R Notebook

# Question 1

#a

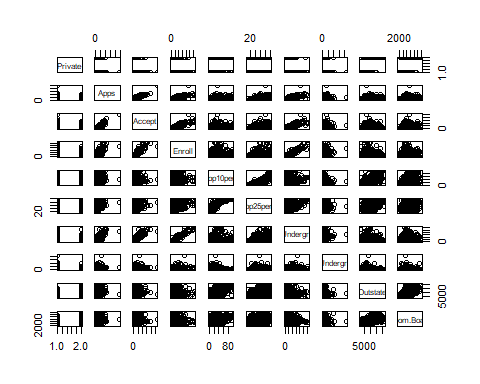
college <- read.csv("College.csv")

#b

rownames (college) <- college[, 1]  
#View (college)  
college <- college[, -1]  
View (college)

#c

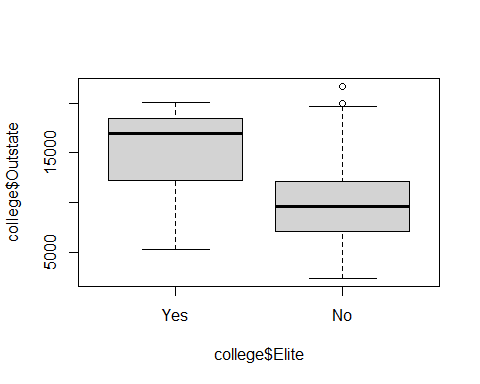
#summary(college)  
college$Private <- as.factor(college$Private)  
pairs(college[,1:10])



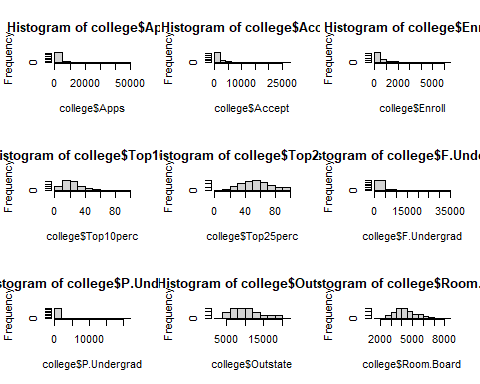
Elite <- rep ("No", nrow (college))  
Elite[college$Top10perc > 50] <- " Yes "  
Elite <- as.factor(Elite)  
college <- data.frame (college, Elite)  
summary(college)

## Private Apps Accept Enroll Top10perc   
## No :212 Min. : 81 Min. : 72 Min. : 35 Min. : 1.00   
## Yes:565 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242 1st Qu.:15.00   
## Median : 1558 Median : 1110 Median : 434 Median :23.00   
## Mean : 3002 Mean : 2019 Mean : 780 Mean :27.56   
## 3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902 3rd Qu.:35.00   
## Max. :48094 Max. :26330 Max. :6392 Max. :96.00   
## Top25perc F.Undergrad P.Undergrad Outstate   
## Min. : 9.0 Min. : 139 Min. : 1.0 Min. : 2340   
## 1st Qu.: 41.0 1st Qu.: 992 1st Qu.: 95.0 1st Qu.: 7320   
## Median : 54.0 Median : 1707 Median : 353.0 Median : 9990   
## Mean : 55.8 Mean : 3700 Mean : 855.3 Mean :10441   
## 3rd Qu.: 69.0 3rd Qu.: 4005 3rd Qu.: 967.0 3rd Qu.:12925   
## Max. :100.0 Max. :31643 Max. :21836.0 Max. :21700   
## Room.Board Books Personal PhD   
## Min. :1780 Min. : 96.0 Min. : 250 Min. : 8.00   
## 1st Qu.:3597 1st Qu.: 470.0 1st Qu.: 850 1st Qu.: 62.00   
## Median :4200 Median : 500.0 Median :1200 Median : 75.00   
## Mean :4358 Mean : 549.4 Mean :1341 Mean : 72.66   
## 3rd Qu.:5050 3rd Qu.: 600.0 3rd Qu.:1700 3rd Qu.: 85.00   
## Max. :8124 Max. :2340.0 Max. :6800 Max. :103.00   
## Terminal S.F.Ratio perc.alumni Expend   
## Min. : 24.0 Min. : 2.50 Min. : 0.00 Min. : 3186   
## 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00 1st Qu.: 6751   
## Median : 82.0 Median :13.60 Median :21.00 Median : 8377   
## Mean : 79.7 Mean :14.09 Mean :22.74 Mean : 9660   
## 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00 3rd Qu.:10830   
## Max. :100.0 Max. :39.80 Max. :64.00 Max. :56233   
## Grad.Rate Elite   
## Min. : 10.00 Yes : 78   
## 1st Qu.: 53.00 No :699   
## Median : 65.00   
## Mean : 65.46   
## 3rd Qu.: 78.00   
## Max. :118.00

plot(college$Outstate~college$Elite)



par(mfrow = c(3, 3))  
hist(college$Apps)  
hist(college$Accept)  
hist(college$Enroll)  
hist(college$Top10perc)  
hist(college$Top25perc)  
hist(college$F.Undergrad)  
hist(college$P.Undergrad)  
hist(college$Outstate)  
hist(college$Room.Board)



# Based on the data I can confirm and support multiple claims  
# Apps: there are more lower numbered applicants vs higher number applicant schools   
# The graph between applicants to accept makes sense. Ther are a lot less acceptance  
# Same thing for enroll, there are even less enrollments than acceptances   
# Based on the boxplots, more elite schools tend to have more students from out of state which makes since.

