Exercise 10 Aida M.

R Markdown

Housing Values in Suburbs of Boston

Description
The Boston data frame has 506 rows and 14 columns.

Usage
Boston

Format
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.

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

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.6.3
```

```
library("knitr")
```

```
## Warning: package 'knitr' was built under R version 3.6.3
```

```
?Boston
```

```
## starting httpd help server ... done
```

Rows and Columns using dim command

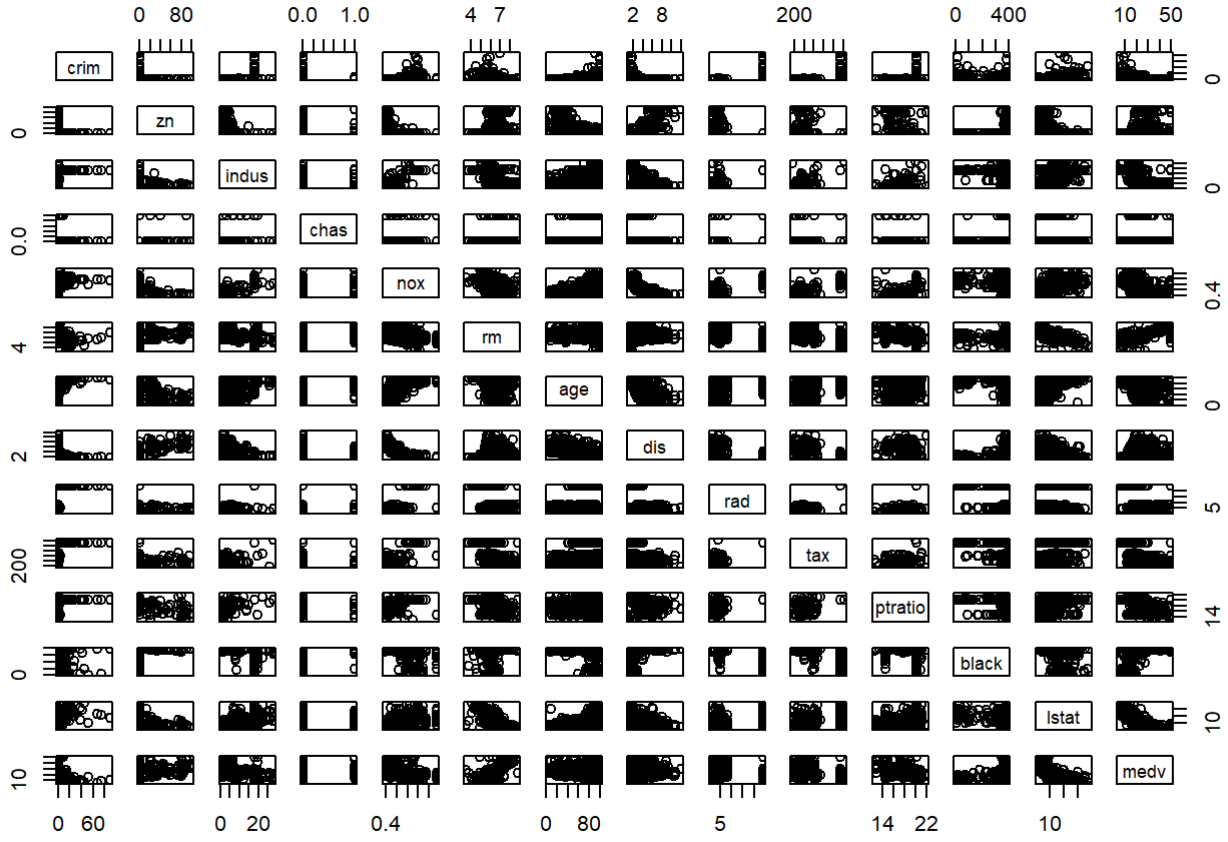
```
dim(Boston)
```

```
## [1] 506  14
```

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

Including pairwise scatterplots

```
pairs(Boston[, 1:14])
```



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

