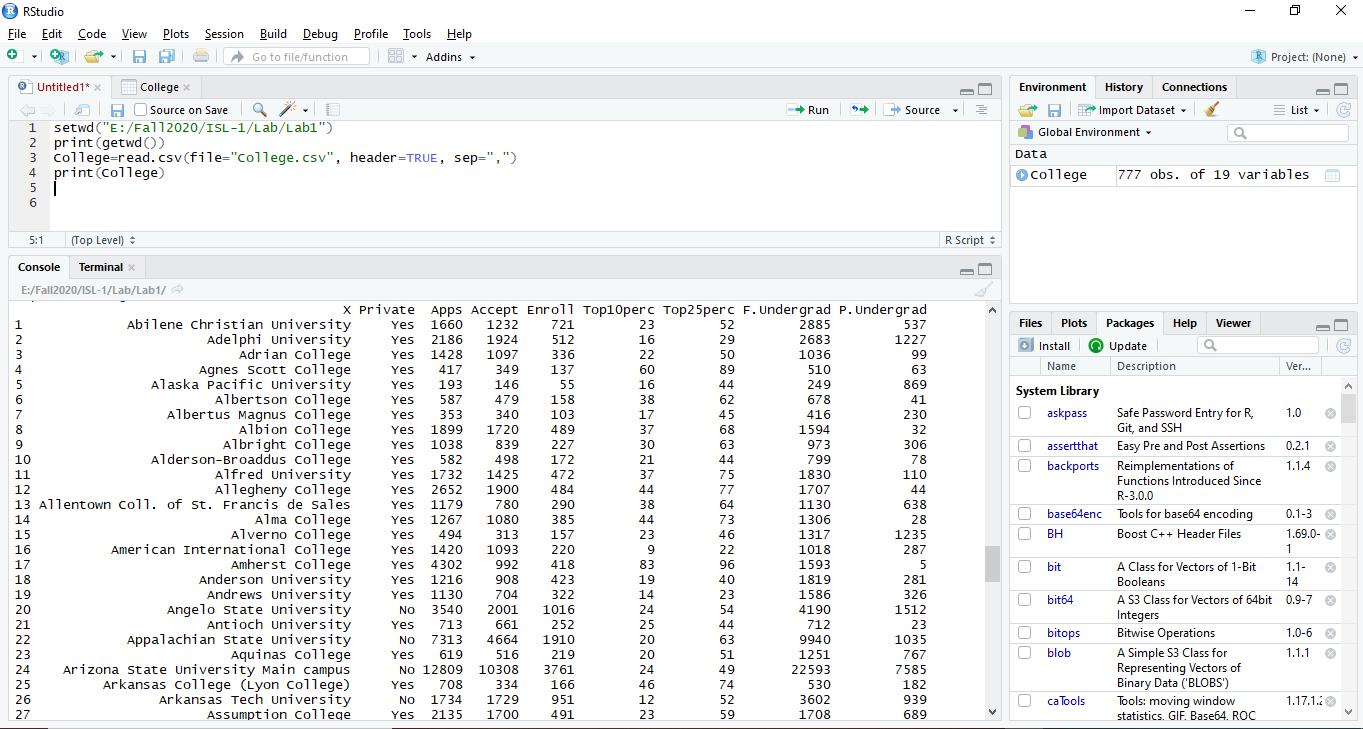
**Introduction to Statistical Learning- Lab1**

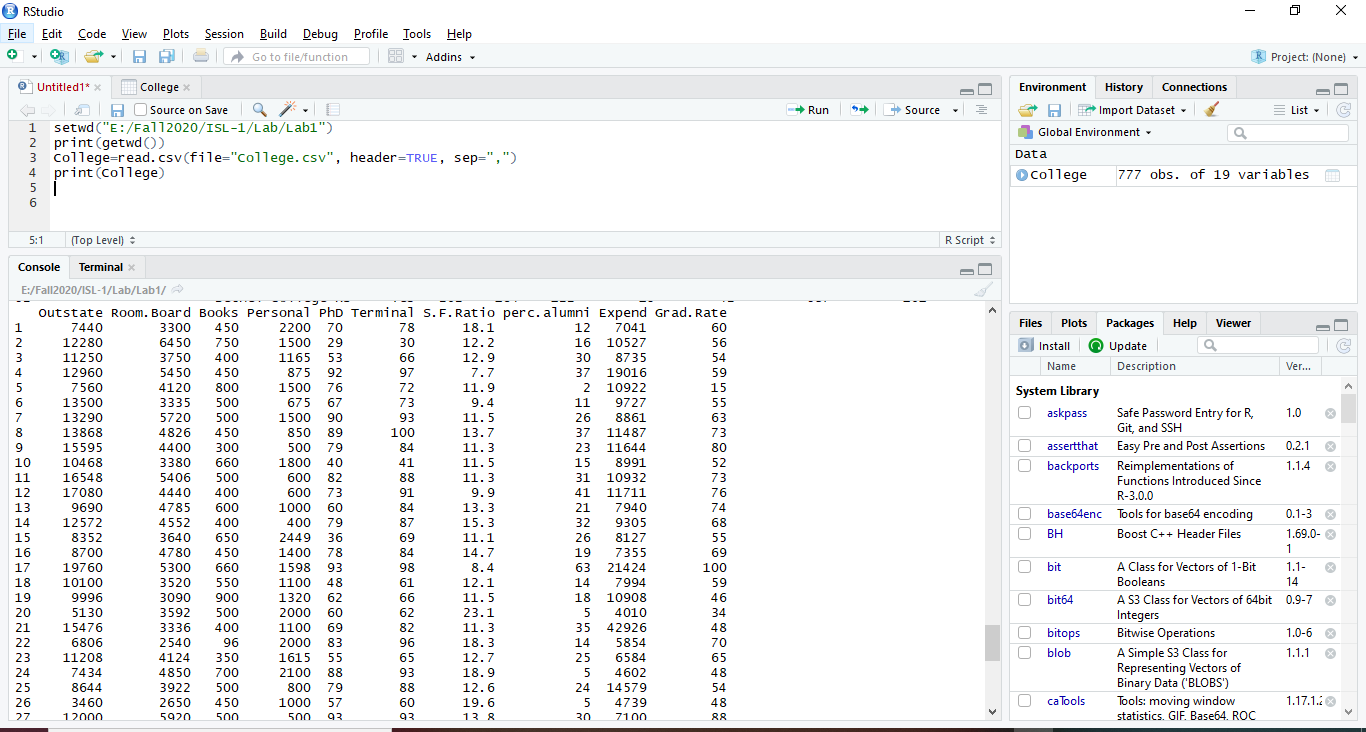
**Name: Sandeep Reddy Salkuti**

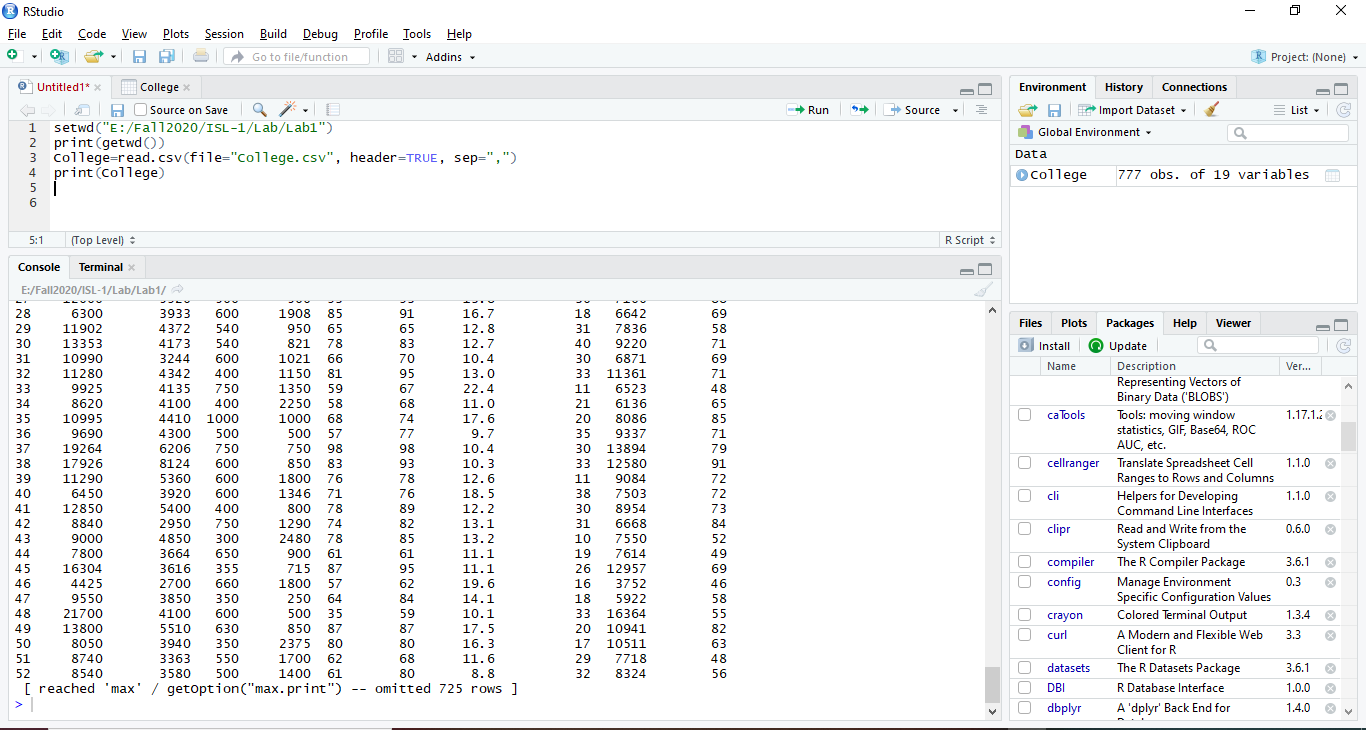
**Student id: 16296868**

**Email:** [**sswf7@umsystem.edu**](mailto:sswf7@umsystem.edu)

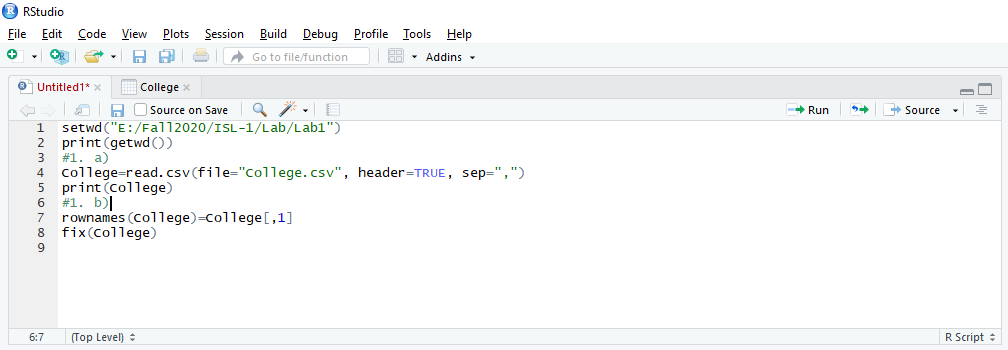
1. **Use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.**

