CptS575 Hw2

Mengxiao

Part 1

a. Read the data into R

```
data = read.csv('https://scads.eecs.wsu.edu/wp-content/uploads/2017/09/College.csv')
```

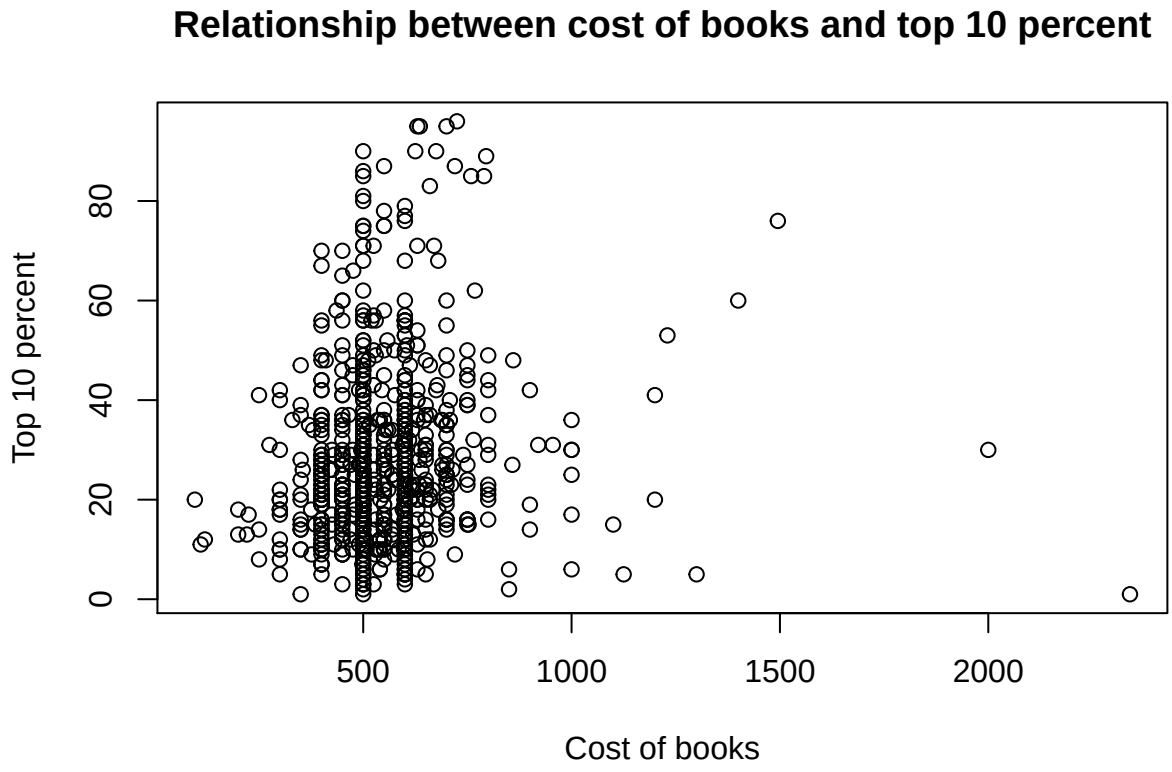
b. Find the median cost of books

```
books_median = median(data['Books'][, 1])
```

```
## [1] 500
```

