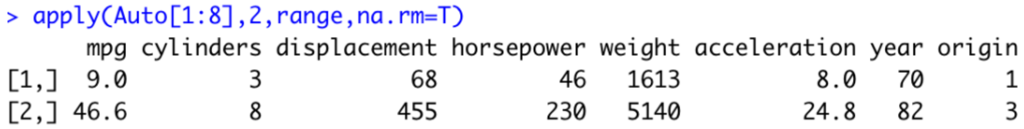
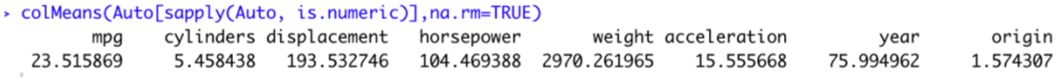
1.
   a.
      i. Regression: We want to create a model with response element – CEO salary which is a single quantitative output.
      ii. Inference: We want to comment on (infer) which predictor variables effect the CEO salary.
      iii. n -> 500, p -> 3
   b.
      i. Classification: The response element is qualitative with two possible categorical outputs {success, failure}.
      ii. Prediction: We want to use the classifier that we have built based on our training data of the previous 20 similar products to predict whether our new product will be a success or a failure.
      iii. n -> 20, p -> 13
   c.
      i. Regression: We want to create a model with response element - % change in US dollar which is a single quantitative output.
      ii. Prediction: We want to use our regression model which has been created using the training data from 2012 to make predictions on what the % change in dollar will be in the week(s) after 2012.
      iii. n -> 52, p -> 3
2.
   a.
      i. Quantitative -> mpg, cylinders, displacement, horsepower, weight, acceleration, year, origin
      ii. Qualitative -> name
   b.



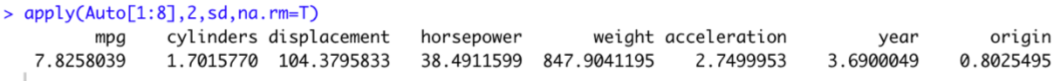
```
> apply(Auto[1:8],2,range,na.rm=T)
      mpg cylinders displacement horsepower weight acceleration year origin
[1,]  9.0         3           68         46   1613          8.0   70      1
[2,] 46.6         8          455        230   5140         24.8   82      3
```

   c. Mean’s:



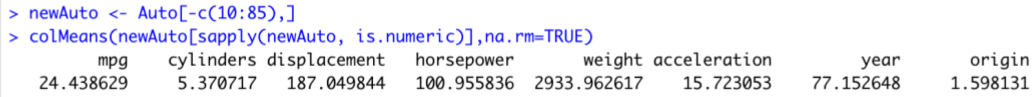
```
> colMeans(Auto[sapply(Auto, is.numeric)],na.rm=TRUE)
        mpg    cylinders displacement   horsepower       weight acceleration         year       origin
  23.515869     5.458438   193.532746   104.469388  2970.261965    15.555668    75.994962     1.574307
```

Standard Deviation’s:



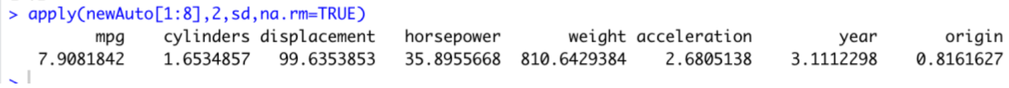
```
> apply(Auto[1:8],2,sd,na.rm=T)
        mpg    cylinders displacement   horsepower       weight acceleration         year       origin
  7.8258039    1.7015770  104.3795833   38.4911599  847.9041195    2.7499953    3.6900049    0.8025495
```

   d. Mean’s:



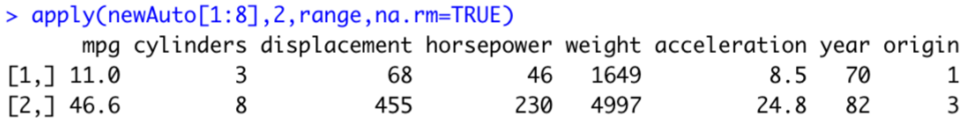
```
> newAuto <- Auto[-c(10:85),]
> colMeans(newAuto[sapply(newAuto, is.numeric)],na.rm=TRUE)
        mpg    cylinders displacement   horsepower       weight acceleration         year       origin
  24.438629     5.370717   187.049844   100.955836  2933.962617    15.723053    77.152648     1.598131
```

Standard Deviation’s:



```
> apply(newAuto[1:8],2,sd,na.rm=TRUE)
        mpg    cylinders displacement   horsepower       weight acceleration         year       origin
  7.9081842    1.6534857   99.6353853   35.8955668  810.6429384    2.6805138    3.1112298    0.8161627
```

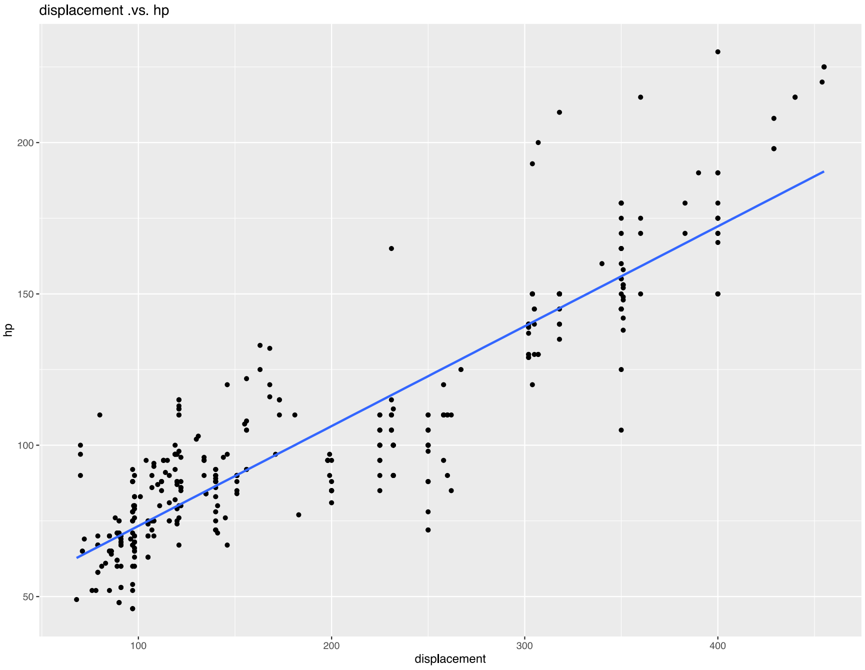
Range's:



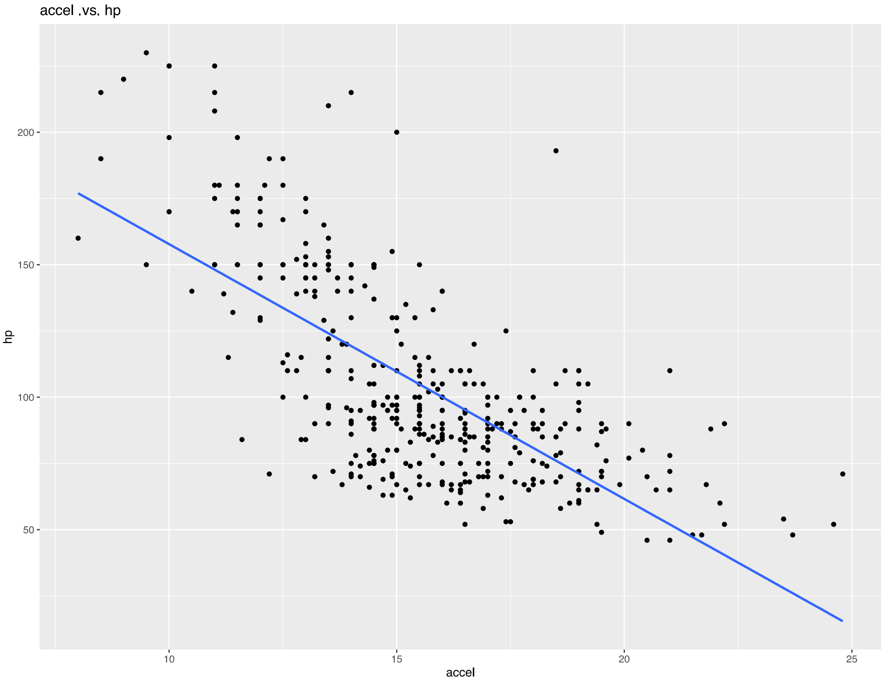
```
> apply(newAuto[1:8],2,range,na.rm=TRUE)
      mpg cylinders displacement horsepower weight acceleration year origin
[1,] 11.0         3           68         46   1649          8.5   70      1
[2,] 46.6         8          455        230   4997         24.8   82      3
```

e.

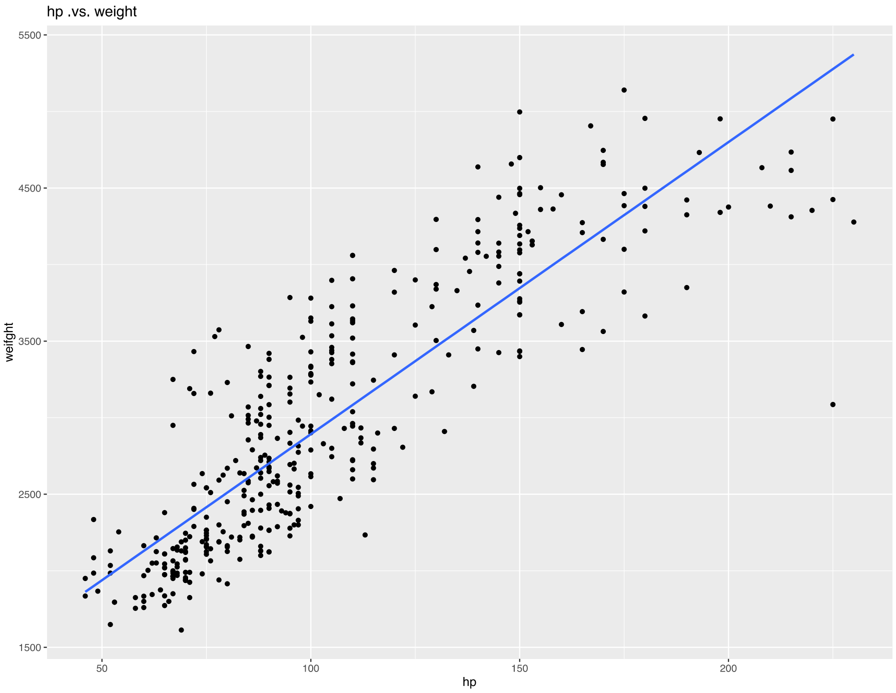
i. The scatter plot of displacement .vs. horsepower below shows that these variables are strongly positively correlated.



ii. The scatter plot of acceleration .vs. horsepower shows that these variables are mildly positively correlated (taking a lower acceleration as a positive increase in acceleration)



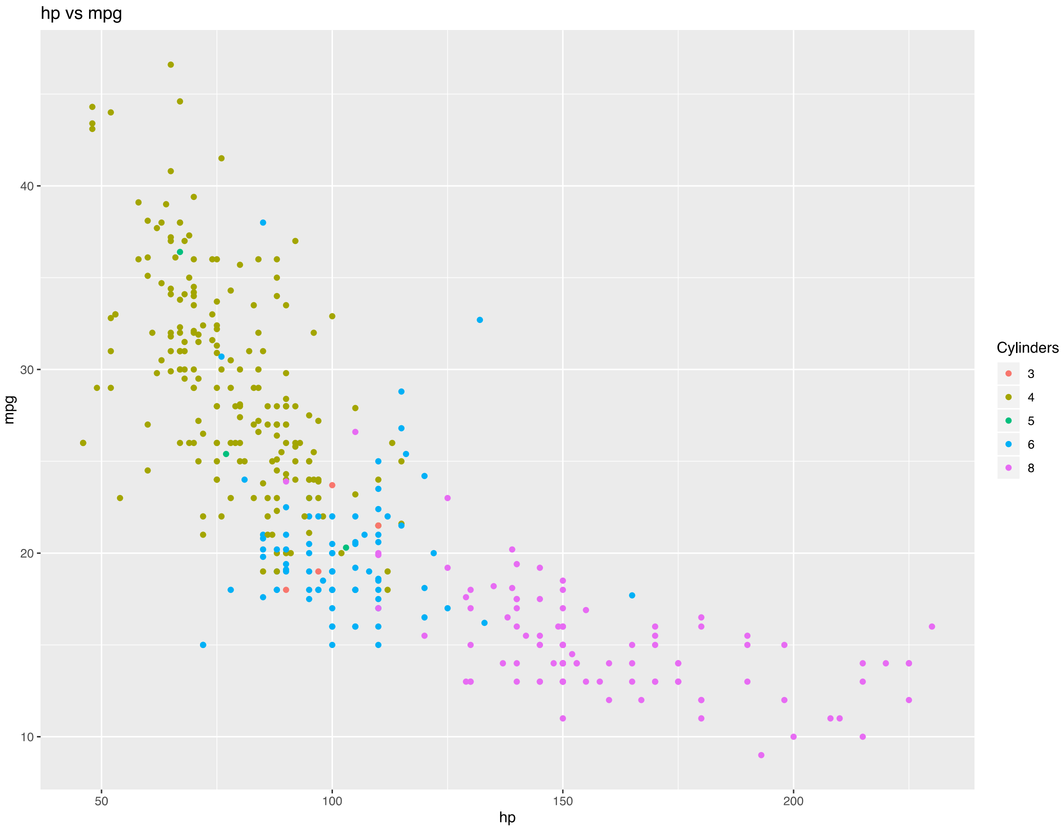
iii. The scatter plot of horsepower .vs. weight shows that there is a strong positive correlation between these variables.



