Name: Hyundoo Jeong

PSID: 2212332

MATH 4323, Fall 2023, Homework # 1.

Instructions: Submit the solutions as a file (type it up and save as a *.pdf* or a *Word*- file, no hand-written solutions) via UH Blackboard. Keep responses brief and to the point. For code & output: include only pieces that are of utmost relevance to the question.

Conceptual.

1. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide the sample size and the number of variables for each scenario.

   (a) We are considering launching a new product and wish to know whether it will be a success or failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

   Answer: It is classification because it will predict either success or failure which will be discrete value.

   It is interested in inference

   Sample size is 20 because it collect data on 20 similar products.

   Number of variable is 13 because of 10 other variables + price + marketing budget + competition price.

   (b) We are trying to figure out the factors potentially leading to cancer. For this, we collect data on 200 patients that either had or didn't have cancer. In particular, we record their body measurements (weight, height etc), heart rate, blood sugar level, family history of disease - measuring twelve variables. On top of that, we conduct a survey on exercise, eating and drinking habits, adding another three variables.

   Answer: It is classification because it will figure out the factors that lead to the cancer(Outcome).

It is interesting in inference.

Sample size is 200 because we collect data on 200 patients.

Number of variables is 15 because measuring 12 variables + 3 different varirables

(c) We dig into UH student database and obtain data on 1500 students. We are interested in figuring out the factors affecting the final year GPA of a student depending on the data from the entrance exams, high school and their first year at UH. We extract students' SAT scores, high school GPA, first year GPA at UH, major, age and other five variables.

Answer: It is regression because it will figure out the factors that are affecting your final year GPA.

It is interested in inference.

Sample size is 1500 because we obtain data on 1500 students.

Number of variable is 8 because 3 GPA variables + other 5 variables.

(d) We would like to predict the outcome of a college football game (not the exact score, but simply who wins, team *A* or team *B*?) depending on various factors. Assuming that it is mid-season, we obtain the data on all 300 games that have already been played and study such variables as: yards gained/allowed, touch- downs scored/allowed, turnovers committed/forced, whether the game is at home or away, whether it rains or not (all-in-all eight variables).

Answer: It is classification because it will predict outcome of football game.

It is interessted in prediction.

Sample size is 300 because we obtain the data on 300 games.

Number of variables is 8.

2. (a) What type of statistical learning do all data examples in Problem 1 correspond to -

supervised or unsupervised

Answer: It is supervised learning

(b) We know that general model formula is

$$Y = f(X) + \epsilon, \tag{1}$$

and we try to estimate true f with $\hat{f}$. For each of examples (a), (b), (c) from Problem 1, proceed to answer the following:

i. Can our estimate $\hat{f}$ be treated as a black box?

Answer: No, In inference, the estimate can't be treated as a black box.

ii. Why/Why not?

Answer: this problem is interested in inference, so f hat needs to know its exact form.

(c) When estimating Y from equation (1) with $\hat{Y} = \hat{f}(X)$, what two errors can we commit? Which one of them can be improved via a better statistical learning technique? Which one of them can't be improved & why?

Answer: reducible and irreducible errors can be committed.
Reducible error is better for improving statistical learning.
Irreducible error can't be improved because it will always provide upper bound on of accuracy in prediction Y.

3. Provide three data examples of unsupervised learning task (on your own, and you can't use the ones already mentioned in class, including the intro lecture), with all of them being from different application areas. Formulate what are the subjects (doesn't have to be people) of interest you are trying to group/cluster, and according to what potential predictors/characteristics. Areas may include, but not limited to, medicine, finance, economics, sociology, education, marketing, journalism, sports, oil industry, meteorology, etc.

Answer: 1. Speech Analysis: It could be applied for tasks like speaker diarizing, where the algorithm segments an audio recording into different speaker segments without prior information about who the speakers are.

2.Fraud Detection: By analyzing traction data or user's behavior pattern, clustering algorithms can identify unusual or anonymous behavior which may indicate fraud activity.

