HW1

Ben Hertzberg

2022-10-01

##Q1 Supervised learning uses a labeled data set to train an algorithm. Unsupervised learning does not us labeled data, and instead finds patterns and clusters in the data provided.

##Q2 Regression models in ML predict a numerical, qualitative variable. Classification models in ML predict a qualitative/categorical variable.

##Q3 Regression ML metrics: Training MSE, expected Test MSE, Classification ML metrics: Training Error Rate, Bayes Error Rate

##Q4 Descriptive models are used visually empahsize a trend in data. Predictive models attempt to make a prediction with as little reducible error as possible. Inferential models attempt to test theories and find the relationship between the outcome and the predictors, which may be causal. (Slide 39, Lec 1)

##Q5 Mechanistic models assume the statistical model follows a parametric form, following some theory to make predictions. More parameters will make the model more flexible. Empirically driven models make no assumptions about the structure of the statistical model and analyze the data without a base theory. These models require more observations than parametric models, and are more flexible by nature. Both models have to avoid overfitting.

Mechanistic models are easier to understand because it is clear to see how the predictor variables impact the outcome in a parametric form.

The bias-variance trade describes the inverse relationship between lowering variance and lowering bias. Finding a method to minimize both of these values as much as possible will result in the lowest expected test MSE, which is the sum of the variance, bias squared, and irreducible error. The lower the expected test MSE, the better the model is espected to perform.

##Q6 The first scenario is a predictive model. You are given some information which can be used as predictors for an outcome. The goal is to see how well you can predict a voter's choice given their data.

The second scenario is an inferential model. In this case, the goal is to analyze how voter contact with the candidate affects their voting (analyzes the relationship between the predictors and the outcome).