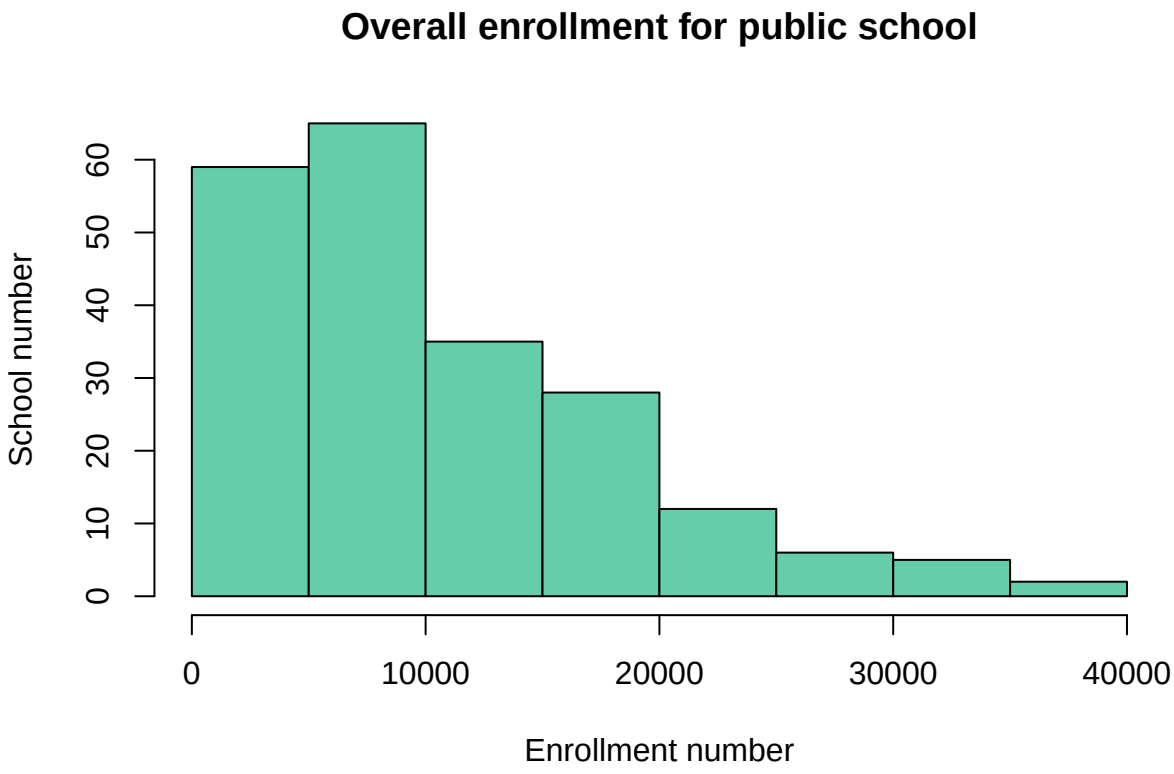
c. Produce a scatterplot to show the relationship between the cost of books and Top 10 percent students.

```
plot(x = data$Books,
     y = data$Top10perc,
     xlab = "Cost of books",
     ylab = "Top 10 percent",
     main = "Relationship between cost of books and top 10 percent"
     )
```

