Zixiao Wu-Individual Assignment 1

2022-09-09

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

#(a)

```
library(MASS)
Boston
```

	crim	**zn**	**indus**	**chas**	**nox**	**rm**	**age**	**dis**	**rad**	▸
	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	
6	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	
7	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	
8	0.14455	12.5	7.87	0	0.5240	6.172	96.1	5.9505	5	
9	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	
10	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	

1-10 of 506 rows | 1-10 of 15 columns Previous **1** 2 3 4 5 6 ... 51 Next

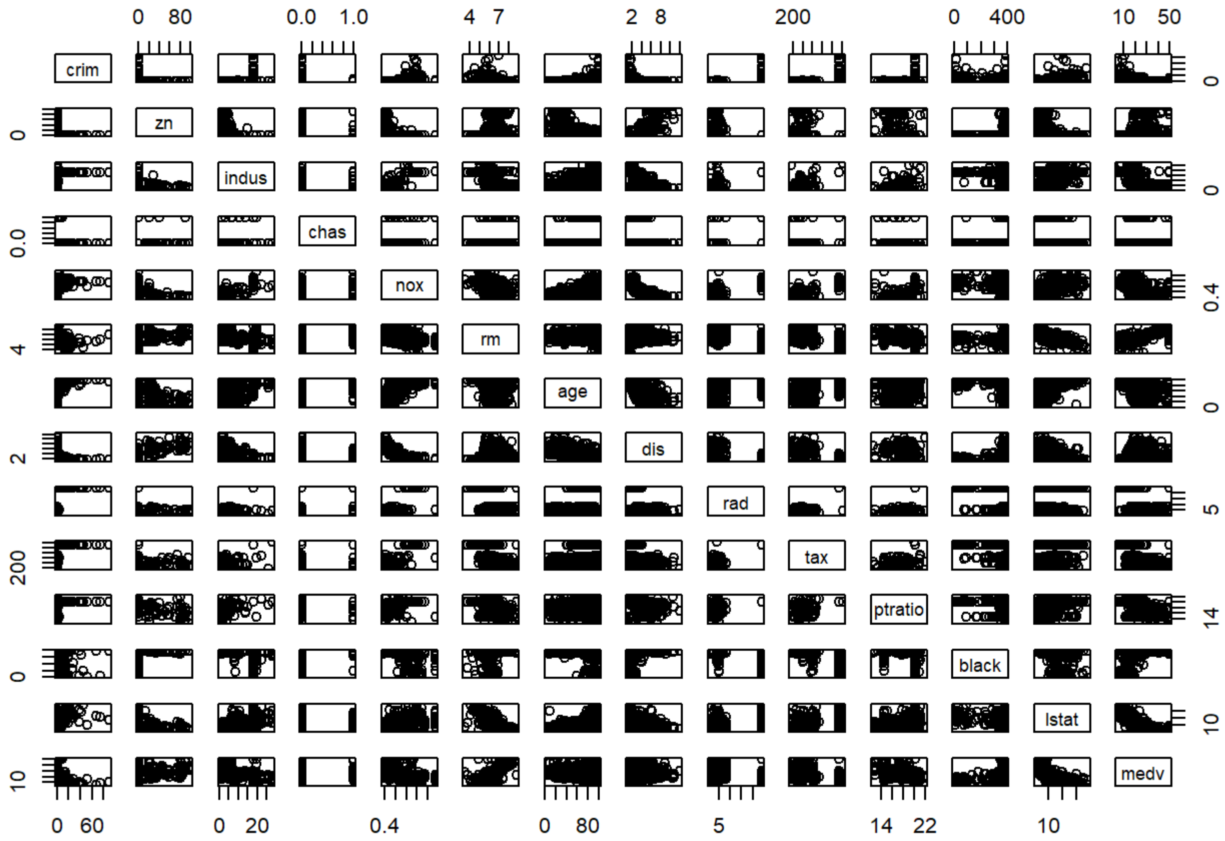
```
dim(Boston)
```

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

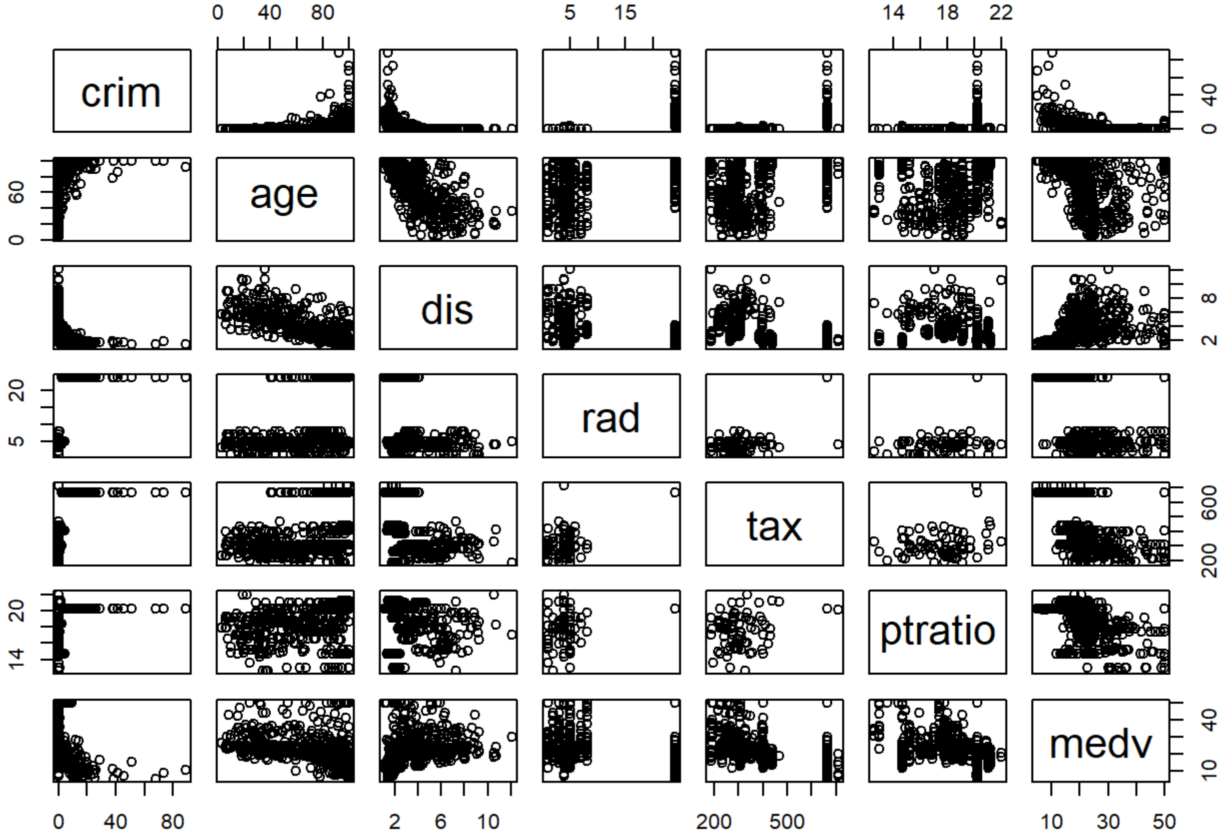
From the result above, we can know that there are 506 rows and 14 columns in the data. Each row is different area in Boston, and each column is factors that relates to house price.

#(b)

