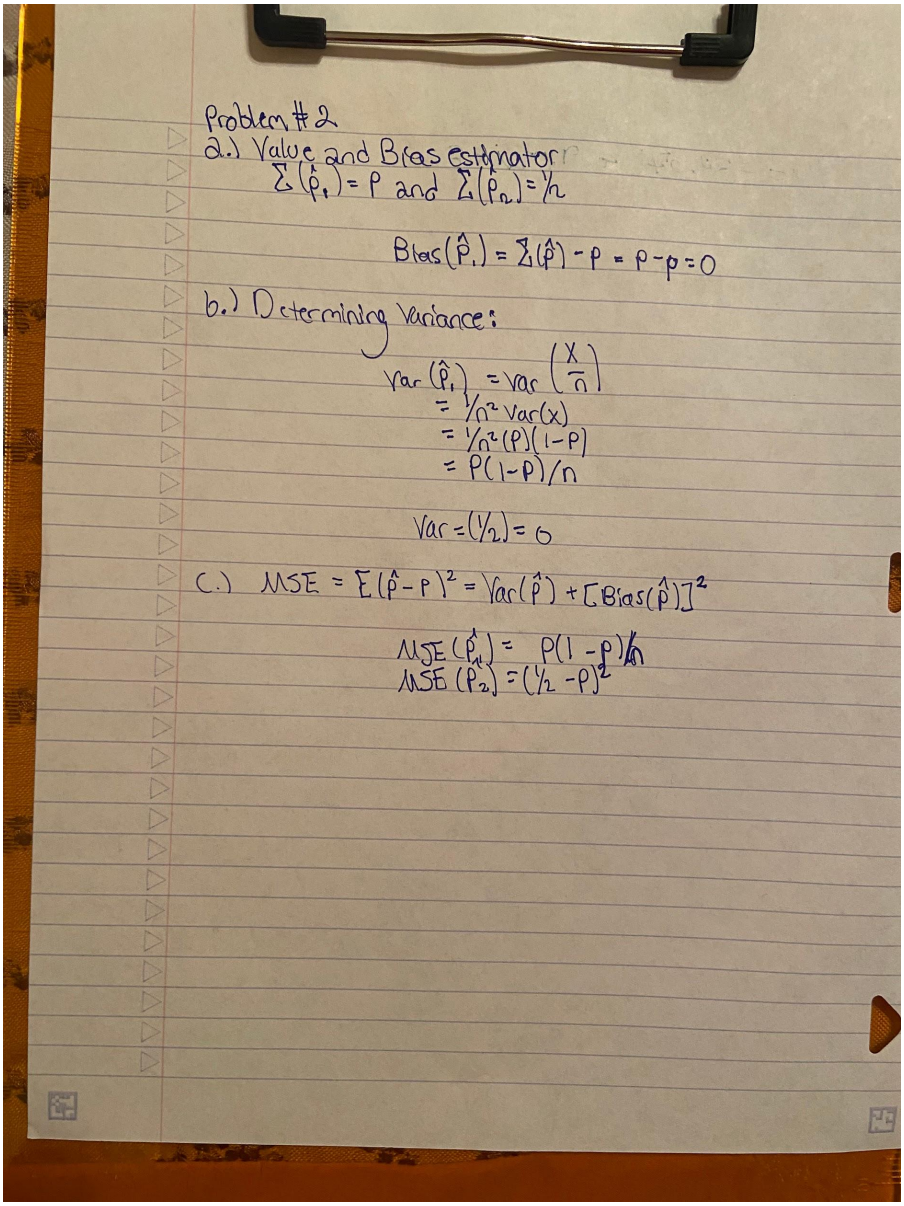
- **Problem 1: Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.**
- n = population size
- p = sample size
  - **We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.**
    - This is a regression problem because we're trying to return a specific numerical value.
    - n = 52(52 weeks in a year)
    - p = 4(4 weeks in a month)
  - **An online store is determining whether or not a customer will purchase additional items. This online store collected data from 1500 customers and looked at the cost of initial purchase, if there was a special offer, type of item purchased, number of times the customer logged into their account, and if they purchased additional items.**
    - Classification problem because its a yes or no problem
    - n = 1500(total amount of customers)
    - p = 5(total number of potential parameters)
- **Problem 2:This is an exercise about bias, variance and MSE. Suppose we have n independent Bernoulli trials with true success probability p. Consider two estimators of p: 1 pˆ1 = pˆ where pˆ is the sample proportion of successes and pˆ2 = 2 , a fixed constant.**



- **Problem 3: Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?**
  - **Definitions:**
    - **Parametric:** In parametric statistics, the information about the distribution of the population is known and is based on a fixed set of parameters.
    - **Non-Parametric:** In nonparametric statistics, the information about the distribution of a population is unknown, and the parameters are not fixed, which makes is necessary to test the hypothesis for the population
  - **Advantages of Parametric:** One advantage of parametric statistics is that they allow one to make generalizations from a sample to a population; this cannot

necessarily be said about nonparametric statistics. Another advantage of parametric tests is that they do not require interval- or ratio-scaled data to be transformed into rank data.