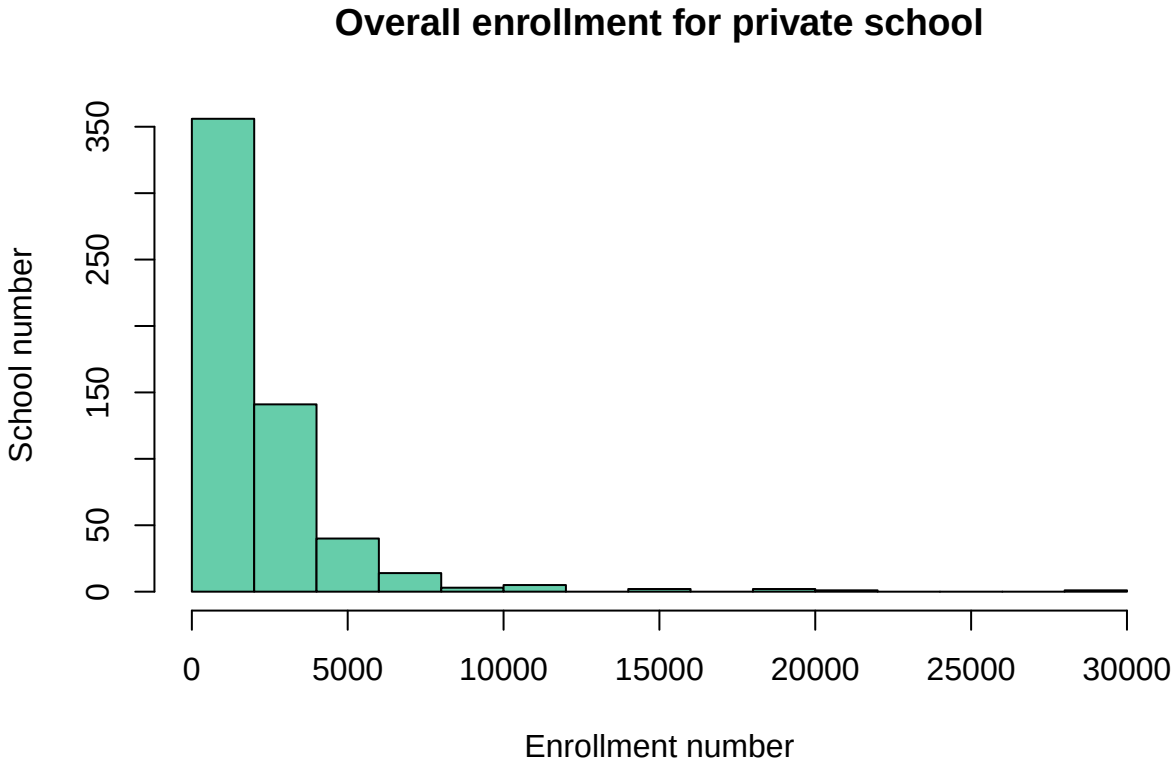


d. Produce a histogram showing the overall enrollment numbers for both public and private schools.

```
OverallEnroll <- list(data[data$Private == 'No',]$P.Undergrad
                  +data[data$Private == 'No',]$F.Undergrad,
                  data[data$Private == 'Yes',]$P.Undergrad
                  +data[data$Private == 'Yes',]$F.Undergrad)
hist(unlist(OverallEnroll[1]), main = "Overall enrollment for public school",
     ylab = "School number", xlab = "Enrollment number", col="aquamarine3"
    )
```

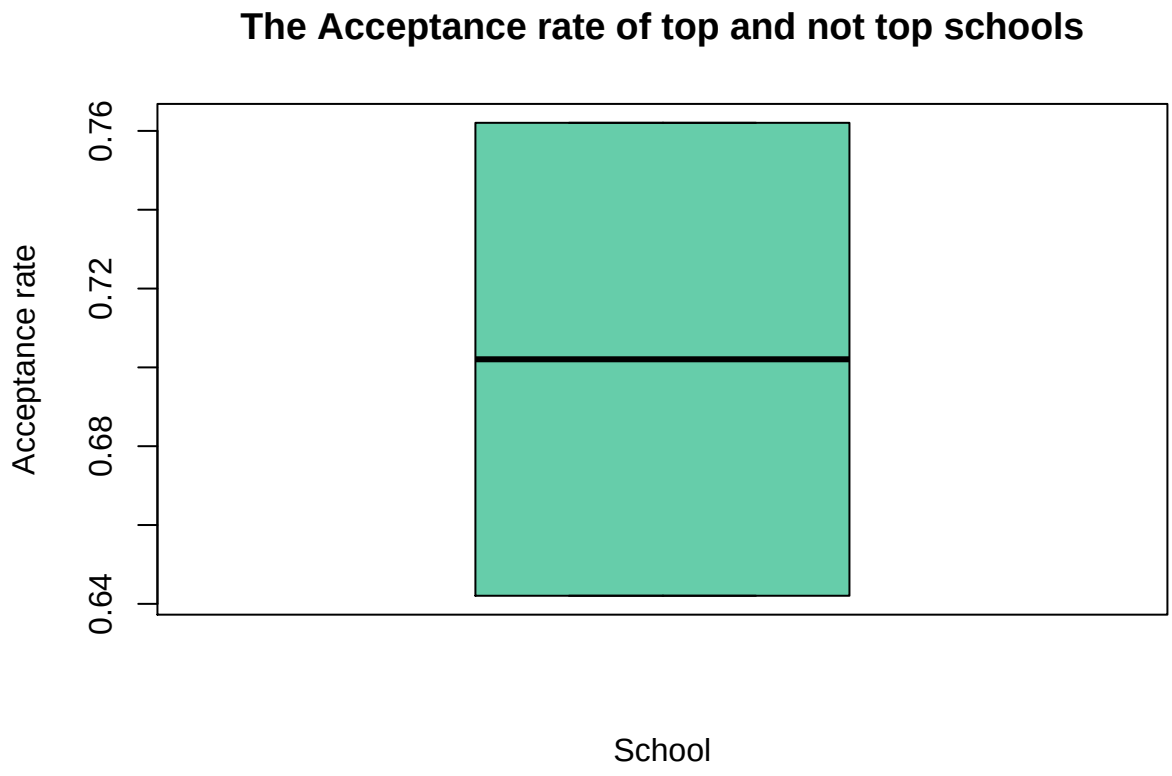



```
hist(unlist(OverallEnroll[2]), main = "Overall enrollment for private school",
     ylab = "School number", xlab = "Enrollment number", col="aquamarine3"
    )
```



e. Seperate the schools to two parts, top and ntop(not top)

```
data$Top <- data$Top25perc > 50
Accept_rate = c(sum(data[data$Top=='TRUE',]$Accept)/sum(data[data$Top=='TRUE',]$Apps),
                sum(data[data$Top!='TRUE',]$Accept)/sum(data[data$Top!='TRUE',]$Apps))
boxplot(Accept_rate,
        ylab="Acceptance rate", xlab="School",
        col=c("aquamarine3", "coral"),
        main="The Acceptance rate of top and not top schools"
        )
```