```
Boston.corr = cor(Boston)
Boston.corr.crim = Boston.corr[-1,1]
print(
  Boston.corr.crim[order(abs(Boston.corr.crim), decreasing = T)]
)
```

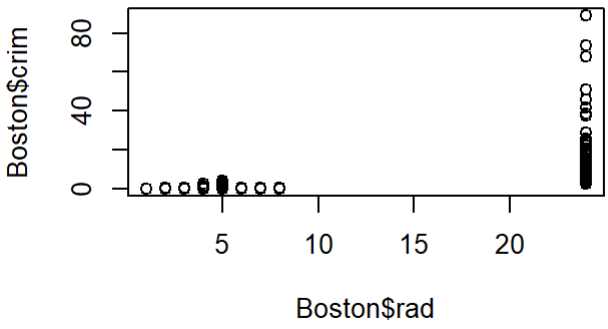
```
##         rad         tax       lstat         nox       indus        medv
##  0.62550515  0.58276431  0.45562148  0.42097171  0.40658341 -0.38830461
##       black         dis         age     ptratio          rm          zn
## -0.38506394 -0.37967009  0.35273425  0.28994558 -0.21924670 -0.20046922
##        chas
## -0.05589158
```

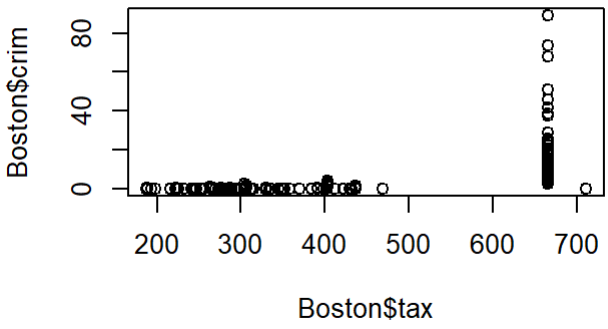
The four greatest correlation values have a positive relationship, plotting them:

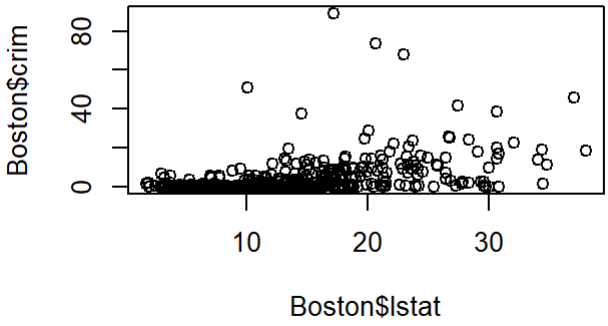
```
par(mfrow=c(2,2))

most_corr = names(Boston.corr.crim[order(abs(Boston.corr.crim), decreasing = T)][1:4])
print(most_corr)
```

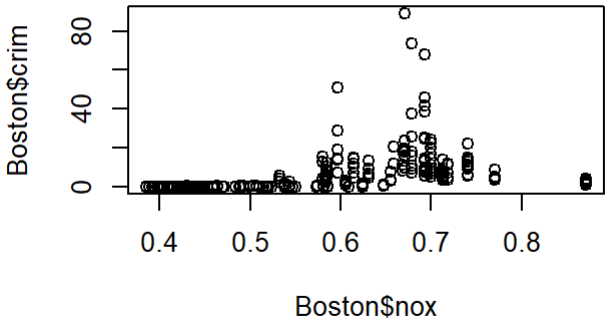
```
## [1] "rad"   "tax"   "lstat" "nox"
```