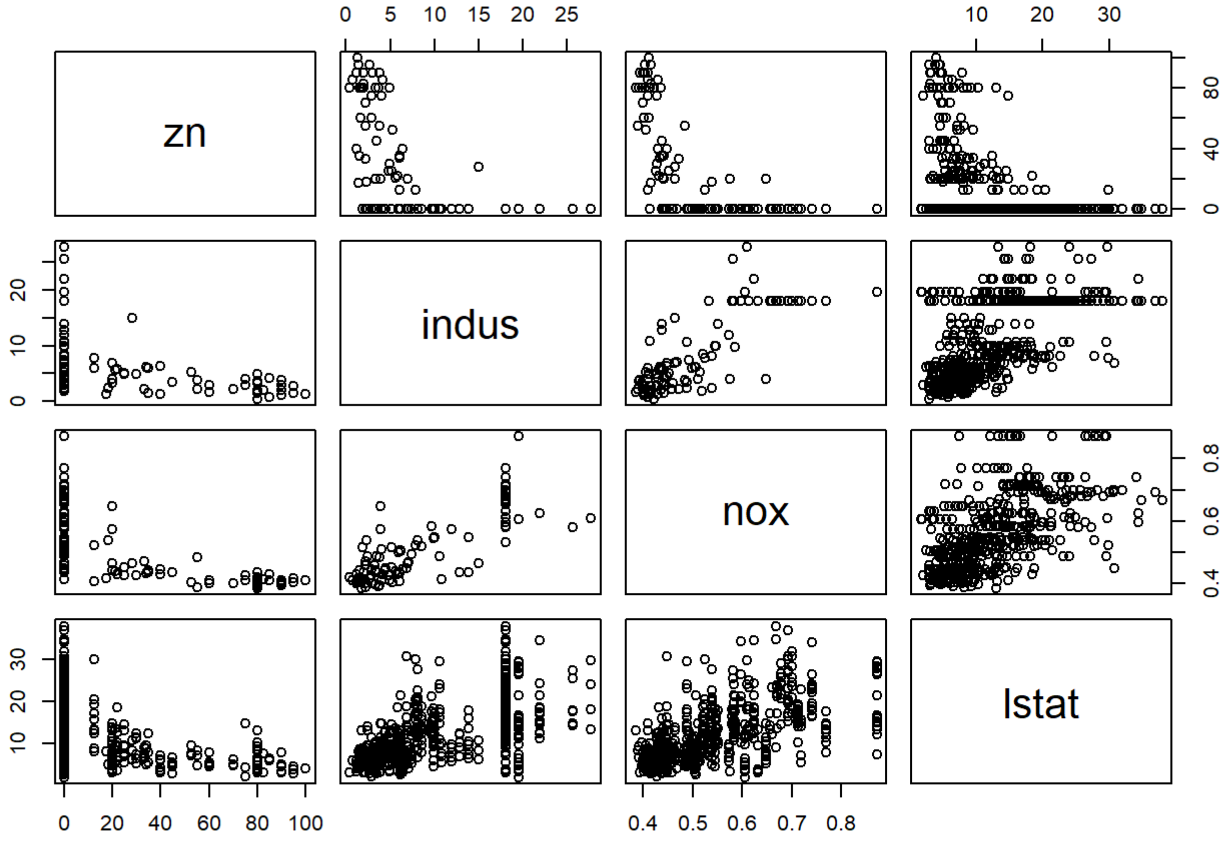
```
pairs(Boston)
```



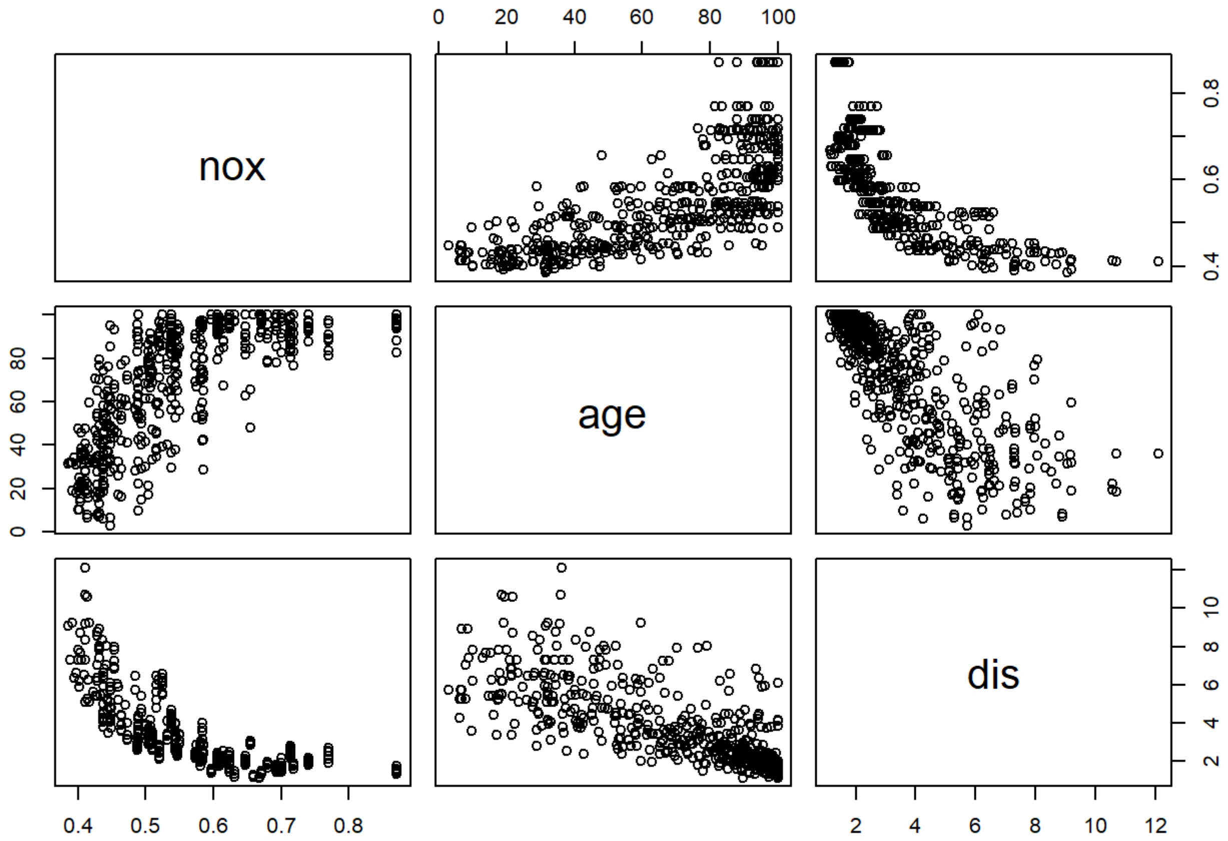
```
pairs(~crim + age + dis + rad + tax +  ptratio + medv, data = Boston)
```



