Homework #1 Pstat 131

Branson Enani

2022-09-27

Question 1

Supervised learning is when we wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference). (Textbook, pg. 26).

Unsupervised learning differs in the sense that there is a vector of measurement (X_i), but there isn't an associated response variable. This means that linear regression doesn't apply to this type of learning.

Question 2

In the context of machine learning, the general rule is that regression models tend to be representations of quantitative data (with some exceptions such as the logistic regression model) and classification models represent qualitative data that is more categorical as oppposed to numerical.

Question 3

When it comes to regression problems, two commonly used metrics to assess the accuracy of the model would be the Training MSE (mean squared error) as well as the Test MSE. By using these two metrics, comparisons can be ade about how well the ML model is doing. For classification models the metrics are similar however they are called the Training Error Rate and the Testing error Rate. These two assess incorrect classifications withing the model.

Question 4 -A descriptive model tends to use visual representations like a scatterplots with an emphasis on the visual aspect.

-A predictive model aims to find the features of a model that can predict Y with the minimum error (most accuracy). Hypthesis tests are not as important for predictive models.

-Inferential models tend to want to test theories and find a relationship between the outcome and the predictors. It aims to find which features are significant in this relationship.

Question 5

-Mechanistic models generally assume a parametric form, can add more parameters for more flexibility and can run into the problem of overfitting from too many parameters. Empirically-driven models do not assume a particular form and are thus also known as non-parametric, such models require a lot more observations and generally have a high degree of flexibility once they are initially created.

-Generally I would assume that mechanistic models are easier to understand because they have a familiar structure with particular variables. People are generally very familiar with this form.

-Because of the fact that simple models tend to have high bias and low variance, and flexible models tend to have low bias and high variance, finding a model where the bias and variance are both low is a difficult task

Question 6

-The question of how likely a voter is to vote for a particular candidate would be considered a predictive model because it is aiming to find a particular outcome with a high degree of accuracy

-When it comes to how a voters support of a candidate would change based on contact would be more of an inferential question because this is aiming to find a particular relationship between X and Y and is essentially testing a theory.