```
plot(Boston$rad, Boston$crim)
plot(Boston$tax, Boston$crim)
plot(Boston$lstat, Boston$crim)
plot(Boston$nox, Boston$crim)
```







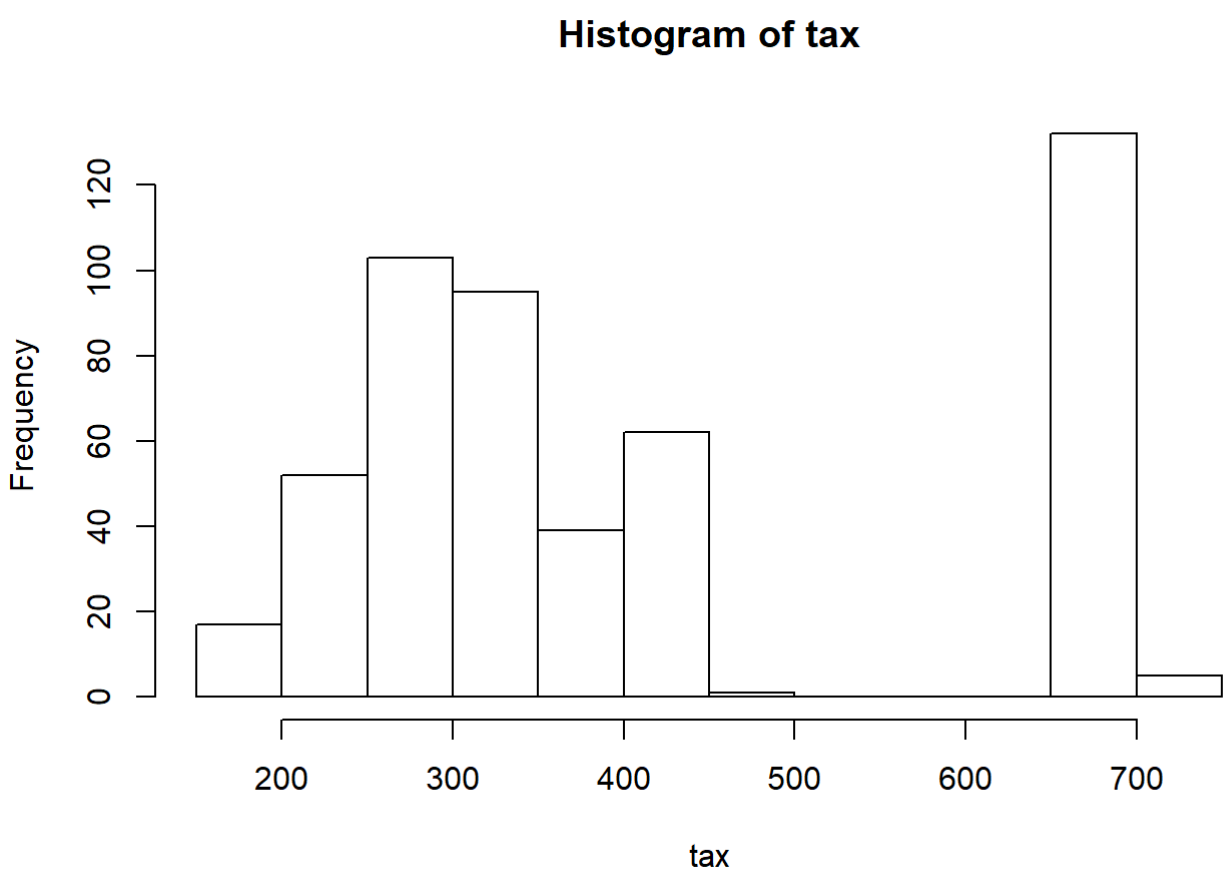


d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

