Solutions to HW 1

Shaoyi Zhang

April 15th, 2016

Question 1.a

```
It is a data mining task. This is a classification problem.
It requires analysis on different attributes of a customer.
```

Question 1.b

```
It is NOT a data mining task. This is just calculation.
There's not much knowledge discovered in this problem.
```

Question 1.c

```
It is a data mining task.
This problem uses previous data to predict new data.
```

Question 1.d

```
It is NOT a data mining task.
There's no knew information or knowledege discovered in this activity.
```

Question 1.e

```
It is NOT a data mining task.
If the dies fair, then all sides have same probability.
```

Question 2.a

From the data description we know that this data set contains 506 observations and 13 variables.

The description of the variables and their corresponding unit is listed bellow:

```
1. CRIM      per capita crime rate by town

2. ZN        proportion of residential land zoned for lots over
             25,000 sq.ft.

3. INDUS     proportion of non-retail business acres per town

4. CHAS      Charles River dummy variable (= 1 if tract bounds
             river; 0 otherwise)

5. NOX       nitric oxides concentration (parts per 10 million)
```

```
6. RM        average number of rooms per dwelling

7. AGE       proportion of owner-occupied units built prior to 1940

8. DIS       weighted distances to five Boston employment centres

9. RAD       index of accessibility to radial highways

10. TAX      full-value property-tax rate per $10,000

11. PTRATIO  pupil-teacher ratio by town

12. B        1000(Bk - 0.63)^2 where Bk is the proportion of blacks
             by town

13. LSTAT    % lower status of the population

14. MEDV     Median value of owner-occupied homes in $1000's
```

```
library(data.table)
library(ggplot2)
library(gridExtra)
library(class)
library(ISLR)
setwd("/Users/Shawn/Desktop/PSTAT 231/assign1/")
getwd()
```

```
## [1] "/Users/Shawn/Desktop/PSTAT 231/assign1"
```

```
houseData = read.table("housing.data")
Boston.Housing = as.data.table(houseData)
setnames(Boston.Housing,c("Crime.Rate","ResiLand.Zoned","NonRetail.Bus","Charles.River","Nitr.Oxide","A

##  [1] "Crime.Rate"   "ResiLand.Zoned" "NonRetail.Bus"  "Charles.River"
##  [5] "Nitr.Oxide"   "Avg.Rooms"      "Age"            "Wigh.Dist"
##  [9] "Access.Idex"  "Tax"            "Pupil.Teacher"  "Blck"
## [13] "Lower.Sts"    "Med.Value"
nrow(Boston.Housing)
```