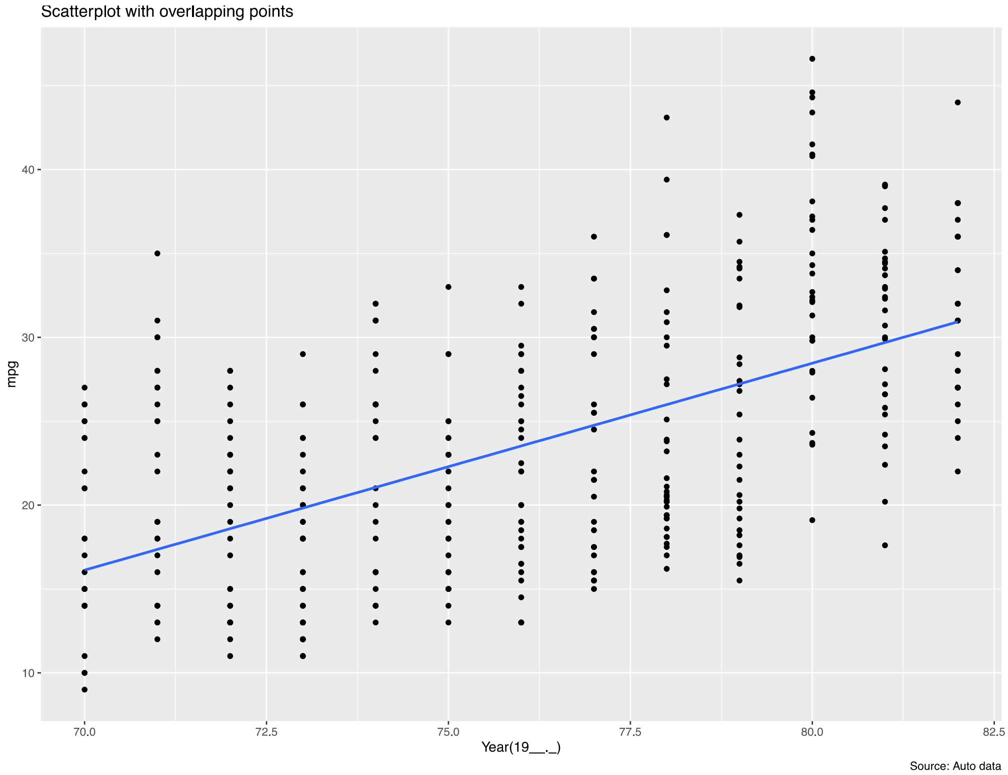
From these graphs it can be seen that the horsepower of the car is positively related to the acceleration, displacement and weight of the car so it will be useful to only look at the effect of the horsepower on the mpg and this will give us an idea of the effect of the other 3 factors on the mpg as well.

Conclusions:

The cylinders of each car are colour coded, so we can see that the number of cylinders a car has and it's horsepower are positively correlated. The scatter plot below shows that the horsepower and mpg of a car have a strong non-linear negative correlation (something like that of an exponential decay).



The scatter plot with overlapping points below shows that the year the car was produced and the mpg of the given car are positively correlated.



These are the two main factors on the mpg of the car.

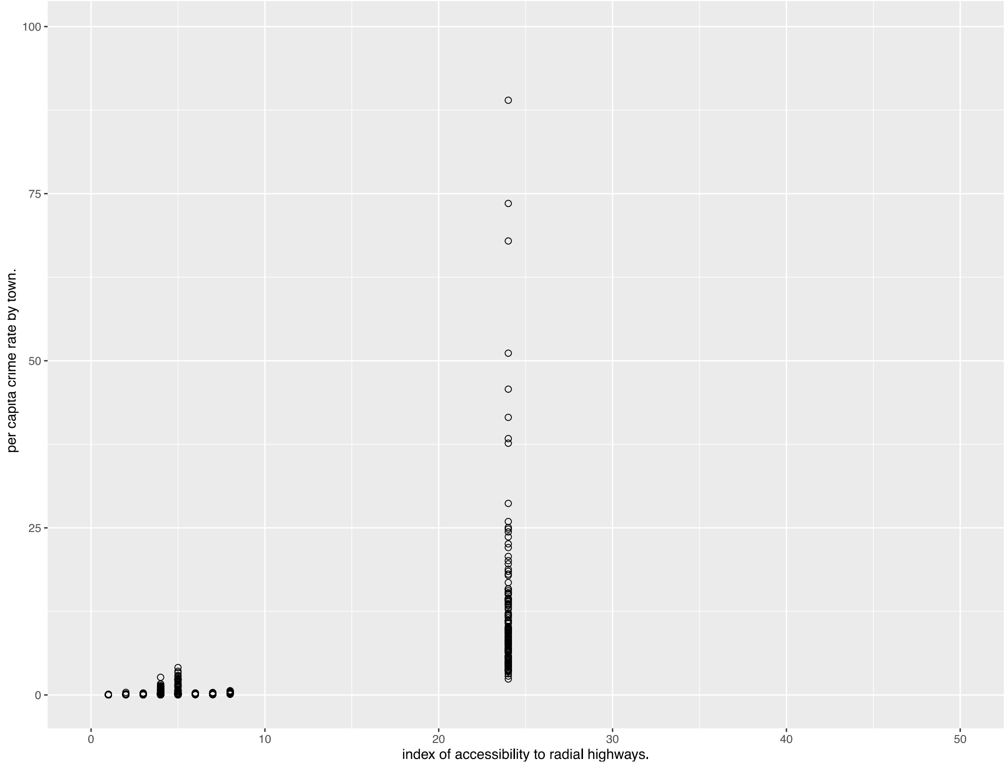
3.

a.

