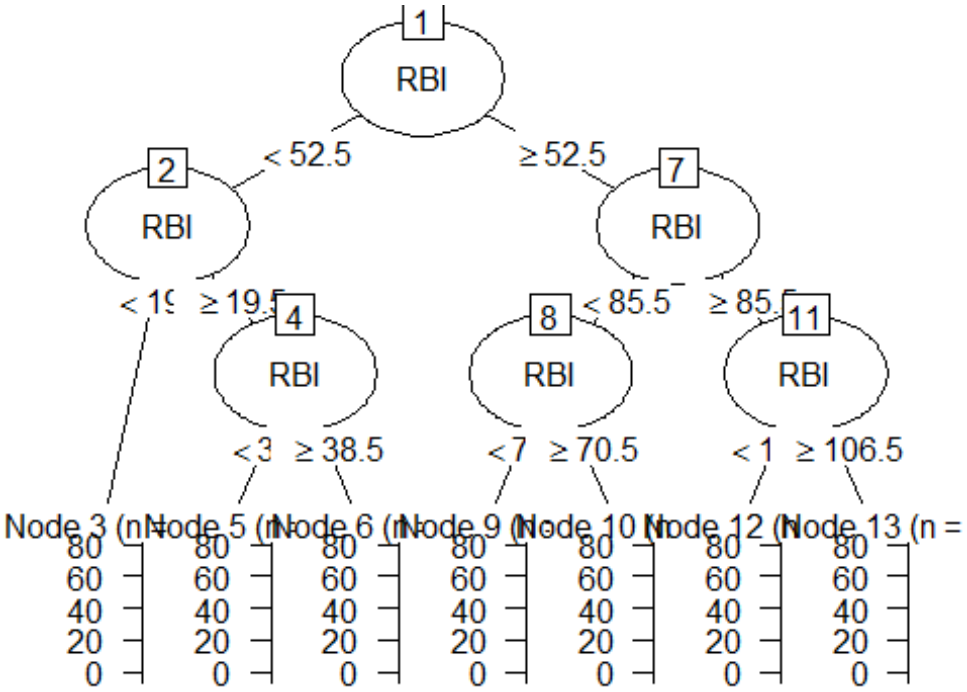
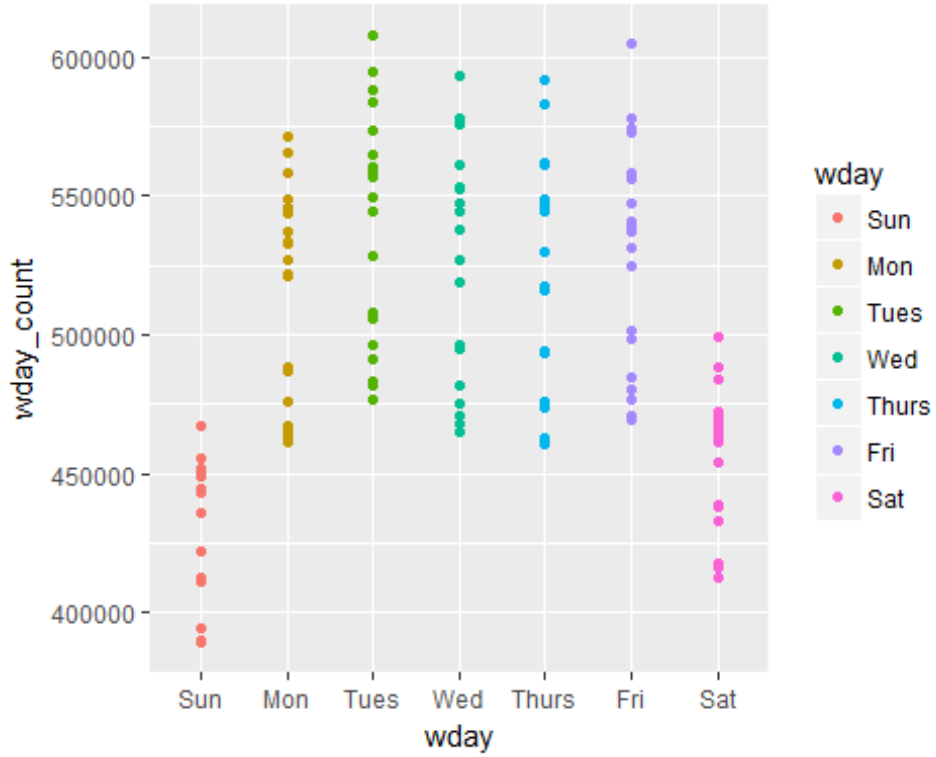
R Notebook

```
library(mdsr)
library(Lahman)
library(NHANES)
library(nycflights13)
library(rpart)
library(partykit)

#1a
batterData <- Batting %>%
  select(HR, AB, R, H, X2B, X3B, RBI, SO, IBB) %>%
  filter (!(IBB %in% NA)) %>%
  filter (!(SO %in% NA))
View(batterData)

#1b

HR_Factor = as.factor(batterData$HR);

#1c

hitHRMaybe <- rpart(HR ~ AB + R + H + X2B + X3B + RBI + SO + IBB, data =
batterData)
  #rpart(HomeOwn ~ Age + Gender + HHIncomeMid + MaritalStatus +
                   #Work + Education, data = people, method = "class")
hitHRMaybe

## n= 66191
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 66191 3334430.00  3.426856
##    2) RBI< 52.5 58958  517854.10  1.423233
##      4) RBI< 19.5 48955   52483.91  0.410193 *
##      5) RBI>=19.5 10003  169253.30  6.381086
##       10) RBI< 38.5 6458   64075.56  5.023072 *
##       11) RBI>=38.5 3545   71571.47  8.855007 *
##    3) RBI>=52.5 7233  650593.50 19.758880
##      6) RBI< 85.5 4972  199658.30 15.600970
##       12) RBI< 70.5 3207   96125.59 13.561580 *
##       13) RBI>=70.5 1765   65959.18 19.306520 *
##      7) RBI>=85.5 2261  175955.40 28.902260
##       14) RBI< 106.5 1561   77537.21 25.957720 *
##       15) RBI>=106.5 700   54702.31 35.468570 *

plot(as.party(hitHRMaybe))
```



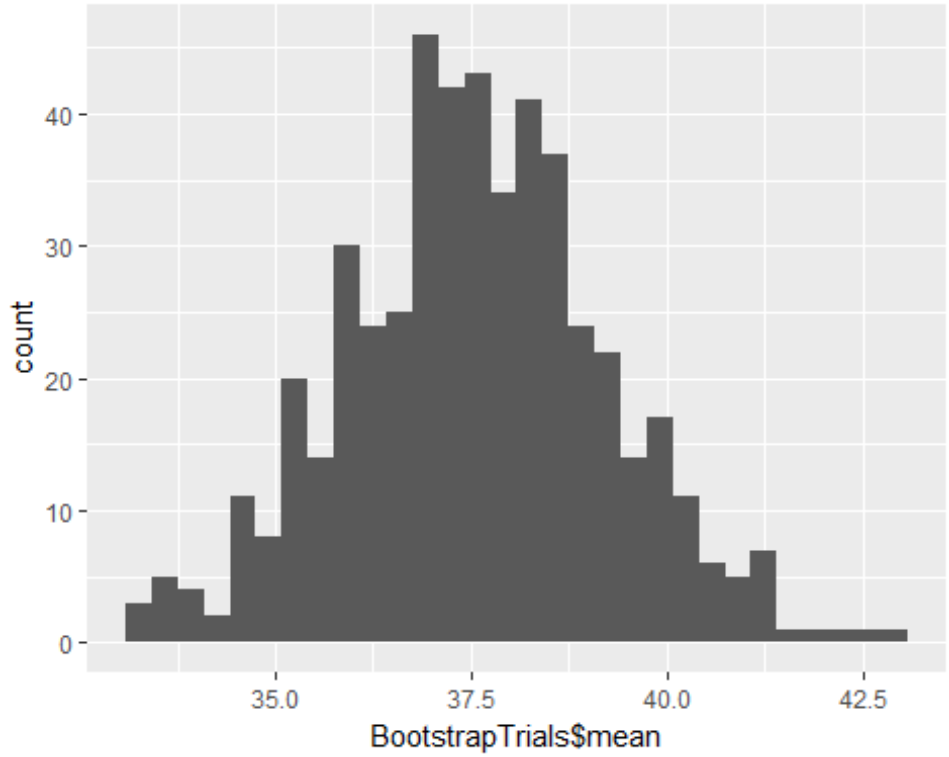

```
#1d
print("The confusion matrix... is so confusing that I have no clue how to
even make one! :D")

## [1] "The confusion matrix... is so confusing that I have no clue how to
even make one! :D"

#1e
print("The model predicts that with those stats, it is unlikely that a player
hit HRs that year, given the very low RBI.")

