131 Homework1

Scott Shang (8458655)

April 04, 2022

Question 1: Supervised learning: Machine learning that including prediction, estimation, model selection, and inference. To be specific, supervised learning can accurately predict future response given predictors, understand how predictors affect response, select the best model, and assess the quality of predictions and estimation. Unsupervised learning don't have a supervisor, and the model work on its own. (From Lec1 pg30)

Question 2: In the context of machine learning, the regression model is where the response variable Y is quantitative. To be specific, Y are numerical values. For the classification model, the response variable Y is qualitative, i.e. categorical values. (Lec1 31)

Question 3: Didn't covered.

Question 4: Descriptive models: Choose model to best visually emphasize a trend in data. Inferential models focus on those features are essential. It is aimed to test theories and possibly causal claims. It also state the relationship between outcome and predictors. Predictive models answer what combo of features fits the model best. And it aims to predict Y with minimum reducible error. It doesn't focus on hypothesis tests. (Lec2 pg7)

Question 5: Mechanistic assume a parametric form. It relies on known formula and parameters. It can add parameters to add more flexibility, but too many parameters may overfit the model. It won't match true unknown f. Empirically-driven means the model has no assumptions about f. It require a large number of observations. By default, it has much more flexibility than mechanistic models. It also has the risk of overfitting. (Lec2 pg6) bias-variance tradeoff didn't covered in class.

Question 6: The first question is predictive, since it using a set of information to predict outcome(who will they vote). The second question is inferential. It want to test whether the feature 'having contact with the candidate' is significant regarding the outcome. We want to observed the pattern with that feature. (Lec2 pg7)

Ex1:

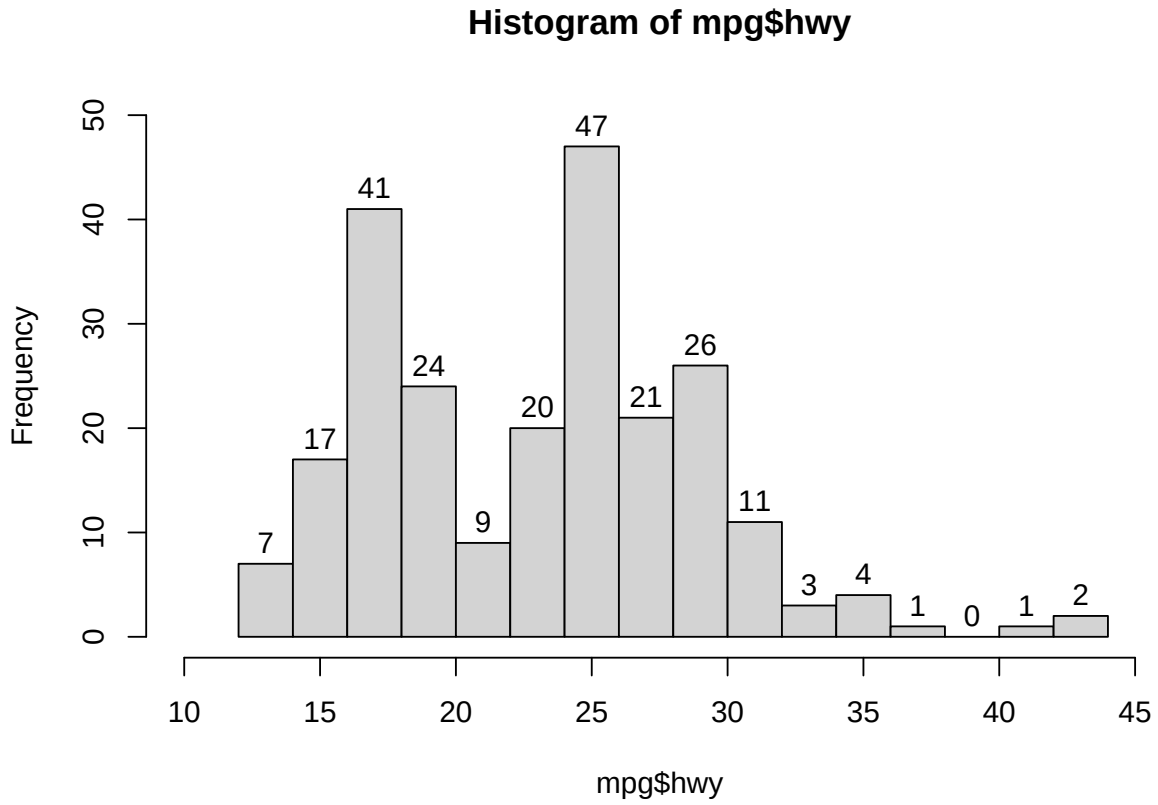
```
library(tidyverse)
```

```
## -- 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()
```

```
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
```