3.Recommendation system: Recommendation systems, as seen in platforms like Netflix, Amazon, and Spotify, often use unsupervised learning to suggest products, movies, or music to users.

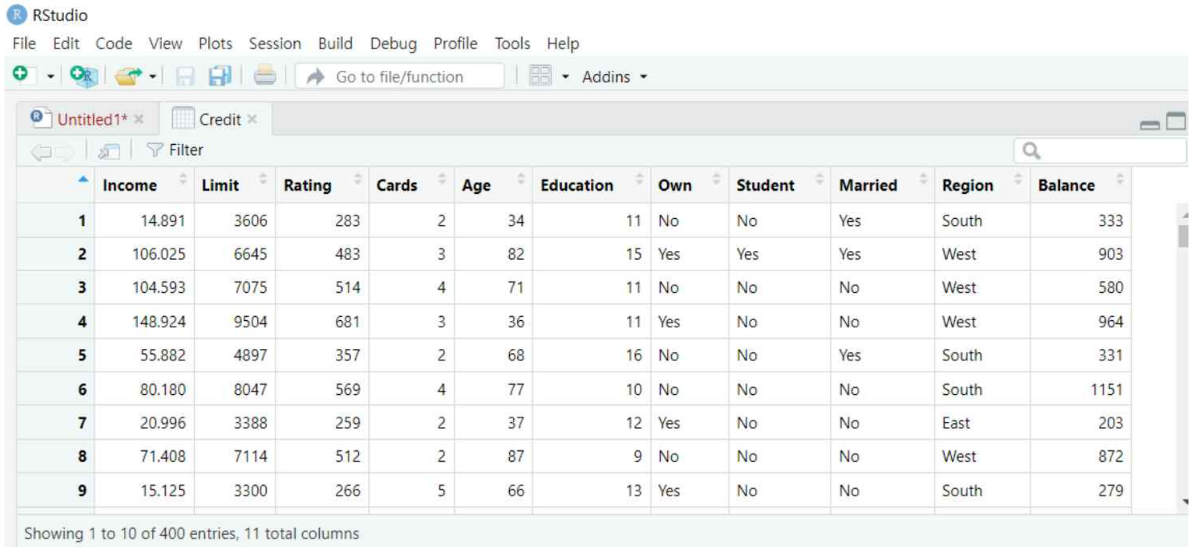
Applied

4. This exercise relates to the *Credit* data set, which can be found in the file *Credit.csv*. It contains information about credit card debt for 10,000 credit card holders. The variables are

- **Income**: in thousands of dollars
- **Limit**: Credit limit
- **Rating**: Credit rating
- **Cards**: Number of credit cards
- **Age**: Age of each card holder
- **Education**: Year of education
- **Own**: House ownership
- **Student**: Student status
- **Married**: Marital status
- **Region**: East, West or South
- **Balance**: Average credit card debt for each card holder

Before reading the data into *R*, it can be viewed in *Excel* or a text editor.

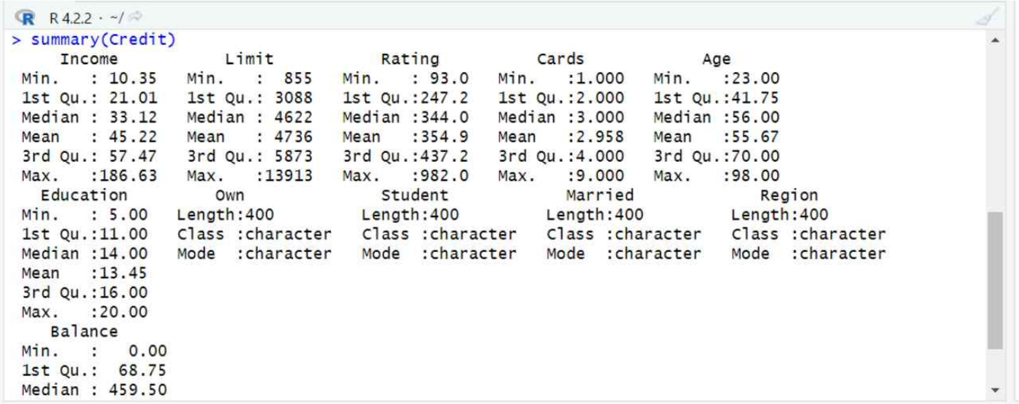
(a) Use *RStudio*'s drop-down menu (Environment → Import Dataset → From Text (base) ...) to read the data into *R*. Make sure the *Heading* is set to Yes. Call the loaded data *Credit*.