- ➢ **Disadvantages of Parametric**:The biggest disadvantage of parametric methods is that the assumptions we make may not always be true. For instance, you may assume that the form of the function is linear, whilst it is not. Therefore, these methods involve less flexible algorithms and are usually used for less complex problems.

❖ **Problem 4: This exercise involves the Auto data set in ISLR package. Make sure that the missing values have been removed from the data.**

- ➢ **(a) Which of the predictors are quantitative, and which are qualitative?**
  - ■ **Quantitative:** mpg, cylinders, displacement, horsepower, weight, acceleration, year.
  - ■ **Qualitative:** Origin, name
- ➢ **(b) What is the range of each quantitative predictor? You can answer this using the summary() function.**
  - ■

```
> summary(Auto)
      mpg          cylinders      displacement     horsepower        weight      acceleration        year           origin
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
 Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804   Median :15.50   Median :76.00   Median :1.000
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000

                 name
 amc matador       :  5
 ford pinto        :  5
 toyota corolla    :  5
 amc gremlin       :  4
 amc hornet        :  4
 chevrolet chevette:  4
 (Other)           :365
```

    - Mpg:(9-46.6)
    - Cylinders:(3-8)
    - Displacement:(68-455)
    - Horsepower:(46-230)
    - Weight:(1613-5140)
    - Acceleration:(8-24.80)
    - year:(70-82)
- ➢ **(c) What is the mean and standard deviation of each quantitative predictor?**
  - ■ **Means:**
    - **Mpg:**23.445918 , **Cylinders:** 5.471939, **Displacement:** 194.411990, **Horsepower:** 104.469388, **Weight:** 2977.584184, **Acceleration:** 15.541327, **Year:** 75.979592
  - ■ **STDev:**

- **Mpg:**7.805007, **Cylinders:** 1.705783, **Displacement:** 104.644004, **Horsepower:** 38.491160, **Weight:** 849.402560, **Acceleration:** 2.758864, **Year:** 3.683737

➢ **(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?**

```
      mpg          cylinders  displacement     horsepower      weight      acceleration
 Min.   :14.00   Min.   :8   Min.   :302.0   Min.   :130   Min.   :3433   Min.   : 8.5
 1st Qu.:14.00   1st Qu.:8   1st Qu.:307.0   1st Qu.:150   1st Qu.:3449   1st Qu.:10.0
 Median :15.00   Median :8   Median :350.0   Median :165   Median :3693   Median :10.5
 Mean   :15.67   Mean   :8   Mean   :373.2   Mean   :177   Mean   :3883   Mean   :10.5
 3rd Qu.:17.00   3rd Qu.:8   3rd Qu.:440.0   3rd Qu.:215   3rd Qu.:4341   3rd Qu.:11.5
 Max.   :18.00   Max.   :8   Max.   :455.0   Max.   :225   Max.   :4425   Max.   :12.0
      year
 Min.   :70
 1st Qu.:70
 Median :70
 Mean   :70
 3rd Qu.:70
 Max.   :70
> colMeans(auto.quant)#The mean
```

- **Range:**
  - Mpg:(14-18)
  - Cylinders:(8-8)
  - Displacement:(302-455)
  - Horsepower:(130-225)
  - Weight:(3433-4425)
  - Acceleration:(8.5-12.0)
  - year:(70-70)

```
> colMeans(auto.quant)#The mean
       mpg    cylinders displacement   horsepower       weight acceleration
  23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327
      year
  75.979592
> sqrt(diag(var(auto.quant.2)))#StdDEv
       mpg    cylinders displacement   horsepower       weight acceleration
   1.658312     0.000000    69.481612    37.446629   458.232474     1.250000
      year
   0.000000
>
```

- **Mean: Mpg:** 23.445918, **Cylinders:** 5.471939 , **Displacement:** 194.411990 , **Horsepower:** 104.469388, **Weight:** 2977.584184, **Acceleration:** 15.541327, **Year:** 75.979592
- **STDev:Mpg:**1.658312, **Cylinders:**0.000000 , **Displacement:** 69.481612, **Horsepower:** 37.446629, **Weight:**458.232474, **Acceleration:**1.250000 , **Year:** 0.000000