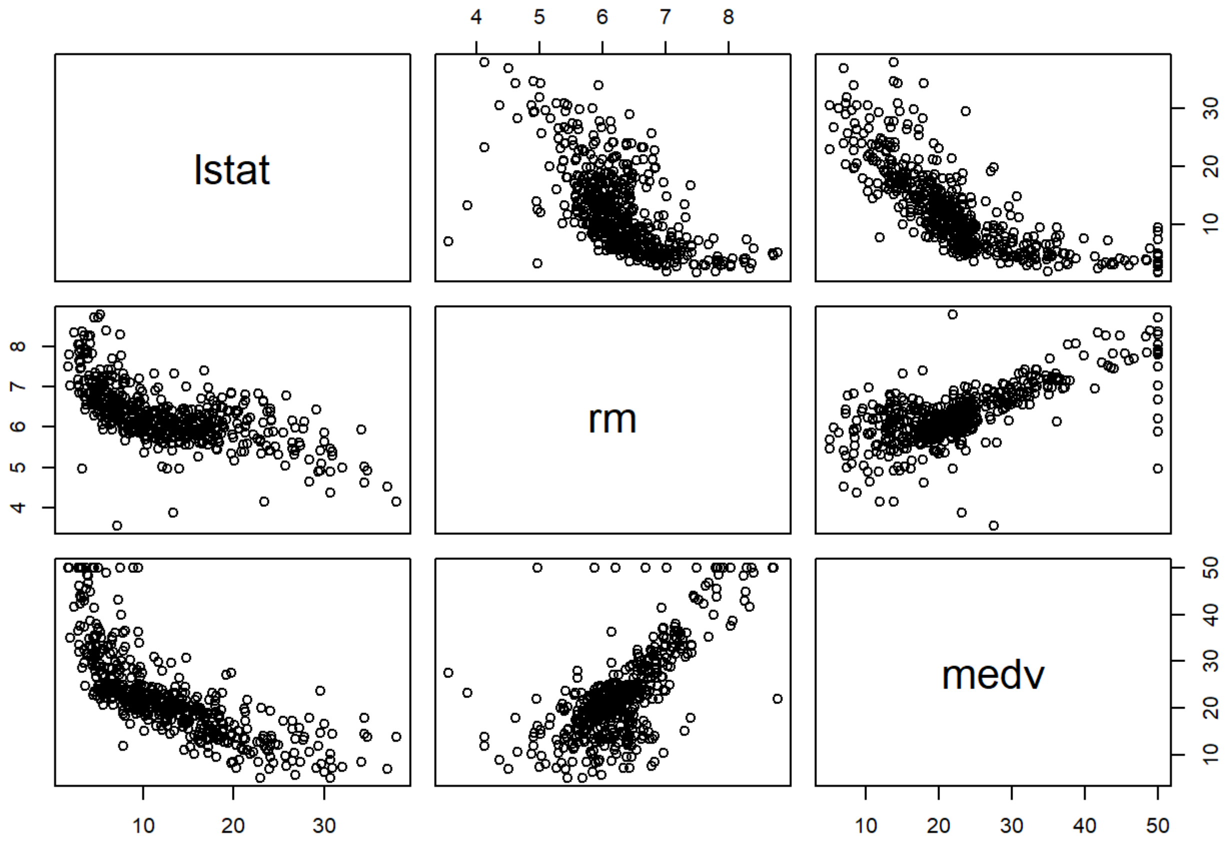
```
pairs(~zn + indus + nox + lstat, data = Boston)
```



```
pairs(~nox + age + dis, data = Boston)
```



```
pairs(~lstat + rm + medv, data = Boston)
```



From the plot we can assume that some variables are correlated with other variables: 1. crim has a negative realtionship with dis and medv, and has a positive relationship with age; 2. zn has a negative realtionship with indus and nox; 3. nox has a negative realtionship with dis, and has a positive relationship with age; 4. lstat has a positive relationship with rm and medv.

