**Assignment 1**

**Talish Barmare**

**IDS 575**

**Q5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred ?**

One advantage of a flexible model is that it may give a better fit for non-linear models like SVM and reduces the bias. The disadvantage of such an approach is the need for a larger number of parameters, it follows the noise too closely (overfit) and thus leading to a high the variance.

A more flexible approach is ideal for prediction compared to a restrictive model where the agenda would be inference and the interpretability of the results

A less flexible approach or restrictive model would be preferred when we are interested in inference and the interpretability of the results.

| **Ob** | **X1** | **X2** | **X3** | **Y** |
| --- | --- | --- | --- | --- |
| 1 | 0 | 3 | 0 | Red |
| 2 | 2 | 0 | 0 | Red |
| 3 | 0 | 1 | 3 | Red |
| 4 | 0 | 1 | 2 | Green |
| 5 | −1 | 0 | 1 | Green |
| 6 | 1 | 1 | 1 | Red |

**7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.**

**Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.**

**Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.**

**(a) Compute the Euclidean distance between each observation and the test point,** X1 = X2 = X3 = 0.

d = sqrt[(X1-X2)^2 + (X2-X3)^2 + (X1-X3)^2}

Obs 1

d1 = sqrt[(0 - 0)^2 + (3 - 0)^2 + (0 - 0)^2] = 3.

Obs 2

d2 = sqrt[(2 - 0)^2 + (0 - 0)^2 + (0 - 0)^2] = 2.

Obs 3

d3 = sqrt[(0 - 0)^2 + (1 - 0)^2 + (3 - 0)^2] = sqrt[0 + 1 + 9] = sqrt[10] = ~3.16.

Obs 4

d4 = sqrt[(0 - 0)^2 + (1 - 0)^2 + (2 - 0)^2] = sqrt[1 + 4] = sqrt[5] = ~2.24.

Obs 5

d5 = sqrt[(-1 - 0)^2 + (0 - 0)^2 + (1 - 0)^2] = sqrt[1 + 1] = sqrt[2] = ~1.41.

Obs 6

d6 = sqrt[(1 - 0)^2 + (1 - 0)^2 + (1 - 0)^2] = sqrt[1 + 1 + 1] = sqrt[3] = ~1.73.

**(b) What is our prediction with K = 1? Why?**

The nearest point to test point Y when X1 = X2 = X3 = 0 is Obs 5 (-1, 0, 1) with euclidean distance = ~1.41. Since Obs 5 was Green, when (K = 1) the prediction for the test point will also be Green.

**(c) What is our prediction with K = 3? Why?**

The nearest three points to test point Y when X1 = X2 = X3 = 0 are Obs 5 with euclidean distance = ~1.41, Obs 6 with distance ~1.73, and Obs 2 with distance 2.

Since Obs 5 was Green, Obs 6 was Red, and Obs 2 was Red, we predict (K = 3) the test point will be the majority Red with a probability of 2/3 or 66.66%

**(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?**

As K becomes larger, the boundary becomes restrictive (linear). So in this case we would expect the best value for K to be small.

**Q8. This exercise relates to the “College” data set, which can be found in the file “College.csv”. It contains a number of variables for 777 different universities and colleges in the US.**

**a) Use the read.csv() function to read the data into R.**

**Call the loaded data “college”. Make sure that you have the directory set to the correct location for the data.**

getwd()

setwd("~/Desktop/IDS 575")

library(ISLR)

data(College)

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

**b) Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don’t really want R to treat this as data. However, it may be handy to have these names for later.**

head(college)

rownames <- college[, 1]

college <- college[, -1]

head(college[, 1:5])