Exercise 1

```
library(tidyverse)
```

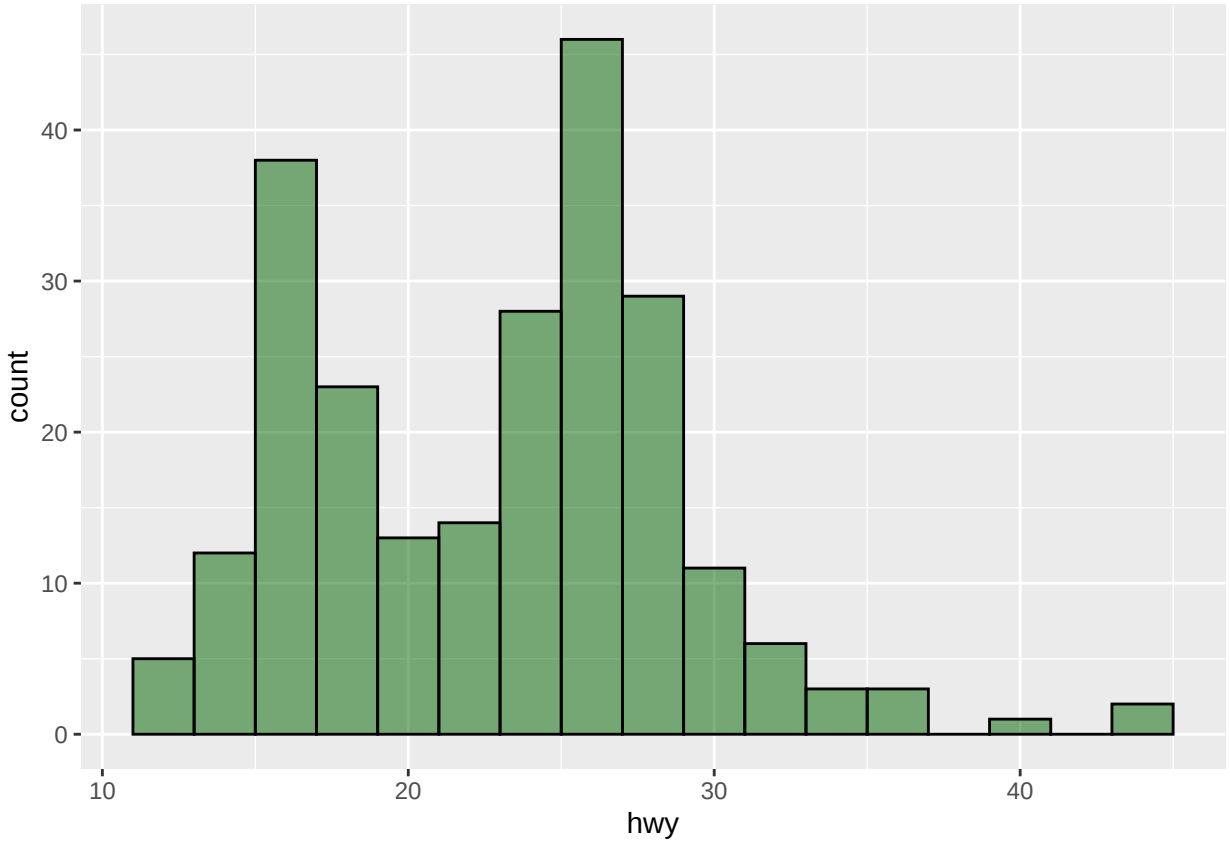
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5      v purrr   0.3.4
## v tibble  3.1.6      v dplyr   1.0.10
## 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()
```

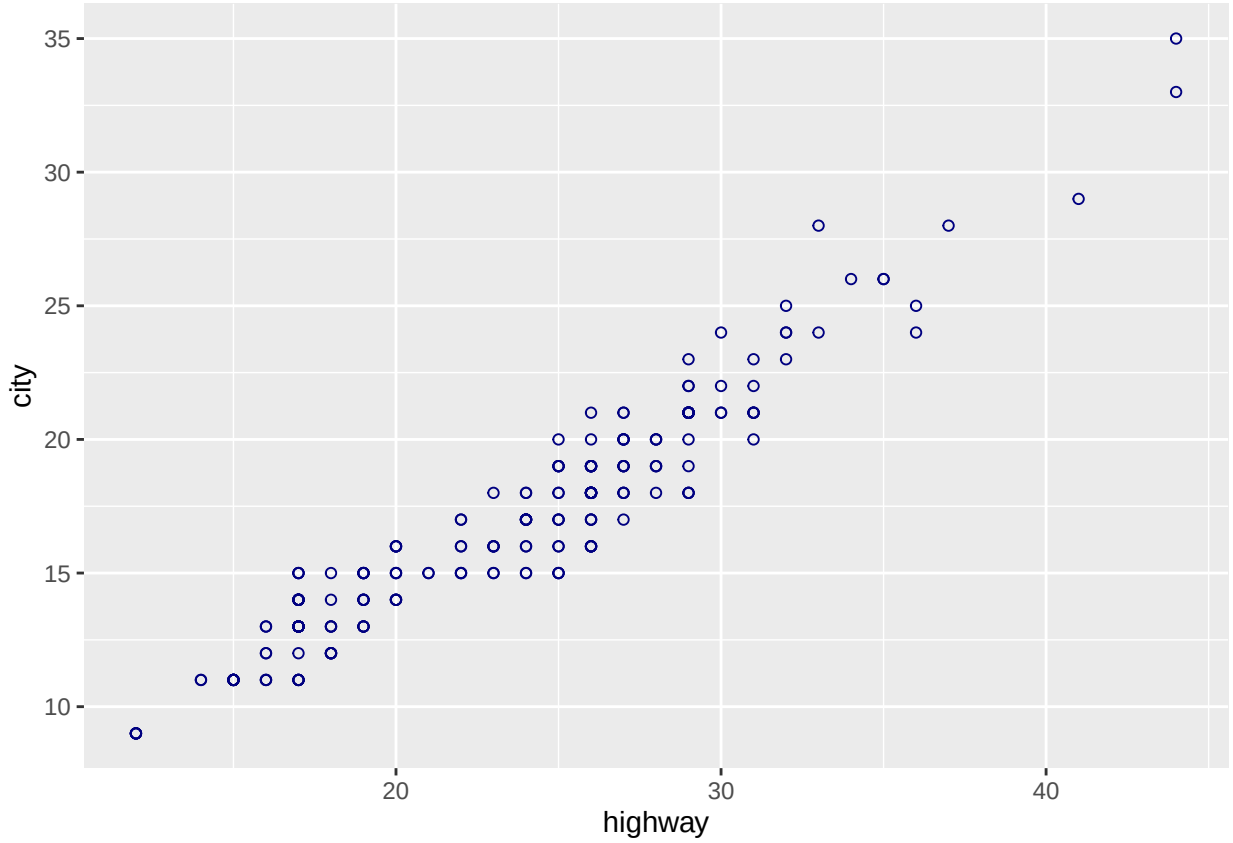
```
library(ggplot2)
```