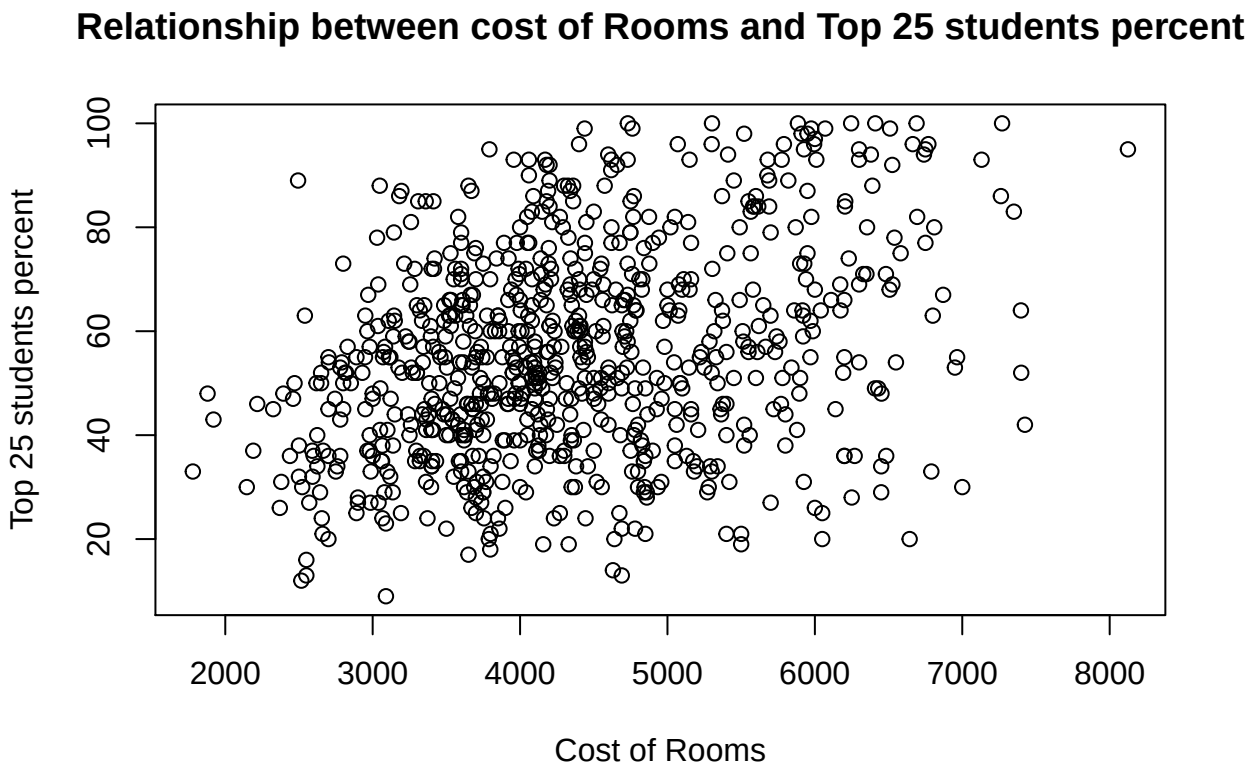

```
## [1] "The number of top universities:"
```

```
## [1] 449
```

f. Produce two new plots.

1. I want to explore the relationship between cost of Rooms and Top 25 students percentage.

```
plot(x = data$Room.Board,
     y = data$Top25perc,
     xlab = "Cost of Rooms",
     ylab = "Top 25 students percent",
     main = "Relationship between cost of Rooms and Top 25 students percent")
```



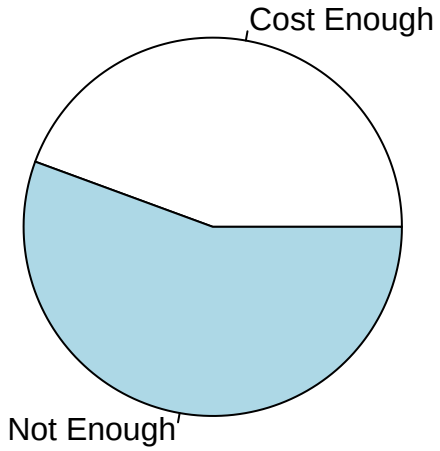
From this chart, we can get the conclusion that the cost of Rooms would let the colleges have more good students.

