Doug Woodward

CS613 HW 1

|  |  |
| --- | --- |
| # | Answer |
| 1 | There are 78 Elite Colleges. See below for rest of answer |
| 2 | 1. Quantitative: mpg cylinders displacement horsepower weight acceleration Qualitative: year origin name (year could be said to also be Quantitative) 2. | mpg | cylinders | displacement | horsepower | weight | acceleration   r| 9.0-46.6 | 3-8 | 68-455 | 46-230 | 1613-5140 | 8.0-24.8   1. | mpg | cylinders | displacement | horsepower | weight | acceleration   v |6.091814e+01 | 2.909696e+00 | 1.095037e+04 | 1.481569e+03 | 7.214847e+05 | 7.611331e+00  sd|7.805007 | 1.705783 | 104.644004 | 38.491160 | 849.402560 | 2.758864   1. | mpg | cylinders | displacement | horsepower | weight | acceleration   r | 11-46.6 | 3-8 | 68-455 | 46-230 | 1649-4997 | 8.5-24.8  v |6.189414e+01 | 2.736307e+00 | 9.935777e+03 | 1.275122e+03 | 6.582080e+05 | 7.256131e+00  sd| 7.867283 | 1.654179 | 99.678367 | 35.708853 | 811.300208 | 2.693721|  e) Note that MPG is negatively correlates to MPG, as are cylinders. Additionally, horsepower and acceleration are strongly positively correlated, with engine displacement being nearly identical to horsepower in this case. Japanese clearly make the most efficient cars, with American cars being largely inefficient.  f) Yes, Weight is a heavily contributing factor, as the loess line shows that MPG decreases by about 1 for each 100 pounds gained. The origin of the car also matters as noted above, as do cylinders. Indeed, it would be very possible to create a model to predict MPG based on other factors. |
| 3 | 1. The ‘Boston’ data frame has 506 rows and 14 columns. Each row is a suburb and columns are:    1. ‘crim’ per capita crime rate by town.    2. ‘zn’ proportion of residential land zoned for lots over 25,000 sqft    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’ nitrogen 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 mean of distances to five Boston employment centers    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. ‘black’ 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town    13. ‘lstat’ lower status of the population (percent).    14. ‘medv’ median value of owner-occupied homes in \$1000s. 2. Notable findings from pairs: Median value of homes is closely tied to room number with a few interesting outliers. Lower class neighborhoods have less rooms per house. NO2 ppm is highest near urban centers and taxes seem to have low correlation with any other predictor. 3. Yes - the data shows that per capita crime rate is highest in areas with older buildings and in urban centers. It also can be seen to be lowest in neighborhoods where the population is well mixed, being highest when the black population is either on the extremes of high or low. 4. PT ratios are mostly standardized across boston at 20:1, with the exception of a single suburb having a 22:1 ratio, and then a smattering of lower ratios, indicating private schools most likely. A similar group of neighborhoods with the 20:1 PT ratio have a much higher tax rate of about 650, indicating these are likely the urban neighborhoods, all similarly run and close together. In this same set(neighborhoods ~370 - 500) we find notably higher crime rates, and one single outlier at nearly 100% crime rate, with a few others above 50%, again indicating the urban center. 5. 35 6. 19.05 7. Neighborhood 399 has the lowest medv.   crim zn indus chas nox rm age dis rad tax ptratio black lstat medv  38.3518 0 18.1 0 0.693 5.453 100 1.4896 24 666 20.2 396.9 30.59 5  This is one of the oldest neighborhoods, and has a high crime rate, but it is not the highest. It is also very urban.  h) 64 avg more than 7 rooms, 13 avg more than 8. These are older neighborhoods, median age of 78.3, they are fairly far from urban areas, Mean 3.430, and they are very low in crime, with median and mean bothe < 1. |
| 4 | 1. The first model will generalize much better. The second is clearly overfit to this particular data set. 2. The rejected model, Model 2, is clearly overfitting. |
| 5 | Average House price of training: $540683.10  Average house price of testing: $538700.10 |
| 6 | 1. See code 2. [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]  X75 1 1 2 11 18 23 18 12 6 8  X0 32 0 0 0 0 0 0 0 0 68  X190 1 0 0 0 4 23 46 17 7 2  X80 1 0 0 1 22 27 27 11 7 4  X91 30 35 21 9 1 0 2 0 0 2 |
| 7 | Code:  data = read.csv("data/RawDataUSCities.csv")  data = data[,-(1:2)]  data = apply(data,2, function(x)(x - mean(x))/sd(x))  data[1:6,] data = apply(data, 2, function(x)(x-min(x))/(max(x)-min(x))) data[1:6,] ~ |

8.