```
data1 <- mpg
ggplot(data= (data1), aes(x=hwy))+geom_histogram(binwidth = 2,colour='black',  fill = 'darkgreen', alpha
```



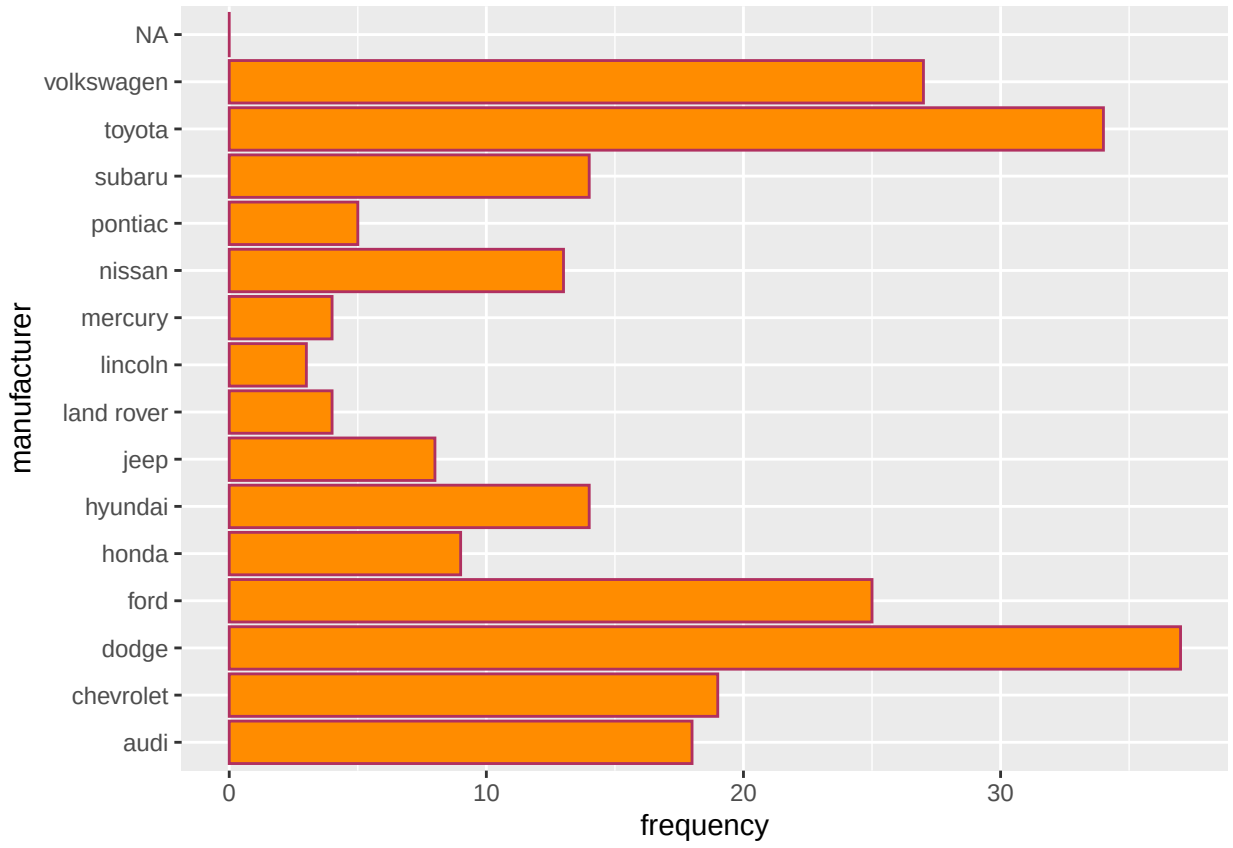
Observation here is that the most common MPG level is between 15-30. This rest of the bins are fairly empty in this histogram.