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled1*
Credit
Filter

	Income	Limit	Rating	Cards	Age	Education	Own	Student	Married	Region	Balance
1	14.891	3606	283	2	34	11	No	No	Yes	South	333
2	106.025	6645	483	3	82	15	Yes	Yes	Yes	West	903
3	104.593	7075	514	4	71	11	No	No	No	West	580
4	148.924	9504	681	3	36	11	Yes	No	No	West	964
5	55.882	4897	357	2	68	16	No	No	Yes	South	331
6	80.180	8047	569	4	77	10	No	No	No	South	1151
7	20.996	3388	259	2	37	12	Yes	No	No	East	203
8	71.408	7114	512	2	87	9	No	No	No	West	872
9	15.125	3300	266	5	66	13	Yes	No	No	South	279

Showing 1 to 10 of 400 entries, 11 total columns

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



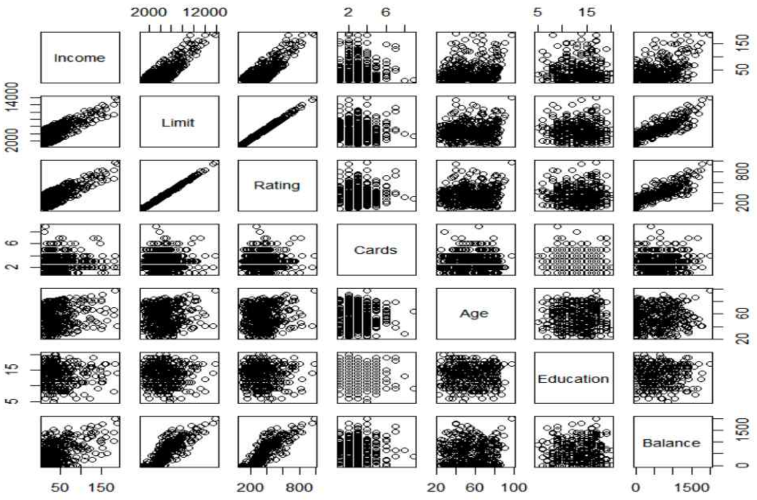
```
> summary(Credit)
     Income           Limit           Rating          Cards            Age       
 Min.   : 10.35   Min.   :  855   Min.   : 93.0   Min.   :1.000   Min.   :23.00  
 1st Qu.: 21.01   1st Qu.: 3088   1st Qu.:247.2   1st Qu.:2.000   1st Qu.:41.75  
 Median : 33.12   Median : 4622   Median :344.0   Median :3.000   Median :56.00  
 Mean   : 45.22   Mean   : 4736   Mean   :354.9   Mean   :2.958   Mean   :55.67  
 3rd Qu.: 57.47   3rd Qu.: 5873   3rd Qu.:437.2   3rd Qu.:4.000   3rd Qu.:70.00  
 Max.   :186.63   Max.   :13913   Max.   :982.0   Max.   :9.000   Max.   :98.00  
   Education          Own              Student            Married         
 Min.   : 5.00   Length:400         Length:400         Length:400        
 1st Qu.:11.00   Class :character   Class :character   Class :character  
 Median :14.00   Mode  :character   Mode  :character   Mode  :character  
 Mean   :13.45                                                           
 3rd Qu.:16.00                                                           
 Max.   :20.00                                                           
    Balance       
 Min.   :   0.00  
 1st Qu.:  68.75  
 Median : 459.50  
```

```
    Region         
 Length:400        
 Class :character  
 Mode  :character  
```