Code:

-----------

library(Hmisc)

college <- read.csv("data/College.csv")

rownames(college)=college[,1]

college = college[,-1]

latex(summary(college))

pairs(college[,1:10])

# force new plot window

dev.new()

# split plot window

par(mfrow=c(3,3))

attach(college)

plot(as.factor(Private), Outstate, xlab="Private", ylab="OutState", main="Private/Out of State Tuition")

Elite = rep("No", nrow(college))

Elite[college$Top10perc > 50] = "Yes"

Elite = as.factor(Elite)

college = data.frame(college, Elite)

eliteSummary = summary(college)

latex(eliteSummary)

plot(Elite, Outstate, xlab="Elite", ylab="Outstate", main="Elite Colleges Out of State Tuition")

hist(Apps)

hist(Grad.Rate, breaks=50)

hist(Books, breaks=50)

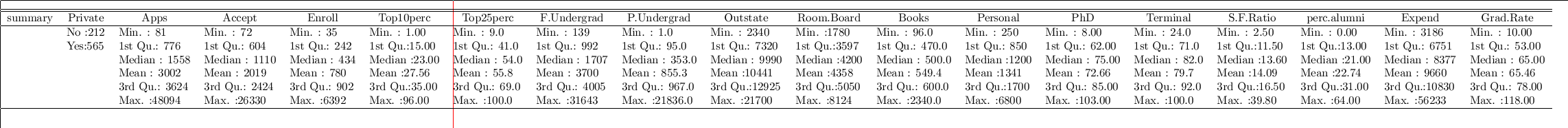
hist(Accept, breaks=200)