```
dat <- data.frame(highway = data1$hwy, city = data1$cty)
ggplot(dat, aes(x=highway, y=city), ) +
    geom_point(shape=1, color = 'navy')
```

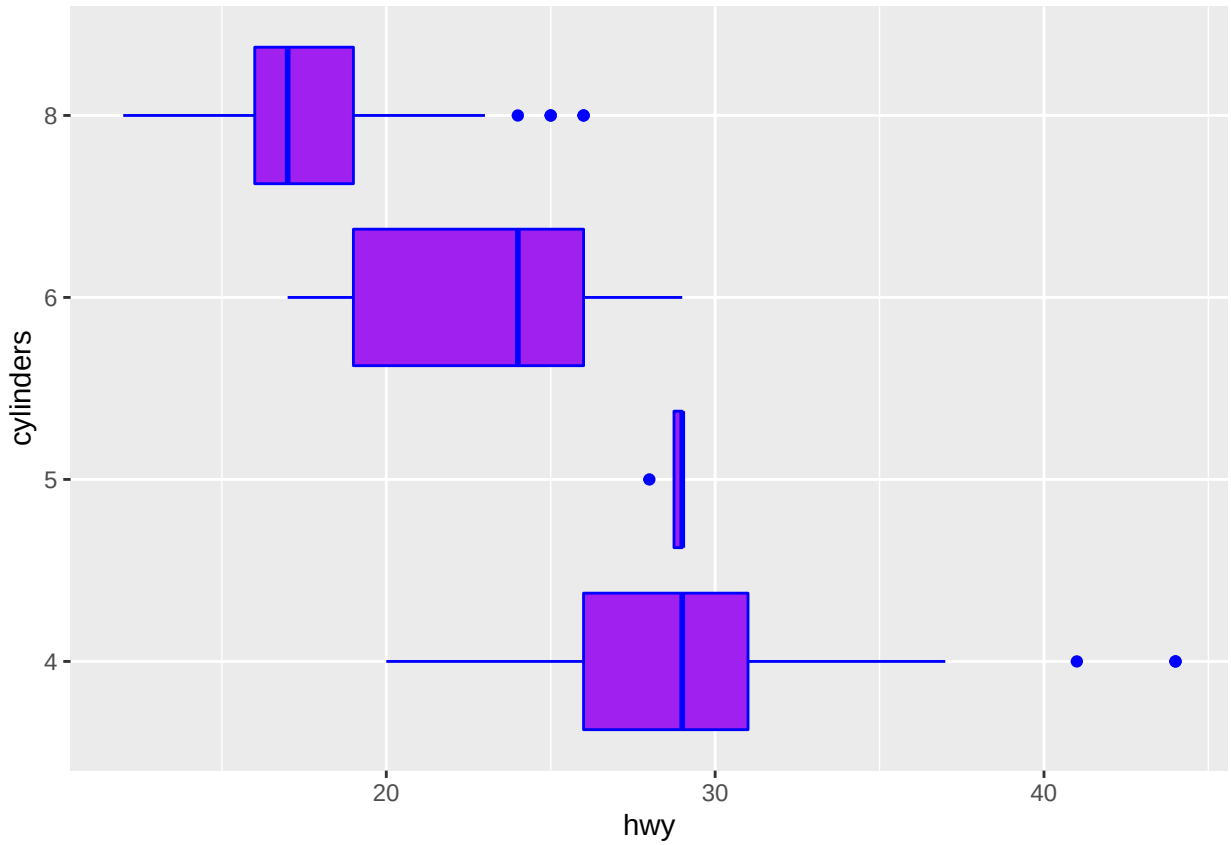


There is a strong linear positive correlation between highway and city mpg meaning that a higher highway mpg translates into a higher city mpg.