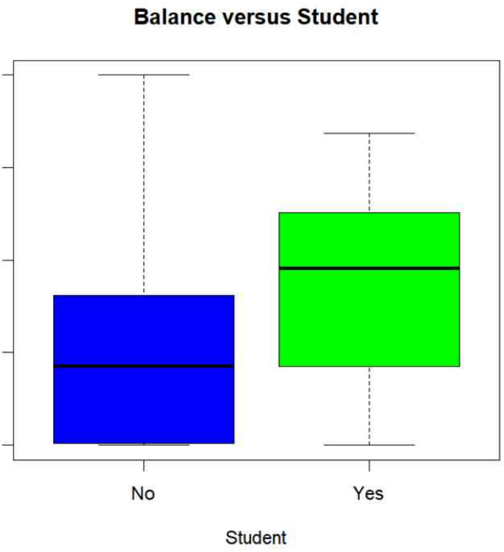
ii. Which columns contain numerical values? Which columns contain categori- cal values?

Answer: Numerical values are income, Limit, Rating, card, age, education, balance. Categorical values are own, student, married, region.

iii. Use the *pairs*() function to produce a scatterplot matrix of the quantitative variables in the dataset. Note that the *pairs*() function requires the input as numeric values, so you have to think about how to select the numerical columns from the dataset.



iv. Use the *plot*() function to produce side-by-side boxplots of *Balance* versus *Student*.

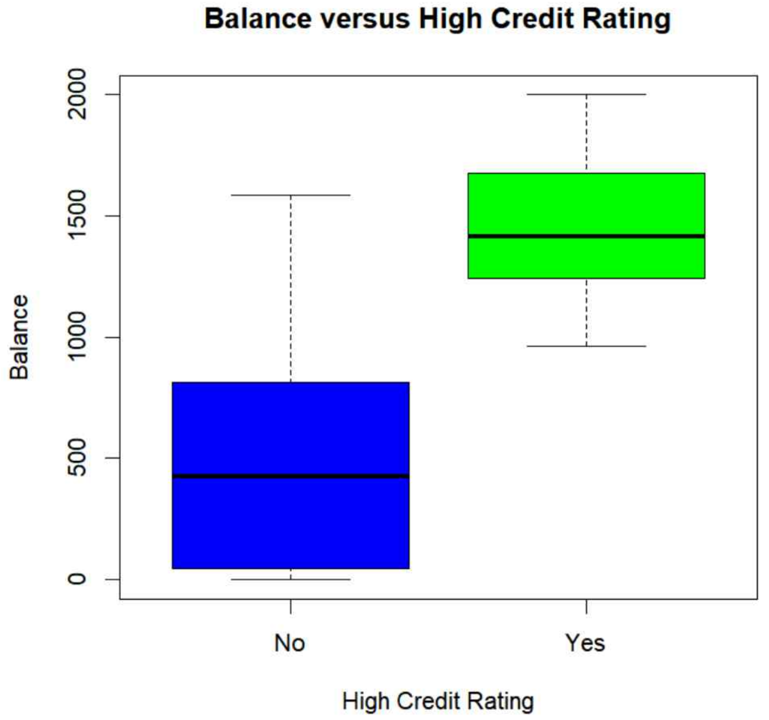


v. Create a new qualitative variable, called *high*, by binning the *Rating* variable. We are going to divide the card holders into two groups based on whether their credit ratings exceed 680.