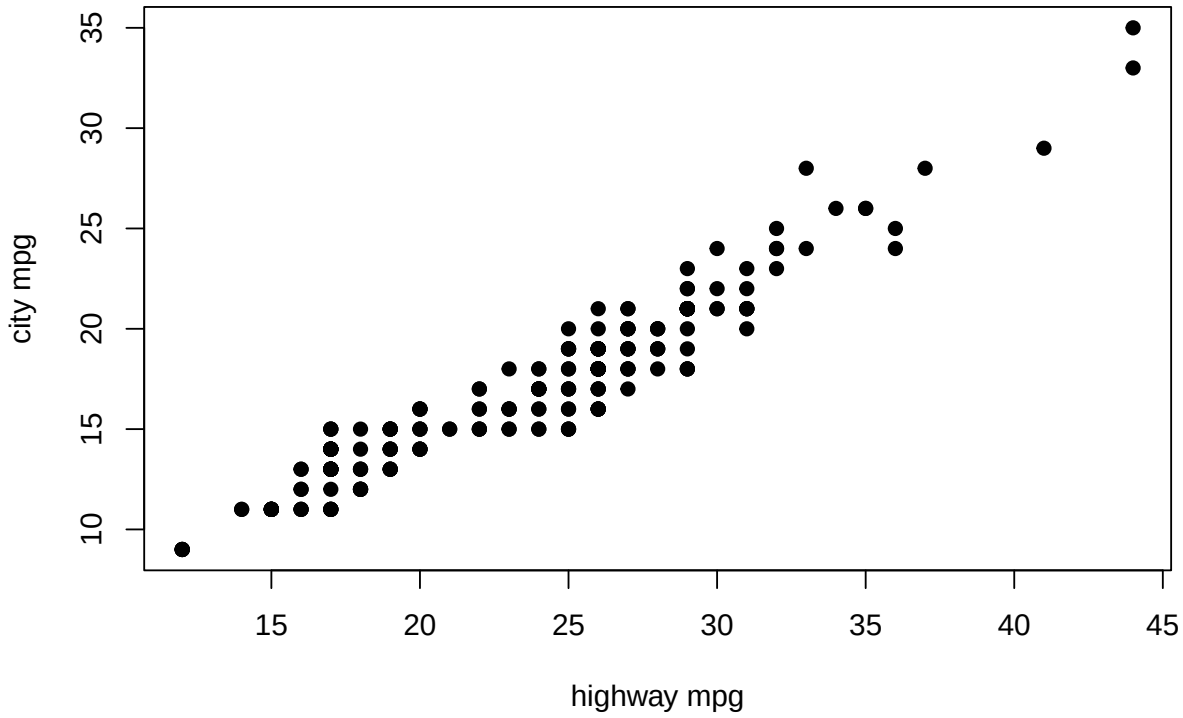
```
h=hist(mpg$hwy,xlim=c(10,45),ylim=c(0,50),breaks=20)
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))
```



We notice that most vehicles' highway mpg are between 15-30 mpg. It's kind of like a normal distribution but not close.