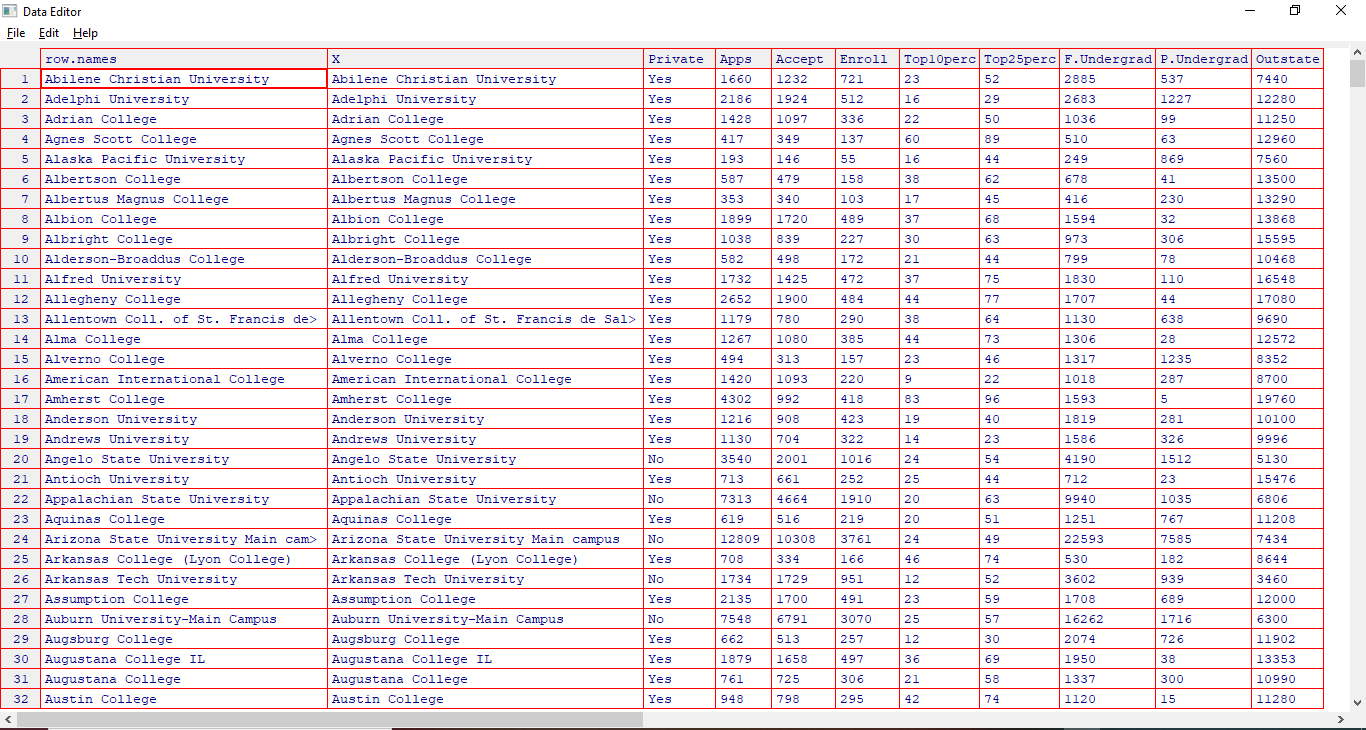




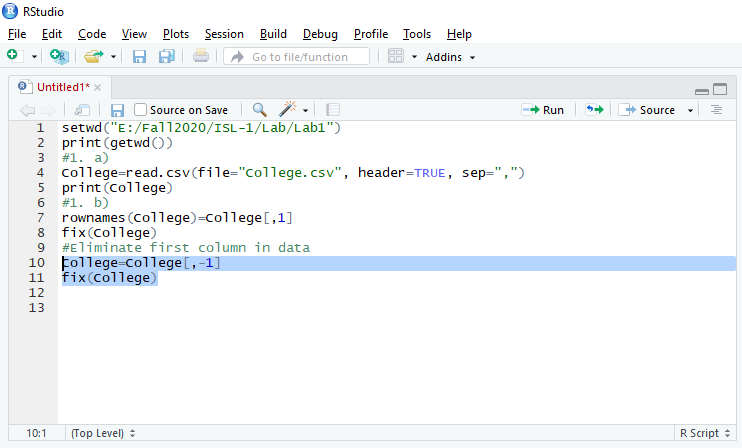


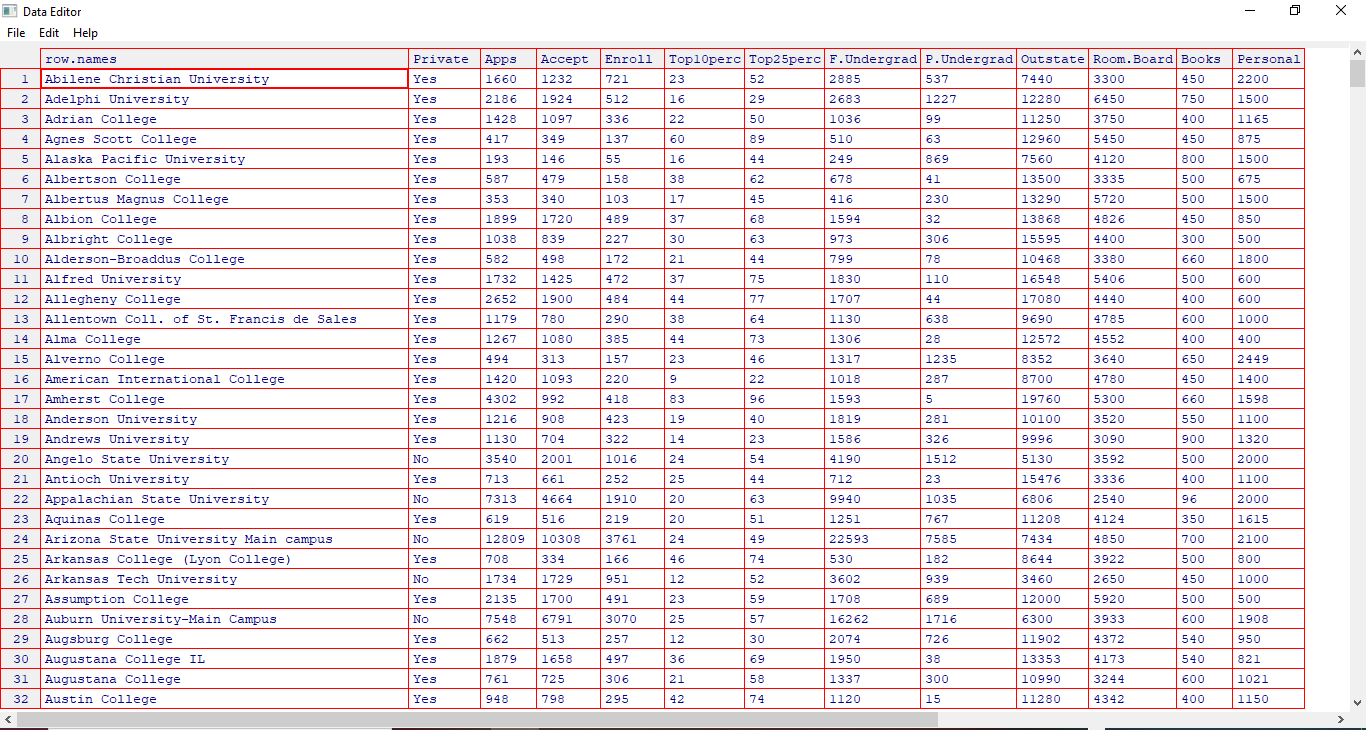
**b) Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don’t really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands: > rownames (college )=college [,1] > fix (college ) You should see that there is now a row.names column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try > college =college [,-1] > fix (college ) Now you should see that the first data column is Private. Note that another column labeled row.names now appears before the Private column. However, this is not a data column but rather the name that R is giving to each row.**





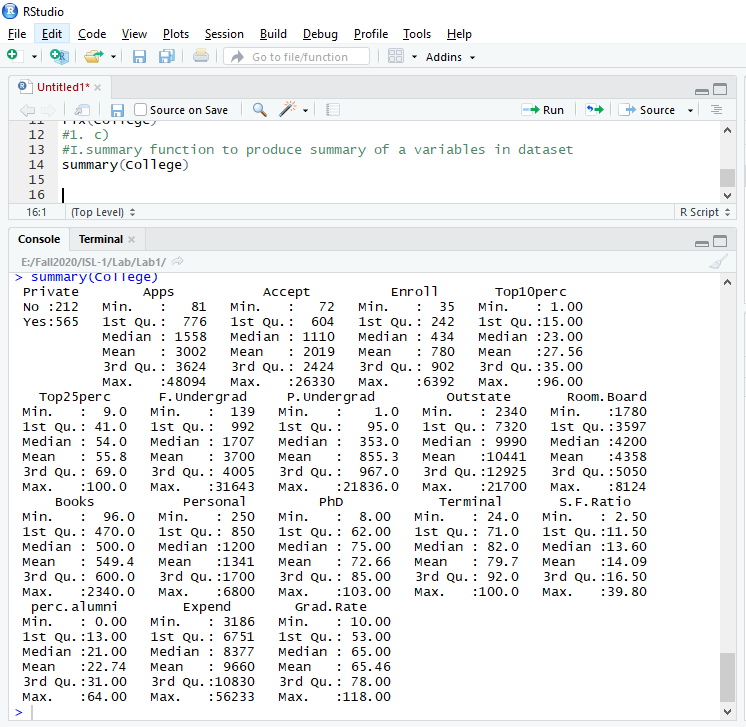
* To eliminate the first column in the data where the names are stored below is the command typed and screenshot attached.