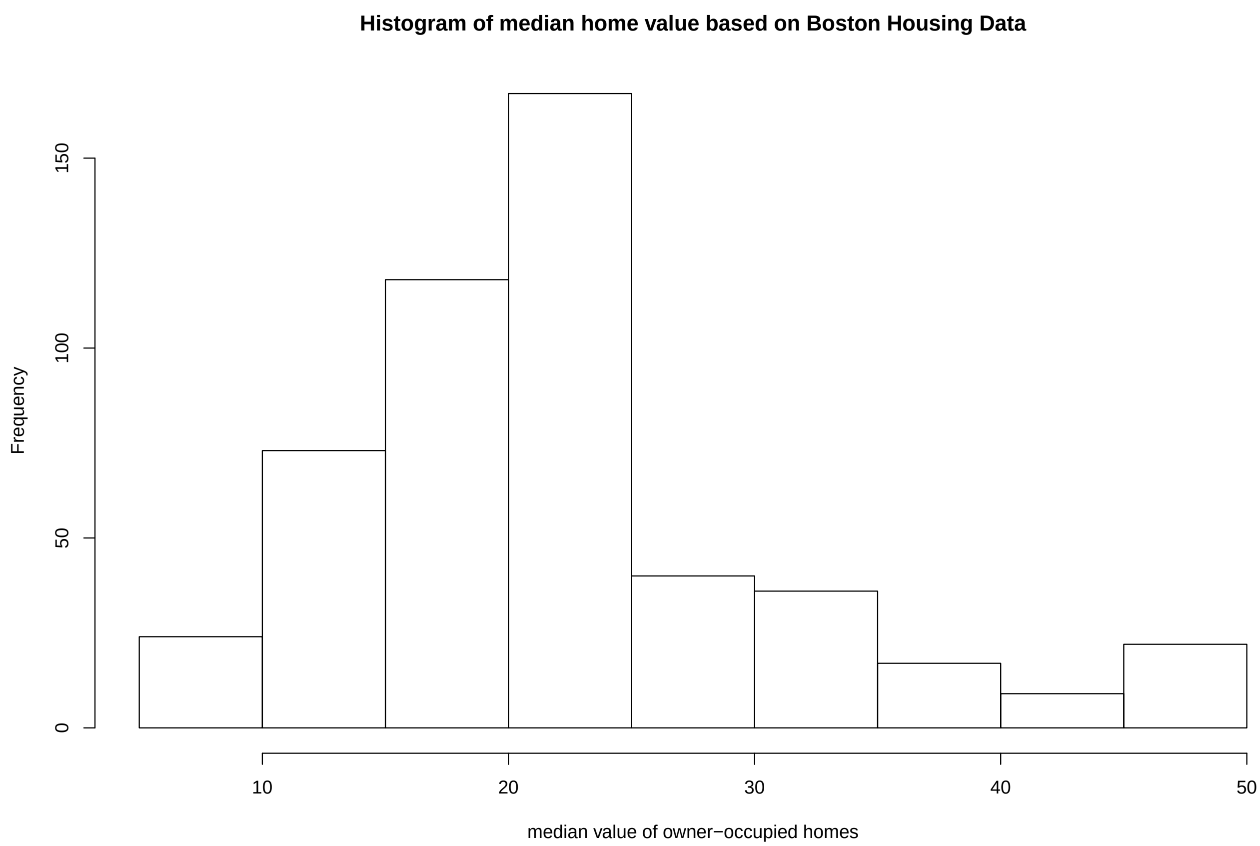
```
## [1] 506
```

```
summary(Boston.Housing)
```

```
##    Crime.Rate       ResiLand.Zoned   NonRetail.Bus   Charles.River
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##    Nitr.Oxide       Avg.Rooms          Age            Wigh.Dist
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
```

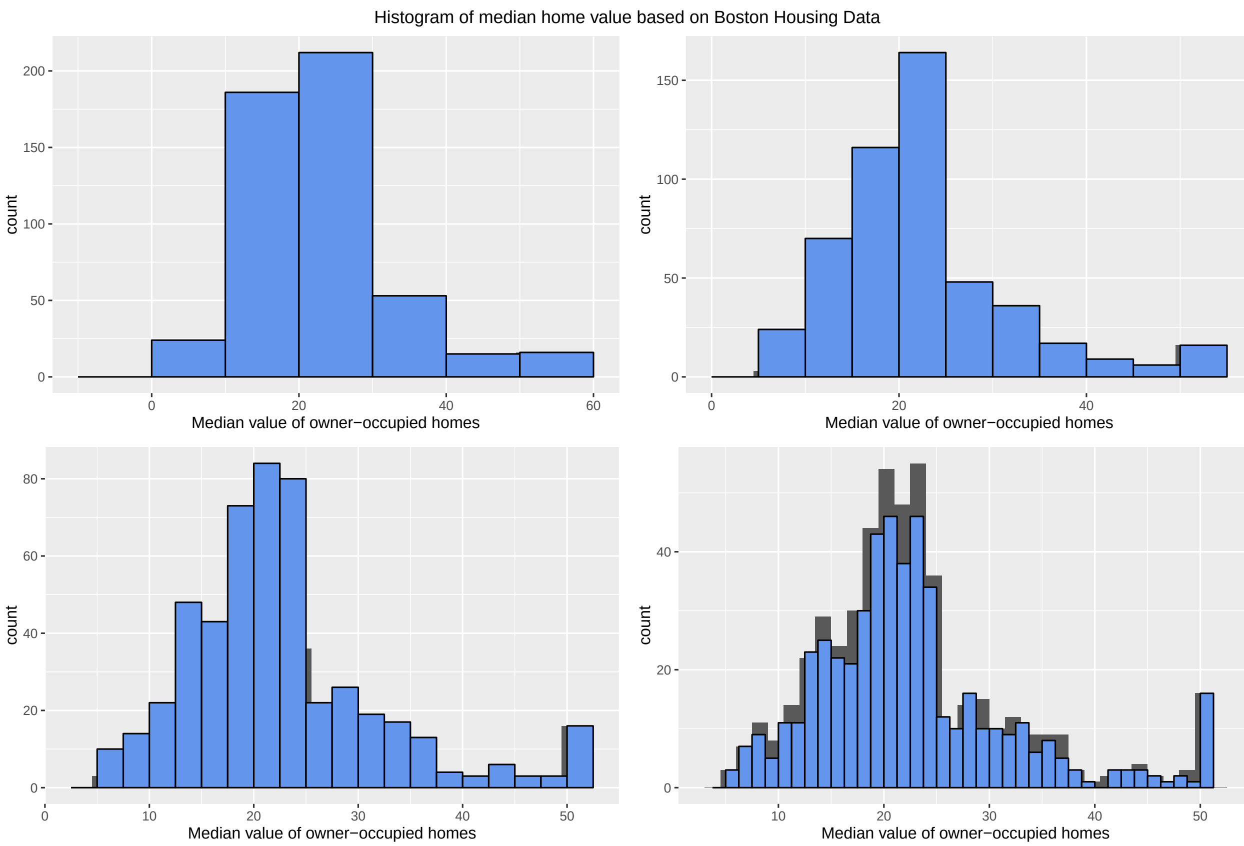
```
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##   Access.Idex          Tax        Pupil.Teacher        Blck
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##    Lower.Sts        Med.Value
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