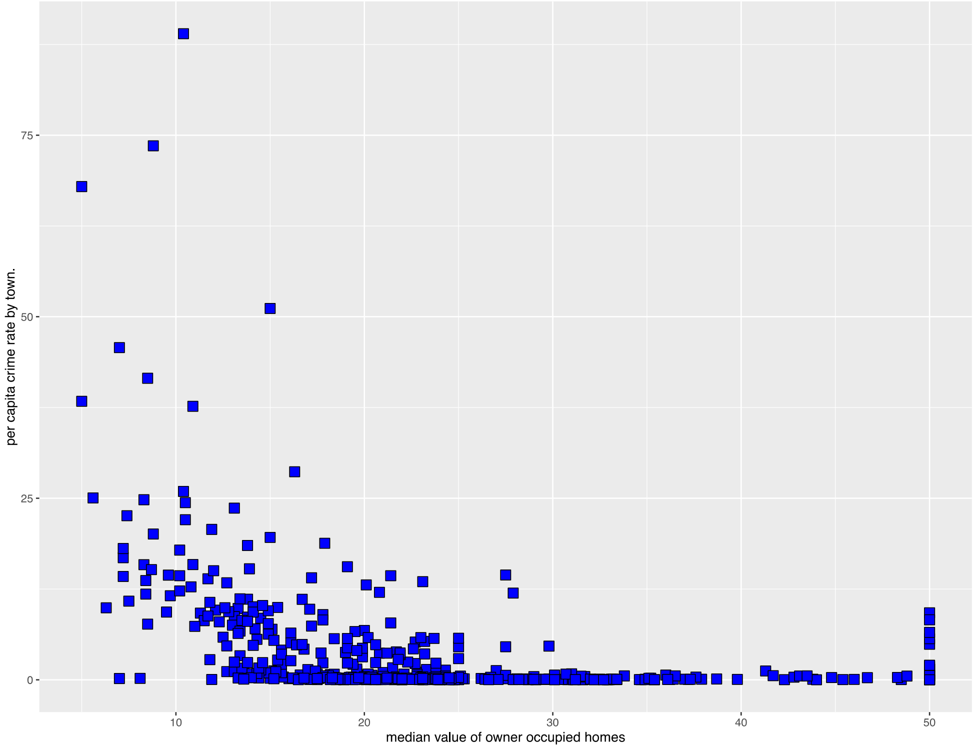
i. There are 506 rows which correspond to the data of each suburb of Boston.

ii. There are 14 columns which correspond to the variables that are collected on the suburbs of Boston.

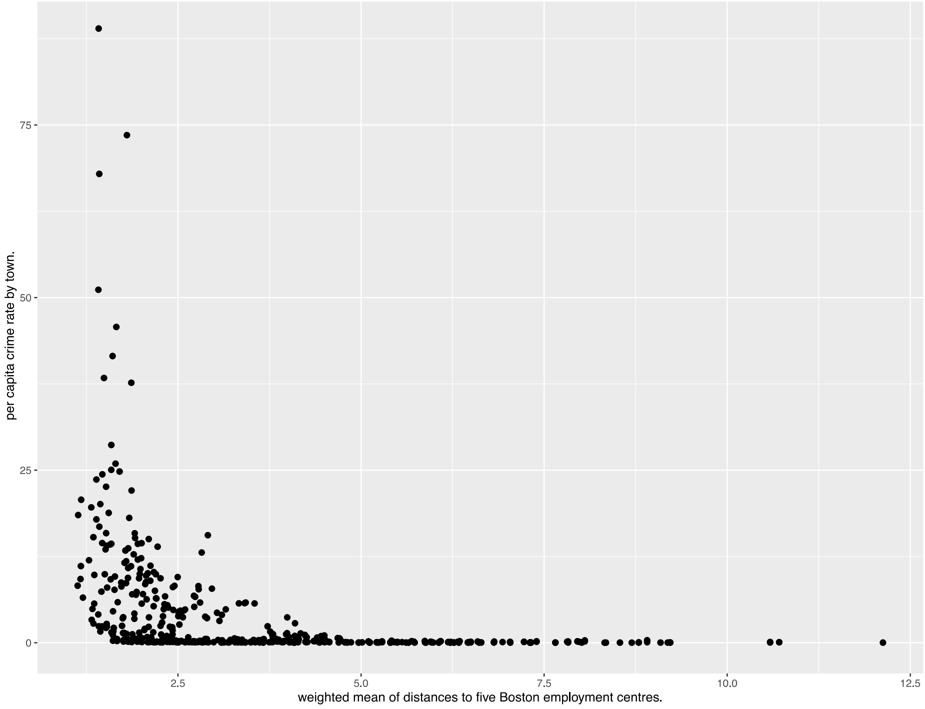
b.



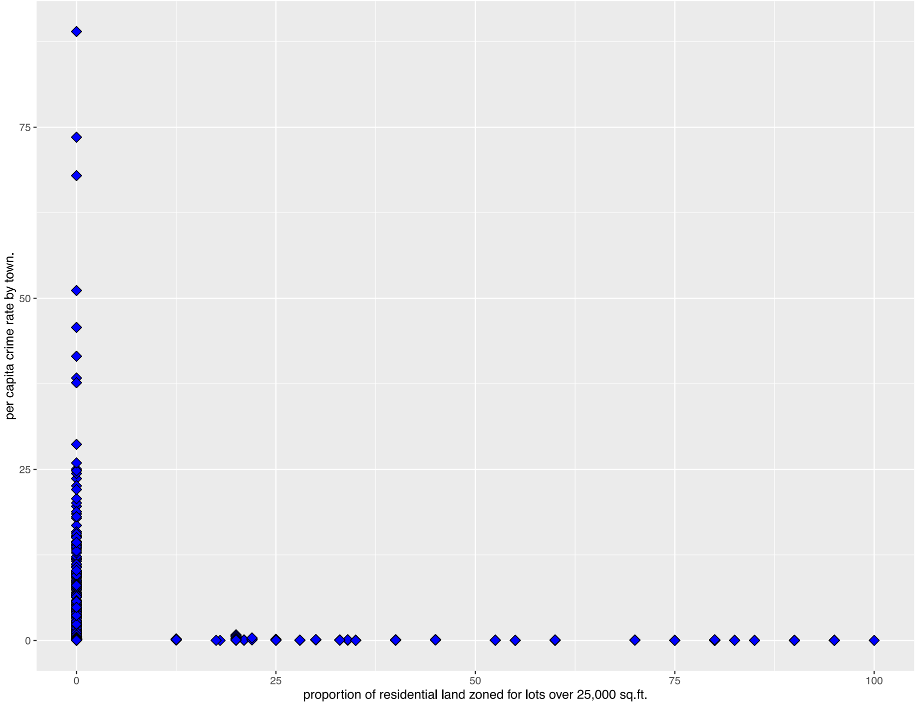
In the above scatter plot it can be seen that there is a strong non-linear positive correlation (something like that of exponential growth) between access to radial highways and crime rates of a suburb), I have adjusted the x,y scales to try and make this more obvious.