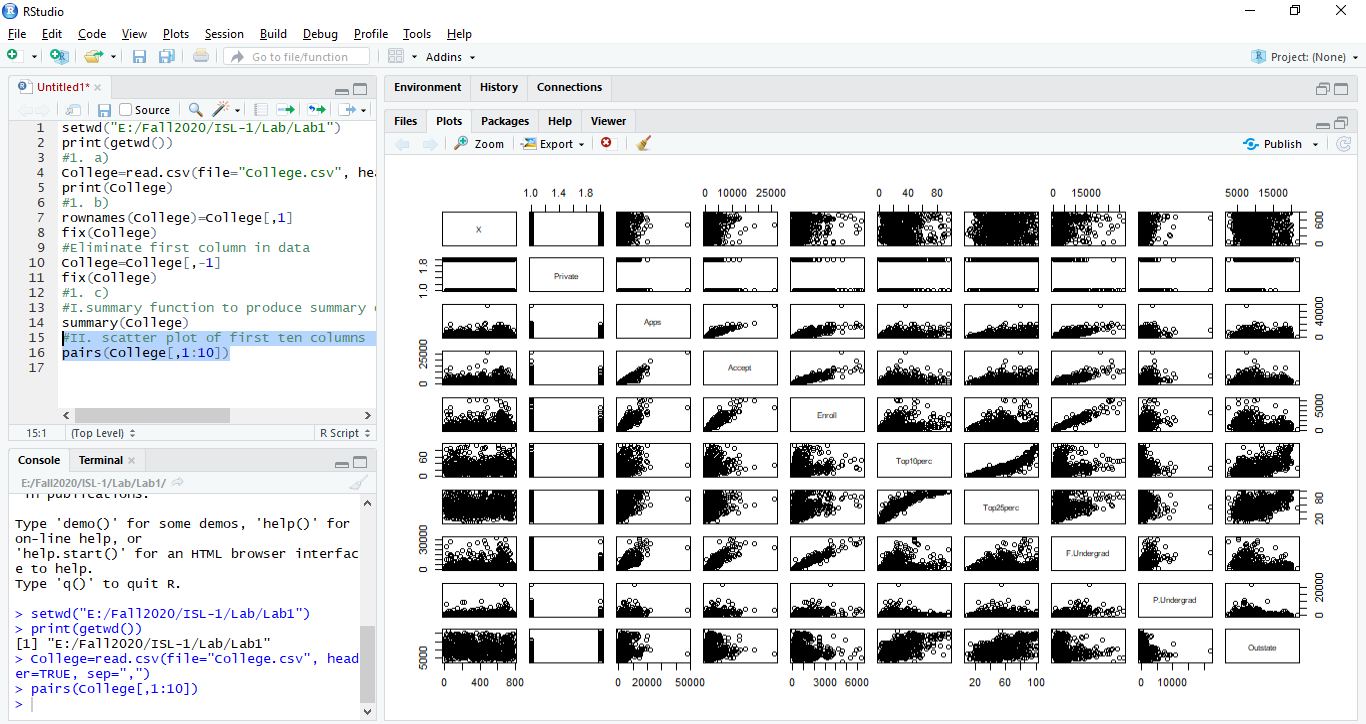


**c)**

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



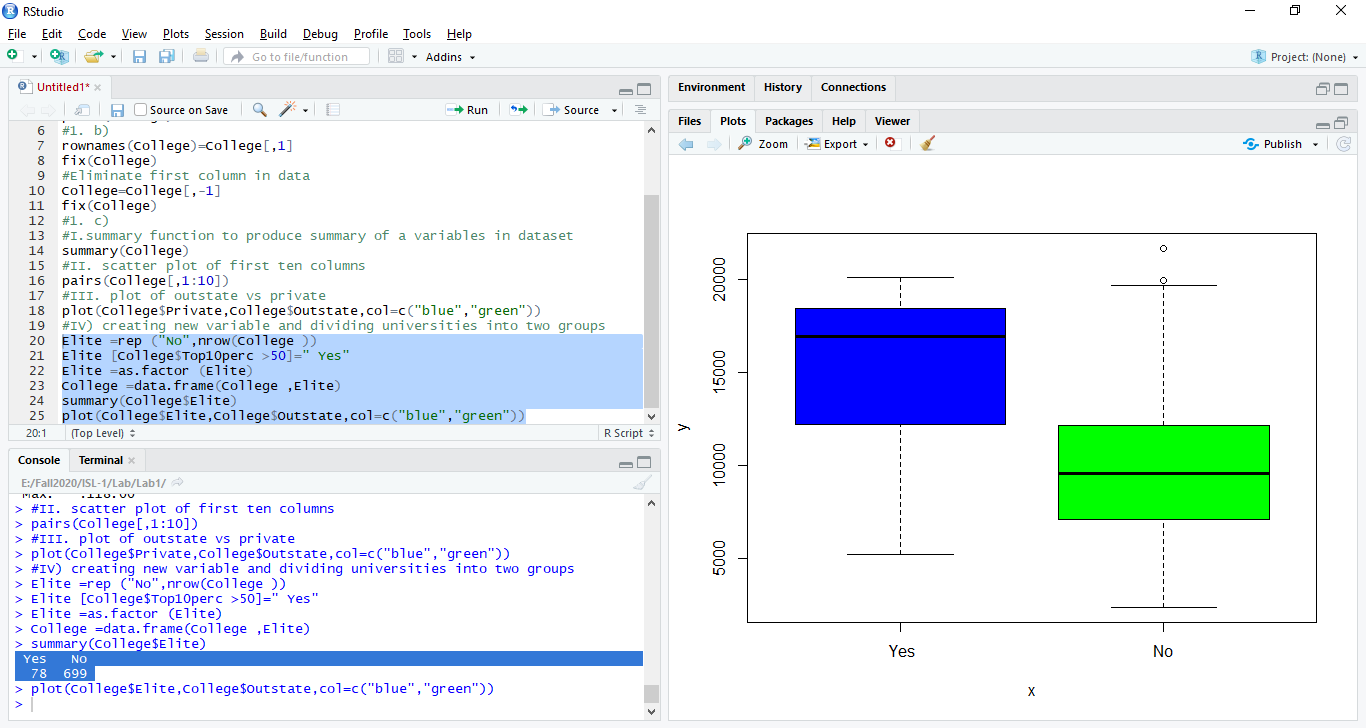
**II) Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].**



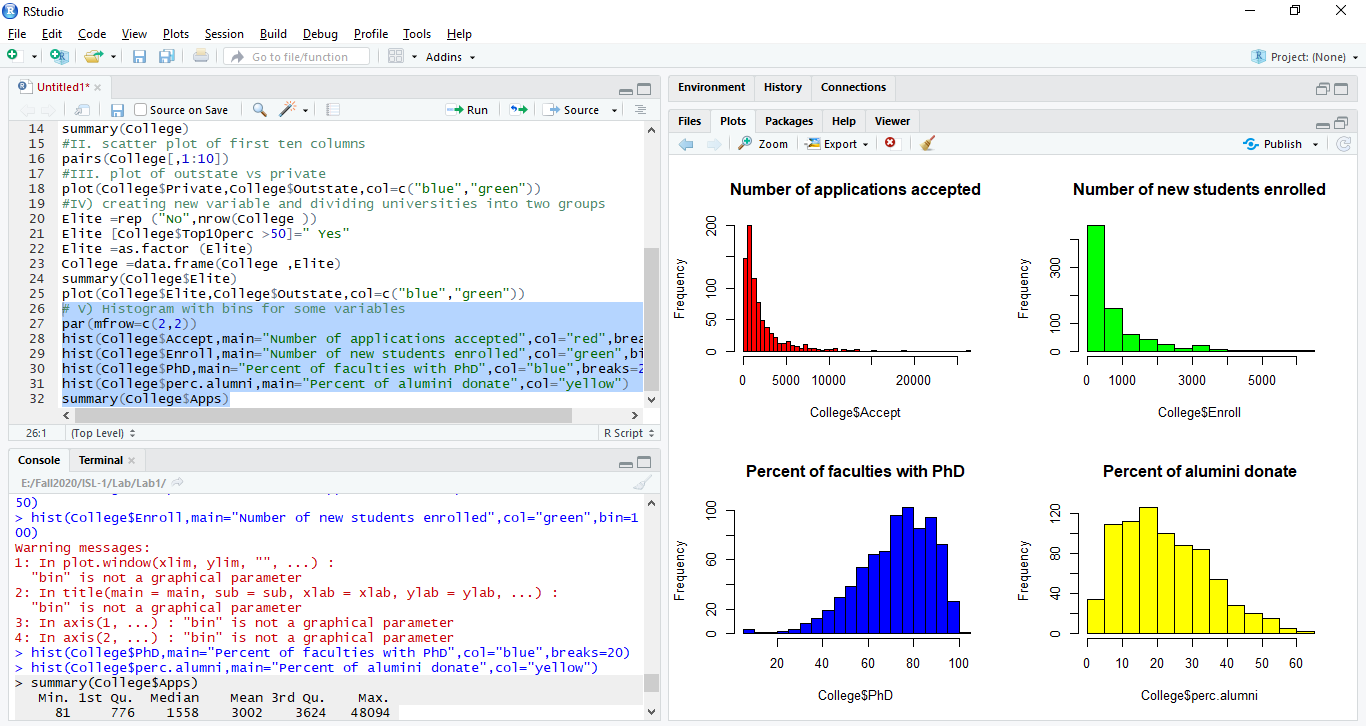
**III) Use the plot() function to produce side-by-side boxplots of Outstate versus Private.**



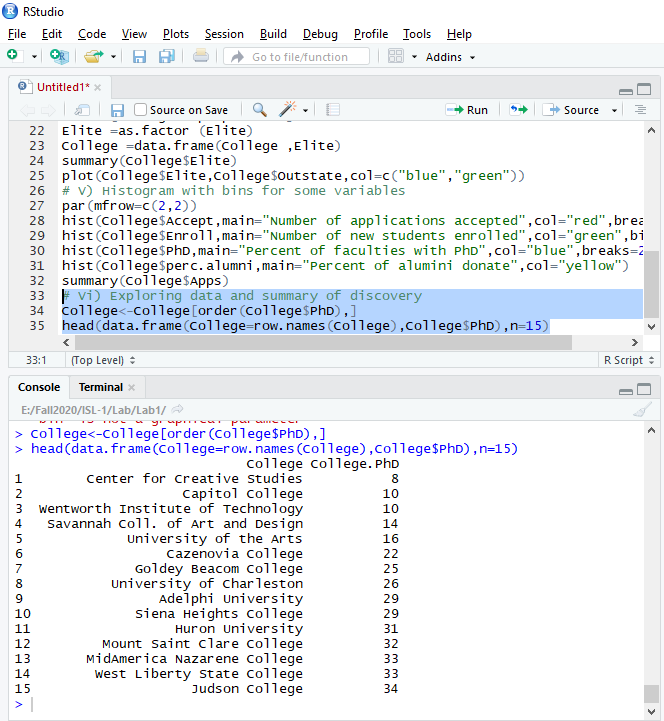
**IV) Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%. > Elite =rep ("No",nrow(college )) > Elite [college$Top10perc >50]=" Yes" > Elite =as.factor (Elite) > college =data.frame(college ,Elite) Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite. v. Use the hist() function to produce some histograms with differing**