```r
hist(Boston.Housing$Med.Value,xlab = "median value of owner-occupied homes", main = "Histogram of media
```




```r
hist1 = qplot(Boston.Housing$Med.Value)+geom_histogram(binwidth=10,color="black",fill="cornflowerblue")

hist2 = qplot(Boston.Housing$Med.Value)+geom_histogram(binwidth=5,color="black",fill="cornflowerblue")+
```

```r
hist3 = qplot(Boston.Housing$Med.Value)+geom_histogram(binwidth=2.5,color="black",fill="cornflowerblue")

hist4 = qplot(Boston.Housing$Med.Value)+geom_histogram(binwidth=1.25,color="black",fill="cornflowerblue

grid.arrange(hist1,hist2,hist3,hist4,top="Histogram of median home value based on Boston Housing Data")
```



Question 2.a

As we gradually increase the number of bins, binwidth decrease and the histgram looks more similar to the probablity distribution graph of median value of owner-occupied homes.

Question 2.e

```r
mean(Boston.Housing$Med.Value)
```

```
## [1] 22.53281
```

```r
median(Boston.Housing$Med.Value)
```

```
## [1] 21.2
```

```r
sd(Boston.Housing$Med.Value)
```

```
## [1] 9.197104
```

```r
IQR(Boston.Housing$Med.Value)
```

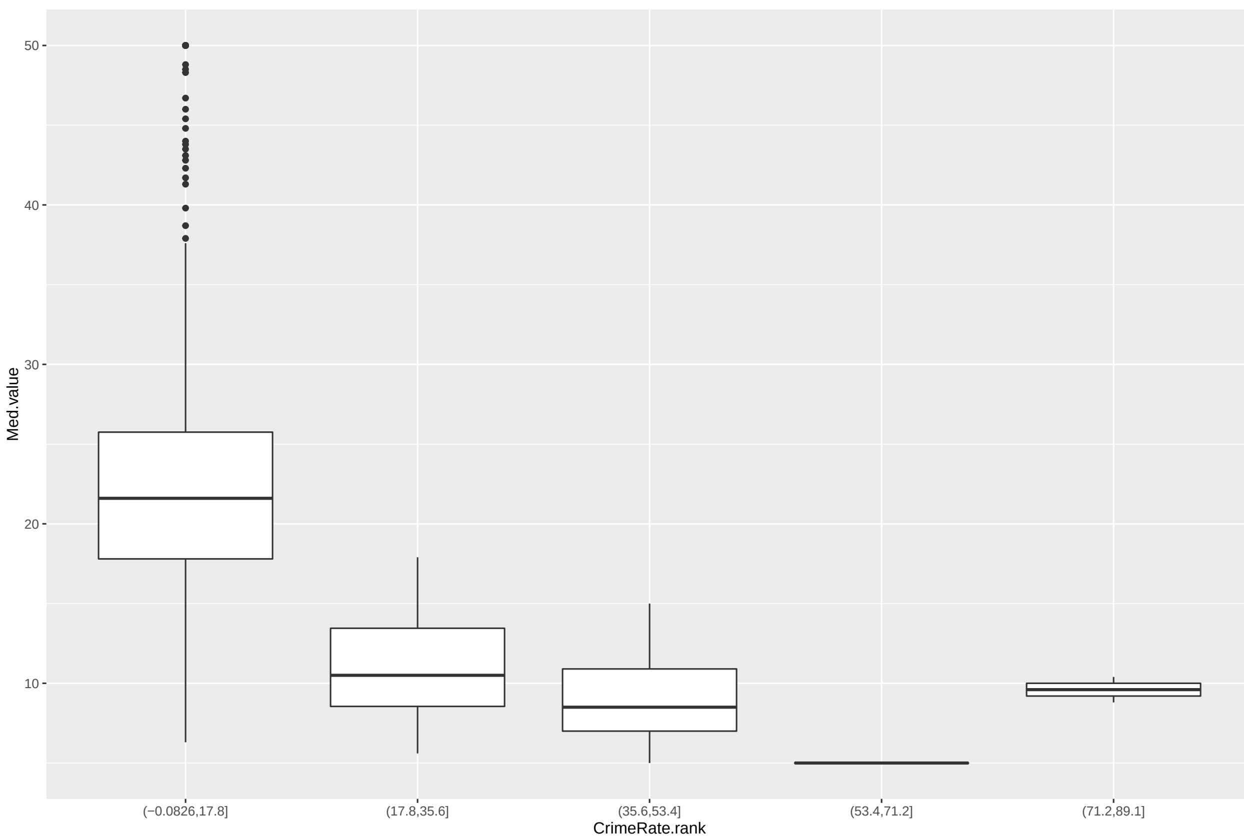
```
## [1] 7.975
```

Median probably better decribes the median home value. Since the standard deviation is quite large, mean may not be useful.

Take the median of median make sense. Because the original median is the median of a particular town. The new median is the median of those town's median.

Question 2.f

```r
CrimeRate.rank = cut(x = Boston.Housing$Crime.Rate,breaks = 5)
df = data.frame(rank = CrimeRate.rank, Med.value = Boston.Housing$Med.Value)
ggplot(data = df, aes(x = CrimeRate.rank, y = Med.value))+geom_boxplot()
```



Question 3.a

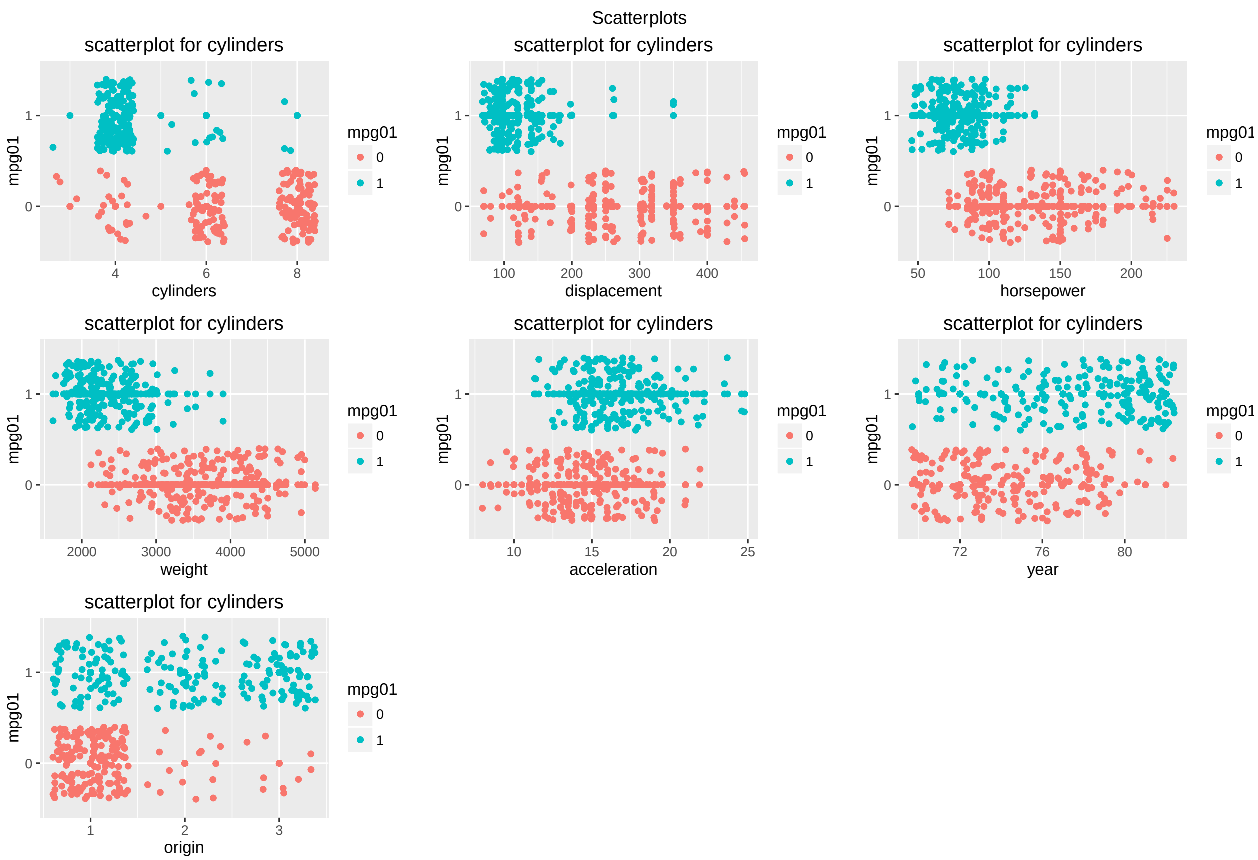
```r
auto = data.table(as.data.frame(Auto))
medMPG = median(auto$mpg)
#mpgtemp = Auto$mpg
#mpgtemp[<medMPG] = 1
```