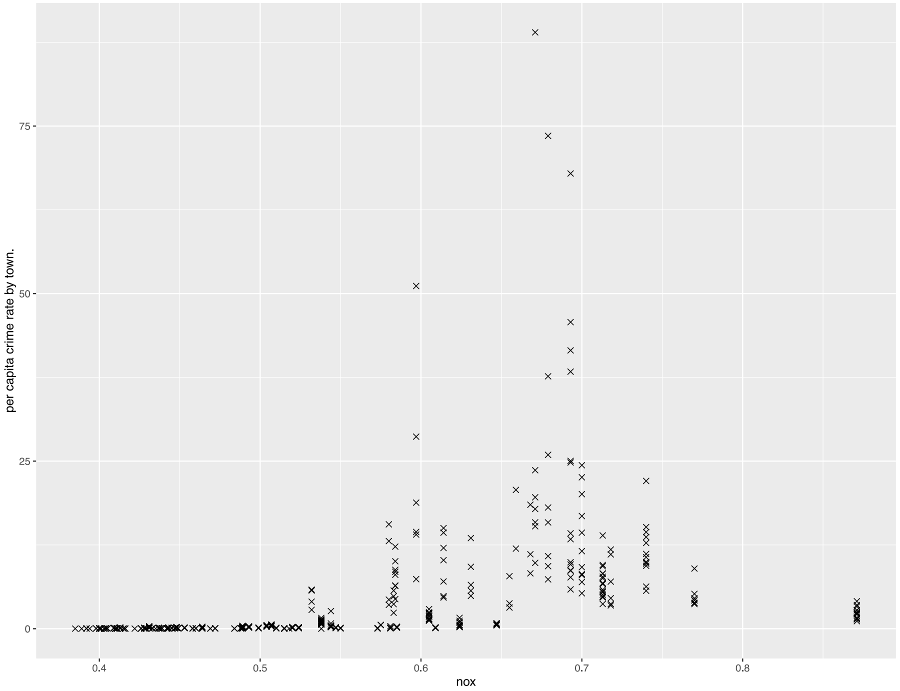
In the above scatter plot it can be seen that there is a strong non-linear negative correlation (something like that of an exponential decay) between median value of owner occupied homes and crime rate of a suburb.



In the above scatter plot it can be seen that there is a strong non-linear negative correlation (something like that of an inverse square relationship) between weighted mean of distances to five Boston employment centres and crime rate of a suburb.

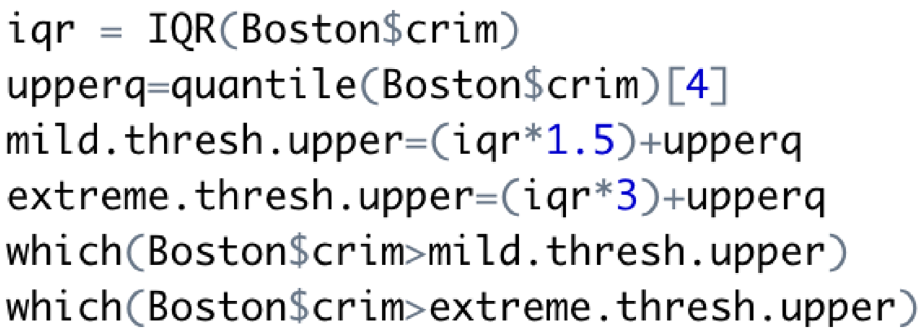


In the above scatter plot it can be seen that at 0 proportion of residential land zoned for lots over 25,00sq.ft the crime rate range is between 0-80 with a dense spread of values in the range of 0-25 crime rate however all other values for the proportion of residential land zoned for lots over 25,00sq.ft have crime rate ~ 0.



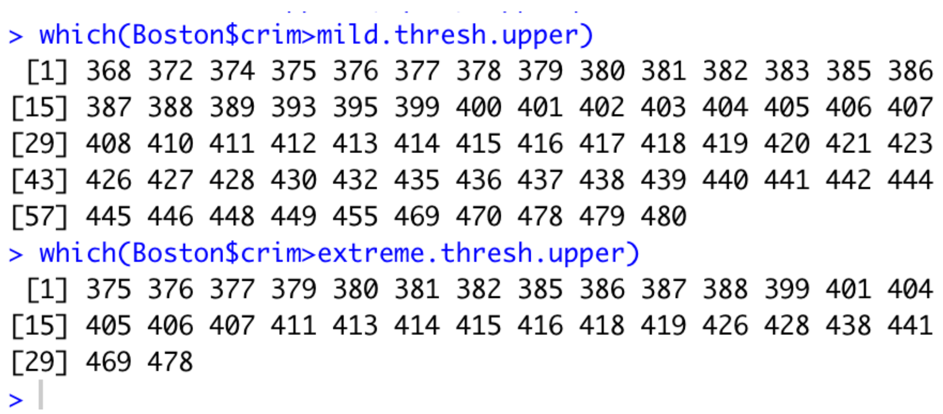
From the above scatter plot there seems to be that the crime rate is normally distributed about the nitrogen oxides concentration with mean ~0.67.

c. To calculate which towns in Boston had a crime rate above a certain threshold I used the following in R:



```
iqr = IQR(Boston$crim)
upperq=quantile(Boston$crim)[4]
mild.thresh.upper=(iqr*1.5)+upperq
extreme.thresh.upper=(iqr*3)+upperq
which(Boston$crim>mild.thresh.upper)
which(Boston$crim>extreme.thresh.upper)
```

Which provided me with the following suburbs:

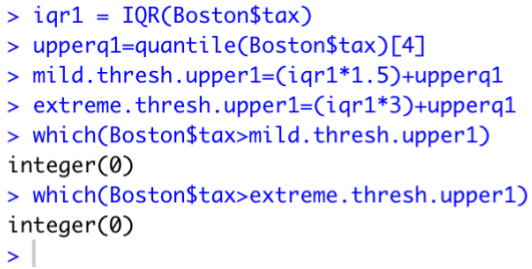


```
> which(Boston$crim>mild.thresh.upper)
 [1] 368 372 374 375 376 377 378 379 380 381 382 383 385 386
[15] 387 388 389 393 395 399 400 401 402 403 404 405 406 407
[29] 408 410 411 412 413 414 415 416 417 418 419 420 421 423
[43] 426 427 428 430 432 435 436 437 438 439 440 441 442 444
[57] 445 446 448 449 455 469 470 478 479 480
> which(Boston$crim>extreme.thresh.upper)
 [1] 375 376 377 379 380 381 382 385 386 387 388 399 401 404
[15] 405 406 407 411 413 414 415 416 418 419 426 428 438 441
[29] 469 478
> 
```

So it is clear to see that there are a lot of suburbs with particularly high crime rates above a mild and extreme threshold.