**c)**

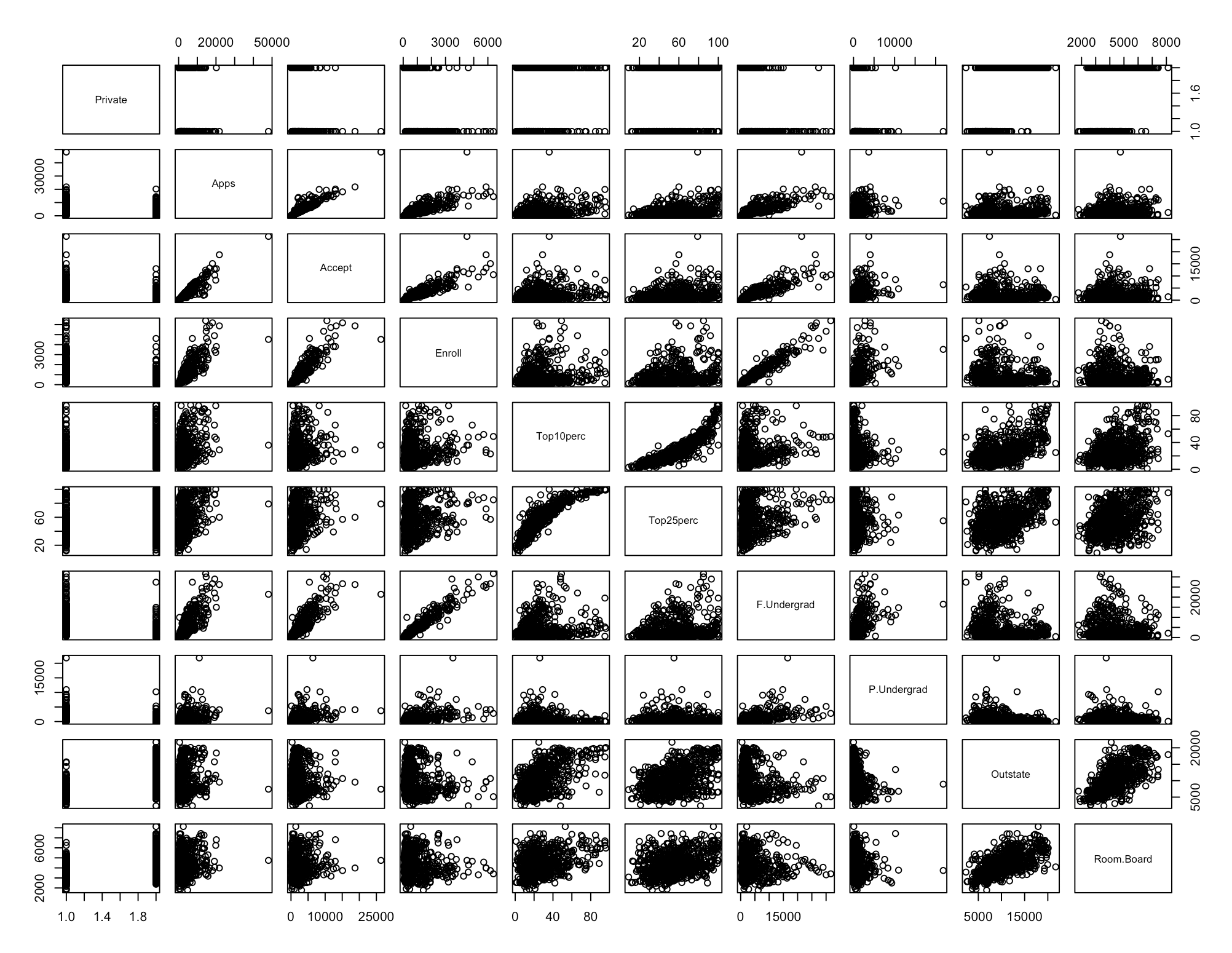
**i)Use the summary() function to produce a numerical summary of the variables in the data set.**