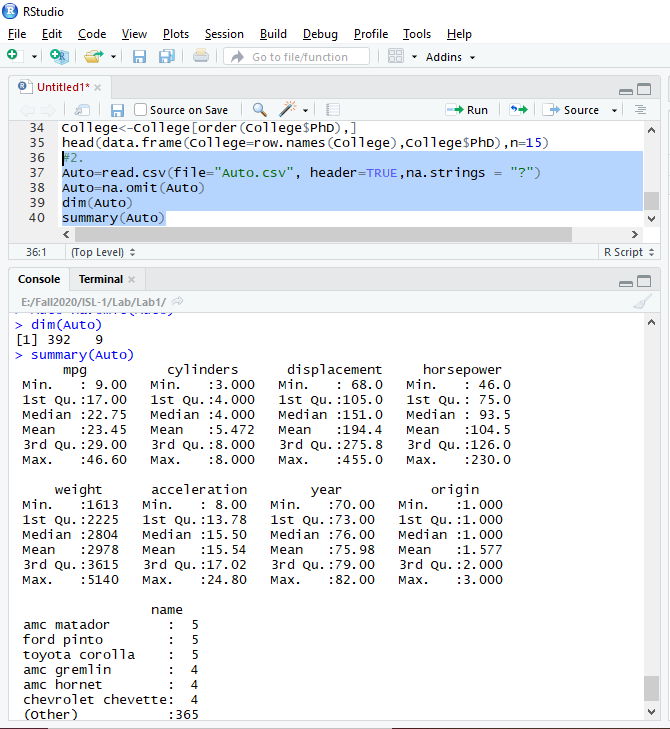
**V) Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to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



**VI) Continue exploring the data, and provide a brief summary of what you discover**



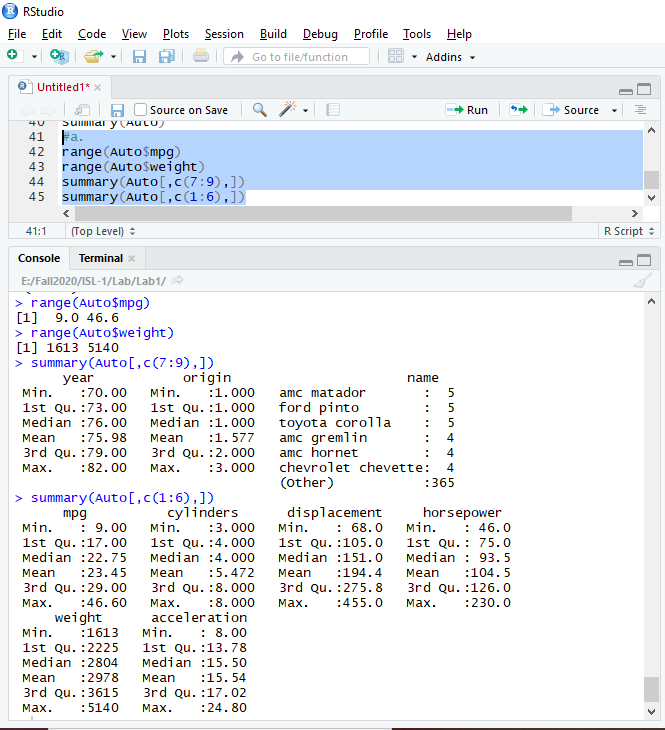
2)

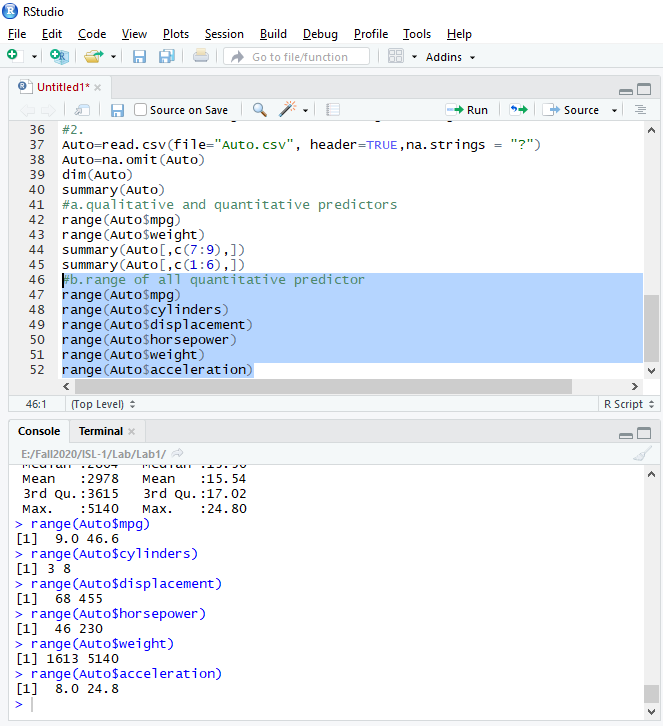


1. **Which of the predictors are quantitative, and which are qualitative?**

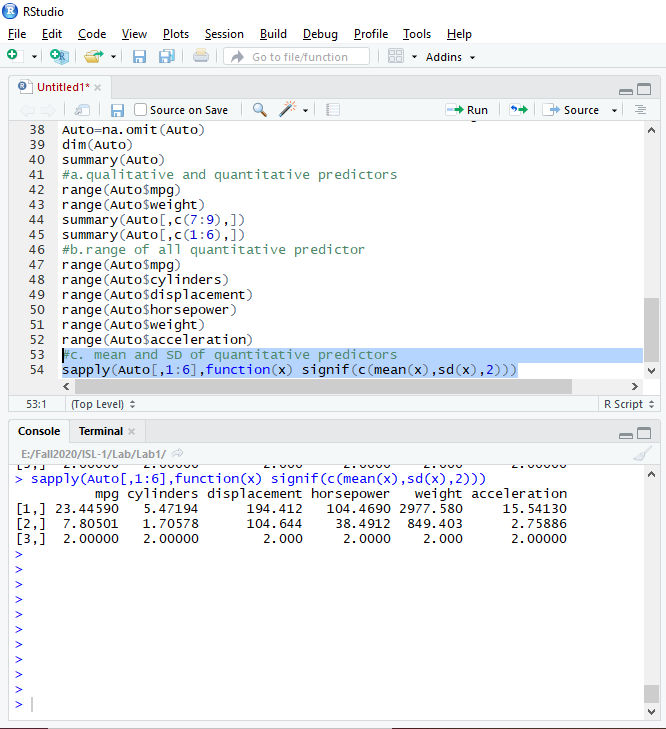
**quantitative variables:** mpg, cylinders, displacement, horsepower, weight, acceleration

**qualitative variables:** year, origin, name

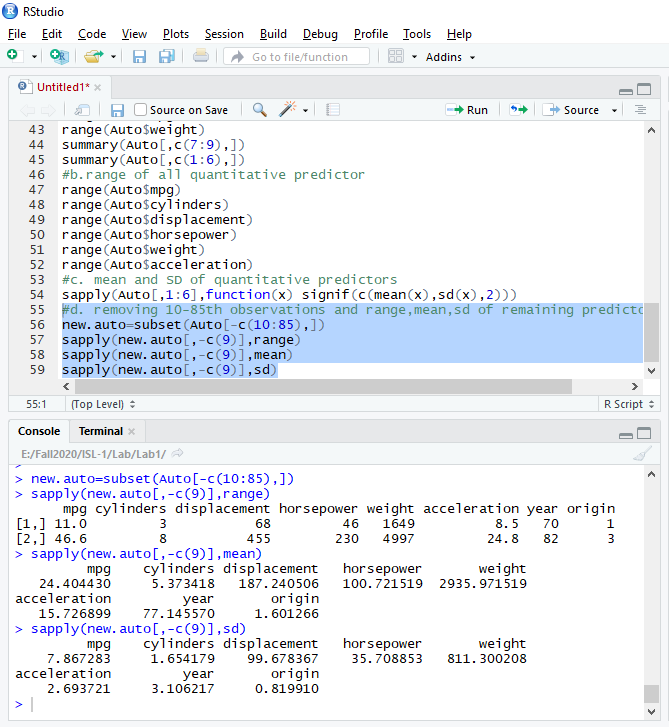


**b) What is the range of each quantitative predictor? You can answer this using the range() function.** 

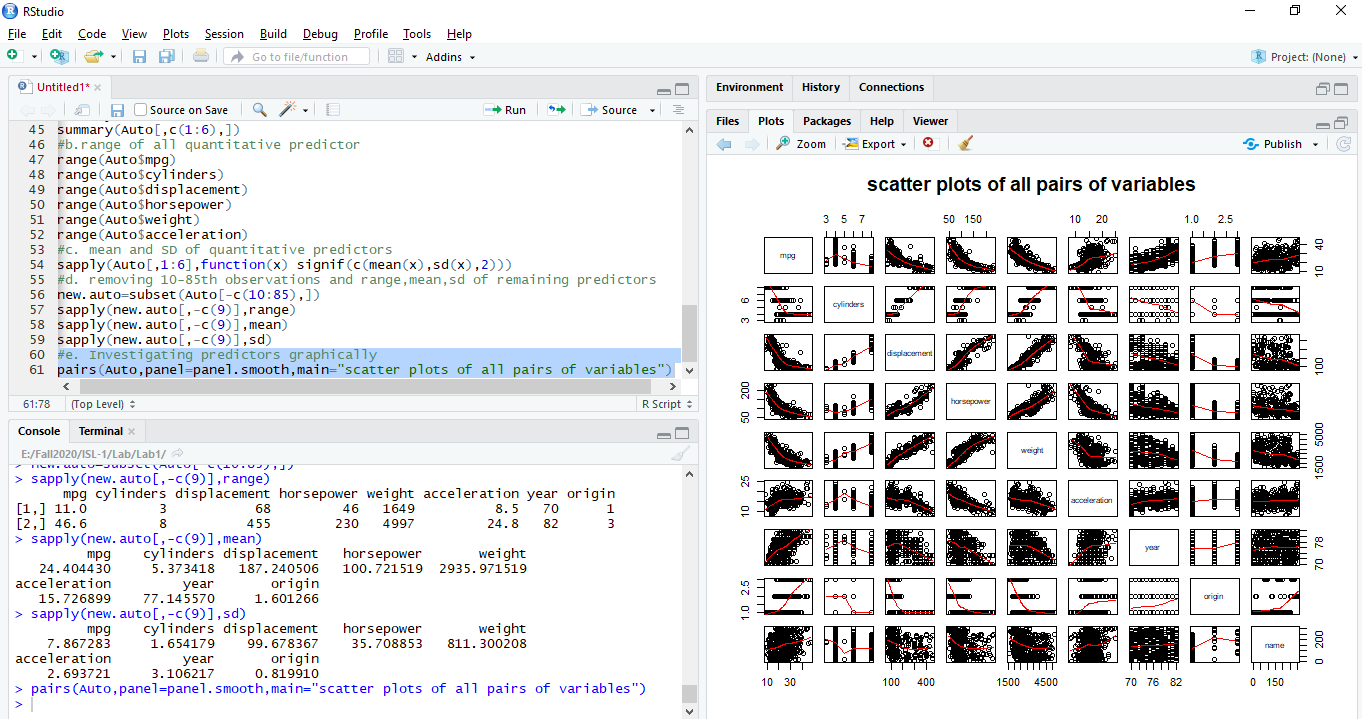
**c) What is the mean and standard deviation of each quantitative predictor?**