# 10

# a)

library(ISLR2)

## Warning: package 'ISLR2' was built under R version 4.0.5

Boston <- Boston  
?Boston

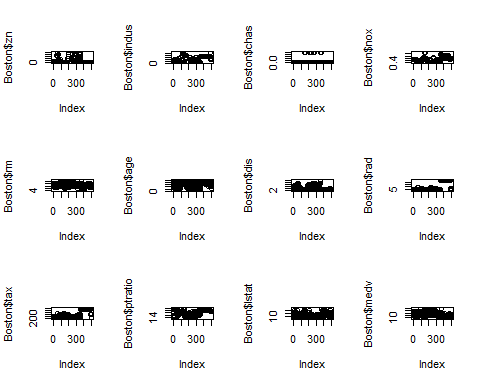
## starting httpd help server ... done

dim(Boston)

## [1] 506 13

# b)

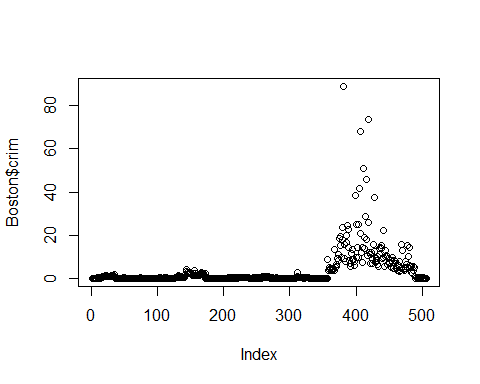
par(mfrow = c(3, 4))  
plot(Boston$zn)  
plot(Boston$indus)  
plot(Boston$chas)  
plot(Boston$nox)  
plot(Boston$rm)  
plot(Boston$age)  
plot(Boston$dis)  
plot(Boston$rad)  
plot(Boston$tax)  
plot(Boston$ptratio)  
plot(Boston$lstat)  
plot(Boston$medv)



# From what I can tell from the plots, there seems to be some sort of relationship between dis, age, medv, lstat  
# another relationship between nox and tax maybe

# c

plot(Boston$crim)



# Boston crim has a potential direct relationship between tax and rad. We can tell based off the plots

# d

cat(("Crime\n"))

## Crime

summary(Boston$crim)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00632 0.08204 0.25651 3.61352 3.67708 88.97620

cat("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\n")

## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

cat(("Tax\n"))

## Tax

summary(Boston$tax)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 187.0 279.0 330.0 408.2 666.0 711.0

cat("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\n")

## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

cat(("Pupil-Teacher\n"))

## Pupil-Teacher

summary(Boston$ptratio)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.60 17.40 19.05 18.46 20.20 22.00

cat("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\n")

## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

summary(Boston)

## crim zn indus chas   
## Min. : 0.00632 Min. : 0.00 Min. : 0.46 Min. :0.00000   
## 1st Qu.: 0.08205 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   
## nox rm age dis   
## 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   
## rad tax ptratio lstat   
## Min. : 1.000 Min. :187.0 Min. :12.60 Min. : 1.73   
## 1st Qu.: 4.000 1st Qu.:279.0 1st Qu.:17.40 1st Qu.: 6.95   
## Median : 5.000 Median :330.0 Median :19.05 Median :11.36   
## Mean : 9.549 Mean :408.2 Mean :18.46 Mean :12.65   
## 3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.20 3rd Qu.:16.95   
## Max. :24.000 Max. :711.0 Max. :22.00 Max. :37.97   
## medv   
## Min. : 5.00   
## 1st Qu.:17.02   
## Median :21.20   
## Mean :22.53   
## 3rd Qu.:25.00   
## Max. :50.00

# there is a huge jump in crime from the 3rd qu to the max meaning there are defineitly some outliers  
# There is not a huge jump in tax   
# there is not a huge jump in pupil- teacher as well.

# e

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.5

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

Boston %>%  
 summarise(Charles\_River = sum(chas), Non\_Charles\_River = (n() - sum(chas)))

## Charles\_River Non\_Charles\_River  
## 1 35 471

# Only 35!

# f

summary(Boston$ptratio)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.60 17.40 19.05 18.46 20.20 22.00

# 19.05!

# g

Boston[order(Boston$medv),][1,]

## crim zn indus chas nox rm age dis rad tax ptratio lstat medv  
## 399 38.3518 0 18.1 0 0.693 5.453 100 1.4896 24 666 20.2 30.59 5

# Suburb 339 has a median of 5000$  
# when comparing suburb 339 with the baseline statistics as we can see in "# d" The distrubition is meaningless! It is random and does not follow a speicifc pattern. Yes there are some information that can be dirived such as it has a very high crime rate, yet that is merely an observavtion, not a finding as the information is laid out right infront of us.

# h

Boston %>%  
 filter(rm >7) %>%  
 summarise(Frequancy = n())

## Frequancy  
## 1 64

Boston %>%  
 filter(rm >8) %>%  
 summarise(Frequancy = n())

## Frequancy  
## 1 13

# there are 64 suburbs with more than 7 rooms  
# there are 13 suburbs with more than 8 rooms