**A close up of a receipt

Description automatically generatedsummary(college)**

**ii) Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using**

**A[,1:10].**

**pairs(college[, 1:10])**

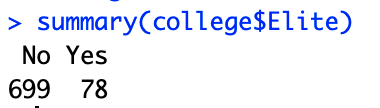
**iii) Use the plot() function to produce side-by-side boxplots of Outstate versus Private.**

plot(college$Private, college$Outstate, xlab = "Private University", ylab ="Out of State tuition USD", main = "Outstate Tuition Vs Private Colleges")

iii) Use the plot() function to produce side-by-side boxplots of Outstate versus Private.A screenshot of a social media post

Description automatically generated

**iv) Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10 % of their high school classes exceeds 50 %.**

Elite <- rep("No", nrow(college))

Elite[college$Top10perc > 50] <- "Yes"

Elite <- as.factor(Elite)

college$Elite <- Elite

summary(college$Elite)

plot(college$Elite, college$Outstate, xlab = "Elite University", ylab ="Out of State tuition in USD", main = "Outstate Tuition Plot")A screenshot of a social media post

Description automatically generated

**v) Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments**

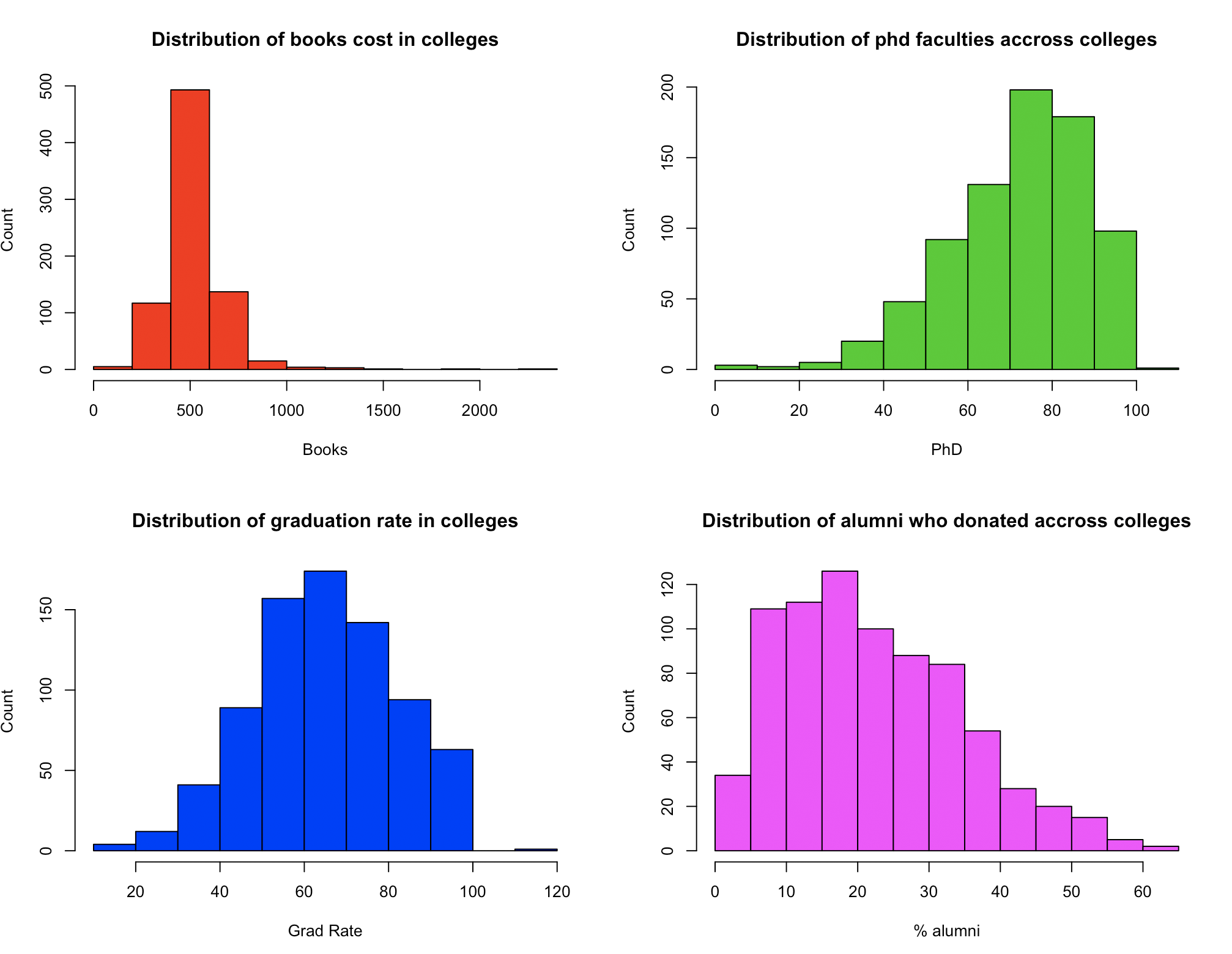
**to this function will divide the screen in other ways.**