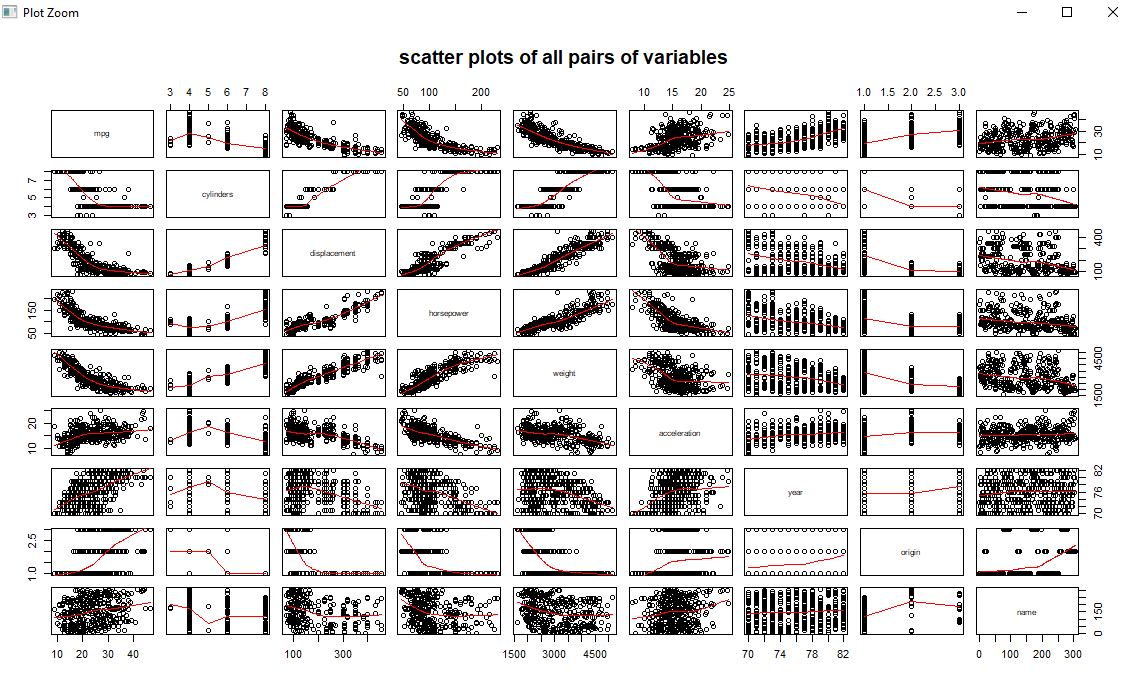


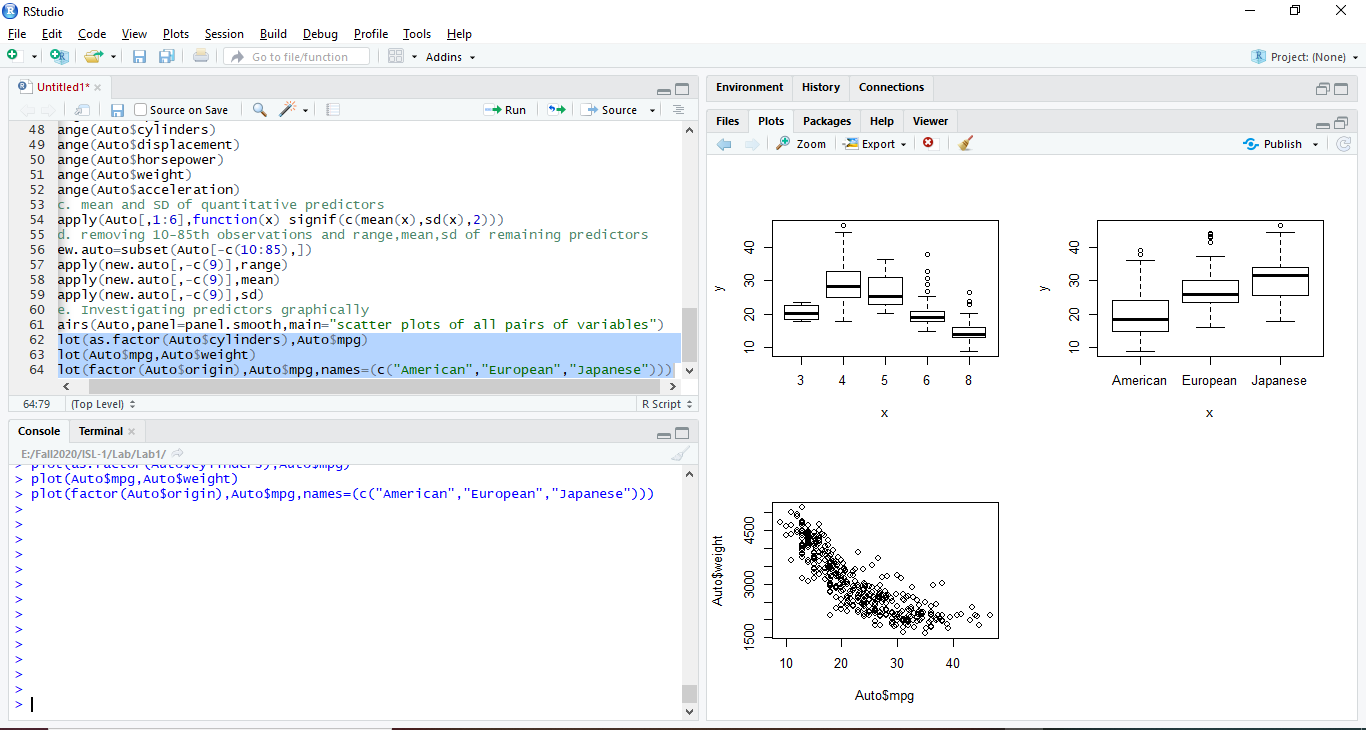
**d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?**