```r
#auto[]
auto$mpg01 = Auto$mpg
auto$mpg01[auto$mpg < medMPG] = "low"
auto$mpg01[auto$mpg > medMPG] = "high"
auto$mpg01 = factor(auto$mpg01, levels = c("low","high"),labels = c(0,1))

# create dataset with mpg01
autoNoLabel = subset(auto,select = -mpg01)
#for (xCol in names(autoNoLabel)){
#  ggplot(data = autoNoLabel, aes(x=xCol,y=mpg01))+geom_boxplot()
#}

#box0 = ggplot(data = auto, aes(x = mpg, y = mpg01))+geom_boxplot()
```

Questiong 3.b

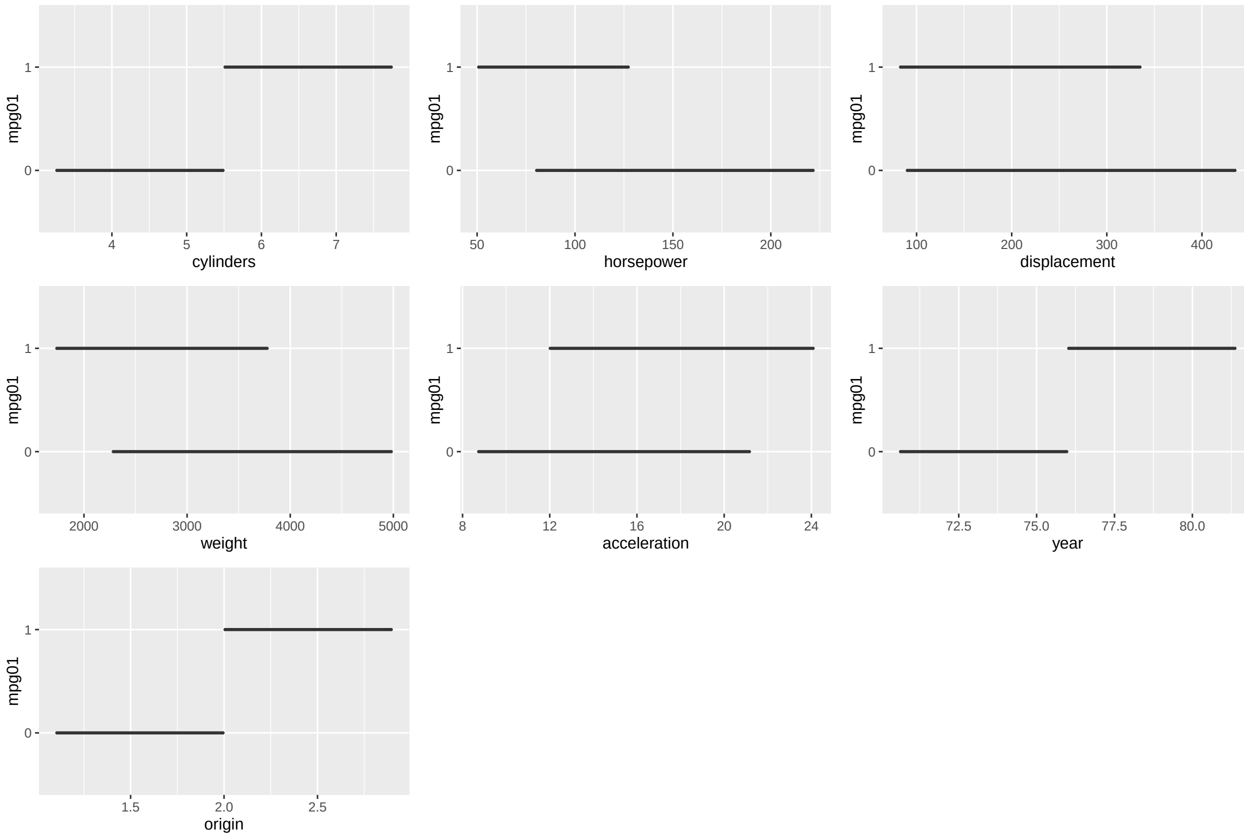