#(c)

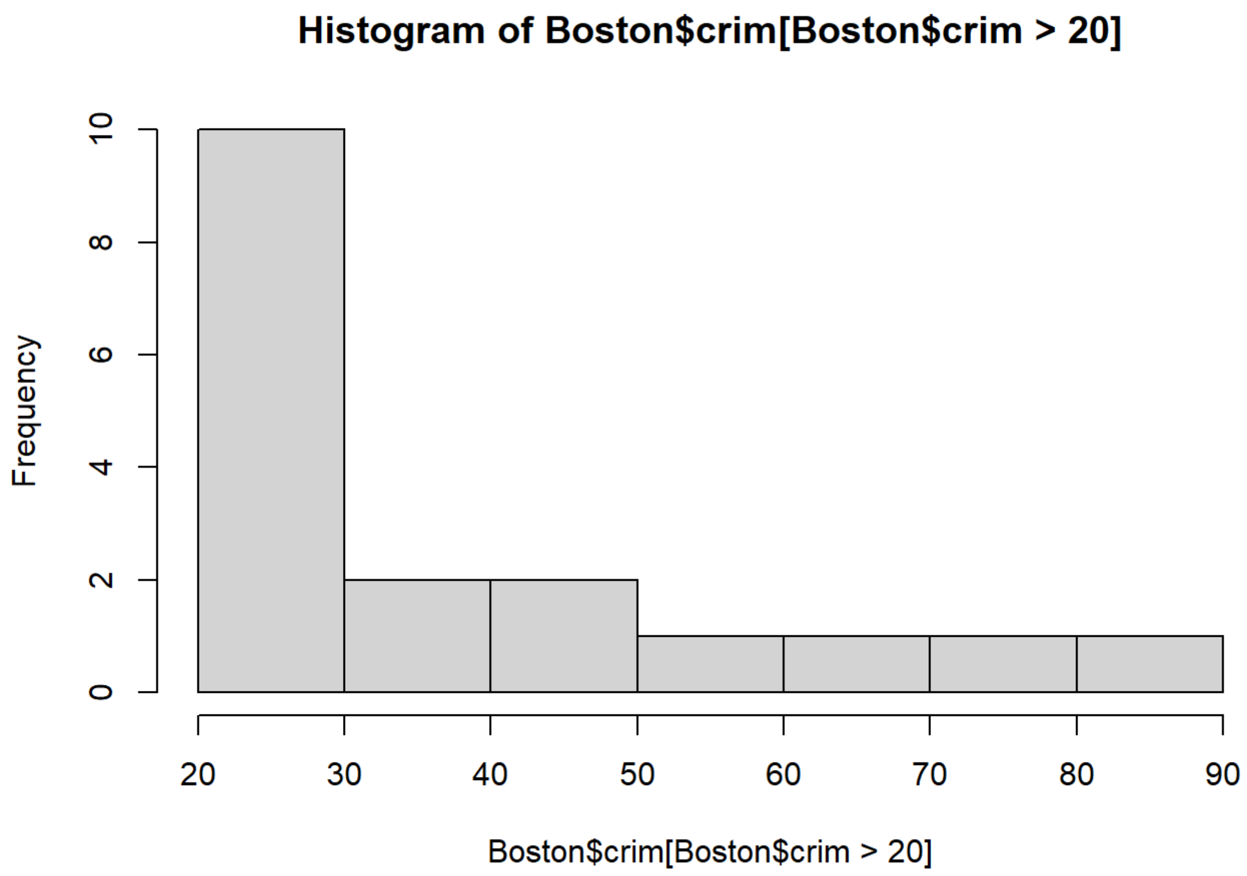
```
cor(Boston)
```

```
##                crim          zn       indus         chas         nox
## crim     1.00000000 -0.20046922  0.40658341 -0.055891582  0.42097171
## zn      -0.20046922  1.00000000 -0.53382819 -0.042696719 -0.51660371
## indus    0.40658341 -0.53382819  1.00000000  0.062938027  0.76365145
## chas    -0.05589158 -0.04269672  0.06293803  1.000000000  0.09120281
## nox      0.42097171 -0.51660371  0.76365145  0.091202807  1.00000000
## rm      -0.21924670  0.31199059 -0.39167585  0.091251225 -0.30218819
## age      0.35273425 -0.56953734  0.64477851  0.086517774  0.73147010
## dis     -0.37967009  0.66440822 -0.70802699 -0.099175780 -0.76923011
## rad      0.62550515 -0.31194783  0.59512927 -0.007368241  0.61144056
## tax      0.58276431 -0.31456332  0.72076018 -0.035586518  0.66802320
## ptratio  0.28994558 -0.39167855  0.38324756 -0.121515174  0.18893268
## black   -0.38506394  0.17552032 -0.35697654  0.048788485 -0.38005064
## lstat    0.45562148 -0.41299457  0.60379972 -0.053929298  0.59087892
## medv    -0.38830461  0.36044534 -0.48372516  0.175260177 -0.42732077
##                  rm         age         dis          rad         tax    ptratio
## crim    -0.21924670  0.35273425 -0.37967009  0.625505145  0.58276431  0.2899456
## zn       0.31199059 -0.56953734  0.66440822 -0.311947826 -0.31456332 -0.3916785
## indus   -0.39167585  0.64477851 -0.70802699  0.595129275  0.72076018  0.3832476
## chas     0.09125123  0.08651777 -0.09917578 -0.007368241 -0.03558652 -0.1215152
## nox     -0.30218819  0.73147010 -0.76923011  0.611440563  0.66802320  0.1889327
## rm       1.00000000 -0.24026493  0.20524621 -0.209846668 -0.29204783 -0.3555015
## age     -0.24026493  1.00000000 -0.74788054  0.456022452  0.50645559  0.2615150
## dis      0.20524621 -0.74788054  1.00000000 -0.494587930 -0.53443158 -0.2324705
## rad     -0.20984667  0.45602245 -0.49458793  1.000000000  0.91022819  0.4647412
## tax     -0.29204783  0.50645559 -0.53443158  0.910228189  1.00000000  0.4608530
## ptratio -0.35550149  0.26151501 -0.23247054  0.464741179  0.46085304  1.0000000
## black    0.12806864 -0.27353398  0.29151167 -0.444412816 -0.44180801 -0.1773833
## lstat   -0.61380827  0.60233853 -0.49699583  0.488676335  0.54399341  0.3740443
## medv     0.69535995 -0.37695457  0.24992873 -0.381626231 -0.46853593 -0.5077867
##               black      lstat       medv
## crim    -0.38506394  0.4556215 -0.3883046
## zn       0.17552032 -0.4129946  0.3604453
## indus   -0.35697654  0.6037997 -0.4837252
## chas     0.04878848 -0.0539293  0.1752602
## nox     -0.38005064  0.5908789 -0.4273208
## rm       0.12806864 -0.6138083  0.6953599
## age     -0.27353398  0.6023385 -0.3769546
## dis      0.29151167 -0.4969958  0.2499287
## rad     -0.44441282  0.4886763 -0.3816262
## tax     -0.44180801  0.5439934 -0.4685359
## ptratio -0.17738330  0.3740443 -0.5077867
## black    1.00000000 -0.3660869  0.3334608
## lstat   -0.36608690  1.0000000 -0.7376627
## medv     0.33346082 -0.7376627  1.0000000
```

