131hw1

2022-04-04

Questions: 1. Unsupervised learning is learning without knowing the answer key. In other words, there are predictor variables but no known response variables. Supervised learning on the other hand has predictor variables and known response variables. The goal of supervised data is to predict outcomes for new data, while unsupervised learning is more used to gain insight from large volumes of data.

2. Regression models are used in machine learning to predict continuous, quantitative values (price, GPA) while classification models are to classify or predict discrete, qualitative values such as M/F or surviving/not.

3. Two metrics for regression: mean squared error and mean absolute error. Two metrics for classification: ROC AUC and confusion matrix.

4. Descriptive models are chosen to best visualize and emphasize a trend. Predictive models decide which features fit best and try to predict the response variable with minimum reducible error, and are not focused on hypothesis tests. Inferential models answer which features are significant, used to test theories, and state relationship between the response and predictors.

5. A mechanistic model assumes a parametric form for f, specify assumptions and attempt to incorporate known factors surrounding data into the model while describing said model. More parameters can be added for more flexibility but there is over fitting. On the other hand, an empirical model describes the data with no assumptions about the analyzed data. It requires a large number of observations, are by default more flexible, and allow over fitting. Both elements are used in many models - empirical models develop a theory, and mechanistic models use a theory to predict something.

A mechanistic model is easier to look at because predicting an outcome with a form should be more simple to obtain information from than looking at correlation and hypothesizing with confidence intervals what the data means.

The bias variance tradeoff (truncating certain predictors, for example) relates to these two models because removing a predictor may make the data fit better but create larger error and variance for future outcomes, or make the predictors be determined to be or not to be correlated to each other or a theory or its opposite as a result.

6. The first question is predictive as it tries to predict Y (how likely to vote) given predictor variables, and doesn't relate to hypothesis testing.

The second question is inferential as it tests the theory if likelihood of voting will change or not, and if so, for the better or worse if there was personal contact, and states the relationship between the outcome and predictor.

Loading packages

```
library(tidyverse)
```

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

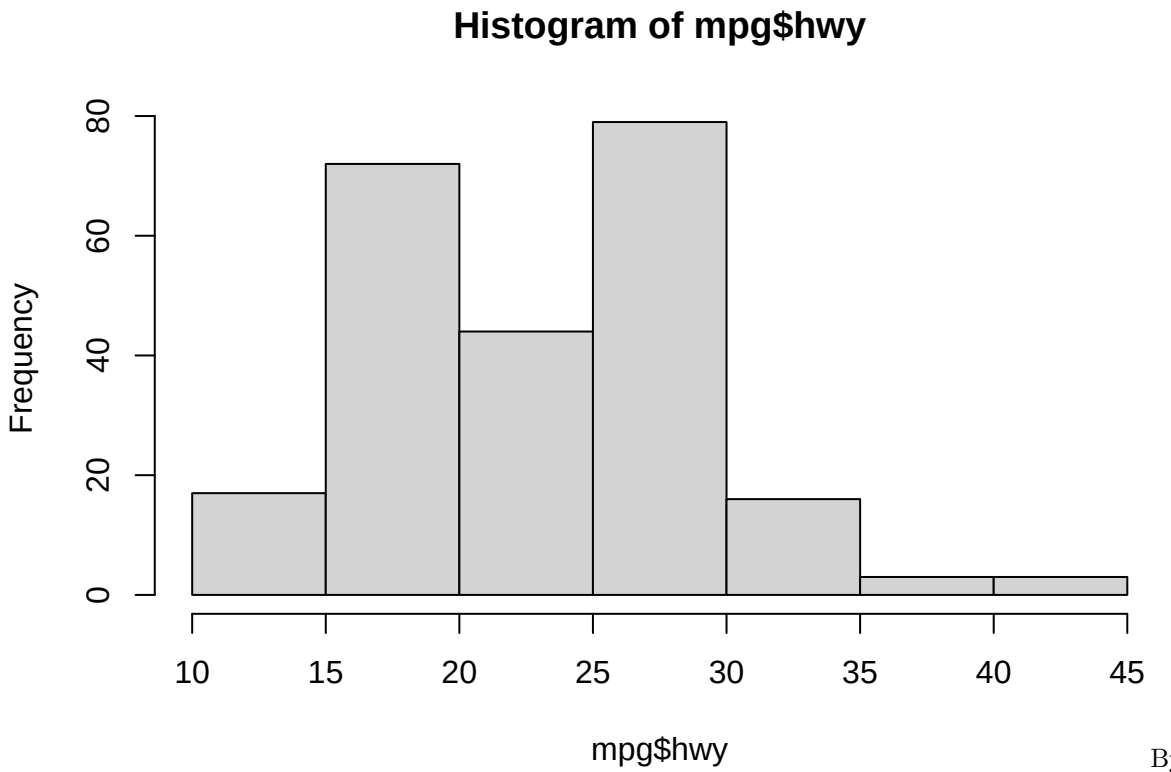
```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

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