```
summary(Boston[,1])
```

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

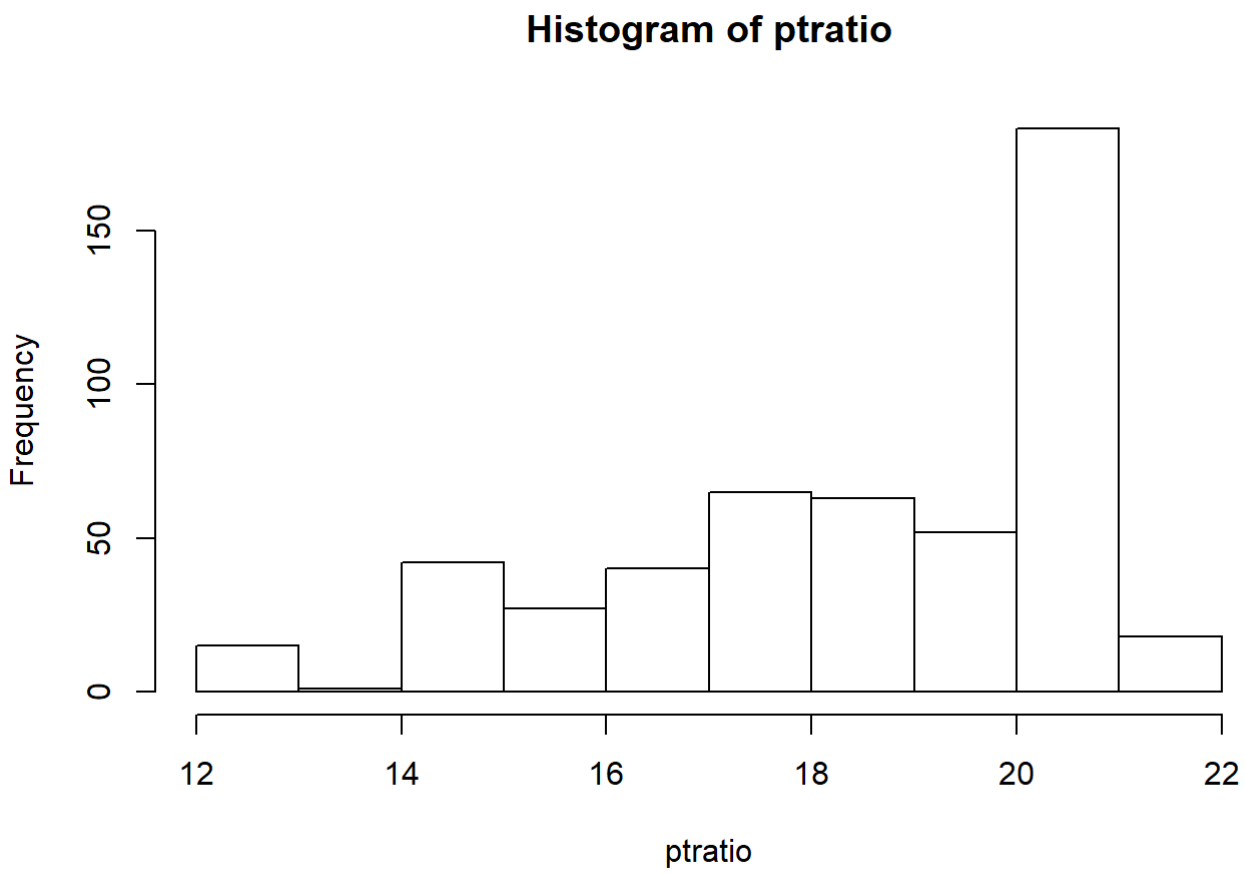
```
## Tax histocram
tax<-Boston[,10]
hist(tax)
```




```
length(tax[tax>500])
```

```
## [1] 137
```

```
## ptratio histocram
ptratio<-Boston[,11]
hist(ptratio)
```



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

```
chas<-Boston[,4]
table(chas)
```

```
## chas
##   0   1
## 471  35
```

The value 1 says that the suburb bounds the Charles Rivers, there are 35 suburbs that bound river.

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