~

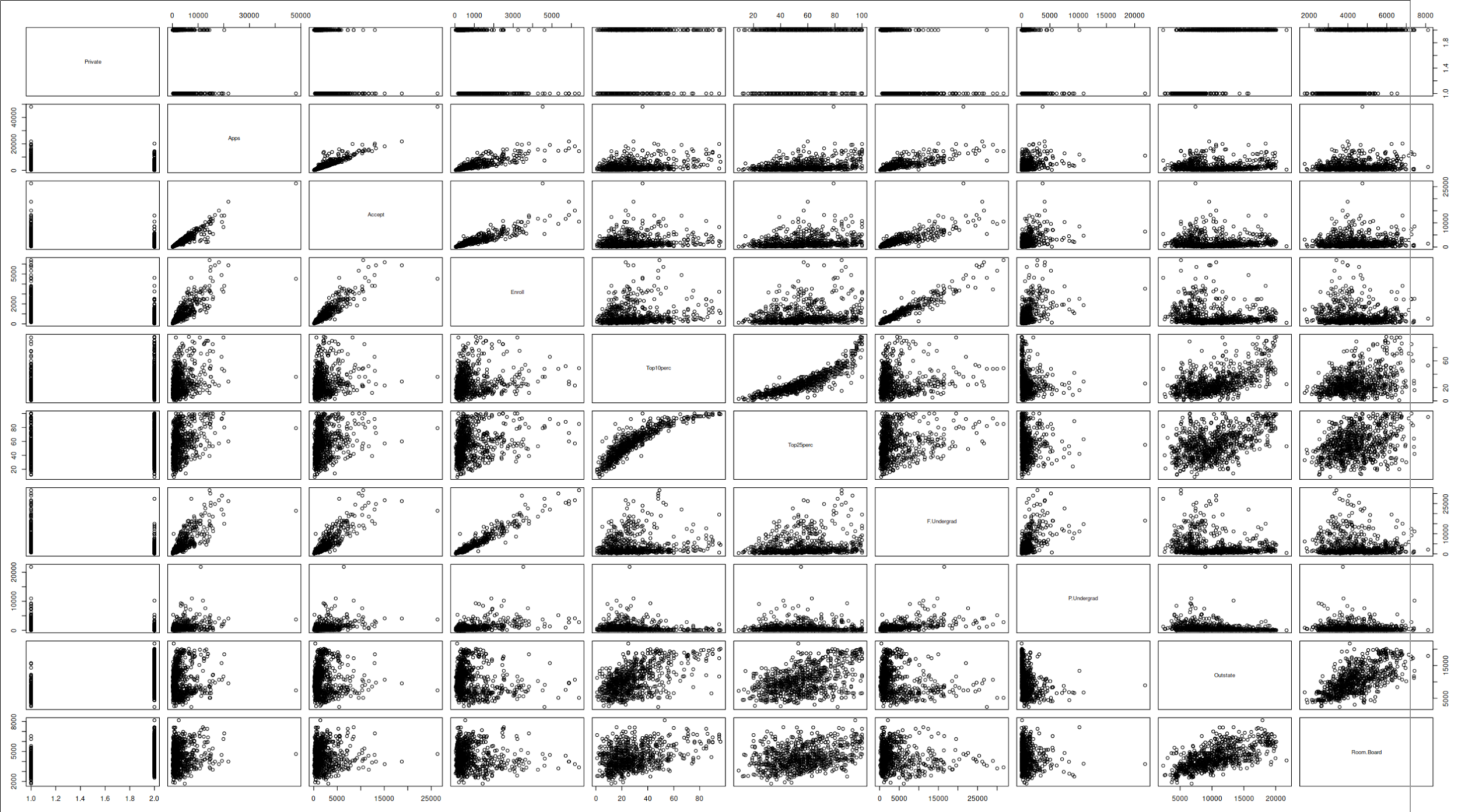
---------

Output:

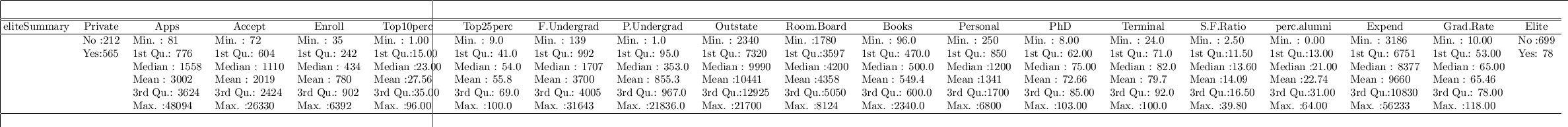
Summary:



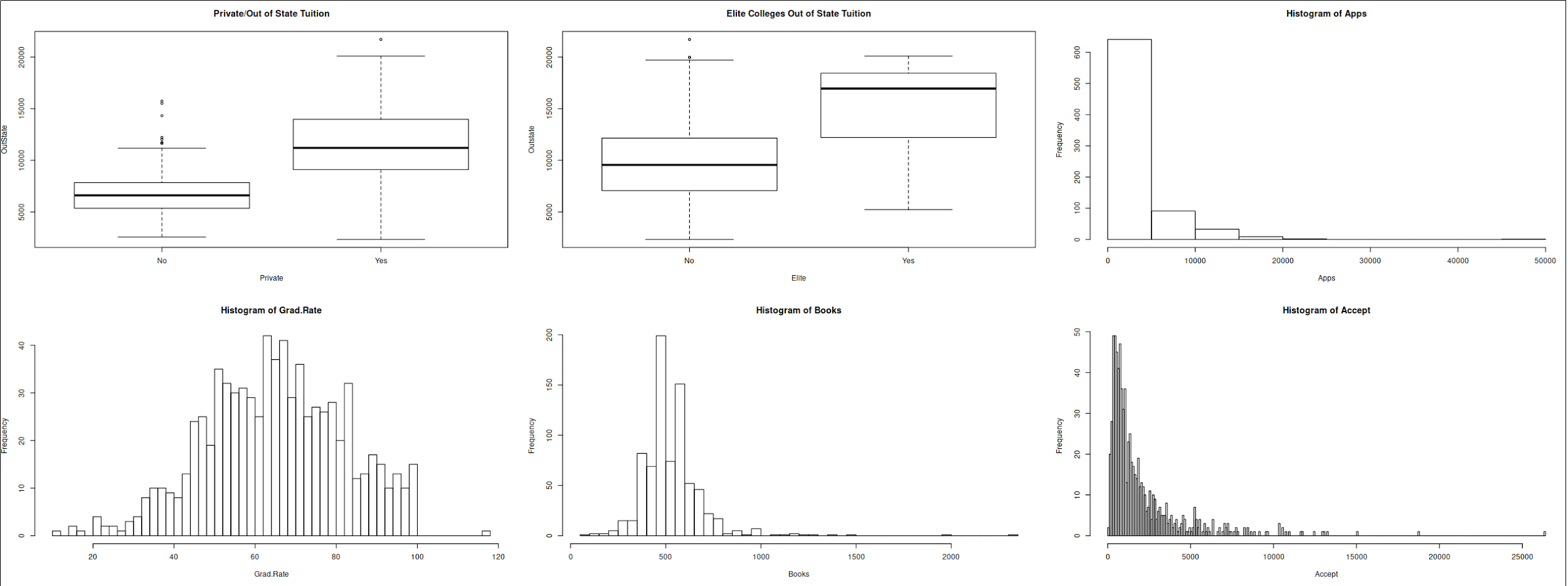
Pairs:



EliteSummary:



Charts and Histograms:



9.

Code:

library(ISLR)

Auto = na.omit(Auto)

summary(Auto)

sapply(Auto[1:6], range)

sapply(Auto[1:6], var)

sapply(Auto[1:6], sd)

Auto = Auto[-(10:85),]

sapply(Auto[1:6], range)

sapply(Auto[1:6], var)

sapply(Auto[1:6], sd)

pairs(Auto)

par(mfrow=c(2,2))

attach(Auto)

plot(weight, mpg, main="Weight to MPG")

lines(lowess(weight,mpg), col="blue") # lowess line (x,y)

plot(cylinders, mpg, main="Cylinders to MPG")

lines(lowess(cylinders, mpg), col="blue") # lowess line (x,y)

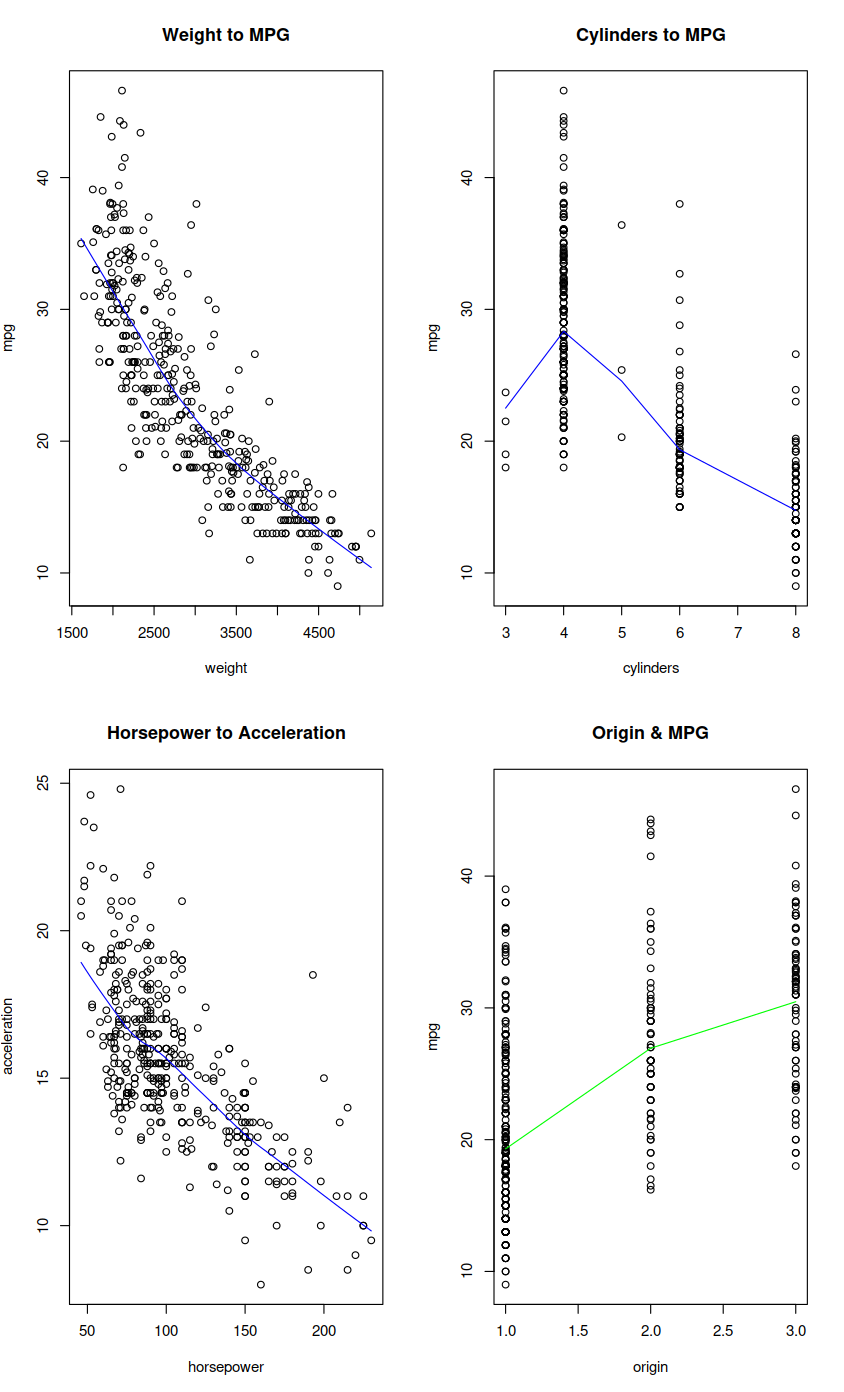
plot(horsepower,acceleration, main="Horsepower to Acceleration")