I don't think name will have any thing to do with mpg. Even if the brand is associated with mpg, brand will highly correlated with origin

Scatterplot



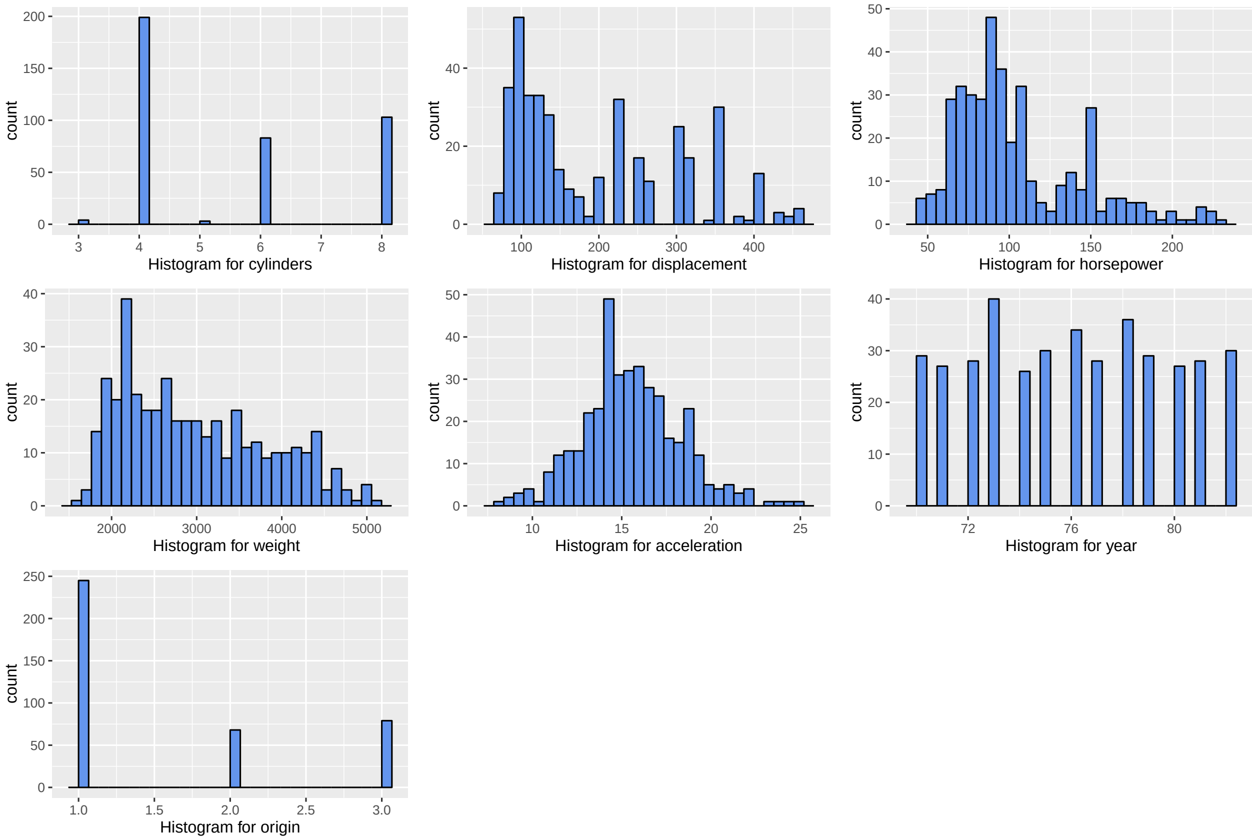
From the scatterplot, we notice that variable year,acceleration cannot classify mpg01 well.

Boxplot



According to boxplots, cylinders, year and origin is highly associated with mpg01

Histogram



In summary, we will use cylinders, horsepower and origin for further association analysis.

Question 3.c