```
median(ptratio)
```

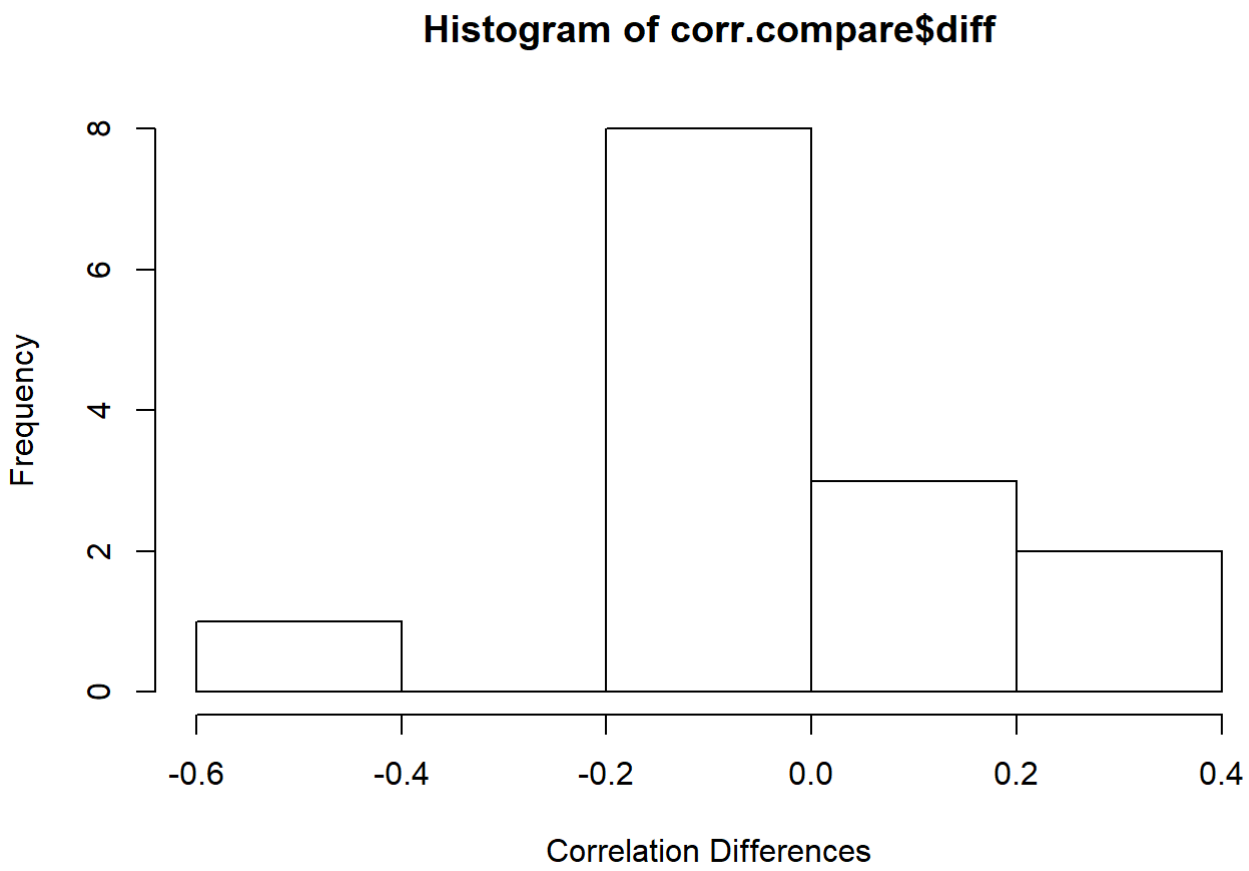
```
## [1] 19.05
```

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? Comment on your findings.

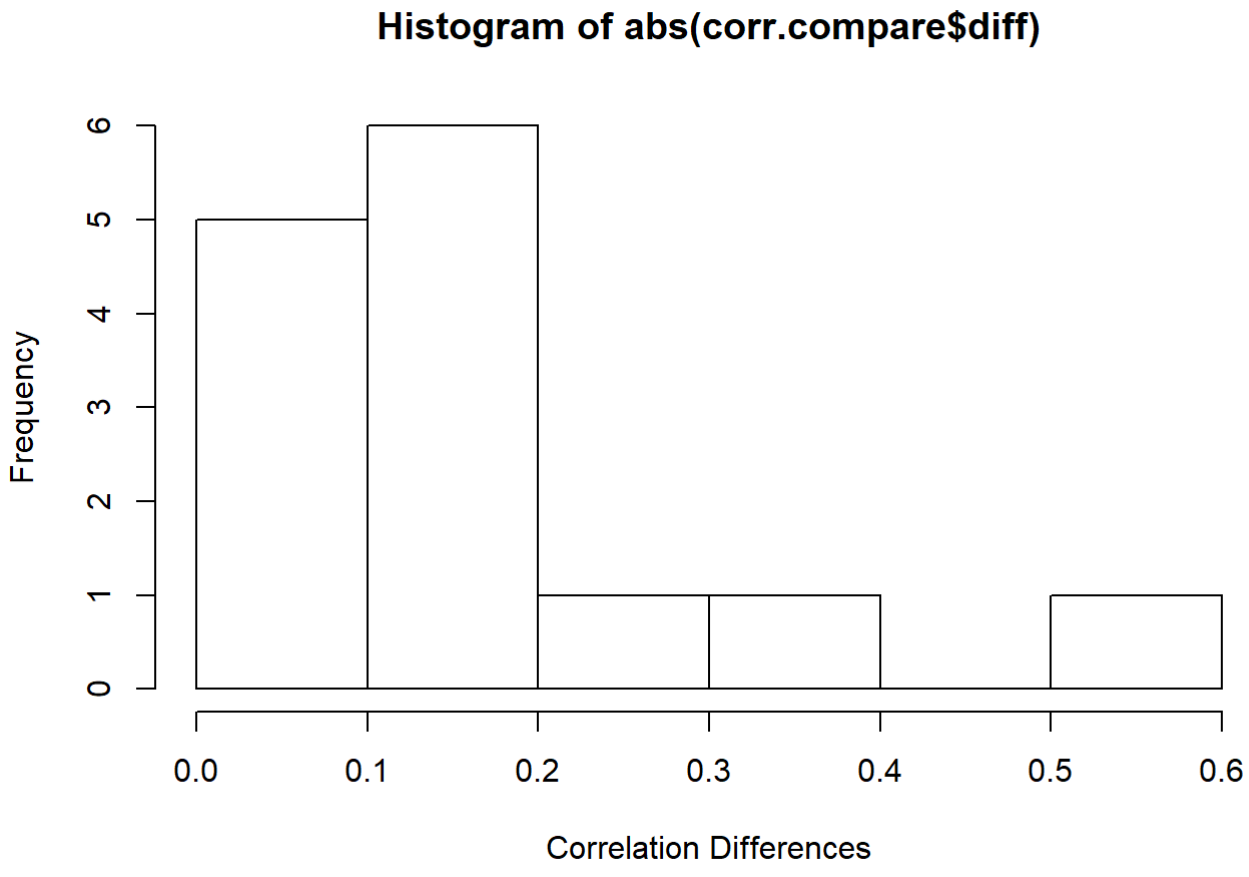
```
medv<-Boston[,14]
subs.lw = which(medv<median(medv))
print(subs.lw)
```