2. I want to build a pie chart to show how many schools have enough cost of rooms.

I want to set the mean of the cost as the bound.

```
bound = mean(data$Room.Board)
CostEnough = sum(data$Room.Board >= bound)
CostNotEnough = sum(data$Room.Board < bound)
pie(c(CostEnough, CostNotEnough), labels=c('Cost Enough', 'Not Enough'),
    main = "How many percentage of schools spend enough money on Rooms")
```

How many percentage of schools spend enough money on Rooms



I don't think the mean of the cost is good enough for bound, but I don't know how to compute the bounday from these data. Also, the cost of room and books must influence the percentage of good students, and I think I can get the relationship with furture study.

Part 2

```
Auto= read.csv("https://scads.eecs.wsu.edu/wp-content/uploads/2017/09/Auto.csv")
Auto = Auto[Auto$horsepower != '?',] #Moving out the missing data
```

a. In my opinion, 'mpg', 'cylinders', 'displacement', 'horsepower', 'weight', 'acceleration' and 'year' are quantitative variables.

'origin' and 'name' are the qualitive variables.

b. Get the range, mean and standard deviation of each quantitative predictor.

```
mpg = Auto$mpg
cylinders = Auto$cylinders
displacement = Auto$displacement
horsepower = Auto$horsepower
weight = Auto$weight
year = Auto$year
quantitative_predictor <- data.frame(
                    data_summary = c('range_down', 'range_up', 'mean', 'standerd_deviation'),
                    mpg = c(range(mpg)[1], range(mpg)[2], mean(mpg), sd(mpg)),
                    cylinders = c(range(cylinders)[1], range(cylinders)[2],
                                  mean(cylinders), sd(cylinders)),
                    displacement =c(range(displacement)[1], range(displacement)[2],
                                    mean(displacement), sd(displacement)),
                    weight = c(range(weight)[1], range(weight)[2],
```

```
                                mean(weight), sd(weight)),
                    year = c(range(year)[1], range(year)[2], mean(year), sd(year))
                    )
quantitative_predictor
```

```
##         data_summary       mpg cylinders displacement    weight      year
## 1         range_down  9.000000  3.000000       68.000 1613.0000 70.000000
## 2           range_up 46.600000  8.000000      455.000 5140.0000 82.000000
## 3               mean 23.445918  5.471939      194.412 2977.5842 75.979592
## 4 standerd_deviation  7.805007  1.705783      104.644  849.4026  3.683737
```

c. Remove 45th through 85th observation

```
Auto_new = rbind(Auto[1:44,],Auto[86:nrow(Auto),])
mpg = Auto_new$mpg
cylinders = Auto_new$cylinders
displacement = Auto_new$displacement
horsepower = Auto_new$horsepower
weight = Auto_new$weight
year = Auto_new$year
quantitative_predictor_new <- data.frame(
                    data_summary = c('range_down', 'range_up', 'mean', 'standerd_deviation'),
                    mpg = c(range(mpg)[1], range(mpg)[2], mean(mpg), sd(mpg)),
                    cylinders = c(range(cylinders)[1], range(cylinders)[2],
                                  mean(cylinders), sd(cylinders)),
                    displacement =c(range(displacement)[1], range(displacement)[2],
                                    mean(displacement), sd(displacement)),
                    weight = c(range(weight)[1], range(weight)[2], mean(weight),
                               sd(weight)),
                    year = c(range(year)[1], range(year)[2], mean(year), sd(year))
                    )
quantitative_predictor_new
```