## [1] "The model predicts that with those stats, it is unlikely that a
player hit HRs that year, given the very low RBI."

#2a
smartBirthCount <- Birthdays %>%
  select(year, wday, births) %>%
  group_by(year, wday) %>%
  summarise(wday_count = sum(births))


View(smartBirthCount)

#2b
smartBirthCountWide <- smartBirthCount %>%
 gather(key=id, value=number, -year, -wday) %>%
  spread(key = wday, value = number) %>%
```

```
  select(year, Sun, Mon, Tues, Wed, Thurs, Fri, Sat)

View(smartBirthCountWide)

#2c

g_smartBirths <- ggplot(smartBirthCount, aes(x = wday, y = wday_count, color
= wday)) + geom_point()
g_smartBirths
```




```
#3a
surveyData <- NHANES %>%
  filter(!(Age %in% NA))
set.seed(364)
View(NHANES)

#3b
surveyDataSamples <- surveyData %>%
  sample_n(200, replace=FALSE)
View(surveyDataSamples)

#3c
BootstrapTrials <- do(500) *
  mean(~ Age, data = sample_n(surveyDataSamples, size = 200, replace = TRUE))
View(BootstrapTrials)
```

```r
#3d
g_surveyAgeMean <- ggplot(BootstrapTrials, aes(x = BootstrapTrials$mean)) +
geom_histogram()
g_surveyAgeMean

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
#3f
qdata(~ mean, p = c(.10, .90), data = BootstrapTrials)

##     quantile   p
## 10%  35.3445 0.1
## 90%  39.7685 0.9

#4a

flightsData <- nycflights13::flights %>%
  filter(dest ==  "SFO")
View(flightsData)

#4b
flightsDataFixed <- flightsData %>%
  mutate(Carrier_Name = "")

x = "haha"

for(i in 1:nrow(flightsDataFixed))
```

```
{
  #if the carrier code on this row matches the carrier code on the airlines
df, then paste the airline name to the data frame

  for(k in 1:nrow(nycflights13::airlines))
  {

    if(flightsDataFixed$carrier[i] == nycflights13::airlines$carrier[k])
    {
      x = nycflights13::airlines$name[k]
    }

    flightsDataFixed$Carrier_Name[i] = x
  }

}
View(flightsDataFixed)
```