```
##   [1]   9  10  11  12  14  15  16  18  19  20  21  22  23  24  25  26  27  28
##  [19]  29  30  31  32  33  34  35  36  37  38  46  47  48  49  50  51  52  55
##  [37]  60  61  62  67  69  70  77  78  80  95 103 104 105 106 107 108 109 110
##  [55] 113 114 115 116 118 119 120 122 123 124 125 127 128 129 130 131 132 134
##  [73] 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 152 153
##  [91] 154 155 156 157 171 172 210 212 242 245 246 248 256 270 271 287 298 310
## [109] 311 313 316 317 318 320 323 324 329 331 332 333 335 336 337 338 339 340
## [127] 341 343 346 347 353 355 356 357 362 363 364 374 375 376 377 378 379 380
## [145] 381 382 383 384 385 386 387 388 389 390 391 393 394 395 396 397 398 399
## [163] 400 401 402 403 404 405 406 407 409 411 412 413 414 415 416 417 418 419
## [181] 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437
## [199] 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455
## [217] 456 457 458 459 460 461 462 463 464 466 467 468 469 470 471 472 475 476
## [235] 477 478 479 485 487 488 489 490 491 492 493 497 498 500 501 503 506
```

```
Boston.corr.subs.lw = cor(Boston[subs.lw, ])
corr.compare = data.frame('lower'=Boston.corr.subs.lw[, "medv"], 'all'=Boston.corr[, "medv"])
corr.compare$diff = corr.compare$lower - corr.compare$all
hist(corr.compare$diff, xlab="Correlation Differences")
```