lines(lowess(horsepower, acceleration), col="blue") # lowess line (x,y)

plot(origin, mpg, main="Origin & MPG")

lines(lowess(origin, mpg), col="green") # lowess line (x,y)

Graphs:



10.

Code:

library(MASS)

?Boston

attach(Boston)

par(mfrow=c(3,3))

plot(dis, medv, main="Distance to Home Values")

plot(rm, medv, main="Rooms to Home Value")

plot(nox, indus, main="NO^2 to Industry")

plot(ptratio, crim, main="P/T v Crime")

plot(ptratio, medv, main="P/T v Home Value")

plot(ptratio, lstat, main="P/T v Lower Class")

pairs(Boston)

plot(crim)

plot(ptratio)

plot(tax)

table(Boston$chas)

summary(Boston$ptratio)

which.min(Boston$medv)

Boston[399,]

beantown <- subset(Boston,rm >= 7)

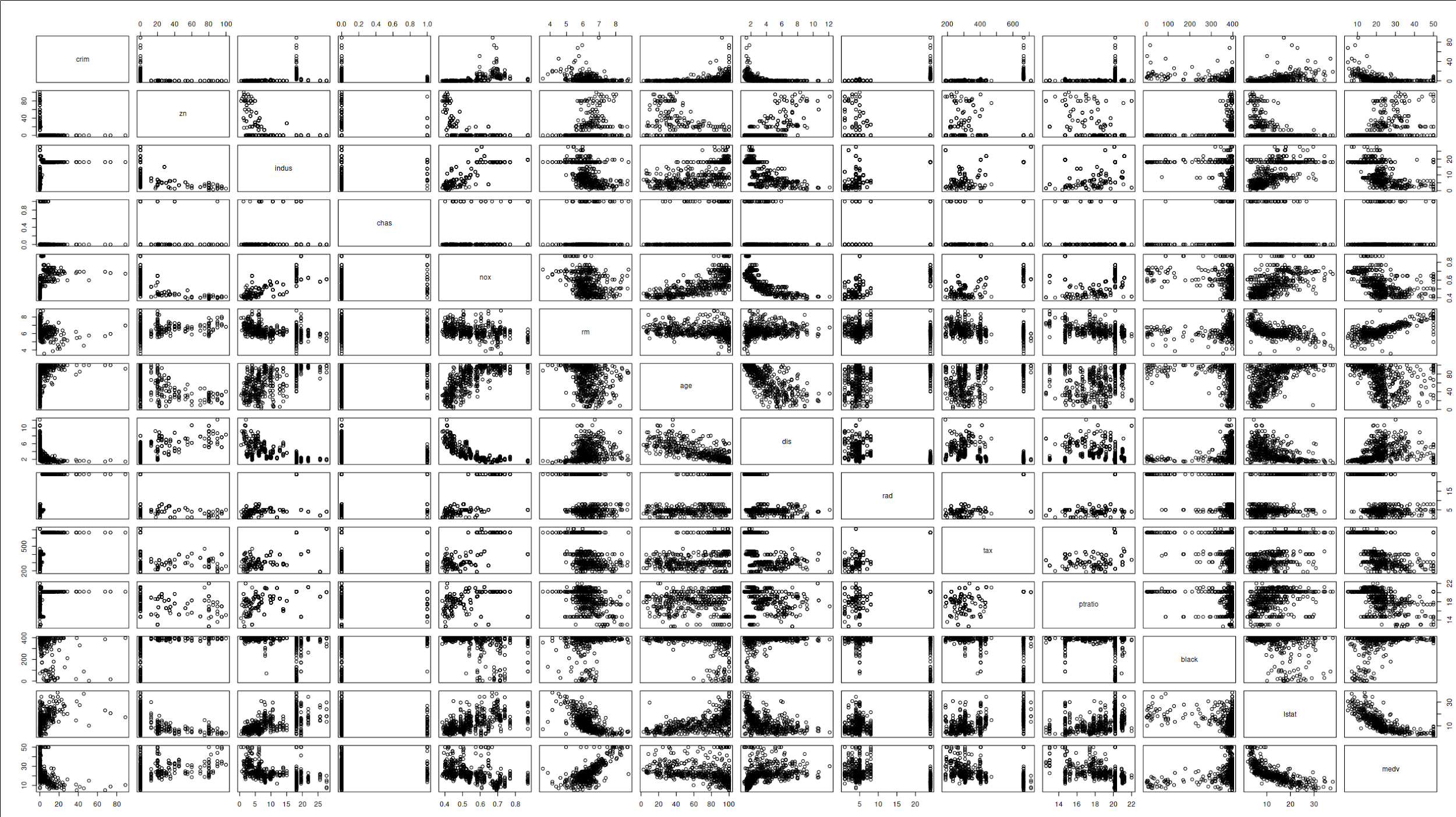
dim(beantown)