```r
set.seed(1)
numOfObs = dim(auto)[1]
numOfObs
```

```
## [1] 392
```

```r
index.train = sample(1:numOfObs,floor(numOfObs*0.75),replace = F)
index.test = setdiff(1:numOfObs,index.train)
train.set = auto[index.train, ]
test.set = auto[index.test, ]
class.train=auto[index.train,"mpg01"]
class.test=auto[index.test,"mpg01"]
dim(train.set)
```

```
## [1] 294  10
```

```r
dim(test.set)
```

```
## [1] 98 10
```

Question 3.d

```r
require(class)
vars = c("cylinders","horsepower","origin")
newdf = data.frame(auto)
X = newdf[,vars]
responseY = as.matrix(as.numeric(newdf[,"mpg01"]))

#X = data.frame(subset(auto,select = vars))
#responseY = as.matrix(auto[,mpg01])
str(responseY)
```

```
##  num [1:392, 1] 1 1 1 1 1 1 1 1 1 1 ...
```

```r
dim(X)
```

```
## [1] 392   3
```

```r
dim(responseY)
```

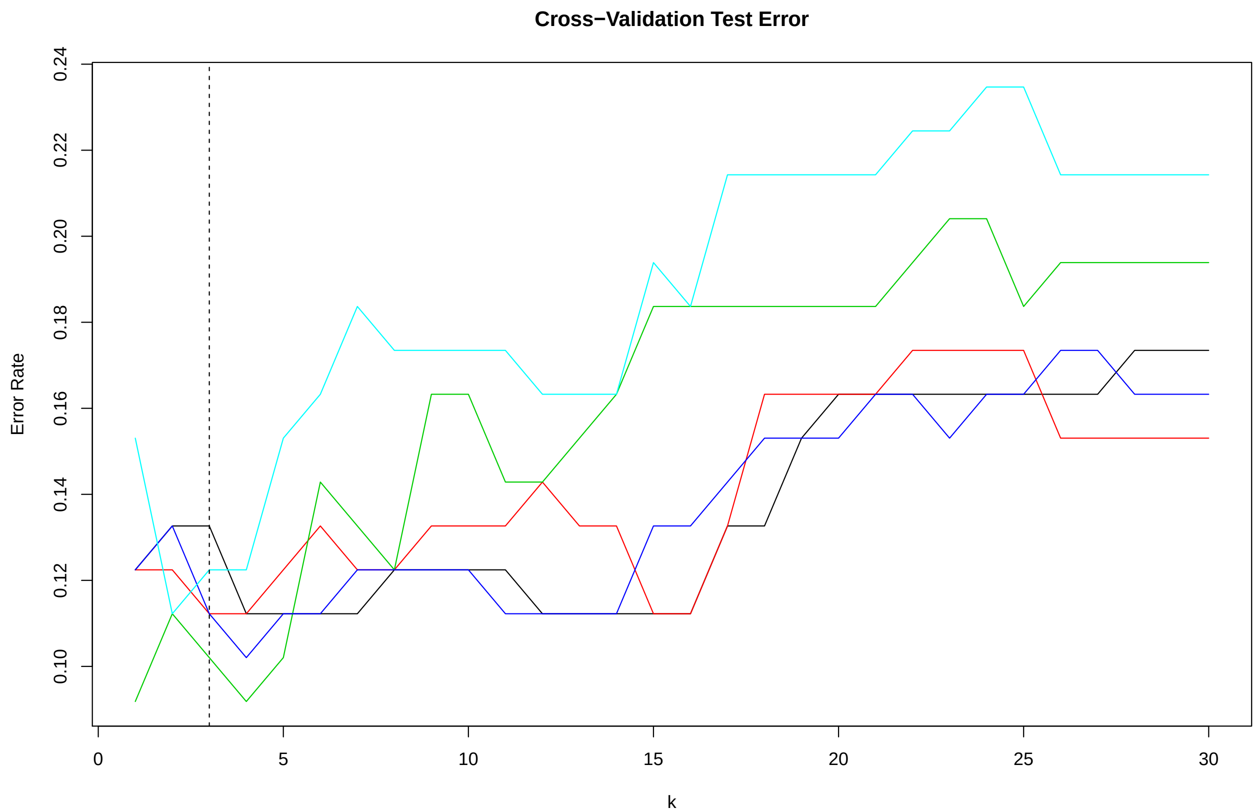
```
## [1] 392   1
```