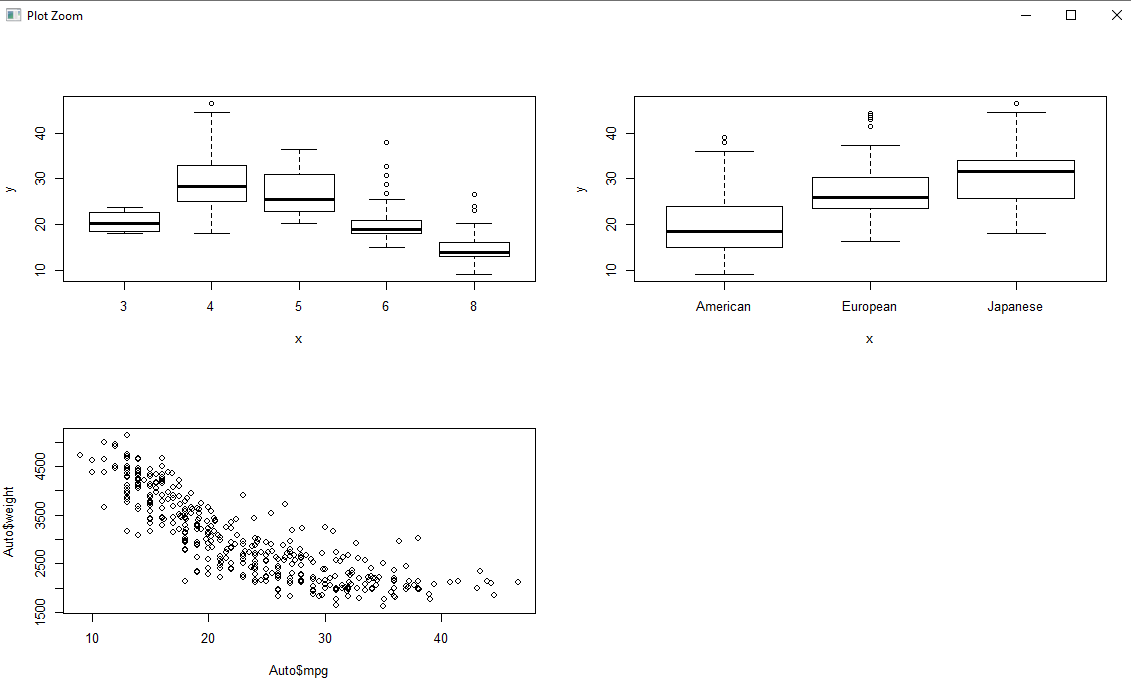


**e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.**



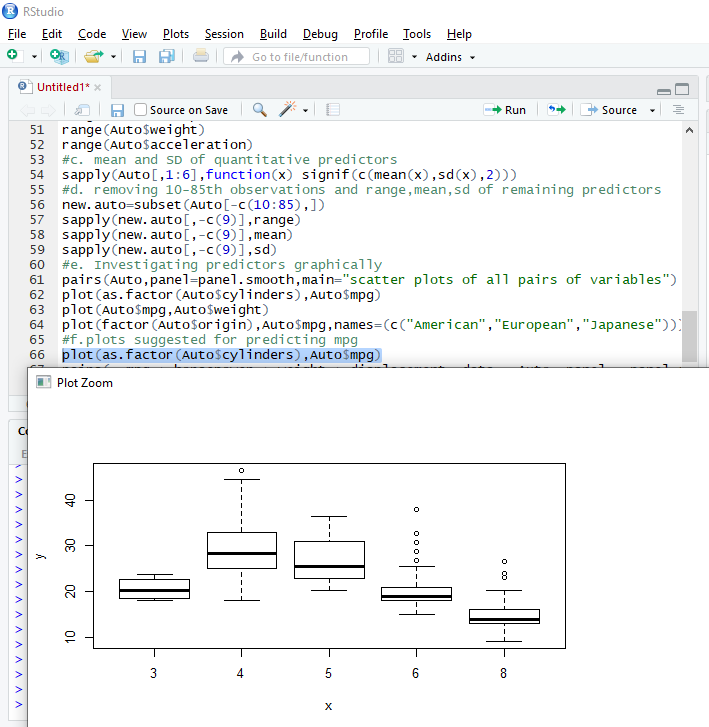


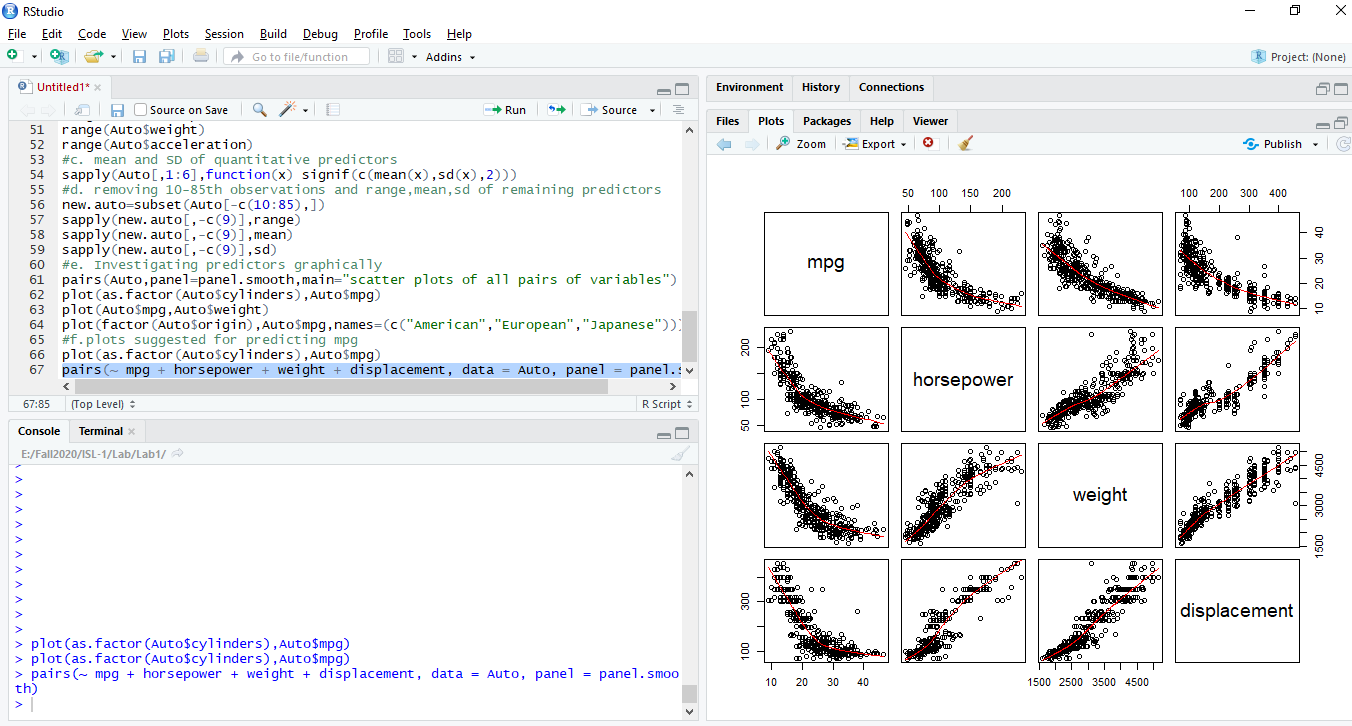




By looking at above graphs Weight, displacement and horse power seems to have an inverse effect with mpg. While displacement with horse power are directly proportional.

**f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.**

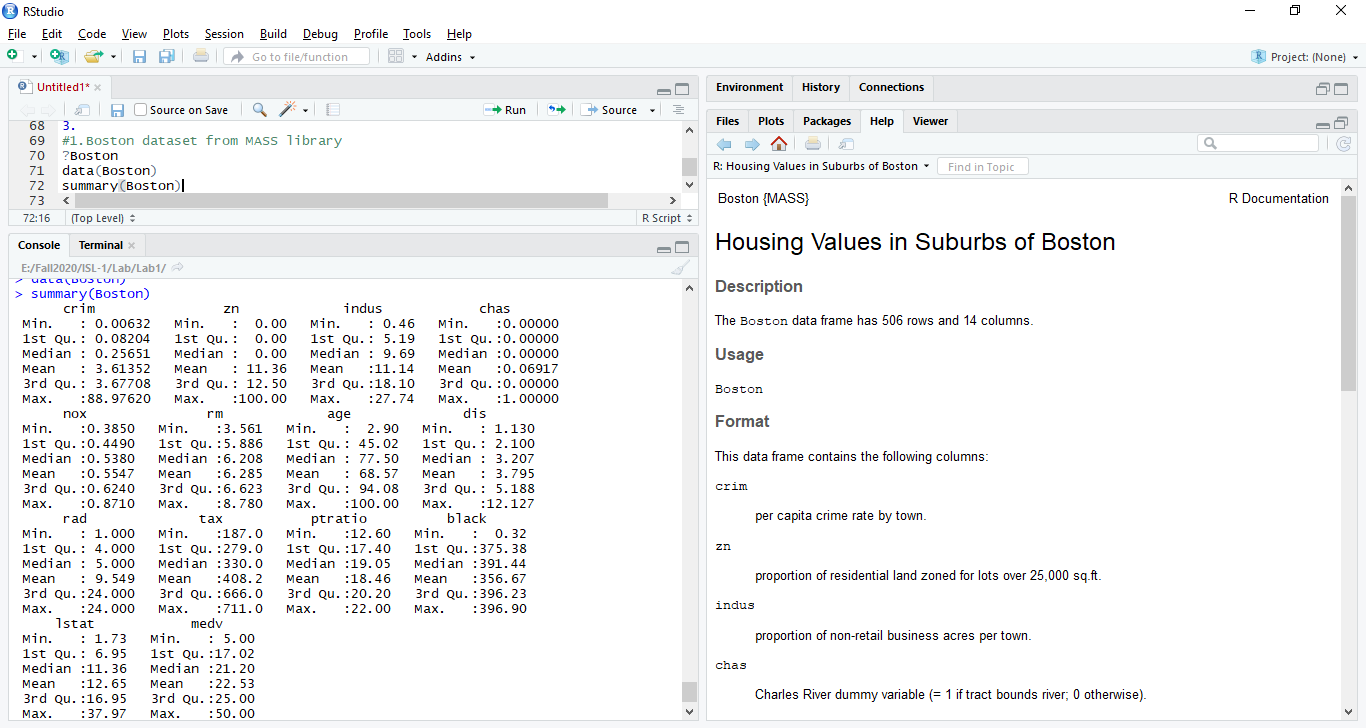


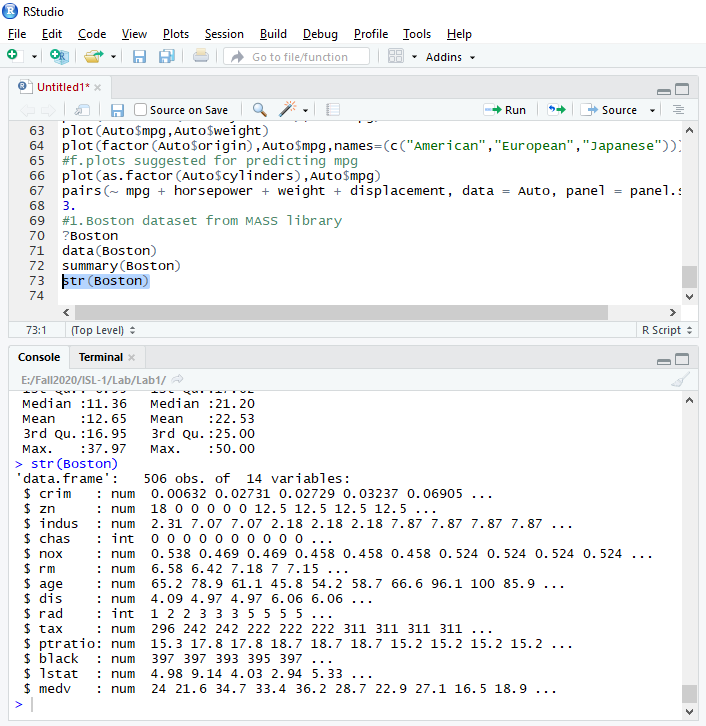


3)

**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?**

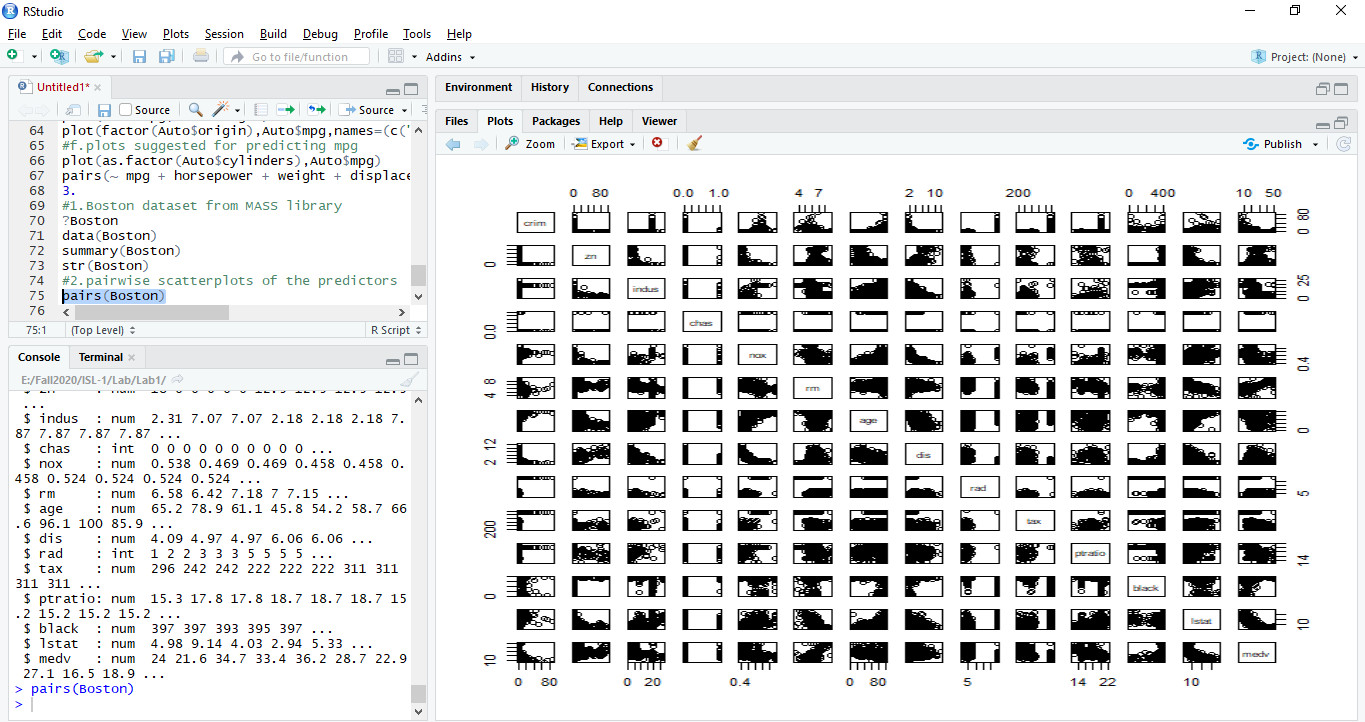
For downloading Boston dataset as it is part of MASS library in R lets install package where dataset exists.