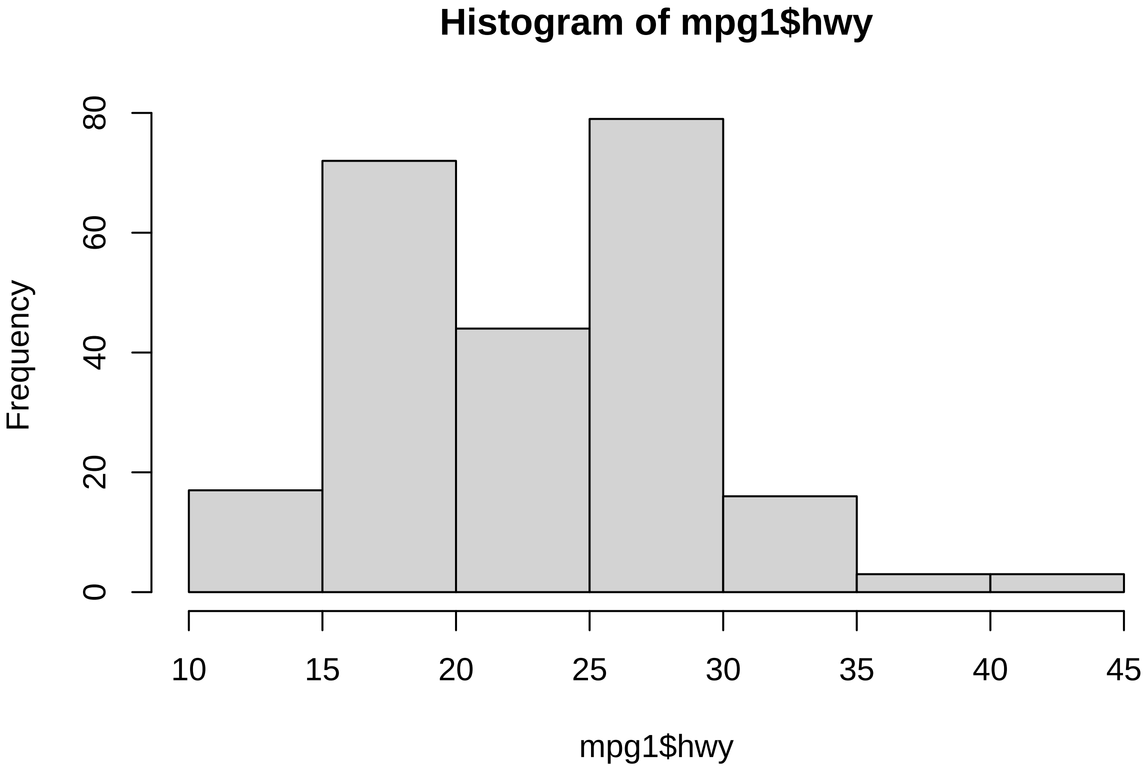
```
library(tidyverse)
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

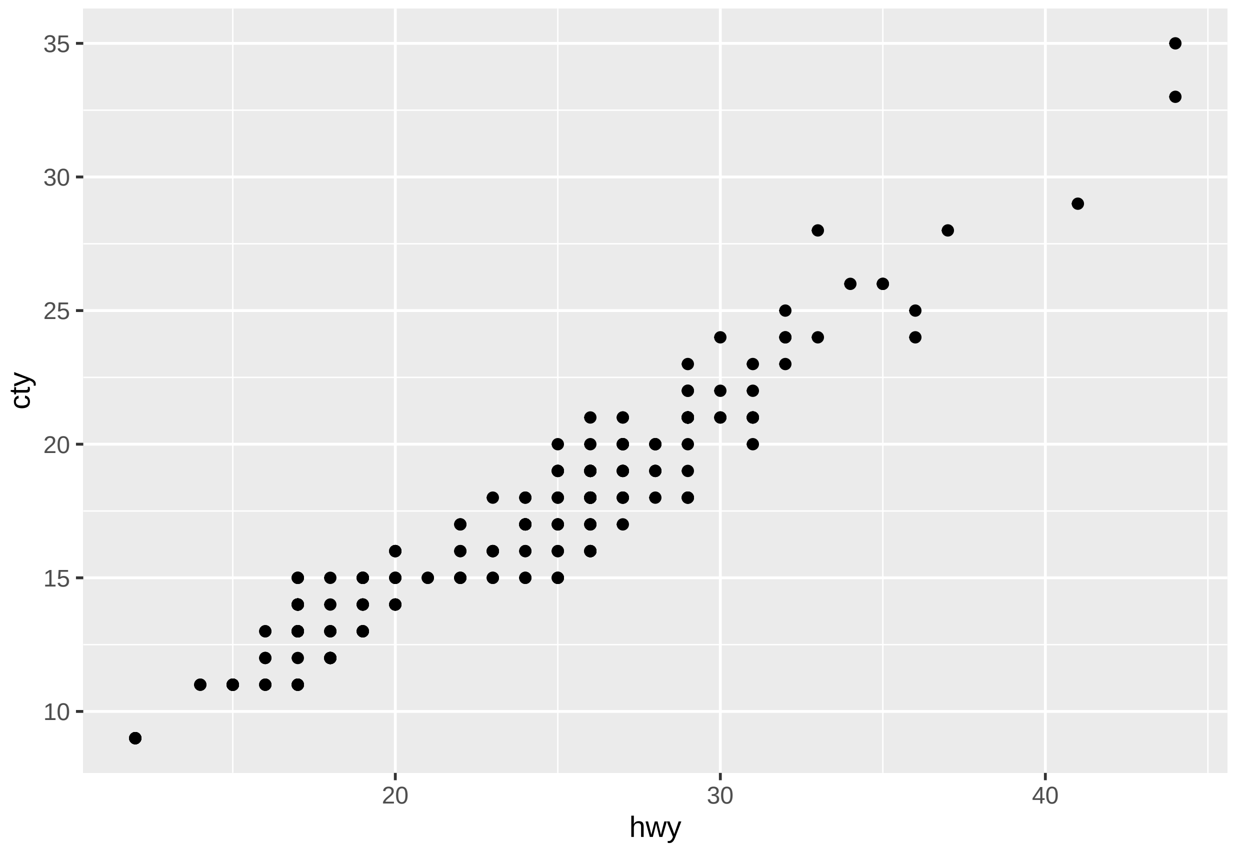
```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#1
mpg1 <- mpg
hist(mpg1$hwy)
```



The data on highway mpg is bimodal with peaks between 15-20 mpg and 25-30 mpg. It is skewed right. The range is 10 to 45 mpg. The median is between 20 and 25.

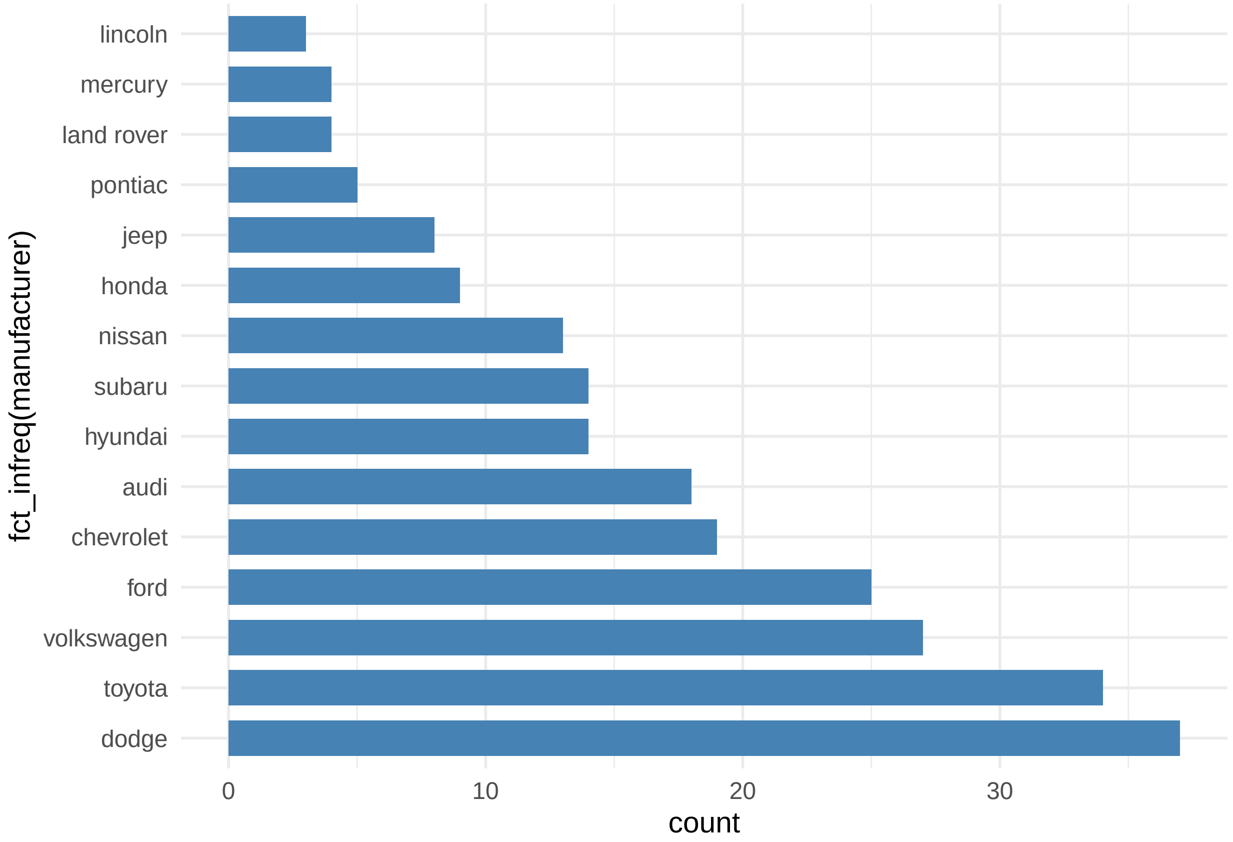
```
#2