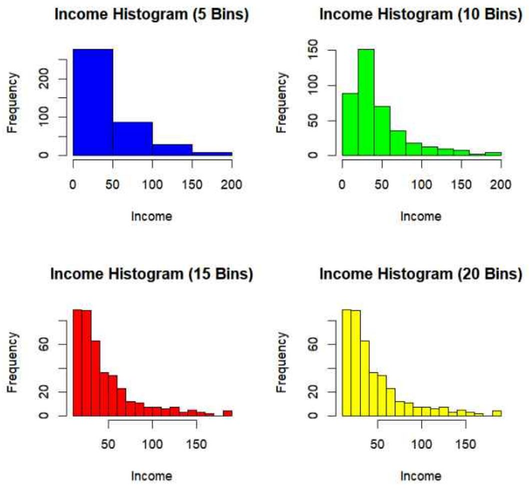
```
> Credit$high=ifelse(Credit$Rating>680,"Yes","No")
```

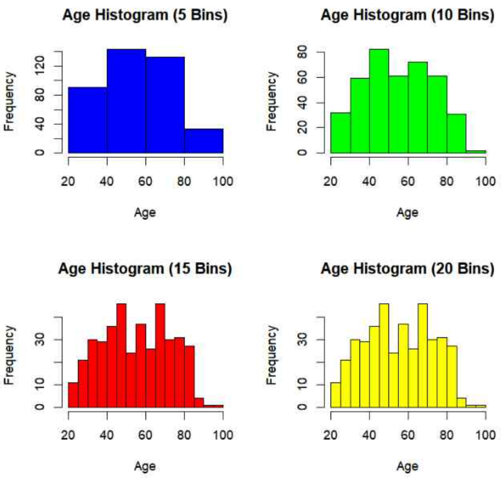
Use the *table*() function to see how many card holders in this dataset have high credit ratings. Now use the *plot*() function to produce side-by-side boxplots of *Balance* versus *high*. What would you comment on the findings based on the boxplot?

Answer: Based on the box diagram, you can observe the relationship between "balance" and "high" (credit rating). For example, you can search for differences between two groups in terms of median and value expansion. If there is a big difference in the boxplot, it may suggest that credit ratings affect credit card balances.



vi. Use the *hist*() function to produce some histograms with differing numbers of bins (e.g. 5, 10, 15, 20) for a few of the quantitative variables (e.g. *Income* and *Age*). 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.





5. This exercise involves the *Boston* 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 ”lstat”, ”ptratio”,”chas”, and ”medv” represent?

Answer: 506 rows and 14 columns.

“lstat” represents lower status of population.

“ptratio” represents pupil-teacher ratio by town.

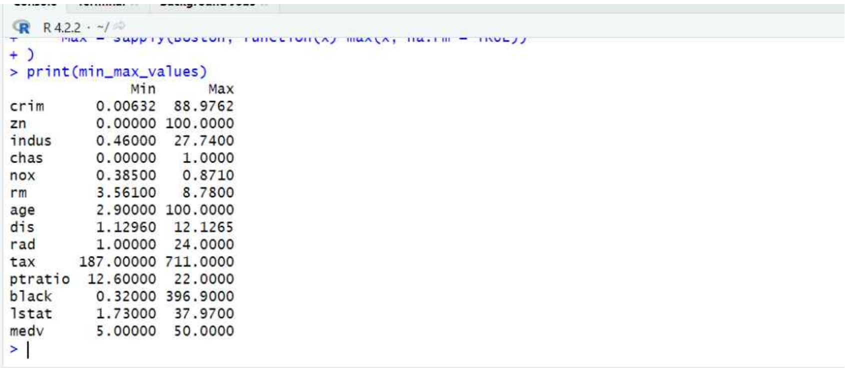
“chas” represents Charles River dummy variable.

“medv” represents median value of owner-occupied homes in 1000$.

(b) Of what type are most of the predictors - quantitative or qualitative?

Answer: All predictors are quantitative because it expressed numerical value, measurement, count.