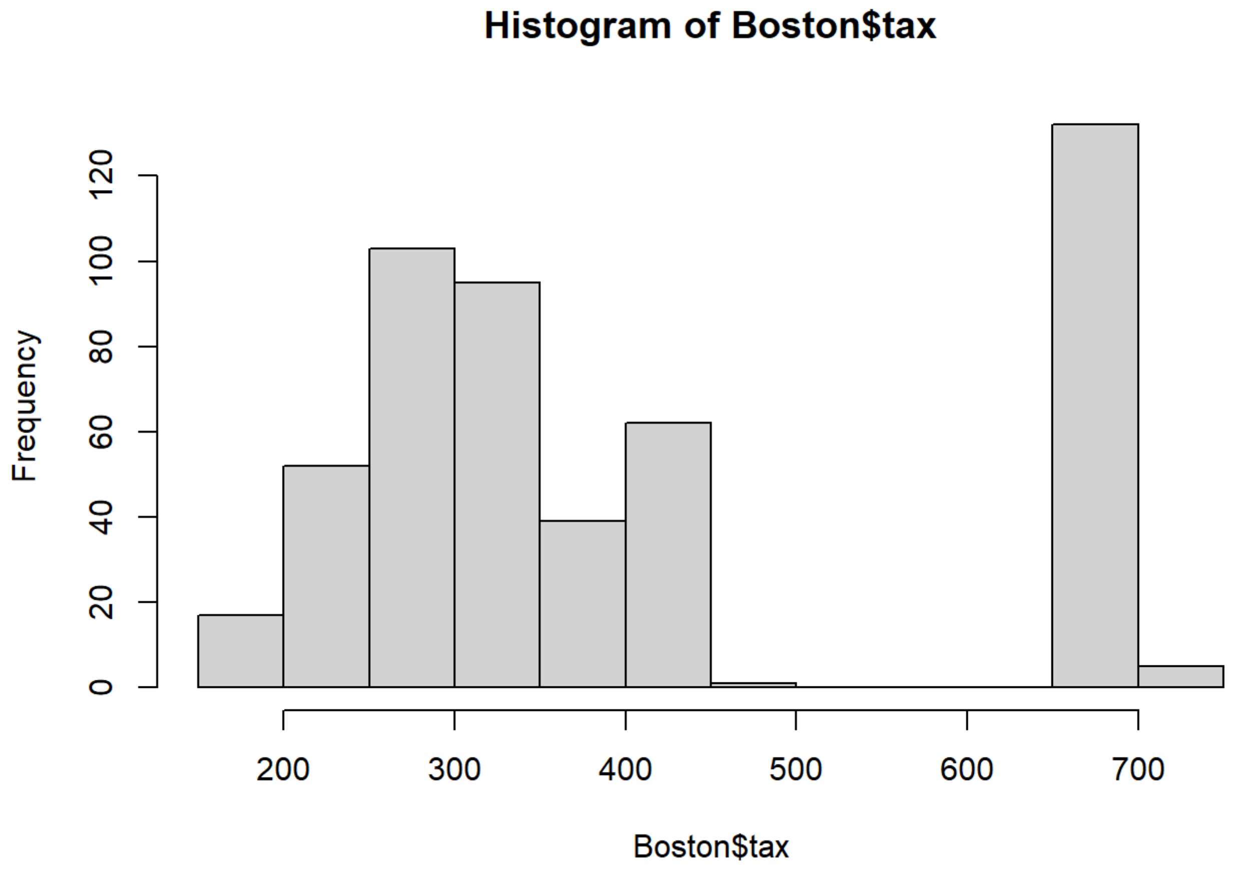
From the data above we can assume that there are a few variables have correlations with crim. crim has a positive relationship with indus, nox, rad, tax and lstat, and the coefficients are all greater than 0.4; crim has a negative relationship with medv, dis and black, and the coefficients are all smaller than -0.3.