par(mfrow = c(2,2))

hist(college$Books, col = 2, xlab = "Books", ylab = "Count", main = "Distribution of books cost in colleges")

hist(college$PhD, col = 3, xlab = "PhD", ylab = "Count",main = "Distribution of phd faculties accross colleges")

hist(college$Grad.Rate, col = 4, xlab = "Grad Rate", ylab = "Count",main = "Distribution of graduation rate in colleges")

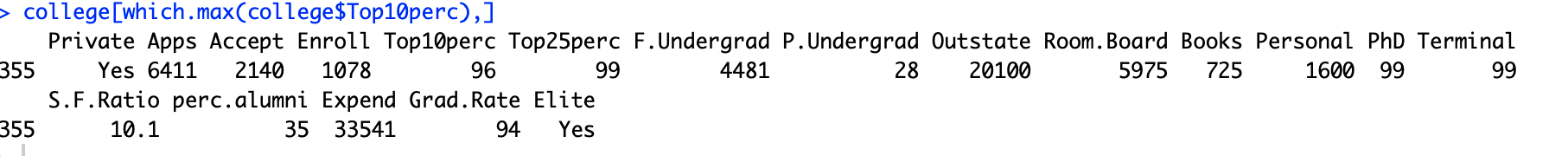
hist(college$perc.alumni, col = 6, xlab = "% alumni", ylab = "Count",main = "Distribution of alumni who donated accross colleges") 

vi)Continue exploring the data, and provide a brief summary of what you discover.

**Some interesting observations:**

**What is the university with the most students in the top 10% of class**

college[which.max(college$Top10perc),]



Thus, on pulling up record 355,

rownames[355]

Massachusetts Institute of Technology

**What university has the smallest & highest acceptance rates**

acceptance\_rate = college$Accept / college$Apps

college[ which.min( acceptance\_rate ), ]

rownames[460]

Princeton University

A screenshot of a cell phone

Description automatically generated

college[ which.max( acceptance\_rate ), ]

rownames[193]

Emporia State UniversityA screenshot of a cell phone

Description automatically generated

**10)**

**This exercise involves the Boston housing data set.**

**(a) To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.**

**> library (MASS)**

**Now the data set is contained in the object Boston.**

**> Boston**

**Read about the data set:**

**> ?Boston**

**How many rows are in this data set? How many columns? What do the rows and columns represent?**

Dimensions of data set

506 \* 14

**Rows represent:** Housing Values in Suburbs of Boston

**This data frame contains the following columns:**

crim: per capita crime rate by town.

zn: proportion of residential land zoned for lots over 25,000 sq.ft.

indus :proportion of non-retail business acres per town.

chas: charles river dummy variable (= 1 if tract bounds river; 0 otherwise).

nox: nitrogen oxides concentration (parts per 10 million).

rm: average number of rooms per dwelling.

age: proportion of owner-occupied units built prior to 1940.

dis: weighted mean of distances to five boston employment centres.

rad: index of accessibility to radial highways.