```
hist(abs(corr.compare$diff), xlab="Correlation Differences")
```



```
main.diffs = head(corr.compare[order(abs(corr.compare$diff), decreasing = T), ], 5)

print(main.diffs)
```

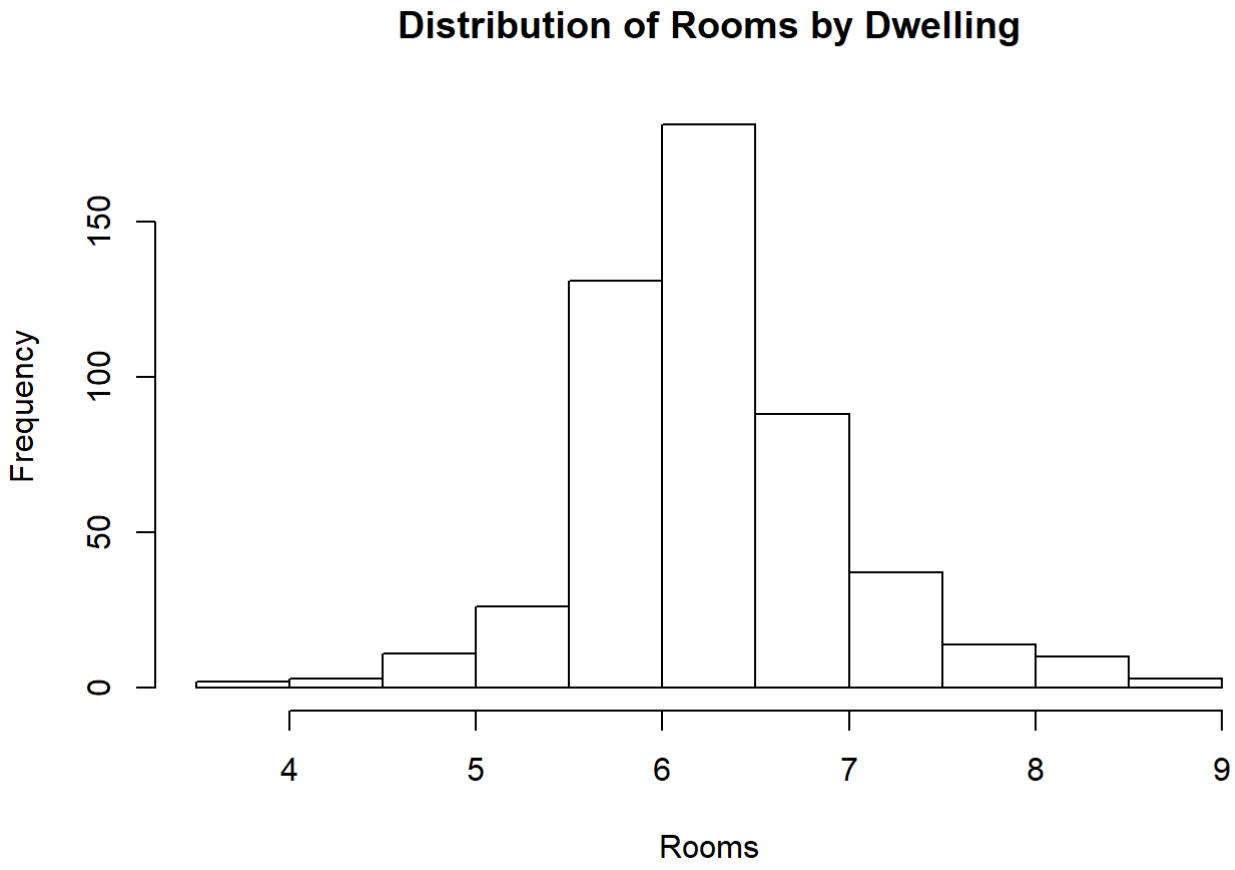
```
##              lower        all       diff
## rm       0.1493689  0.6953599 -0.5459910
## ptratio -0.1971183 -0.5077867  0.3106684
## dis      0.4815516  0.2499287  0.2316229
## crim    -0.5857651 -0.3883046 -0.1974605
## rad     -0.5380354 -0.3816262 -0.1564092
```

```
print(rownames(main.diffs))
```

```
## [1] "rm"      "ptratio" "dis"     "crim"    "rad"
```

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.