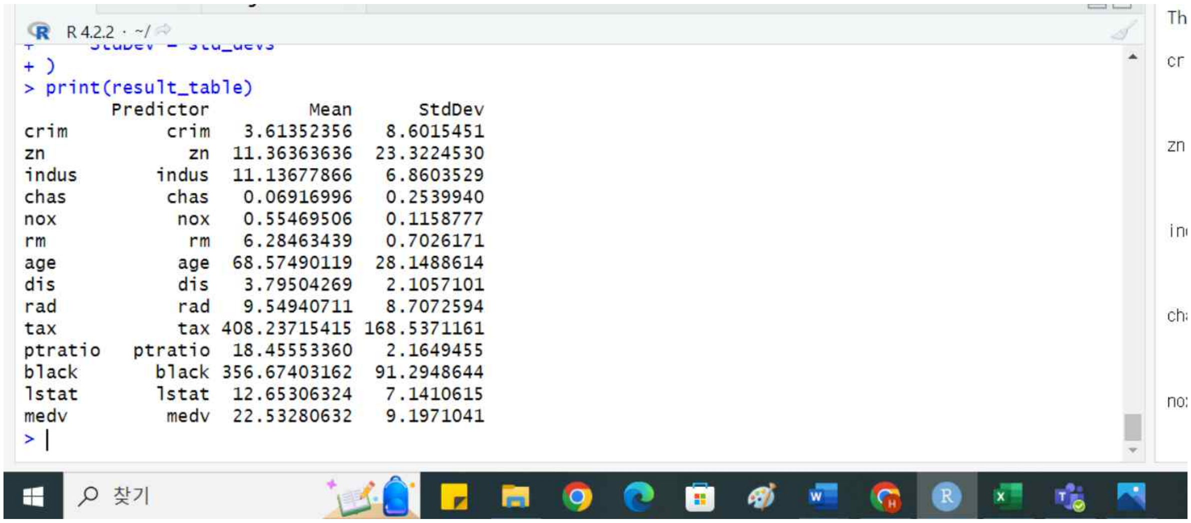
(c) What is the range of each predictor? You can answer this either by applying the *range*() function to each predictor, or by using *summary*() function on the whole data set and extracting the range from there. Please provide a table with ranges for all predictors:



```
+    max = sapply(Boston, function(x) max(x, na.rm = TRUE))
+ )
> print(min_max_values)
              Min      Max
crim      0.00632  88.9762
zn        0.00000 100.0000
indus     0.46000  27.7400
chas      0.00000   1.0000
nox       0.38500   0.8710
rm        3.56100   8.7800
age       2.90000 100.0000
dis       1.12960  12.1265
rad       1.00000  24.0000
tax     187.00000 711.0000
ptratio  12.60000  22.0000
black     0.32000 396.9000
lstat     1.73000  37.9700
medv      5.00000  50.0000
> |
```

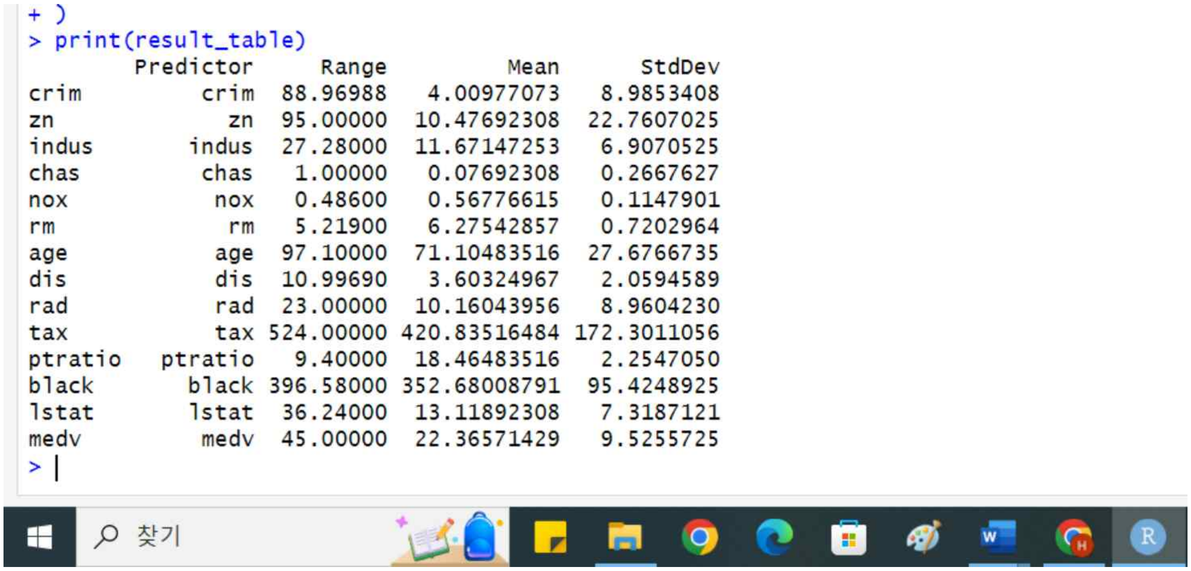
(d) What is the mean and standard deviation of each quantitative predictor?