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

ptratio: pupil-teacher ratio by town.

black :1000(bk - 0.63) ^2 where bk is the proportion of blacks by town.

lstat :lower status of the population (percent).

medv :median value of owner-occupied homes in \$1000s.

b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

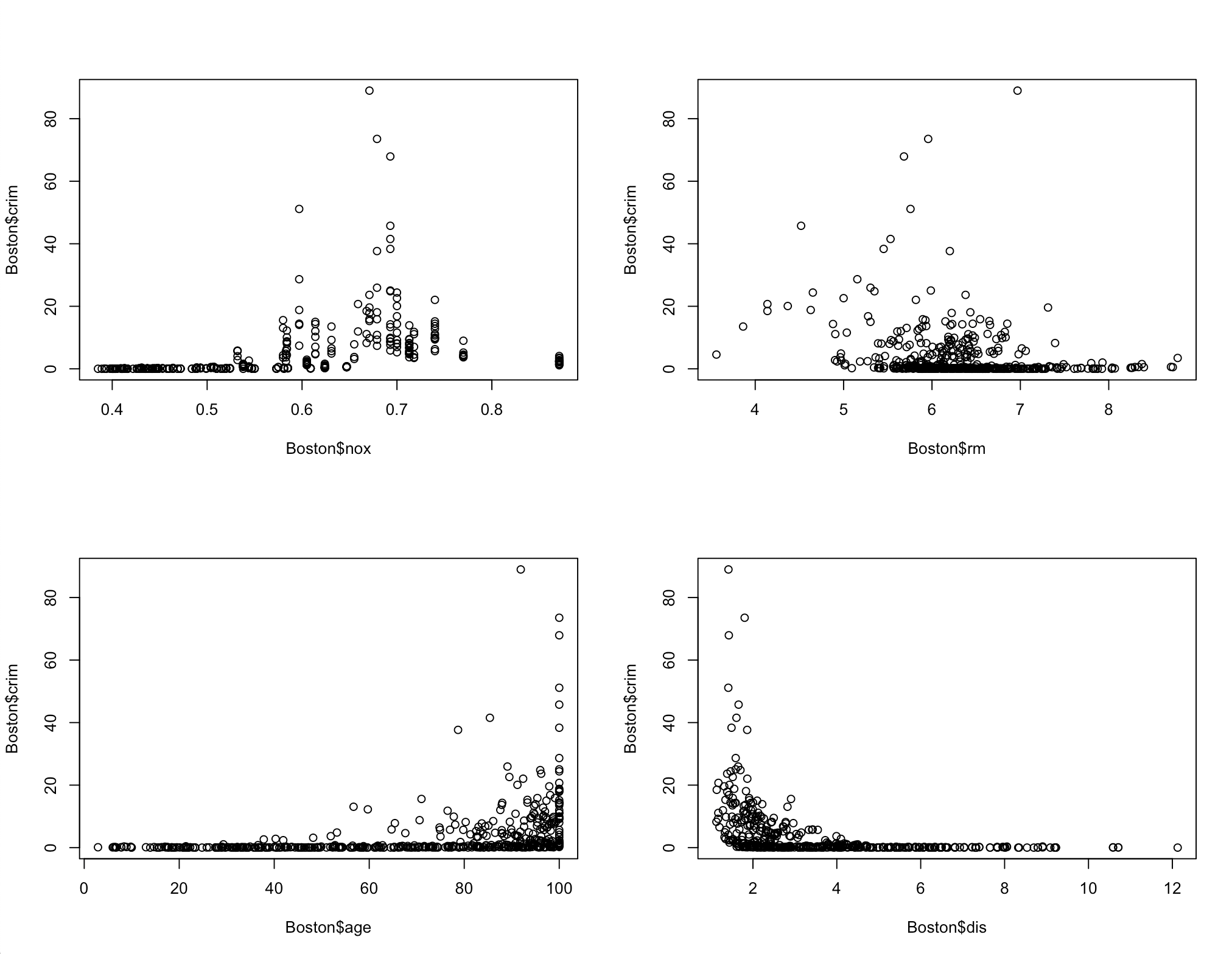
par(mfrow = c(2, 2))

plot(Boston$nox, Boston$crim)

plot(Boston$rm, Boston$crim)

plot(Boston$age, Boston$crim)

plot(Boston$dis, Boston$crim)

