrics for classification machine learning methods are training error rate and test error rate.

Question 4:

Each of the following models prioritizes a goal for the statistical model.

Descriptive Models: This model is to visually emphasize a trend in the data.

Inferential Models: This model seeks to investigate relationships between outcome and predictors, to test theories, and discover potential causal relationships.

Predictive Models: This model aims to predict Y, that is reduce test MSE. This method will look for the best combination of predictors to predict Y.

Question 5:

In a mechanistic model assumptions are made of the form of f, the relationship between the response and the predictors. In empirically driven no such assumption are made about the form of f. Mechanistic models are less flexible than empirically driven models. Mechanistic models tend to be more interpretable, the functional form that is assumed generally comes with easier interpretation. Empirically driven models (being more flexible) tend to fit data better – this causes bias to be low, however the variance can be high. With Mechanistic models, bias is generally higher however variance is low.

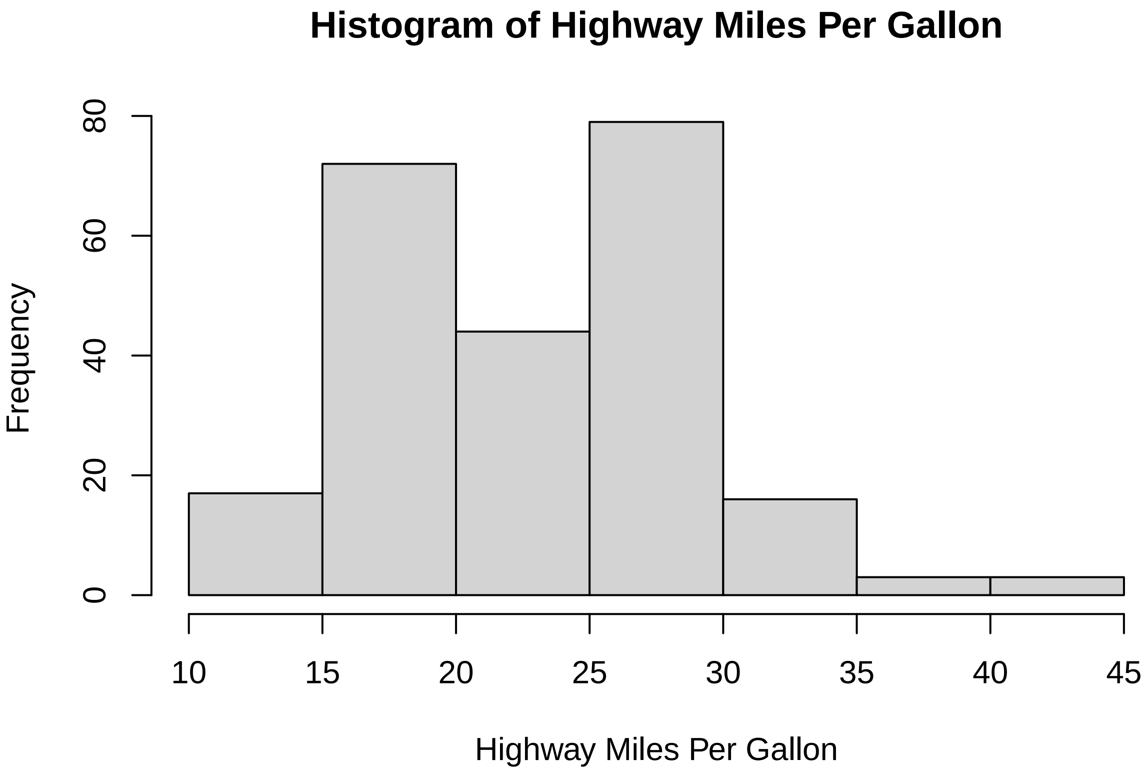
Question 6:

the first question would be predictive, that is from the predictors (voter's profile) we want to predict the likelihood they vote in favor of a given candidate.