Provide the answer in the form of a table:



```
+     StdDev = std_devs
+ )
> print(result_table)
        Predictor         Mean      StdDev
crim         crim   3.61352356   8.6015451
zn             zn  11.36363636  23.3224530
indus       indus  11.13677866   6.8603529
chas         chas   0.06916996   0.2539940
nox           nox   0.55469506   0.1158777
rm             rm   6.28463439   0.7026171
age           age  68.57490119  28.1488614
dis           dis   3.79504269   2.1057101
rad           rad   9.54940711   8.7072594
tax           tax 408.23715415 168.5371161
ptratio   ptratio  18.45553360   2.1649455
black       black 356.67403162  91.2948644
lstat       lstat  12.65306324   7.1410615
medv         medv  22.53280632   9.1971041
> |
```

(e) Now remove the 50th through 100th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains? Provide the answer in the form of a table:



```
+ )
> print(result_table)
        Predictor     Range         Mean      StdDev
crim         crim  88.96988   4.00977073   8.9853408
zn             zn  95.00000  10.47692308  22.7607025
indus       indus  27.28000  11.67147253   6.9070525
chas         chas   1.00000   0.07692308   0.2667627
nox           nox   0.48600   0.56776615   0.1147901
rm             rm   5.21900   6.27542857   0.7202964
age           age  97.10000  71.10483516  27.6766735
dis           dis  10.99690   3.60324967   2.0594589
rad           rad  23.00000  10.16043956   8.9604230
tax           tax 524.00000 420.83516484 172.3011056
ptratio   ptratio   9.40000  18.46483516   2.2547050
black       black 396.58000 352.68008791  95.4248925
lstat       lstat  36.24000  13.11892308   7.3187121
medv         medv  45.00000  22.36571429   9.5255725
> |
```

(f) Investigate graphically (via scatterplots) whether any of the predictors are asso- ciated with per capita crime rate (*crim*)? If so, comment on the relationship.

Answer: What I observe is if rm increased, crim tends to decrease.

(g) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? If yes - please comment on those (e.g. which suburbs are those, and what is the highest or lowest the value gets).