Ex2:

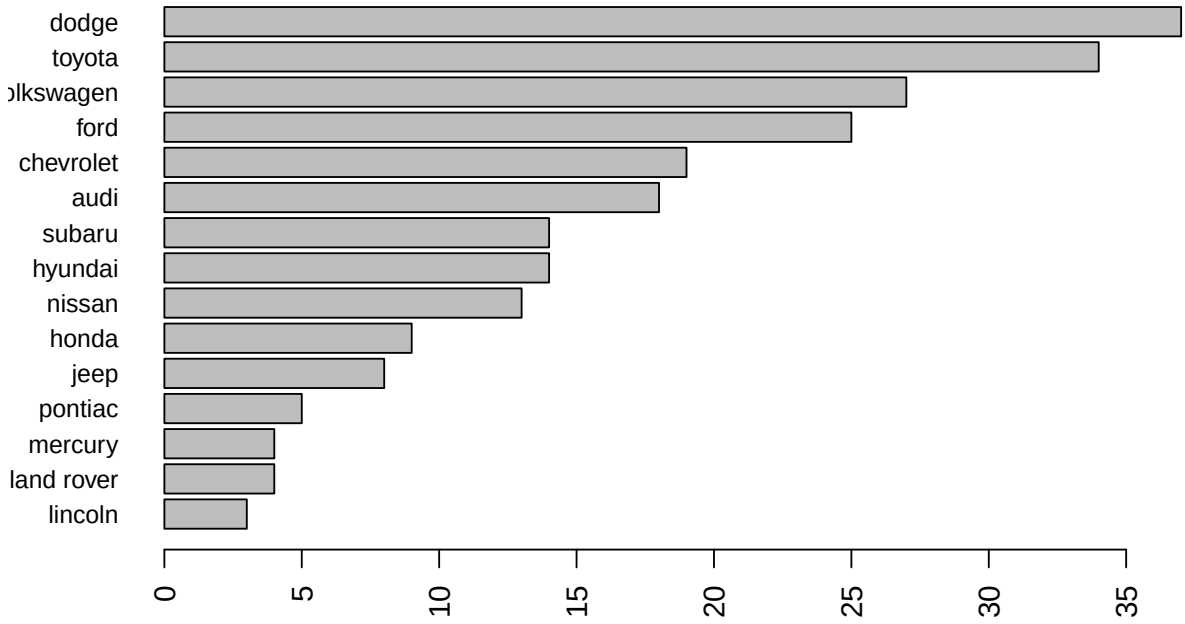
```
plot(mpg$hwy, mpg$cty, xlab="highway mpg", ylab="city mpg", pch=19)
```



We notice that there is a possible relationship between highway mpg and city mpg. That means a car with higher hwy mpg are likely to have higher city mpg. And vice versa.

Ex3:

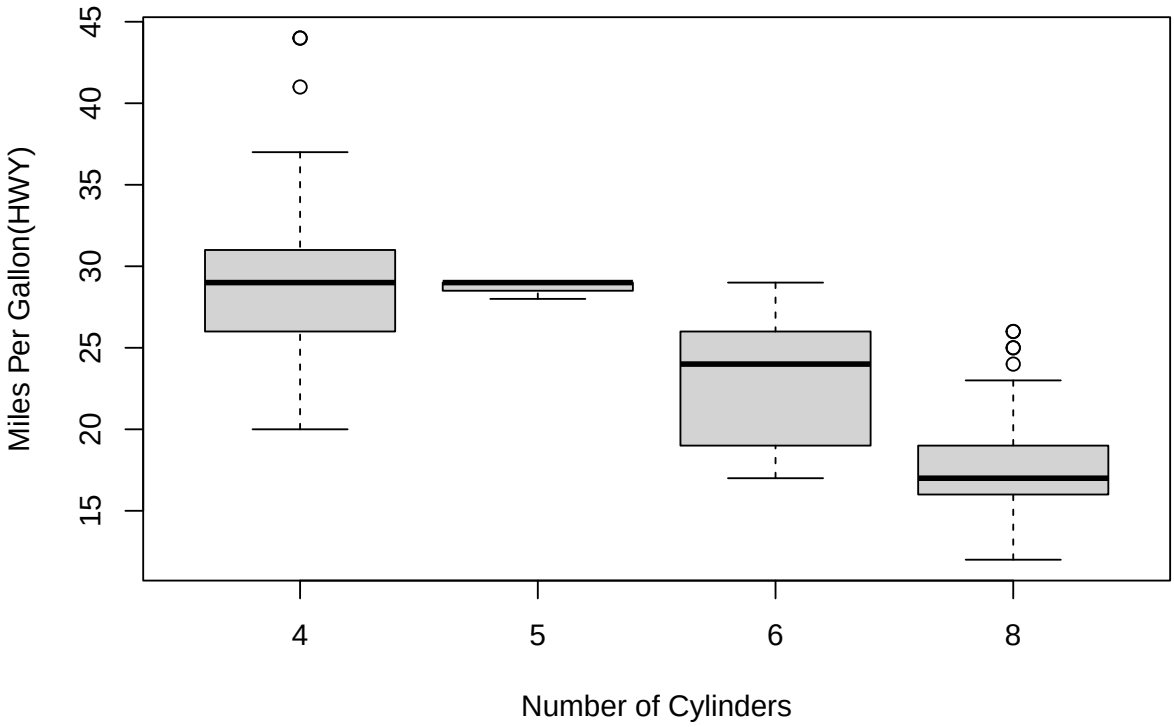
```
count=sort(table(mpg$manufacturer))
barplot(count,horiz=TRUE,las=2,cex.names = 0.8)
```