#(d)

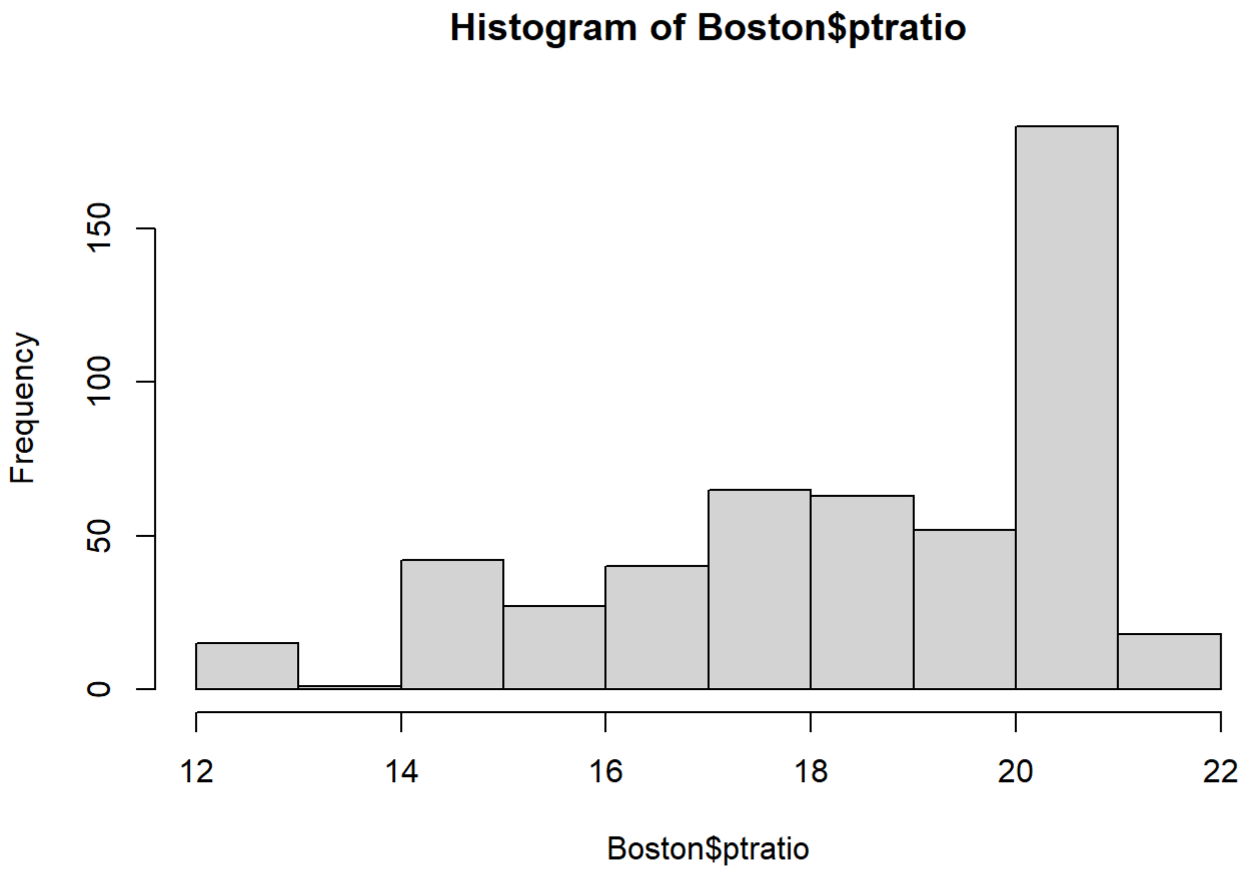
```
hist(Boston$crim[Boston$crim > 20])
```



```
hist(Boston$tax)
```



```
hist(Boston$ptratio)
```



```
sapply(Boston[(1:14)], range)
```

```
##          crim  zn indus chas   nox    rm   age     dis rad tax ptratio  black
## [1,]  0.00632   0  0.46    0 0.385 3.561   2.9  1.1296   1 187    12.6   0.32
## [2,] 88.97620 100 27.74    1 0.871 8.780 100.0 12.1265  24 711    22.0 396.90
##      lstat medv
## [1,]  1.73    5
## [2,] 37.97   50
```

```
sapply(Boston[(1:14)], mean)
```

```
##         crim           zn        indus         chas          nox           rm
##   3.61352356  11.36363636  11.13677866   0.06916996   0.55469506   6.28463439
##          age          dis          rad          tax      ptratio        black
##  68.57490119   3.79504269   9.54940711 408.23715415  18.45553360 356.67403162
##        lstat         medv
##  12.65306324  22.53280632
```

```
sapply(Boston[(1:14)], sd)
```

```
##        crim          zn       indus        chas         nox          rm 
##   8.6015451  23.3224530   6.8603529   0.2539940   0.1158777   0.7026171 
##         age         dis         rad         tax     ptratio       black 
##  28.1488614   2.1057101   8.7072594 168.5371161   2.1649455  91.2948644 
##       lstat        medv 
##   7.1410615   9.1971041
```