```
library(ggplot2)
data(mpg)
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model      displ  year   cyl trans drv     cty   hwy fl    class
##    <chr>        <chr>      <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
##  1 audi         a4           1.8  1999     4 auto~ f        18    29 p     comp~
##  2 audi         a4           1.8  1999     4 manu~ f        21    29 p     comp~
##  3 audi         a4           2    2008     4 manu~ f        20    31 p     comp~
##  4 audi         a4           2    2008     4 auto~ f        21    30 p     comp~
##  5 audi         a4           2.8  1999     6 auto~ f        16    26 p     comp~
##  6 audi         a4           2.8  1999     6 manu~ f        18    26 p     comp~
##  7 audi         a4           3.1  2008     6 auto~ f        18    27 p     comp~
##  8 audi         a4 quattro   1.8  1999     4 manu~ 4        18    26 p     comp~
##  9 audi         a4 quattro   1.8  1999     4 auto~ 4        16    25 p     comp~
## 10 audi         a4 quattro   2    2008     4 manu~ 4        20    28 p     comp~
## # ... with 224 more rows
```

Exercises: 1. Solution

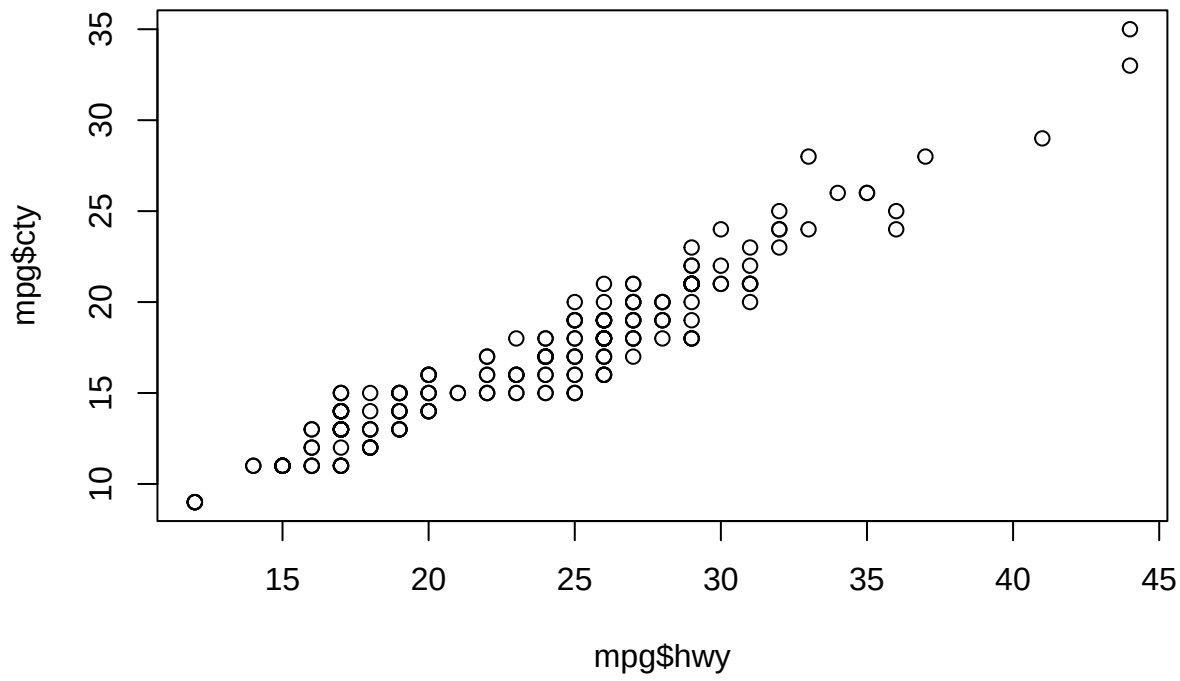
```
hist(mpg$hwy)
```



By creating a histogram of hwy, we see the ranges of mpg (in ranges of 5) of different cars in mpg data set as well as their frequencies in that range.

2. Solution

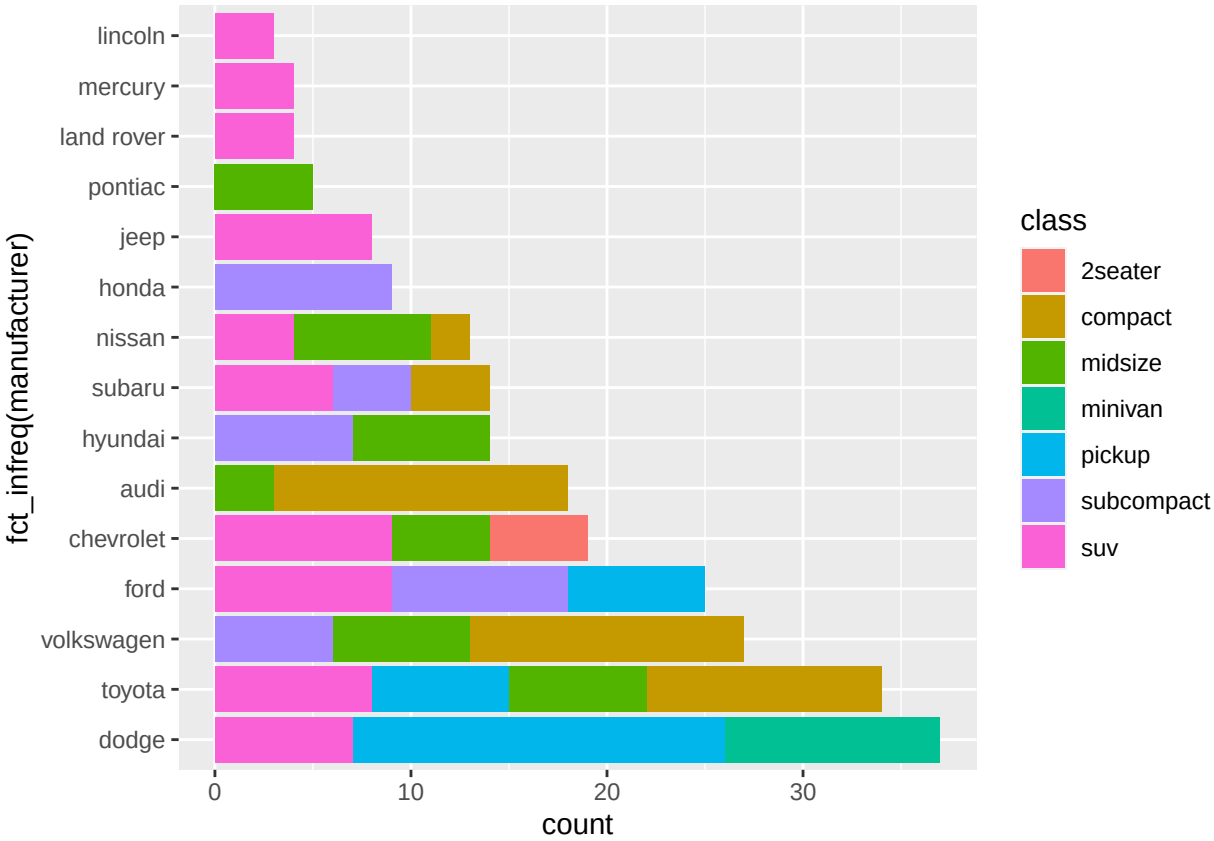
```
plot(mpg$hwy, mpg$cty)
```



There is a linear relationship between mpg in the city and mpg in the highway. This means that if a car gets low/high mpg in the city, it will also get relatively low/high mpg in the highway.

3. Solution

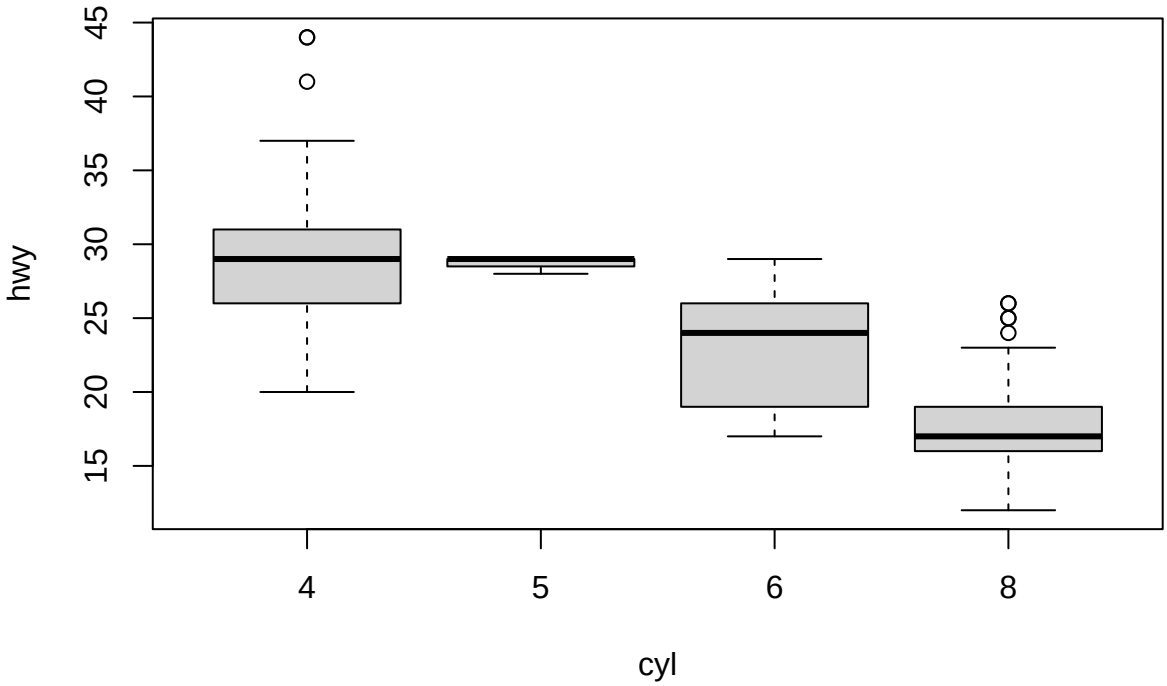
```
ggplot(mpg, aes(fct_infreq(manufacturer), fill = class)) + geom_bar() + coord_flip()
```



Dodge produced the most different type of cars; Lincoln produced the least.

4. Solution