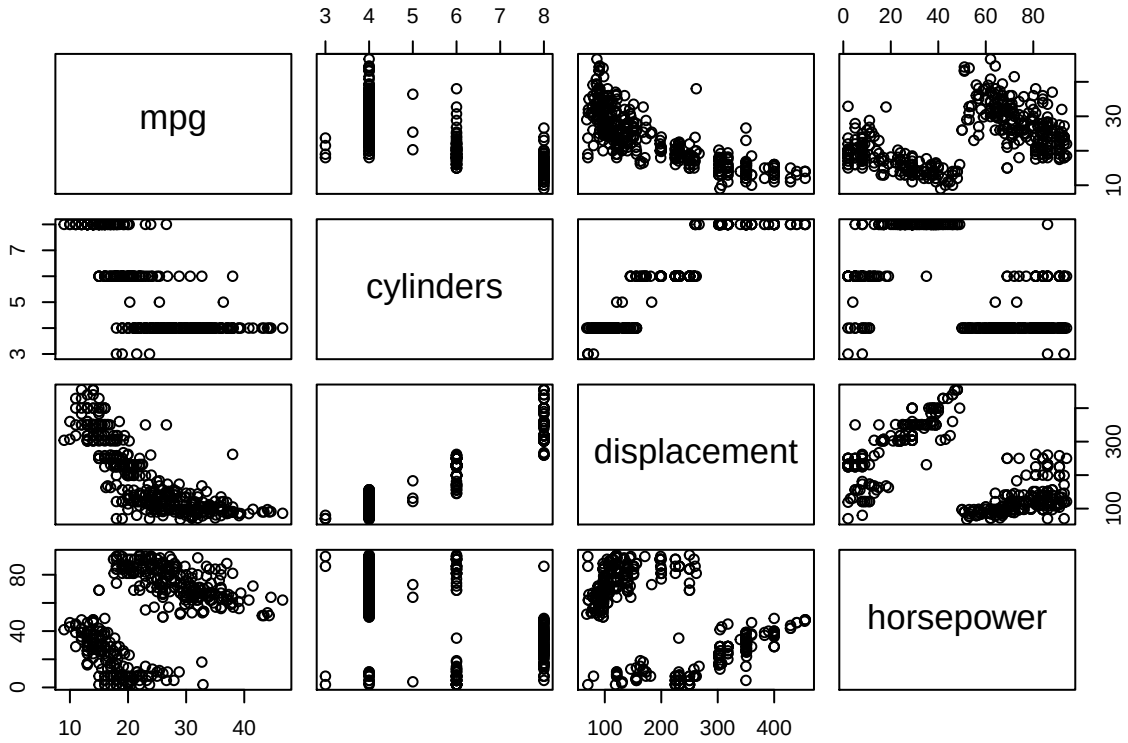
```
##         data_summary       mpg cylinders displacement    weight      year
## 1         range_down  9.000000  3.000000       68.0000 1649.0000 70.000000
## 2           range_up 46.600000  8.000000      455.0000 5140.0000 82.000000
## 3               mean 23.780057  5.470085      194.0484 2977.2336 76.475783
## 4 standerd_deviation  7.900879  1.683055      103.2051  835.3627  3.573531
```

d. Create some plots to show the relationship I find.

1. Build a scatter matrix with 'mpg', 'cylinders', 'displacement' and 'horsepower' to study the relationship between these four variables.

```
AutoGroup = data.frame(
                mpg = Auto$mpg,
                cylinders = Auto$cylinders,
                displacement = Auto$displacement,
                horsepower = Auto$horsepower
```

```
                )
library(graphics)
pairs(AutoGroup)
```



1. We can easily know, more cylinders will let the cars have more displacement. The 8 cylinders almost have twice displacement than 4 cylinders, it's easy to understand, since they have 4 more cylinders to contain the displacement. 2. Also, less cylinders will give the car higher horsepower in general, but not all. I think the horsepower would also be influenced by other variables. 3. According to the plot between displacement and horsepower, we can know the horsepower is increase with the displacement increase. But it is seperated to two part. With the relationship plot of cylinders and horsepower, we can finally know these two part may have different cylinders. The horsepower would increase with higher displacement, but less cylinders is more effective.

e. Which variables can be used to predict mpg.

From the scatter matrix created above, I think cylinders and horsepower can be used to predict mpg effectively. The displacement may have influence but not obviously, the change of mpg during the scatter is more likely caused by cylinders and horsepower. It is obviously, less cylinders can improve the mpg. Cars with 4 cylinders is almost all having higher mpg. In the relationship scatter of mpg and horsepower, we can know the mpg will increase with the decrease of horsepower. But it also not effective as the change of cylinders. Since the point of horsepower-mpg is seperated to two part, we can know the cylinders is the important variable in predictding the mpg.