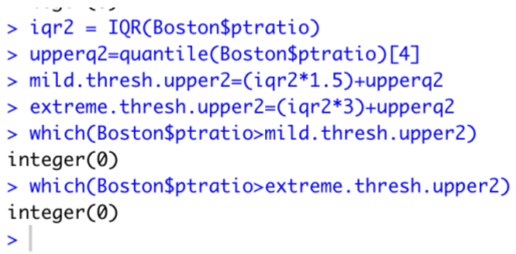
Using the same formula but for tax rates and pupil to teacher ratio I found that there were no towns with particularly high levels of either.

Tax Rate:



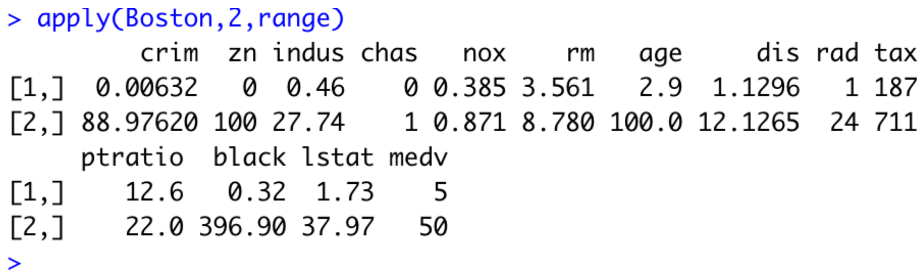
```
> iqr1 = IQR(Boston$tax)
> upperq1=quantile(Boston$tax)[4]
> mild.thresh.upper1=(iqr1*1.5)+upperq1
> extreme.thresh.upper1=(iqr1*3)+upperq1
> which(Boston$tax>mild.thresh.upper1)
integer(0)
> which(Boston$tax>extreme.thresh.upper1)
integer(0)
> 
```

Pupil to Teacher Ratio:



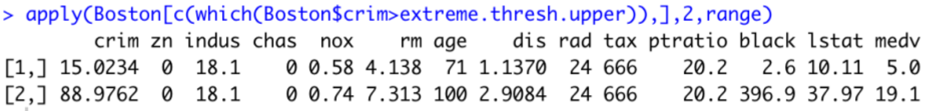
```
> iqr2 = IQR(Boston$ptratio)
> upperq2=quantile(Boston$ptratio)[4]
> mild.thresh.upper2=(iqr2*1.5)+upperq2
> extreme.thresh.upper2=(iqr2*3)+upperq2
> which(Boston$ptratio>mild.thresh.upper2)
integer(0)
> which(Boston$ptratio>extreme.thresh.upper2)
integer(0)
> 
```

The range of each predictor of the entire data set:



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

Range of predictors with crime rate above extreme threshold:



```
> apply(Boston[c(which(Boston$crim>extreme.thresh.upper)),],2,range)
        crim zn indus chas  nox    rm age    dis rad tax ptratio black lstat medv
[1,] 15.0234  0  18.1    0 0.58 4.138  71 1.1370  24 666    20.2   2.6 10.11  5.0
[2,] 88.9762  0  18.1    0 0.74 7.313 100 2.9084  24 666    20.2 396.9 37.97 19.1
```

Comments on ranges of predictors:

proportion of residential land zoned for lots over 25,000 sq.ft: The suburbs with extremely high crime rates all have no land zoned for lots over 25,000 sq ft.

proportion of non-retail business acres per town: The suburbs with extremely high crime rates have a proportion of non-retail business acres per town around the middle of the range for all the suburbs.

Charles River dummy variable: None of the suburbs with extremely high crime rates are bounded by the Charles River dummy variable.

Nitrogen oxides concentration: The suburbs with extremely high crime rates have a reasonably high nitrogen oxides concentration compared to the range for all of the suburbs.

Average number of rooms per dwelling: The range of this predictor for the suburbs with extremely high crime rates is similar to that of the range for all the suburbs.

Proportion of owner-occupied units built prior to 1940: The suburbs with extremely high crime rates are particularly old compared to the range for all the suburbs.

Weighted mean of distances to five Boston employment centres: Compared to all of the suburbs of Boston the suburbs with extremely high crime rates are very close to employment centres.

Index of accessibility to radial highways: Compared to all of the suburbs of Boston the suburbs with extremely high crime rates have the best possible access to radial highways.

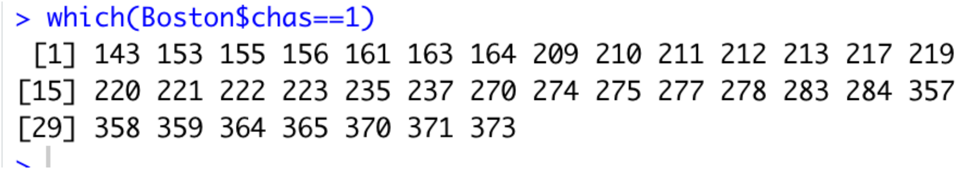
Pupil-teacher ratio by town: Compared to all of the suburbs of Boston the suburbs with extremely high crime rates have very high pupil to teacher ratio.

1000(Bk - 0.63)^2 where *Bk* is the proportion of blacks by town: This range for extremely high crime rates is essentially identical to that of all suburbs.

lower status of the population (percent): Compared to all of the suburbs of Boston the suburbs with extremely high crime rates have slightly higher range of % of lower status of the population.

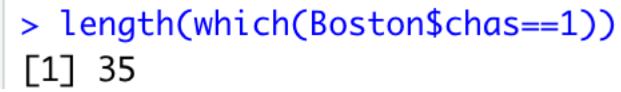
median value of owner-occupied homes in \$1000s: Compared to all of the suburbs of Boston the suburbs with extremely high crime rates have a much lower range of median value of owner-occupied homes.

d. The following suburbs bound the Charles River.