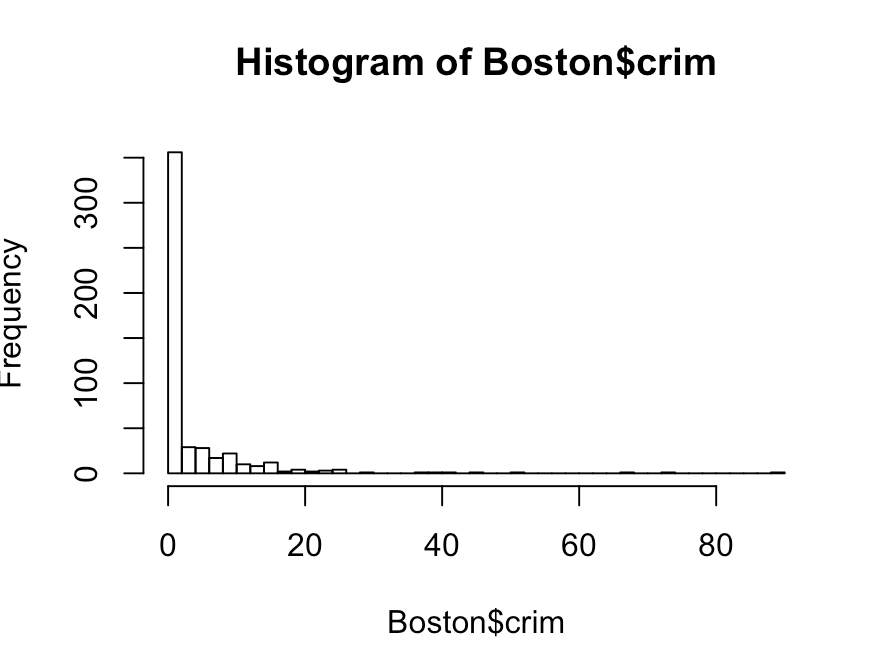


**c)Are any of the predictors associated with per capita crime rate? If so, explain the relationship.**

**There seems to be an association between the median value of owner-occupied homes and crime per capital. Low median value of owner-occupied home indicates a higher crime per capita.**

There is also seem to be a correlation between distance to employment centers and crime. Areas close to employment centers show high crime per capita.