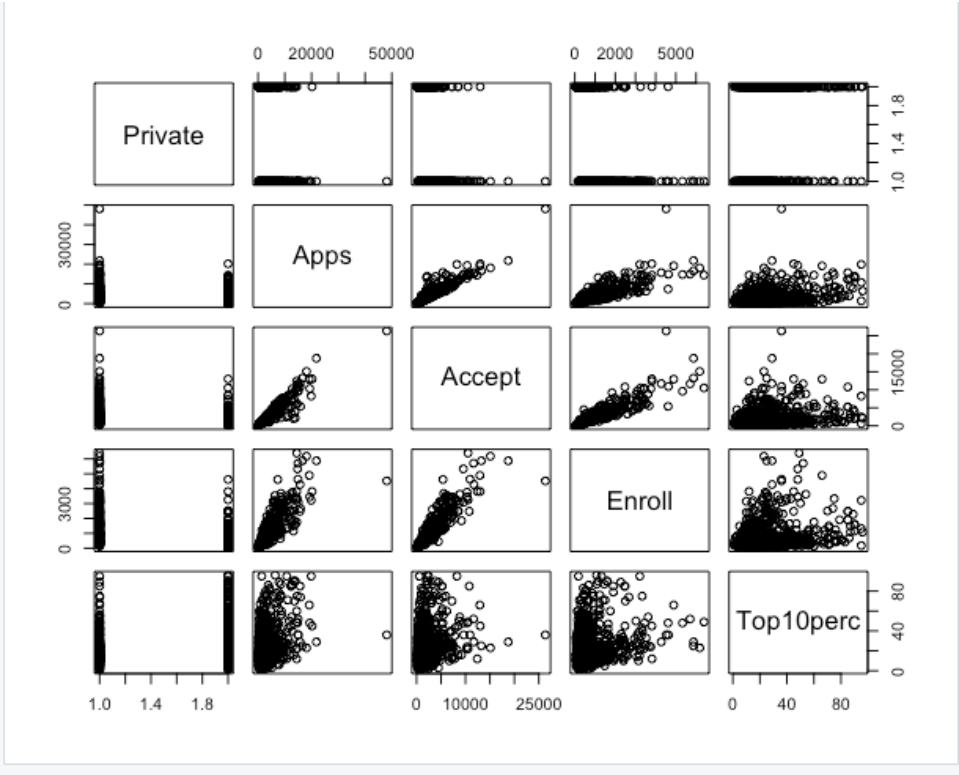
- **(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.**
  -
- **(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer**

- **Problem 5: This exercise relates to the College data set, which can be found in the fle College.csv attached to this homework set in Blackboard. It contains a number of variables for 777 different universities and colleges in the US. The variables are**
  - **Private : Public/private indicator**
  - **Apps : Number of applications received**
  - **Accept : Number of applicants accepted**
  - **Enroll : Number of new students enrolled**
  - **Top 10 Perc : New students from top 10% of high school class**
  - **Top 25 Perc : New students from top 25% of high school class**
  - **F.Undergrad : Number of full-time undergraduates**
  - **P.Undergrad : Number of part-time undergraduates**
  - **Outstate : Out-of-state tuition**
  - **Room.Board : Room and board costs**
  - **Books : Estimated book costs • Personal : Estimated personal spending**
  - **PhD : Percent of faculty with Ph.D.'s**
  - **Terminal : Percent of faculty with terminal degree**
  - **S.F.Ratio : Student/faculty ratio • perc.alumni : Percent of alumni who donate**
  - **Expend : Instructional expenditure per student**
  - **Grad.Rate : Graduation rate Before reading the data into R, it can be viewed in Excel or a text editor**
  -
- C.) Use the summary() function to produce a numerical summary of the variables in the data set. Are there any variables that do not show a numerical summary?

```
  Private                Apps           Accept
Length:777         Min.   :   81   Min.   :   72
Class :character   1st Qu.:  776   1st Qu.:  604
Mode  :character   Median : 1558   Median : 1110
                   Mean   : 3002   Mean   : 2019
                   3rd Qu.: 3624   3rd Qu.: 2424
                   Max.   :48094   Max.   :26330
    Enroll       Top10perc       Top25perc
Min.   :  35   Min.   : 1.00   Min.   :  9.0
1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
Median : 434   Median :23.00   Median : 54.0
Mean   : 780   Mean   :27.56   Mean   : 55.8
3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
Max.   :6392   Max.   :96.00   Max.   :100.0
 F.Undergrad     P.Undergrad         Outstate
Min.   :  139   Min.   :    1.0   Min.   : 2340
1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
Median : 1707   Median :  353.0   Median : 9990
Mean   : 3700   Mean   :  855.3   Mean   :10441
3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
Max.   :31643   Max.   :21836.0   Max.   :21700
  Room.Board       Books           Personal
Min.   :1780   Min.   :  96.0   Min.   : 250
1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
Median :4200   Median : 500.0   Median :1200
Mean   :4358   Mean   : 549.4   Mean   :1341
3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700
Max.   :8124   Max.   :2340.0   Max.   :6800
     PhD            Terminal       S.F.Ratio
Min.   :  8.00   Min.   : 24.0   Min.   : 2.50
1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50
Median : 75.00   Median : 82.0   Median :13.60
Mean   : 72.66   Mean   : 79.7   Mean   :14.09
3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50
Max.   :103.00   Max.   :100.0   Max.   :39.80
 perc.alumni        Expend        Grad.Rate
Min.   : 0.00   Min.   : 3186   Min.   : 10.00
1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
Median :21.00   Median : 8377   Median : 65.00
Mean   :22.74   Mean   : 9660   Mean   : 65.46
3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
Max.   :64.00   Max.   :56233   Max.   :118.00
>
```