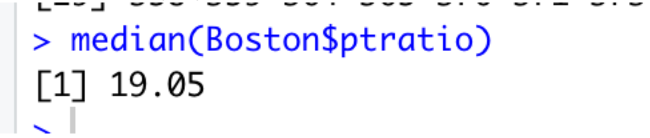
```
> which(Boston$chas==1)
 [1] 143 153 155 156 161 163 164 209 210 211 212 213 217 219
[15] 220 221 222 223 235 237 270 274 275 277 278 283 284 357
[29] 358 359 364 365 370 371 373
>
```

Of which there are a total of:



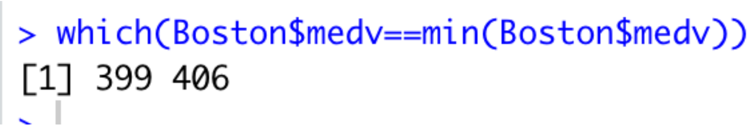
```
> length(which(Boston$chas==1))
[1] 35
```

e. The median pupil to teacher ratio is:



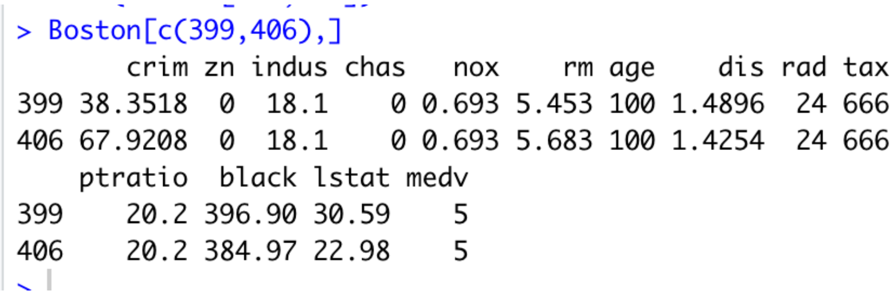
```
> median(Boston$ptratio)
[1] 19.05
>
```

f. The suburbs with the lowest median value of owner occupied homes are:



```
> which(Boston$medv==min(Boston$medv))
[1] 399 406
```

The values of the other predictors of these suburbs are:



```
> Boston[c(399,406),]
       crim zn indus chas   nox    rm age    dis rad tax
399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666
406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666
    ptratio  black lstat medv
399    20.2 396.90 30.59    5
406    20.2 384.97 22.98    5
```

Comments on the predictors:

per capita crime rate by town: The suburbs with lowest median value of owner-occupied homes both have extremely high crime rates (from part c).

proportion of residential land zoned for lots over 25,000 sq.ft: The suburbs with lowest median value of owner-occupied homes are both have no land zoned for lots over 25,000 sq ft.

proportion of non-retail business acres per town: The suburbs with lowest median value of owner-occupied homes both have a proportion of non-retail business acres per town around the middle of the range for all the suburbs.

Charles River dummy variable The suburbs with lowest median value of owner-occupied homes are not bounded by the Charles River dummy variable.

Nitrogen oxides concentration: The suburbs The suburbs with lowest median value of owner-occupied homes both have a reasonably high nitrogen oxides concentration compared to the range for all of the suburbs.

Average number of rooms per dwelling: The suburbs with lowest median value of owner-occupied homes have number of rooms per dwelling around the middle of the range for all the suburbs.

Proportion of owner-occupied units built prior to 1940: The suburbs with lowest median value of owner-occupied homes both are among the oldest suburbs.

Weighted mean of distances to five Boston employment centres: Compared to all of the suburbs of Boston the suburbs with lowest median value of owner-occupied homes both are extremely close to employment centres.

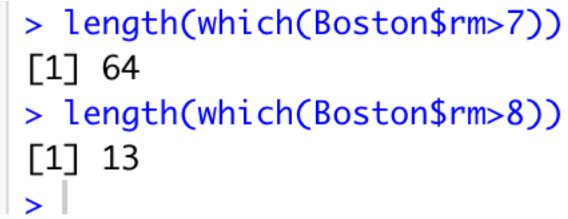
Index of accessibility to radial highways: Compared to all of the suburbs of Boston the suburbs with lowest median value of owner-occupied homes both have the best possible access to radial highways.

Pupil-teacher ratio by town: Compared to all of the suburbs of Boston the suburbs with lowest median value of owner-occupied homes both have very high pupil to teacher ratio.

1000(Bk - 0.63)^2 where *Bk* is the proportion of blacks by town: The suburbs with lowest median value of owner-occupied homes both have values for this predictor on the upper value of range for all the suburbs.

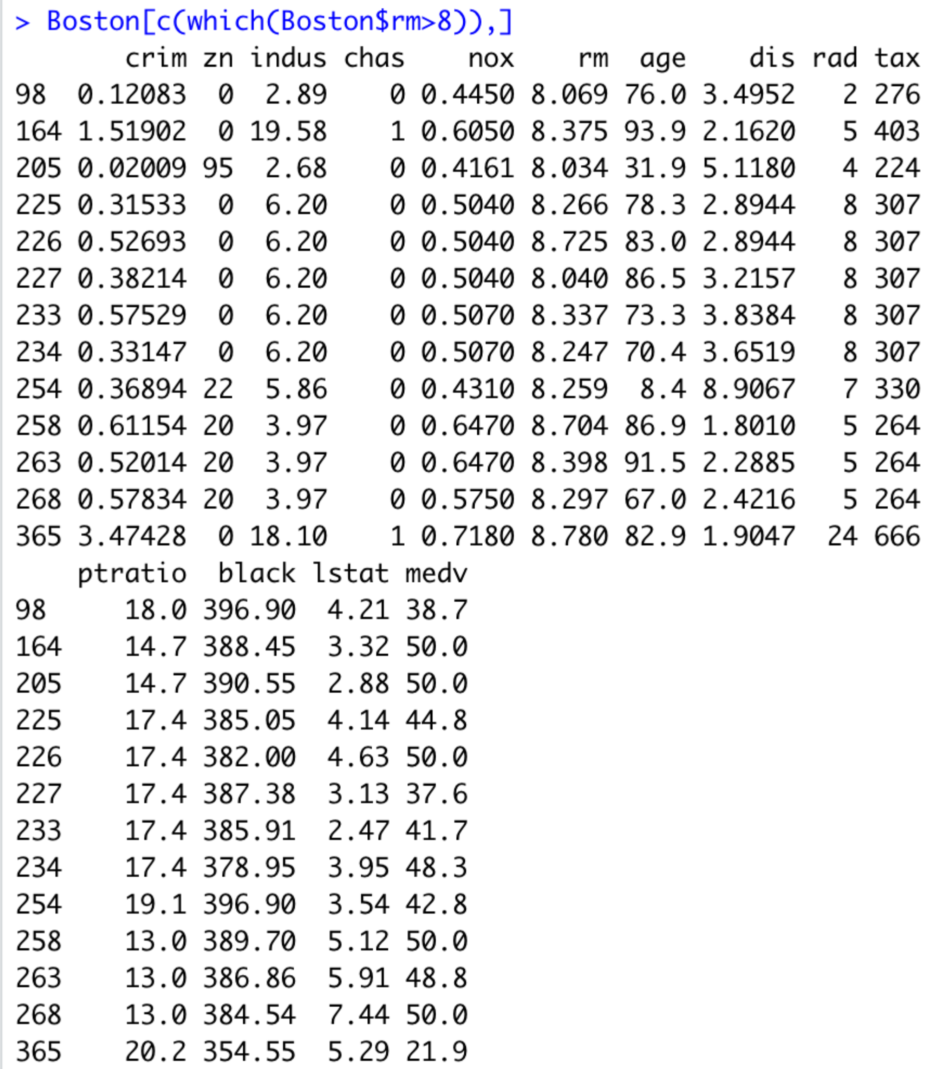
lower status of the population (percent): Compared to all of the suburbs of Boston the suburbs with lowest median value of owner-occupied homes both have particularly high % of lower status of the population.