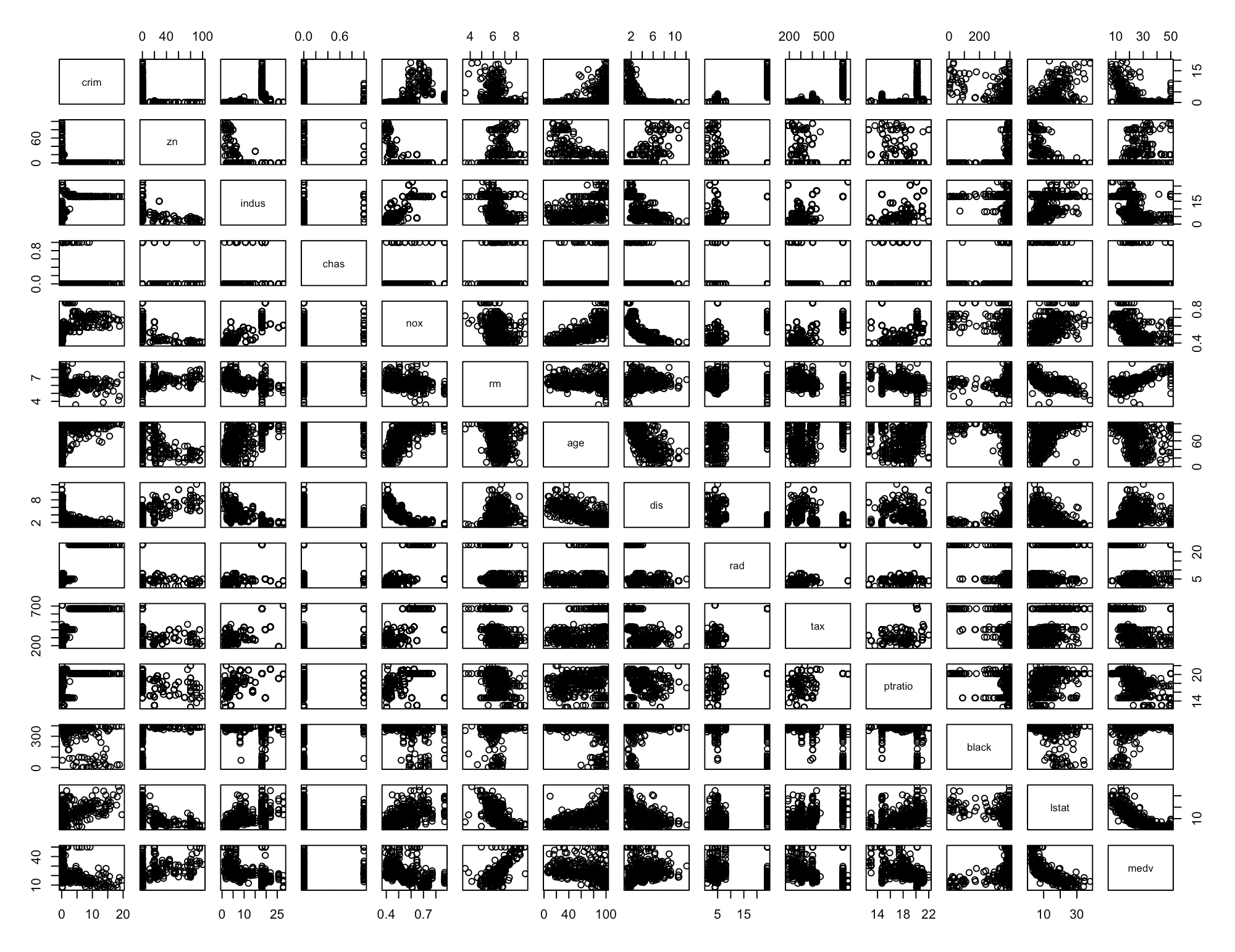
dev.off()

hist(Boston$crim, breaks = 50)

**Most crime falls under 20% of Boston (80% of data falls in crim < 20)**

pairs(Boston[Boston$crim < 20, ])

**There may be a relationship between crim and nox, rad, age, dis, lstat and medv.**



plot(Boston$age, Boston$crim)

**Older homes, more crime**

plot(Boston$dis, Boston$crim)