1. From the data above we can assume that there are 18 suburbs appear to have a crime rate larger than 20, reaching to above 80.
2. Also, there are a few suburbs have a high tax rate, but most suburbs’ tax rate are lower than 450.
3. According to the diagram, most of the suburbs have a Pupil-teacher ratios higher than 14, and the number of suburbs with Pupil-teacher ratios between 20 and 21 is the highest.

#(e)

```
sum(Boston$chas == 1)
```

```
## [1] 35
```

There are 35 suburbs in this data set bound the Charles river.

#(f)

```
summary(Boston$ptratio)
```

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

The median pupil-teacher ratio among the suburbs is 19.05.

#(g)

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

	crim <dbl>	zn <dbl>	indus <dbl>	chas <int>	nox <dbl>	rm <dbl>	age <dbl>	dis <dbl>	rad <int>
399	38.3518	0	18.1	0	0.693	5.453	100	1.4896	24
406	67.9208	0	18.1	0	0.693	5.683	100	1.4254	24

2 rows | 1-10 of 15 columns

```
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          black       
##  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  
##      lstat            medv      
##  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
```

We can know that there are two suburbs, 399 and 406, that have the lowest median property values. Also we can find that black red and age are higher than the mean, zn is much lower than the mean, other predictors are close to the mean.

#(h)

```
sum(Boston$rm > 7)
```

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

```
sum(Boston$rm > 8)
```

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

```
summary(subset(Boston, rm > 8))
```

```
##       crim                zn             indus             chas       
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
##       nox               rm             age             dis       
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
##       rad              tax           ptratio          black      
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2  
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7  
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9  
##      lstat           medv     
##  Min.   :2.47   Min.   :21.9  
##  1st Qu.:3.32   1st Qu.:41.7  
##  Median :4.14   Median :48.3  
##  Mean   :4.31   Mean   :44.2  
##  3rd Qu.:5.12   3rd Qu.:50.0  
##  Max.   :7.44   Max.   :50.0  
```

From above we can know that there are 64 suburbs average more than seven rooms per dwelling, and 13 more than eight rooms per dwelling. Also we can find that these suburbs have a lower crim, tax, lstat and indus, and a higher age and medv.