```
boxplot(hwy ~ cyl, data=mpg)
```



There seems to be a pattern where the higher number of cylinders, the lower MPG a car gets in the highway.

5. Installing packages

```
library(corrplot)
```

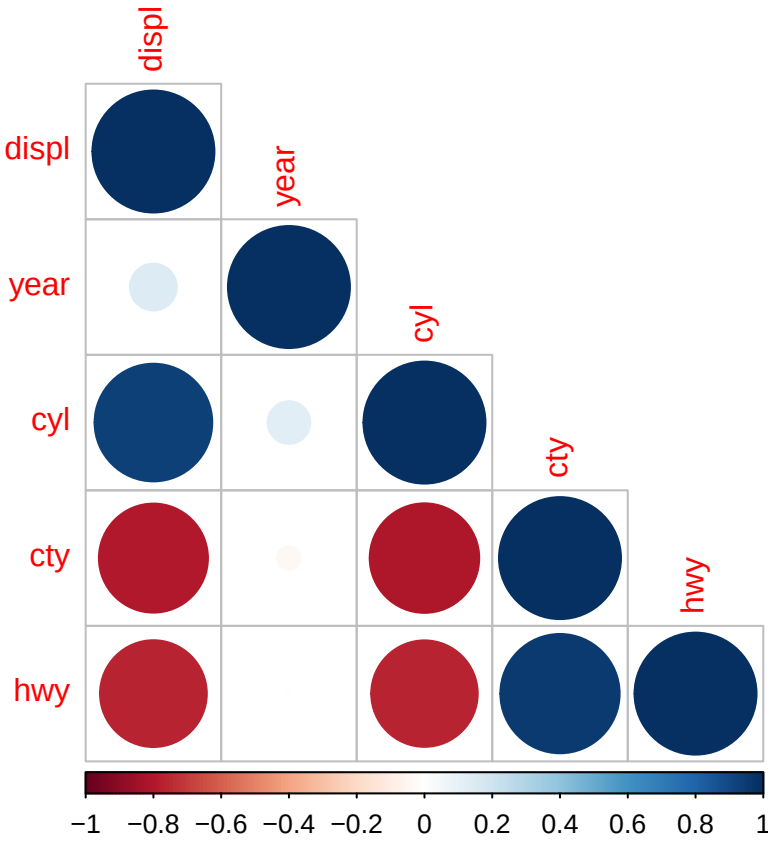
```
## corrplot 0.92 loaded
```

Solution [I omitted any character variables as corrplot accepts only numeric vectors]