**Closer to work-area, more crime**

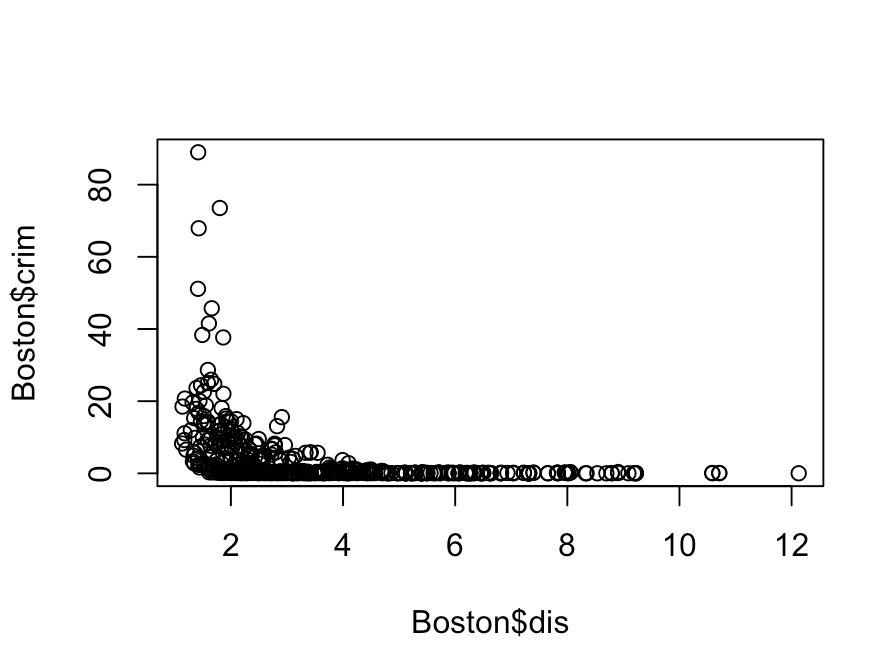
plot(Boston$rad, Boston$crim)

**Higher index of accessibility to radial highways, more crime**

plot(Boston$medv, Boston$crim)

**Higher crime in low median value houses**

A screenshot of a cell phone

Description automatically generated  A screenshot of a cell phone

Description automatically generated A close up of a piece of paper

Description automatically generated

**d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.**

par(mfrow=c(1,3))

hist(Boston$crim[Boston$crim>1], breaks=25)

Most cities have low crime rates, but there is a long tail: 18 suburbs appear to have a crime rate > 20, reaching to above 80

hist(Boston$tax, breaks=25)

There is a large divide between suburbs with low tax rates and a peak at 660

hist(Boston$ptratio, breaks=25)

A skew towards high ratios, but no particularly high ratios

A screenshot of a cell phone

Description automatically generated

**(e)How many of the suburbs in this data set bound the Charles river?**

dim(subset(Boston, chas == 1))

35 suburbs in the data set bound the Charles river

**(f) What is the median pupil-teacher ratio among the towns in this data set ?**

median(Boston$ptratio)

19.05 is the median pupil-teacher ration among the towns in data set

**g) Which suburb of Boston has lowest median value of owner-occupied homes ? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors?**

row.names(Boston[min(Boston$medv), ]) # lowest median value of owner-occupied homes = 5.0

there are two suburbs with this value 399 and 406.

**Stats for 399 & 406 towns**

t(subset(Boston, medv == min(Boston$medv)))

**# 399 406 comparison with overall ranges**

# crim 38.3518 67.9208 above 3rd quartile

# zn 0.0000 0.0000 at min

# indus 18.1000 18.1000 at 3rd quartile

# chas 0.0000 0.0000 not bounded by river

# nox 0.6930 0.6930 above 3rd quartile

# rm 5.4530 5.6830 below 1st quartile

# age 100.0000 100.0000 at max

# dis 1.4896 1.4254 below 1st quartile

# rad 24.0000 24.0000 at max

# tax 666.0000 666.0000 at 3rd quartile

# ptratio 20.2000 20.2000 at 3rd quartile

# black 396.9000 384.9700 at max; above 1st quartile

# lstat 30.5900 22.9800 above 3rd quartile

# medv 5.0000 5.0000 at min

There two towns are not the best places to live, but not too bad either

**h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.**

dim(subset(Boston, rm > 7))

64 suburbs average more than 7 rooms per dwelling

dim(subset(Boston, rm > 8))

13 suburbs average more than 8 rooms per dwelling

summary(subset(Boston, rm > 8))

summary(Boston)

On comparison, we notice relatively lower crime (comparing range) & lower lstat (comparing range) than the overall metrics of Boston