The second question is inferential, they seek to find the relationship of the response (voter's likelihood f support od a candidate) to a predictor (personal contact with candidate).

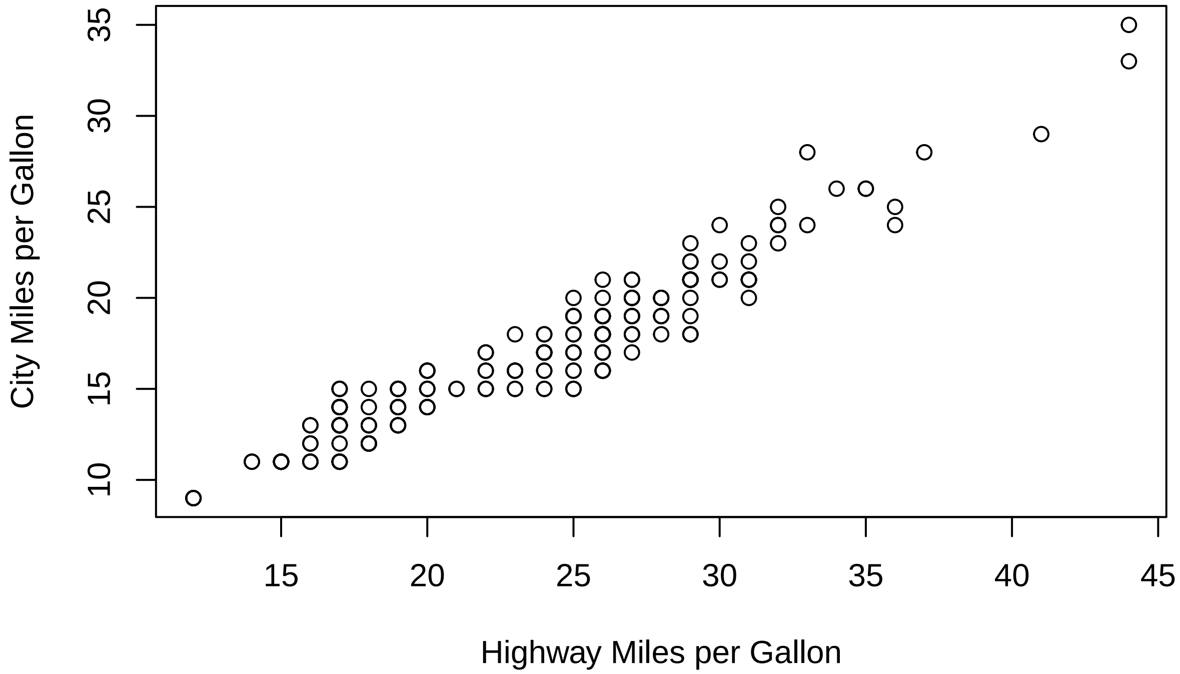
```
## -- Attaching packages -------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1      v recipes      1.0.1
## v dials        1.0.0      v rsample      1.1.0
## v dplyr        1.0.10     v tibble       3.1.8
## v ggplot2      3.3.6      v tidyr        1.2.1
## v infer        1.0.3      v tune         1.0.0
## v modeldata    1.0.1      v workflows    1.1.0
## v parsnip      1.0.2      v workflowsets 1.0.0
## v purrr        0.3.4      v yardstick    1.1.0

## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x purrr::discard() masks scales::discard()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x recipes::step()  masks stats::step()
## * Use suppressPackageStartupMessages() to eliminate package startup messages

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v readr   2.1.3     v forcats 0.5.2
## v stringr 1.4.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard()    masks scales::discard()
## x dplyr::filter()     masks stats::filter()
## x stringr::fixed()    masks recipes::fixed()
## x dplyr::lag()        masks stats::lag()
## x readr::spec()       masks yardstick::spec()
## corrplot 0.92 loaded
```