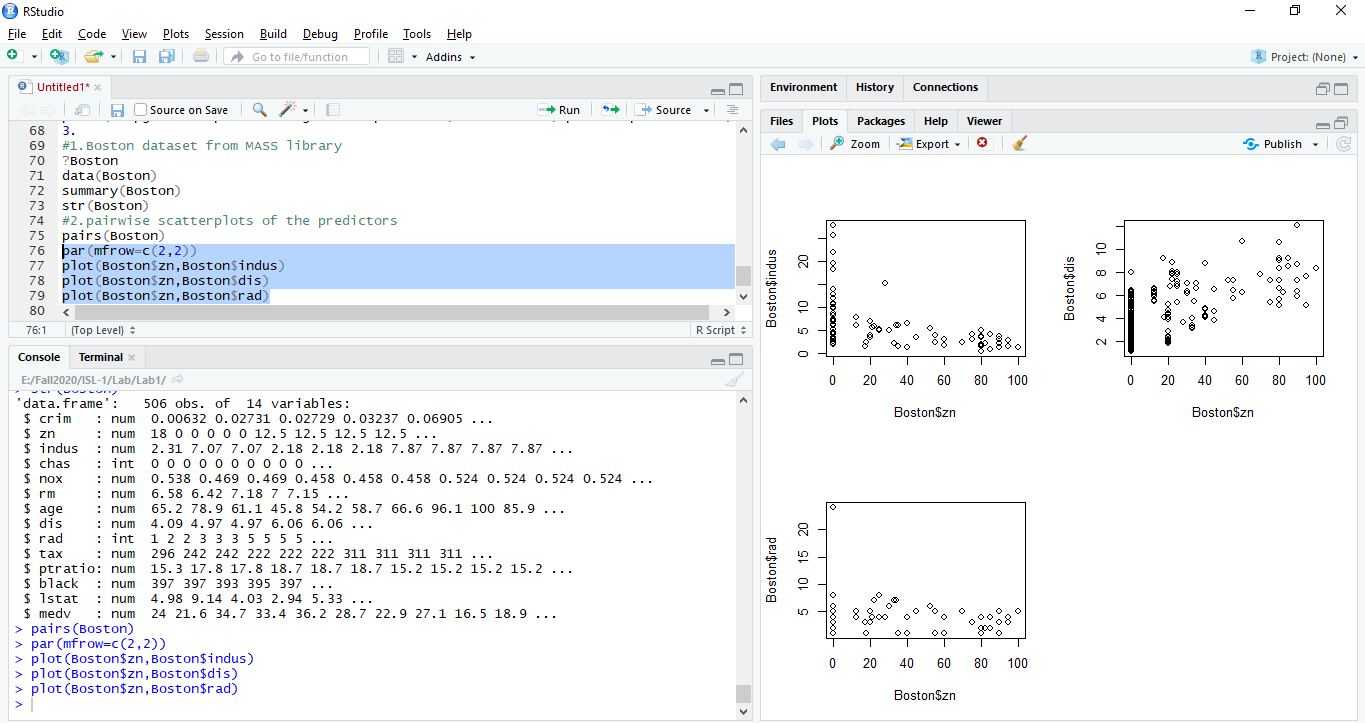




From above there 506 rows and 14 columns

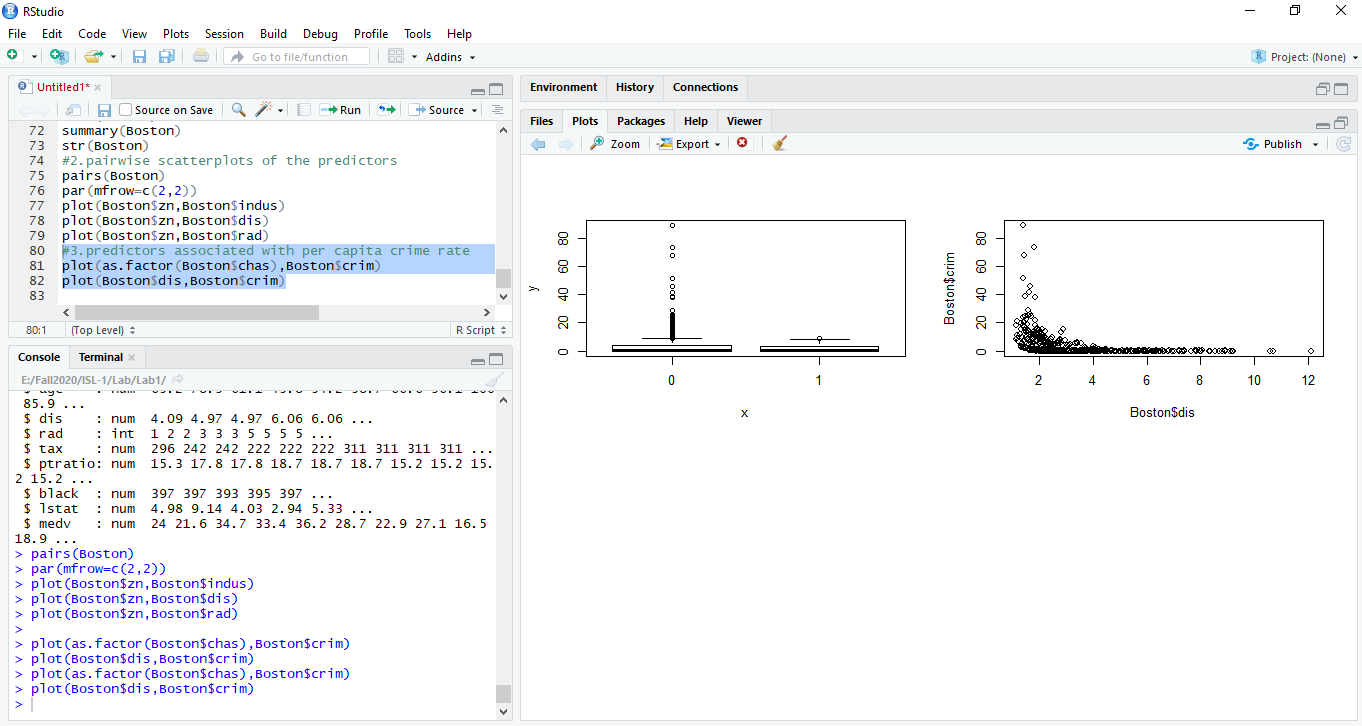
**b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.**



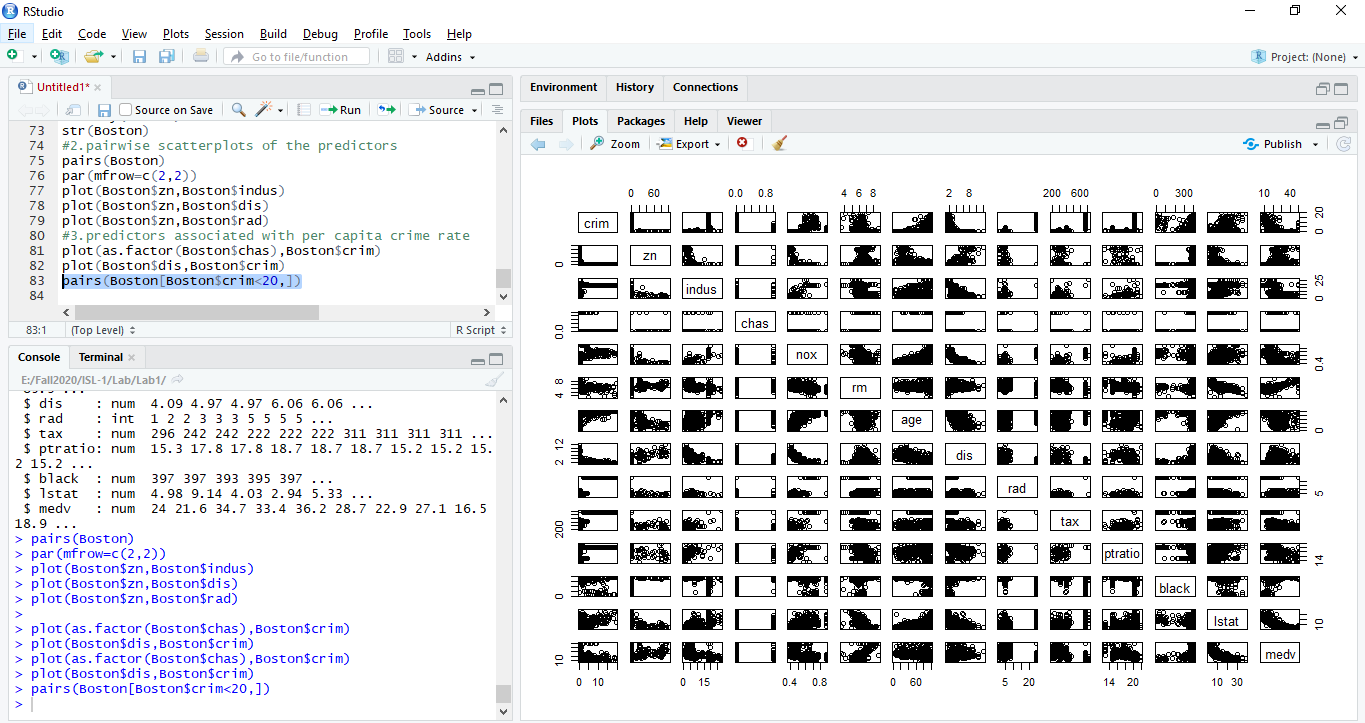


There will be number of scatter plots like above it will become difficult to read all, so maybe a heatmap will be easier to read. Data cleaning is hard.