As we can see, Dodge produced the most, and Lincoln produced the least.

Ex4:

```
boxplot(hwy~cyl, data=mpg, xlab="Number of Cylinders", ylab="Miles Per Gallon(HWY)")
```



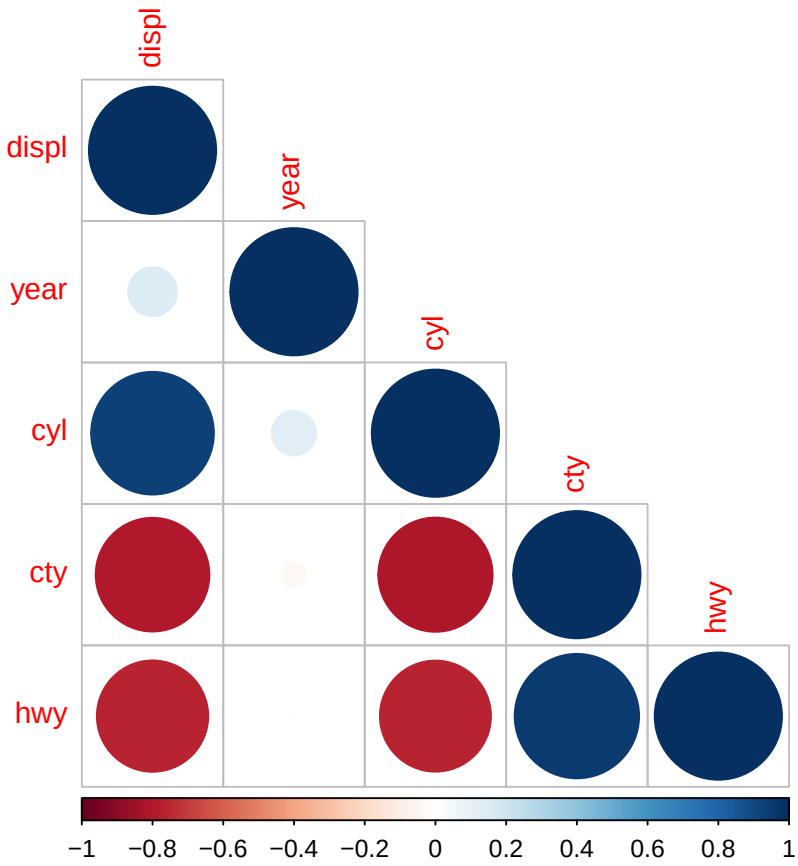
Cars with less cylinders tend to have higher mpg. We can say that cylinder number and hwy mpg are negatively correlated.

Ex5:

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
newmpg=subset(mpg,select=-c(manufacturer, drv, model, trans, fl, class))
cormpg=cor(newmpg)
corrplot(cormpg,type="lower")
```



For example, displacement is positively correlated to cylinder numbers, city mpg is positively correlated to highway mpg, and cylinder number is negatively correlated to city or highway mpg. Those relationships make sense to me. None of them is surprising.