Exercise 1:



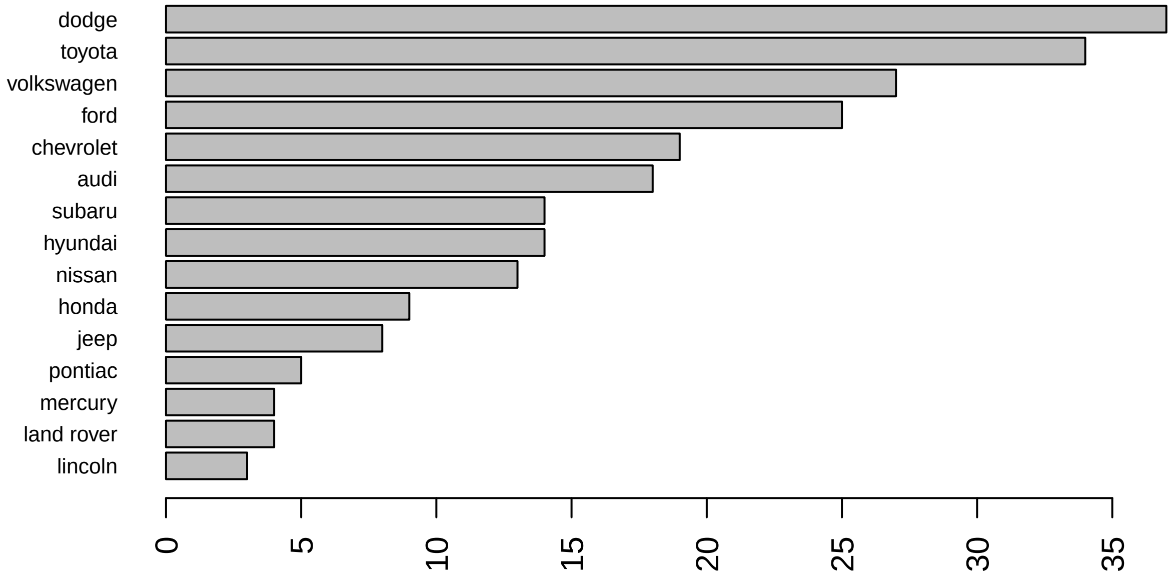
The majority of the data points fall within a 15-30 miles per gallon.

Exercise 2:



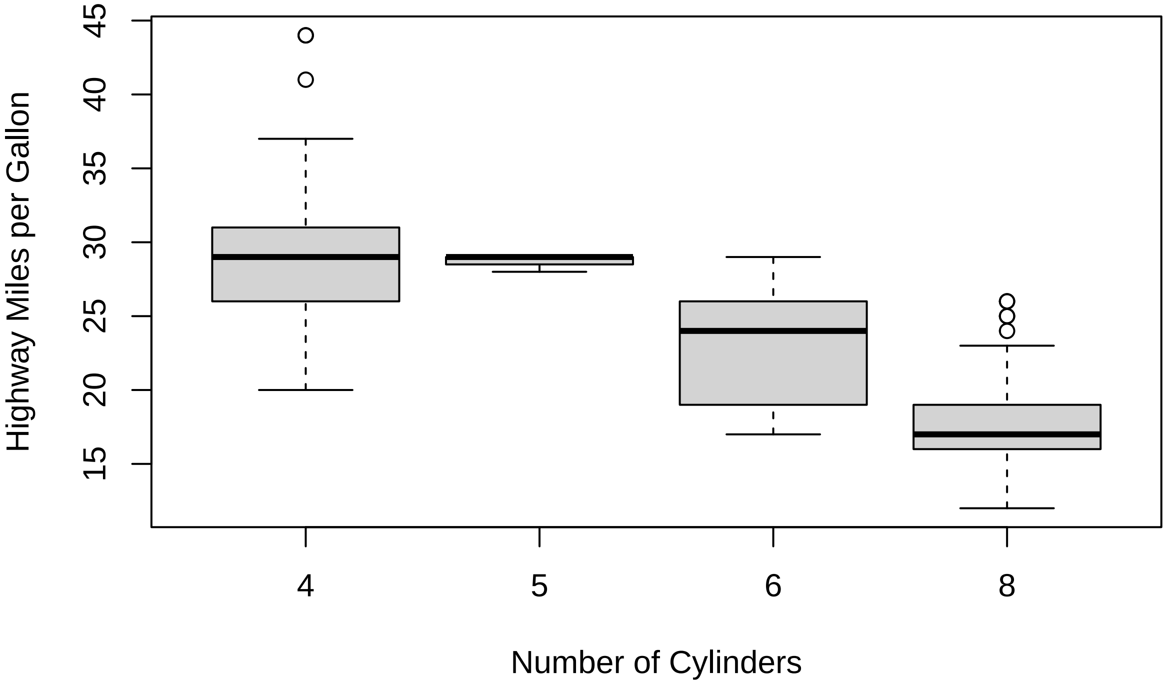
There seems to be some positive relationship, as highway miles per gallon increase city miles per gallon increases.