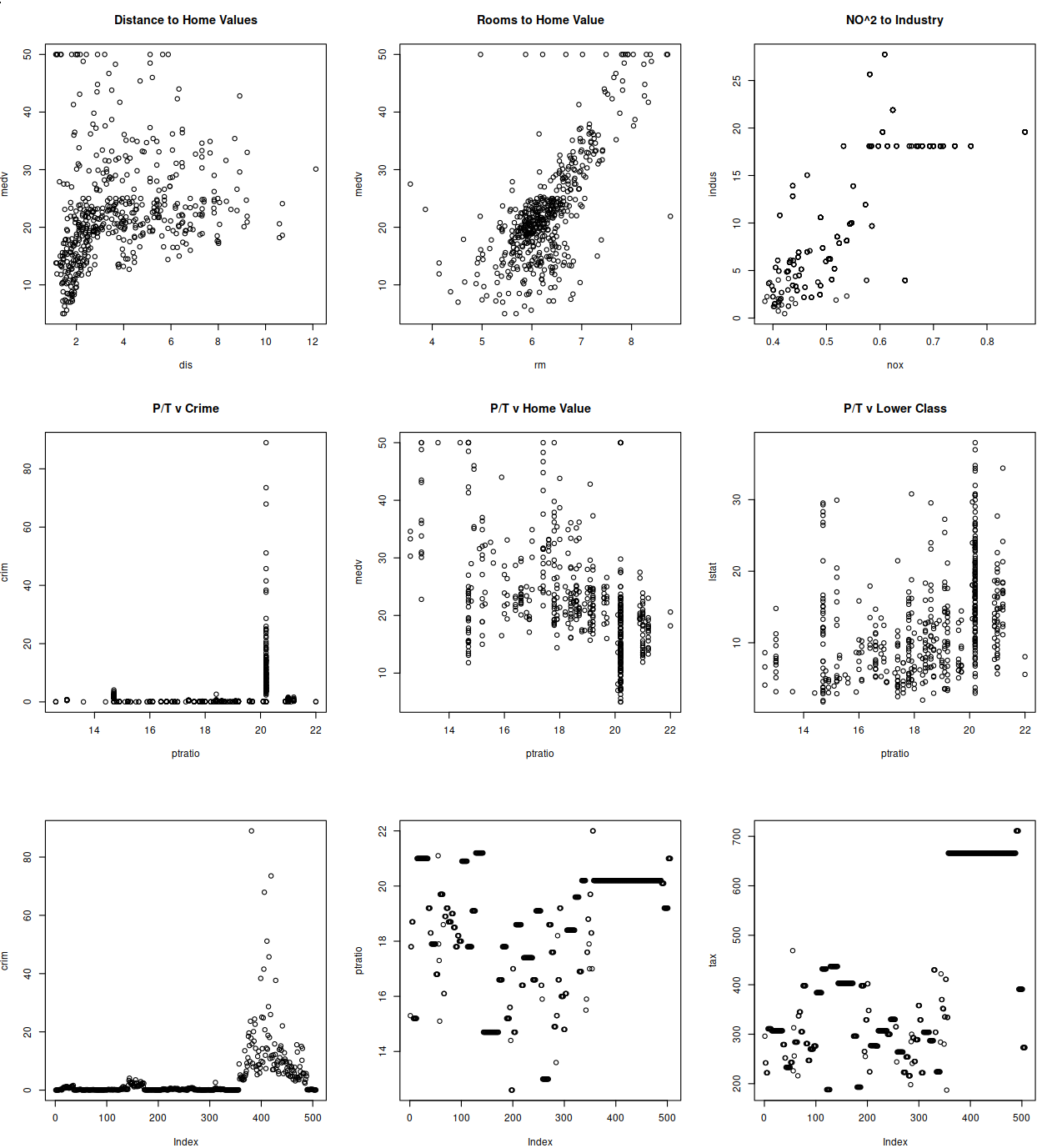
beantown <- subset(Boston,rm >= 8)