g. Suburbs that average more than 7 or 8 rooms per dwelling:



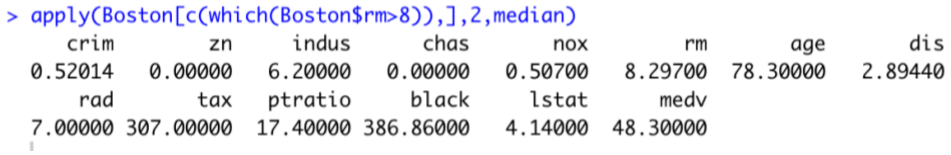
```
> length(which(Boston$rm>7))
[1] 64
> length(which(Boston$rm>8))
[1] 13
> 
```

The data for these suburbs can be seen below:



```
> Boston[c(which(Boston$rm>8)),]
       crim zn indus chas    nox    rm  age    dis rad tax
98  0.12083  0  2.89    0 0.4450 8.069 76.0 3.4952   2 276
164 1.51902  0 19.58    1 0.6050 8.375 93.9 2.1620   5 403
205 0.02009 95  2.68    0 0.4161 8.034 31.9 5.1180   4 224
225 0.31533  0  6.20    0 0.5040 8.266 78.3 2.8944   8 307
226 0.52693  0  6.20    0 0.5040 8.725 83.0 2.8944   8 307
227 0.38214  0  6.20    0 0.5040 8.040 86.5 3.2157   8 307
233 0.57529  0  6.20    0 0.5070 8.337 73.3 3.8384   8 307
234 0.33147  0  6.20    0 0.5070 8.247 70.4 3.6519   8 307
254 0.36894 22  5.86    0 0.4310 8.259  8.4 8.9067   7 330
258 0.61154 20  3.97    0 0.6470 8.704 86.9 1.8010   5 264
263 0.52014 20  3.97    0 0.6470 8.398 91.5 2.2885   5 264
268 0.57834 20  3.97    0 0.5750 8.297 67.0 2.4216   5 264
365 3.47428  0 18.10    1 0.7180 8.780 82.9 1.9047  24 666
    ptratio  black lstat medv
98     18.0 396.90  4.21 38.7
164    14.7 388.45  3.32 50.0
205    14.7 390.55  2.88 50.0
225    17.4 385.05  4.14 44.8
226    17.4 382.00  4.63 50.0
227    17.4 387.38  3.13 37.6
233    17.4 385.91  2.47 41.7
234    17.4 378.95  3.95 48.3
254    19.1 396.90  3.54 42.8
258    13.0 389.70  5.12 50.0
263    13.0 386.86  5.91 48.8
268    13.0 384.54  7.44 50.0
365    20.2 354.55  5.29 21.9
```

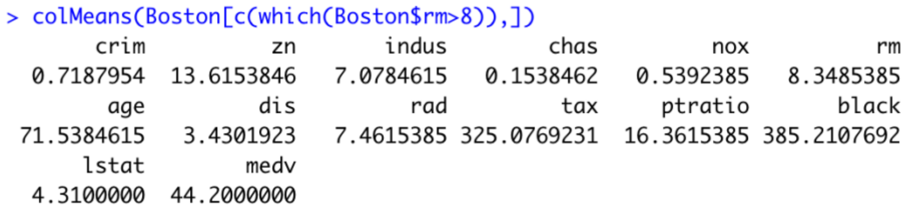
It is clear to see that there are outliers in the zn, indus, chas, age, rad, tax and medv columns so it may be more useful to look at the medians of these columns as an average:



```
> apply(Boston[c(which(Boston$rm>8)),],2,median)
    crim        zn     indus      chas       nox        rm       age       dis 
 0.52014   0.00000   6.20000   0.00000   0.50700   8.29700  78.30000   2.89440 
     rad       tax   ptratio     black     lstat      medv 
 7.00000 307.00000  17.40000 386.86000   4.14000  48.30000 
```

It can be seen that compared to the range of the entire data set, using the median as average, these suburbs have no proportion of land owned for lots over 25,000 sq. ft, have low proportion of non-retail business acres per town, are not bound by the Charles River dummy variable, are relatively old suburbs, and have a high median value of owner occupied homes in \$1000s.

Means of the columns of these suburbs:



```
> colMeans(Boston[c(which(Boston$rm>8)),])
       crim          zn       indus        chas         nox          rm 
  0.7187954  13.6153846   7.0784615   0.1538462   0.5392385   8.3485385 
        age         dis         rad         tax     ptratio       black 
 71.5384615   3.4301923   7.4615385 325.0769231  16.3615385 385.2107692 
      lstat        medv 
  4.3100000  44.2000000 
```

From the means, when comparing to the ranges of the entire data set, notably we can see that the crime rate of these suburbs is relatively low and the proportion of blacks per town is very high.

4.

a. We expect RSS(cubic) > RSS(linear).

   Since the linear model stated is the true relationship between Y and X this means the training data will fit the model close to perfectly. This means the RSS of the linear model will be minimal. The cubic model will not fit the training data as well, so there will be more “error” which entails a larger RSS of the cubic model than that of the linear model.

b. We expect RSS(cubic) > RSS(linear).

   The test data will follow the same probability distribution as the training data however will be independent of the training data. The true relationship between X and Y is the linear model stated so this means the test data should fit this model close to perfectly as in (a). As well as this the model was produced using n=100 and p=2 which implies that there will be minimal overfitting, so we expect the test data to fit the linear model “almost perfectly”. This means the RSS of the linear model using test data will be minimal whereas the cubic model will not fit the test data as well meaning the RSS of the cubic model using test data will be much greater than that of the linear model.

c. There is not enough information to tell. The true relationship between X and Y is non-linear but since we do not know the extent of the non-linearity this leaves a large number of possible non-linear models this could be. If we say each model is as likely as the next, the probability that the relationship is cubic as stated is very unlikely. This means we have no idea how well the cubic model fits the training data. Therefore it is impossible to speculate whether the RSS of the cubic or linear will be greater.

d. There is not enough information to tell. The test data again is independent from the training data however the fact that there will be minimal overfitting will not help us to decide whether the RSS of the cubic or linear will be greater since we do not know which relationship fits the training data better.