Exercise 3:



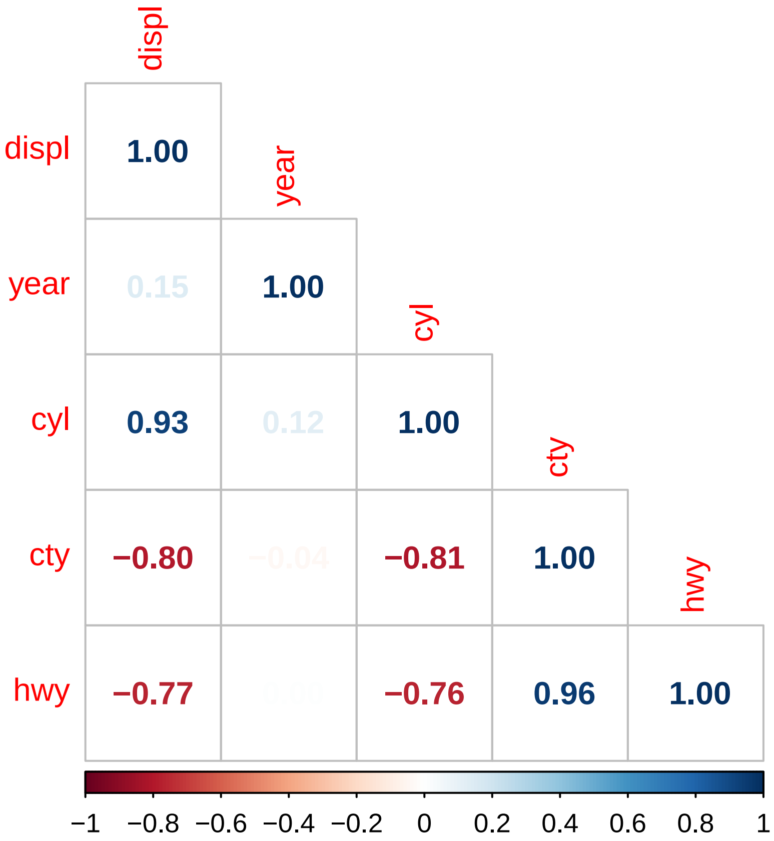
Dodge produced the most cars. Lincoln produced the least amount of cars.

Exercise 4:



As the number of cylinders increases the Highway miles per gallon decreases – the average highway miles per gallon decreases.

Exercise 5:



There exists a positive correlation between: highway miles per gallon and city miles per gallon, number of cylinders and engine displacement. There exists a negative correlation between: city miles per gallon and engine displacement, city miles per gallon and number of cyliners, highway miles per gallon and engine displacement, highway miles per gallon and number of cylinders.