ggplot(mpg1, aes(x=hwy, y=cty)) + geom_point()
```



There is a strong positive correlation between city mpg and highway mpg. This suggests that cars that have better mpg in one area will also have better mpg in the other.

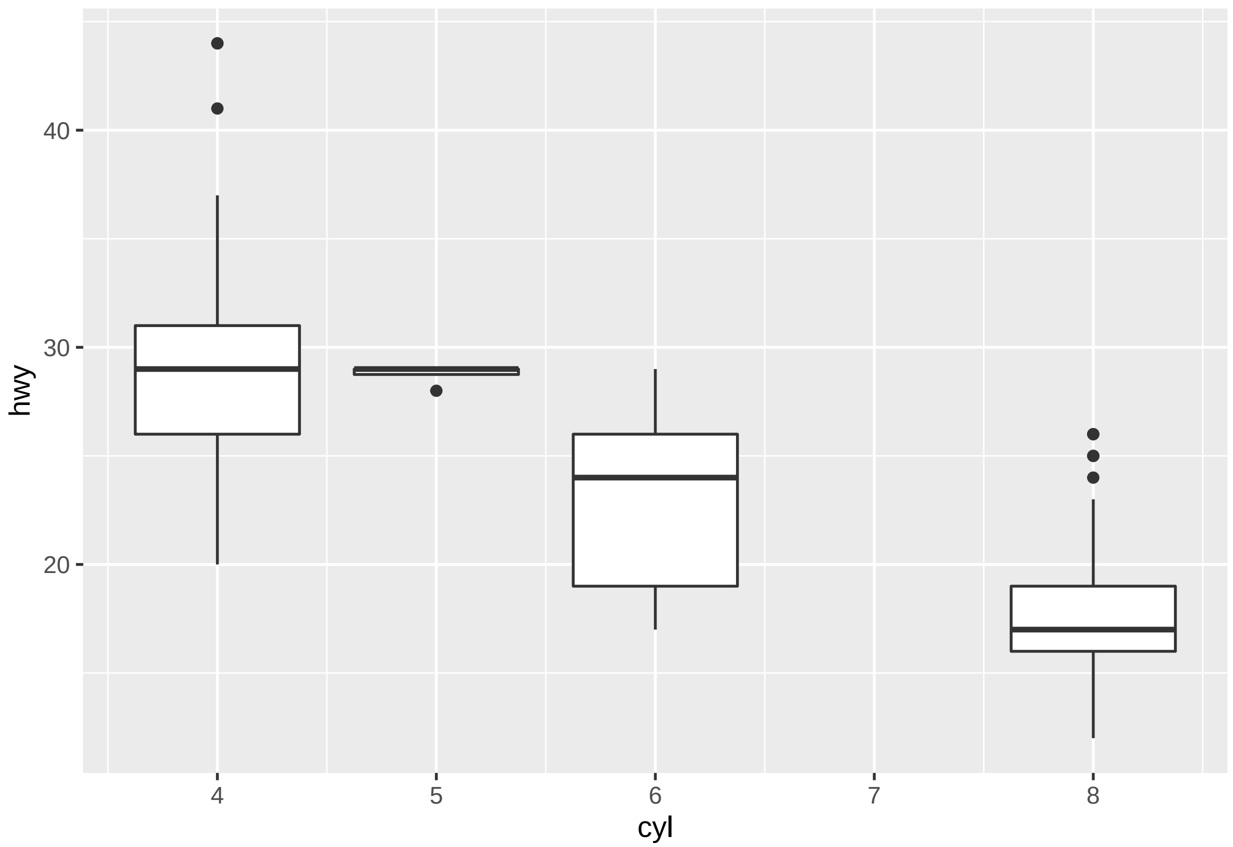
```
#3
p <- ggplot(mpg1, aes(x=fct_infreq(manufacturer)))+
  geom_bar( width=0.7, fill="steelblue")+
  theme_minimal()

p + coord_flip()
```



Dodge produced the most cars. Lincoln produced the least.

```
#4
box1 <- ggplot(mpg1, aes(group = cyl, x = cyl, y=hwy)) +
  geom_boxplot()
box1
```



As the number of cylinders in a model's engine increases, the highway mpg tends to decrease.

```
#5
#library(corrplot)
#corrmpg <- c(mpg$displ, mpg$year, mpg$cyl, #mpg$cty, mpg$hwy)
#M = cor(corrmpg)
#removed qualitative variables
#corrplot(M, method = 'square', order = 'FPC', #type = 'lower', diag = FALSE)


#I can't figure out this error, I had to comment out this code to render it correctly
```