```
df = subset(mpg, select = -c(manufacturer,model,trans,drv,fl,class) )
head(df)
```

```
## # A tibble: 6 x 5
##   displ  year   cyl   cty   hwy
##   <dbl> <int> <int> <int> <int>
## 1   1.8  1999     4    18    29
## 2   1.8  1999     4    21    29
## 3   2    2008     4    20    31
## 4   2    2008     4    21    30
## 5   2.8  1999     6    16    26
## 6   2.8  1999     6    18    26
```

```
M = cor(df)
corrplot(M, type = 'lower')
```



There is positive correlation between: city mpg and highway mpg, cylinder and year (on a small scale), year and displacement (on a small scale), cylinders and displacement.

There is negative correlation between: city and highway mpg with displacement and cylinders.

All of the correlations make sense - we saw that a car with high city/hwy mpg probably has high hwy/city mpg. Cars generally get larger engines over the years (newer models should always have an incentive to be bought), and more cylinders generally makes more displacement. More displacement and cylinders also means the car is more gas hungry.

```
library(usethis)
edit_git_config()
git config --global user.email "johntigerwei@gmail.com"
git config --global user.name "John Wei"
```

R Markdown

```
cd Projects/website

git remote add origin https://github.com/girafang/131hw1.git

git pull origin master

git push -u origin master

install.packages('tinytex')
tinytex::install_tinytex()
```