dim(beantown)

summary(beantown)

Graphs:





5)

Code:

# read data from csv  
data = read.csv("data/kc\_house\_data.csv")  
  
invisible((nData = nrow(data)))  
  
invisible(set.seed(0))  
  
invisible(trainIdx <- sample(seq(1, nrow(data)), floor(nrow(data) \* 0.70)))  
  
mean(trainSet <- data$price[trainIdx])  
mean(testSet <- data$price[-trainIdx])

Output:

[1] 540683.1   
[1] 538700.1

6) Code:

data = read.csv("data/arrythmia.csv")  
subset = data[1:100, 1:5]  
  
print(subset)  
  
# mean -> 0 and stddev = 1  
subset = scale(subset)  
  
t(apply(subset,2,function(x)table(cut(x, 10))))

Output: (sans the large normalized data set)

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]   
X75 1 1 2 11 18 23 18 12 6 8   
X0 32 0 0 0 0 0 0 0 0 68   
X190 1 0 0 0 4 23 46 17 7 2   
X80 1 0 0 1 22 27 27 11 7 4   
X91 30 35 21 9 1 0 2 0 0 2

7) Code:

data = read.csv("data/RawDataUSCities.csv")  
  
data = data[,-(1:2)]  
  
data = apply(data,2, function(x)(x - mean(x))/sd(x))   
data[1:6,]  
data = apply(data, 2, function(x)(x-min(x))/(max(x)-min(x)))  
data[1:6,]

Output:

PercentageBlack PercentageHispanic PercentageAsian MedianAge UnemploymentRate PerC   
[1,] -1.17872113 1.2389537 -0.36257405 0.06134197 -0.7514633   
[2,] 2.35518849 -0.7644344 -0.45230197 -0.43961742 -0.7514633   
[3,] -0.68176509 0.5104489 -0.27284613 -1.44153619 -1.4953360   
[4,] 1.91344978 -0.8251431 -0.45230197 0.56230135 1.4801550   
[5,] 0.09127764 -0.2180558 -0.09339029 -0.94057681 -0.7514633   
[6,] 0.42258167 -0.8251431 -0.36257405 0.06134197 -1.4953360   
 PercentageBlack PercentageHispanic PercentageAsian MedianAge UnemploymentRate PerCap   
[1,] 0.02666667 0.50000000 0.01428571 0.4444444 0.2   
[2,] 0.88000000 0.01470588 0.00000000 0.3333333 0.2   
[3,] 0.14666667 0.32352941 0.02857143 0.1111111 0.0   
[4,] 0.77333333 0.00000000 0.00000000 0.5555556 0.8   
[5,] 0.33333333 0.14705882 0.05714286 0.2222222 0.2   
[6,] 0.41333333 0.00000000 0.01428571 0.4444444 0.0