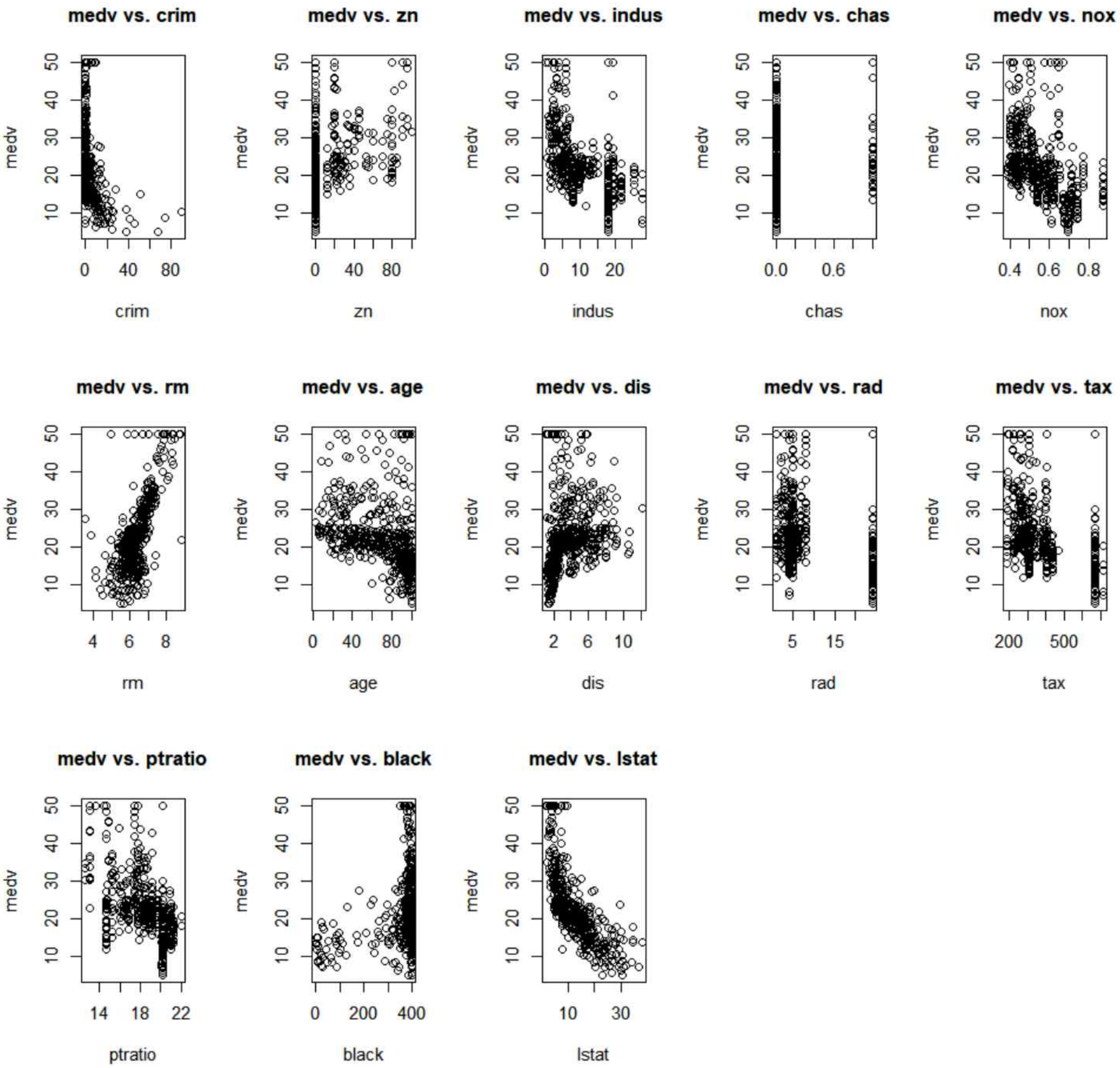
Answer: crime of ratio about suburb is median is 0.256% and max is 89% which means some neighborhoods has really high risks of crime.

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

Answer: It is 35.

(i) Suppose that we wish to predict median value price of the house (*medv*) on the basis of the other variables. Do any of the scatter plots (*medv* vs other predictor) suggest that any of the other variables might be useful in predicting *medv*? Justify your answer.

Answer:


medv vs. crim
medv vs. zn
medv vs. indus
medv vs. chas
medv vs. nox
medv vs. rm
medv vs. age
medv vs. dis
medv vs. rad
medv vs. tax
medv vs. ptratio
medv vs. black
medv vs. lstat
medv
crim
zn
indus
chas
nox
rm
age
dis
rad
tax
ptratio
black
lstat