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

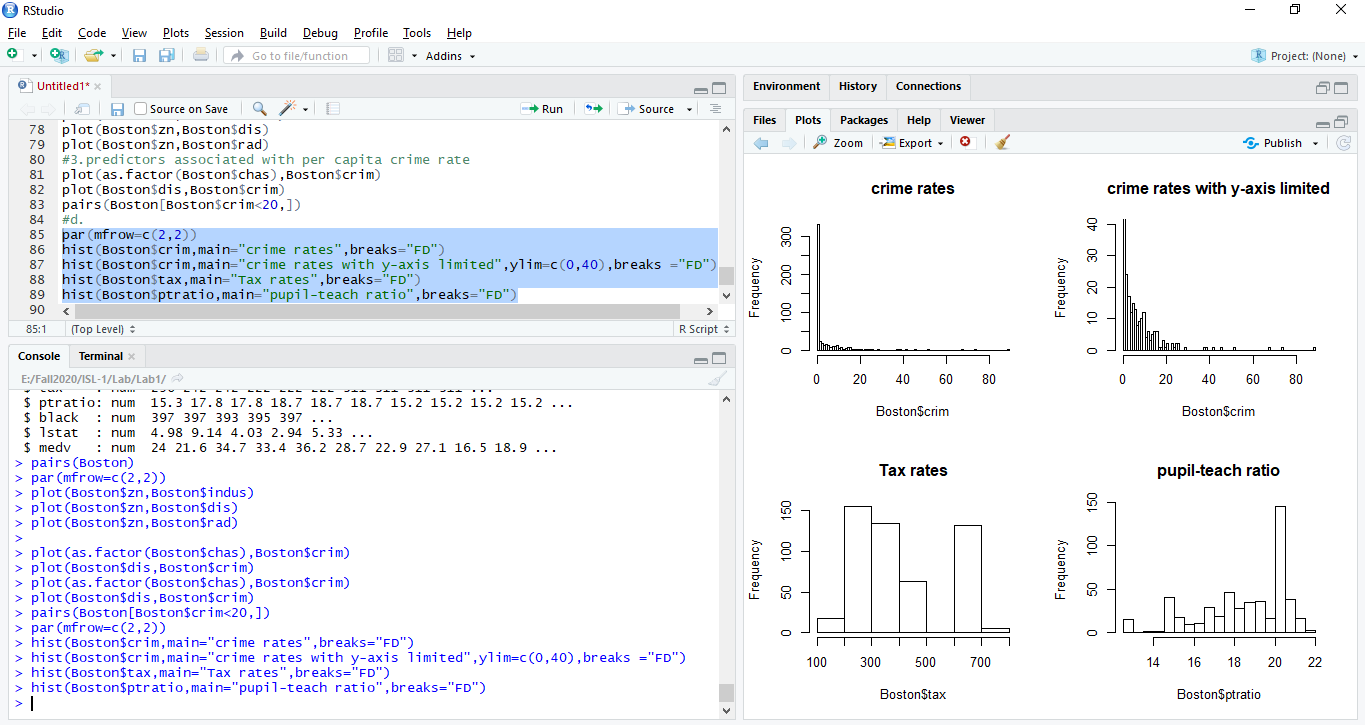


From above diagram crime is large close to five Boston employment centers

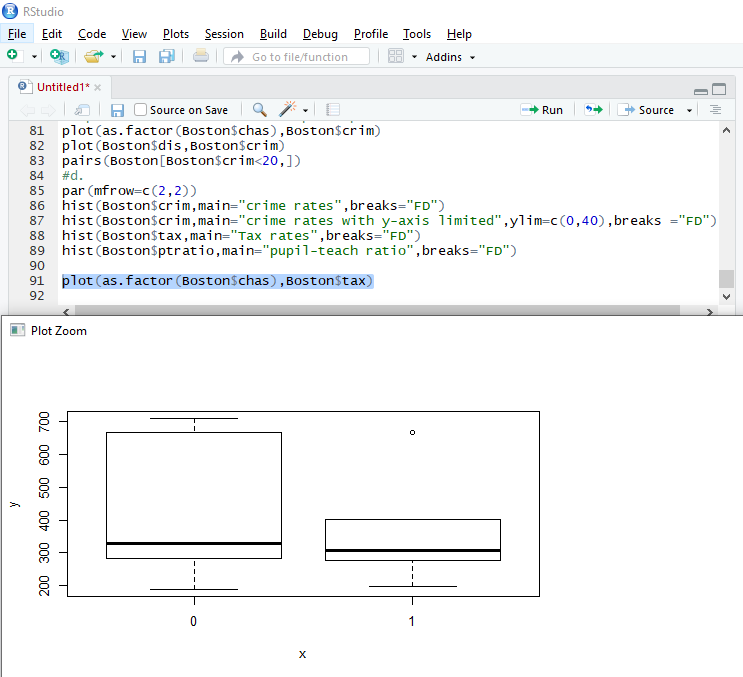


From above Lower status populated areas have the more crime rate.

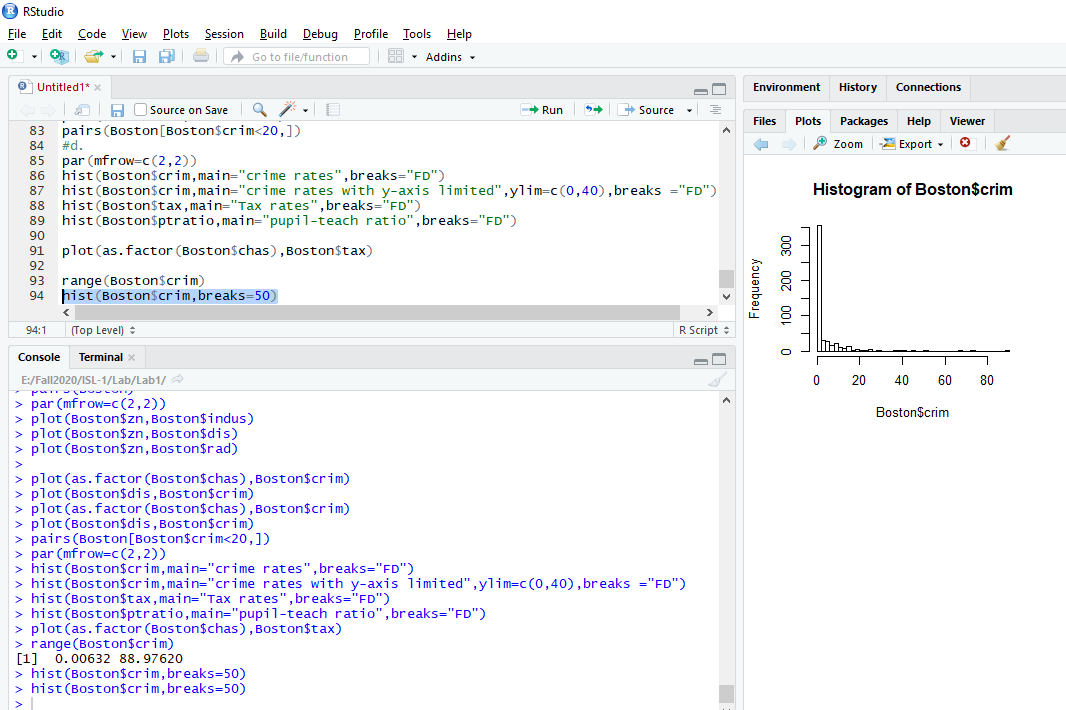
**d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.**



Most of the suburbs do not have any crime rate.



From above surprisingly tax is less near river area



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

> table(Boston$chas)

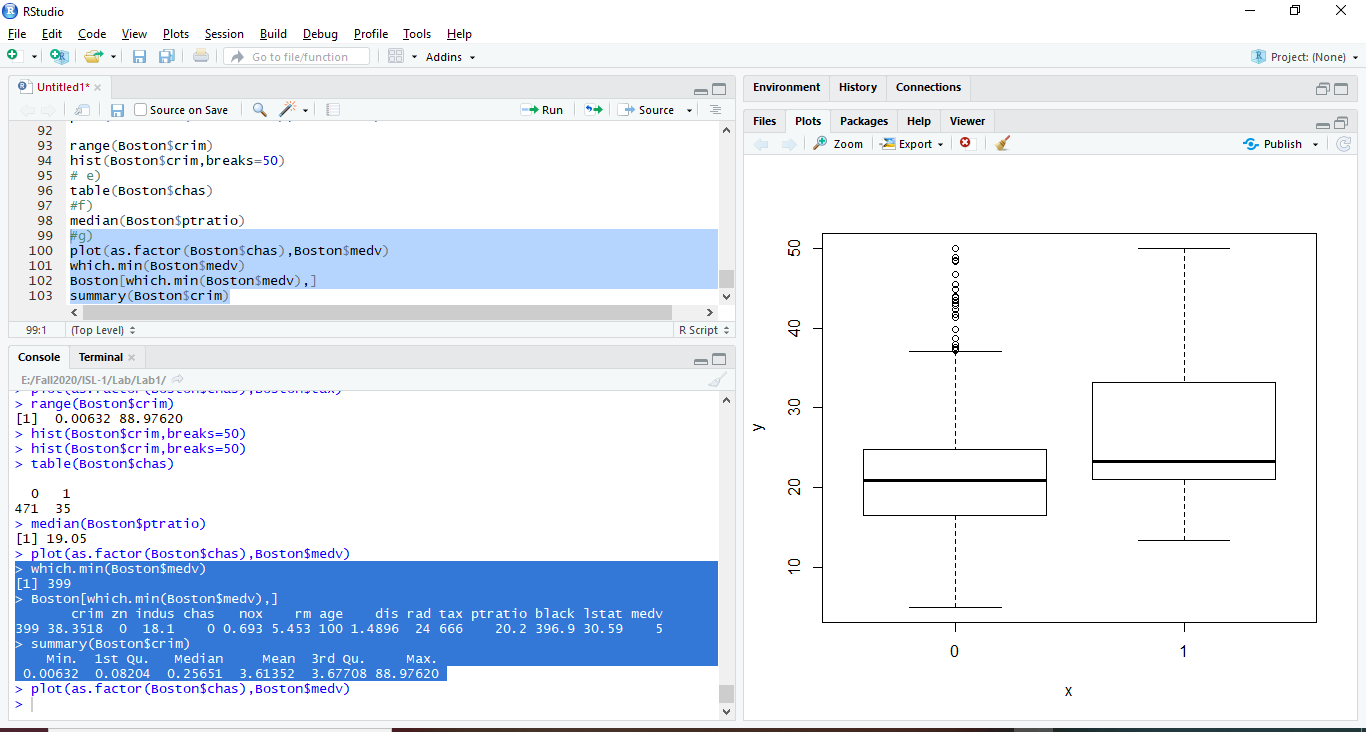
0 1

471 35

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

> median(Boston$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** 

Median 0.25 , Maximum is 88.7 and the crime in the median value of owner occupied homes is 38.3518 and we can see that the crime is larger in this area.. It is far from radial highways and Charles river area.

**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.**

> summary(Boston$rm)

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.561 5.886 6.208 6.285 6.623 8.780

Average is around 6.285 rooms

> table(Boston$rm > 7)

FALSE TRUE

442 64

From above more than 7 rooms is 64 houses

> table(Boston$rm >8)

FALSE TRUE

493 13

More than 8 rooms – 13 houses and these have the lesser crime rate.

> rooms8 = Boston[Boston$rm > 8, ]

> summary(rooms8)

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 rad

Min. :0.4161 Min. :8.034 Min. : 8.40 Min. :1.801 Min. : 2.000

1st Qu.:0.5040 1st Qu.:8.247 1st Qu.:70.40 1st Qu.:2.288 1st Qu.: 5.000

Median :0.5070 Median :8.297 Median :78.30 Median :2.894 Median : 7.000

Mean :0.5392 Mean :8.349 Mean :71.54 Mean :3.430 Mean : 7.462

3rd Qu.:0.6050 3rd Qu.:8.398 3rd Qu.:86.50 3rd Qu.:3.652 3rd Qu.: 8.000

Max. :0.7180 Max. :8.780 Max. :93.90 Max. :8.907 Max. :24.000

tax ptratio black lstat medv

Min. :224.0 Min. :13.00 Min. :354.6 Min. :2.47 Min. :21.9

1st Qu.:264.0 1st Qu.:14.70 1st Qu.:384.5 1st Qu.:3.32 1st Qu.:41.7

Median :307.0 Median :17.40 Median :386.9 Median :4.14 Median :48.3

Mean :325.1 Mean :16.36 Mean :385.2 Mean :4.31 Mean :44.2

3rd Qu.:307.0 3rd Qu.:17.40 3rd Qu.:389.7 3rd Qu.:5.12 3rd Qu.:50.0

Max. :666.0 Max. :20.20 Max. :396.9 Max. :7.44 Max. :50.0

> table(rooms8$chas)

0 1

11 2

Crime seems to be less in the houses which have 8 rooms

> summary(rooms8$black)

Min. 1st Qu. Median Mean 3rd Qu. Max.

354.6 384.5 386.9 385.2 389.7 396.9

11 of the houses with 8 rooms are not near Charles river (only 2 are near Charles river)

> summary(Boston$black)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.32 375.38 391.44 356.67 396.23 396.90

All the rooms8 houses blacks population