```r
MiscError <- function(X,responseY,m,n){
  # Args:
  # X: dataset with explanatory variables
  # responseY : lables
  # m: max value for nearest neighbors
  # n: Number of times cross-validation is conducted

  error.cv <- list()
  for(i in 1:n){

    # Training data
    index <- sample(dim(X)[1],size = floor(dim(X)[1]*0.75), replace = F)
    train.set <- X[index,]

    # Test data
    index.test <- setdiff(1:dim(X)[1],index)
    test.set <- X[index.test,]

    # Vector of classes
    class.train <- responseY[index,]
    class.test <- responseY[index.test,]

    # For this given samples fit k-NN model for several values of k
```

```r
    knn.error <- vector()  # initialize vector
    for (j in 1:m){  # m: Maximum number of values of k
      model.knn <- knn(train = train.set,
                       test = test.set,
                       cl = class.train,
                       k=j,
                       prob=T)  # Fit model
      error <- table(model.knn,class.test)
      # Compute Error
      knn.error[j] <- (error[1,2] + error[2,1])/sum(error)
    }
    error.cv[[i]] <- knn.error
  }
  return(error.cv)
}
CrossValid <- MiscError(X,responseY,m=30,n=5)
class(CrossValid)
```

```
## [1] "list"
```

```r
names(CrossValid) <- paste("Sample",1:5) # Assign names

# Plot error curves
matplot(data.frame(CrossValid), type = "l", lty=1,
        ylab = "Error Rate",
        xlab = "k",
        main = "Cross-Validation Test Error")
abline(v=3, lty=2)
```



K = 3 seems to perform the best on this data set.

Question 4.a

Definition of KDD: KDD stands for "knoledge discovery in databases". KDD covers the overall process of discovering useful knowledge from data. The goal of KDD is to extract high-level, interperable and useful data from low-level, tedious, noisy data. KDD process include but not limited to obtain and process data, exploratory analysis and hypothesis selecting, data mining and interpretation, and finally make use of discovered knowledge. Moreover, KDD emphasis on automating the whole knowledge discovery process.

Definition of Data Mining: Data Mining is a process of discovering knowlege from data. Data Mining involves fitting model, or finding pattern from, observed data. Data obtaining, cleaning and preprocessing are not considered as parts of Data Mining.

Relation: Data Mining is a component of KDD. In my opinion, other parts in KDD process are preparation or results of Data Mining step.

Question 4.b

Since we are in the "Big Data" era, the authors suggests that the amonut of data is overloaded. It is slow, expensive and even impossible for manual data probing of some large data set. Thus, the substantial amount of information(data) explains the necessity of KDD.

Question 4.c

The author mentioned FAIS – Financial Crimes Enforcement Network. The objective of FAIS is to discover previously unknown, potentially high-value leads for possible investigation. In this system, designers integrated

AI algorithm to predict potential crime so that the police and FBI can react faster. This system also has special hardware setting.