```
rm<-Boston[,6]
hist(rm, main="Distribution of Rooms by Dwelling", xlab="Rooms")
```



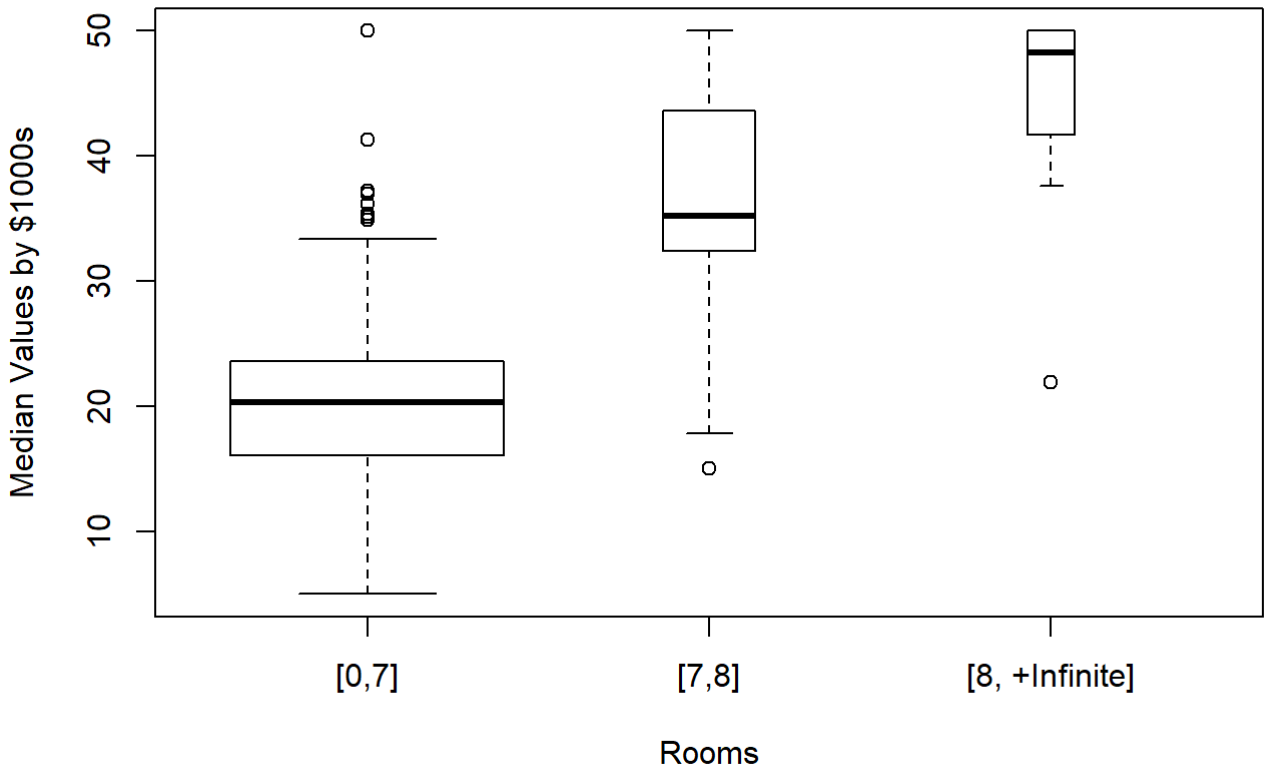

```
##7 rooms per dwelling
length(rm[rm>7])
```

```
## [1] 64
```

```
##8 rooms per dwelling
length(rm[rm>8])
```

```
## [1] 13
```

```
frm =as.factor(as.character(lapply(rm, function(x) ifelse(x>8, "[8, +Infinite]", ifelse(x>7,"
[7,8]","[0,7]")))))
plot(frm, medv, varwidth=T, xlab="Rooms",
     ylab="Median Values by $1000s",
     title="Median Value")
```


Median Values by $1000s
10
20
30
40
50
[0,7]
[7,8]
[8, +Infinite]
Rooms