```
man_data <- (data1$manufacturer)
tab_1 <- table(man_data, useNA = 'always')
dat1 <- as.data.frame(tab_1)
manufacturer <- dat1$man_data
frequency <- dat1$Freq
ggplot(data = dat1, aes(y = manufacturer,x= frequency), horiz = TRUE)+geom_bar(stat = 'identity', colou
```



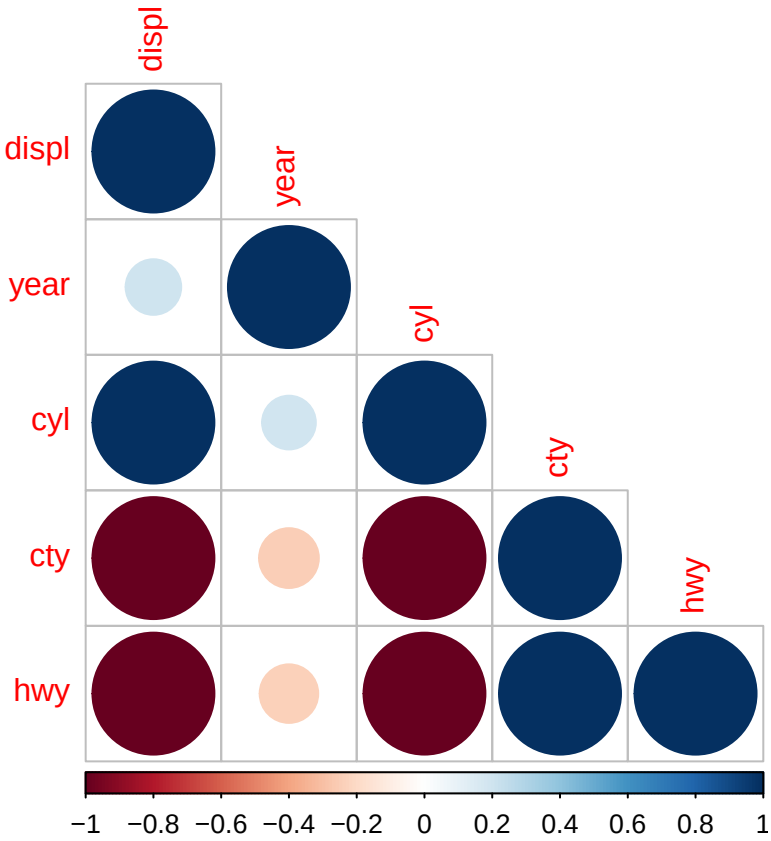

```
cylinders <- factor(data1$cyl)
ggplot(data = data1, aes(hwy, cylinders  ))+geom_boxplot(colour = 'blue', fill = 'purple')
```




```
library('corrplot')
```

```
## corrplot 0.92 loaded
```

```
mpg_matrix <- mpg %>%
    select_if(is.numeric) %>%
    cor(.)
corr1 <- cor(mpg_matrix)
corrplot(corr1, type = 'lower')
```



From this information it is clear that there are some correlations in both the negative and the positive direction. For example, the year has a negative correlation with highway and city mileage and a positive correlation with displacement. These things seem to make sense as to what we would expect. Additionally, highway and city mileage are strongly correlated in a positive direction.