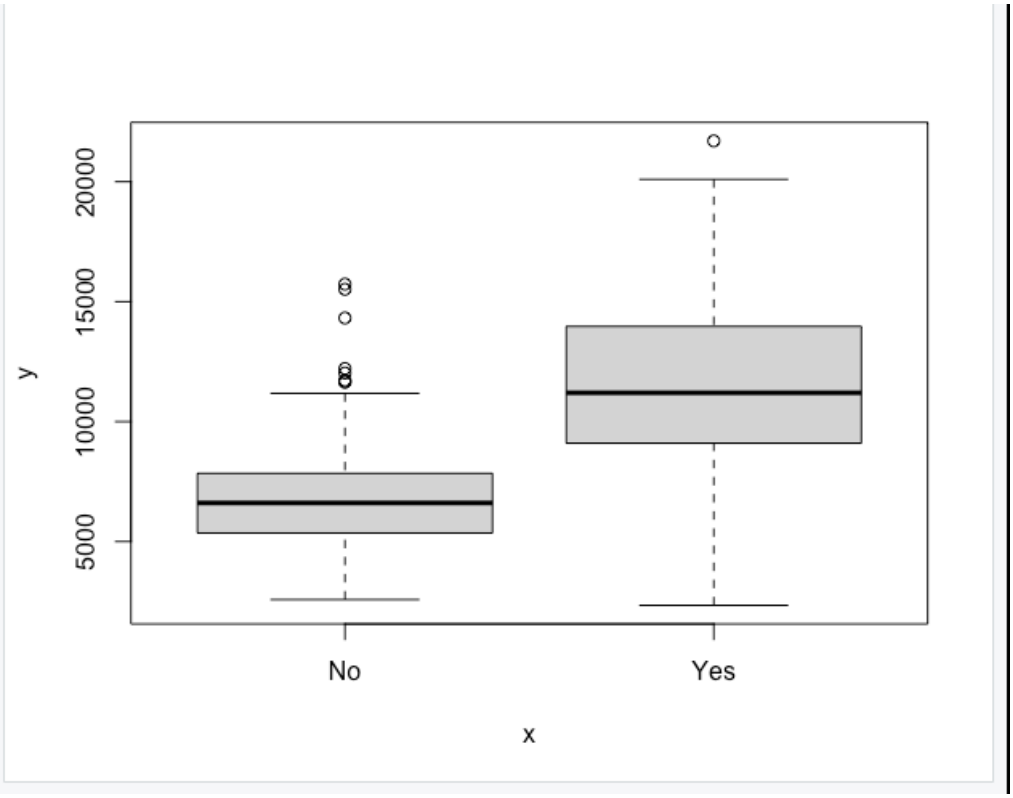
-
- Private doesn't show any numerical data

➢ **D) Use the pairs() function to produce a scatterplot matrix of the first five columns or variable of the dataset. Describe any relationships you see in these plots.**



-
- There is a positive linear relationship between number of applications, number accepted, and number enrolled**.**

➢ **E) Use the plot() function to produce a plot of Outstate versus Private. What type of plot was produced? Give a description of the relationship. Hint: 'Outstate is in the y-axis.**



- A box plot was produced which shows that there seems to be a significantly high out of state tuition for private institutions.

➢ **F) Create a new qualitative variable, called Elite, by binning the Top 10 Perc 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%. Type in the following in R:**

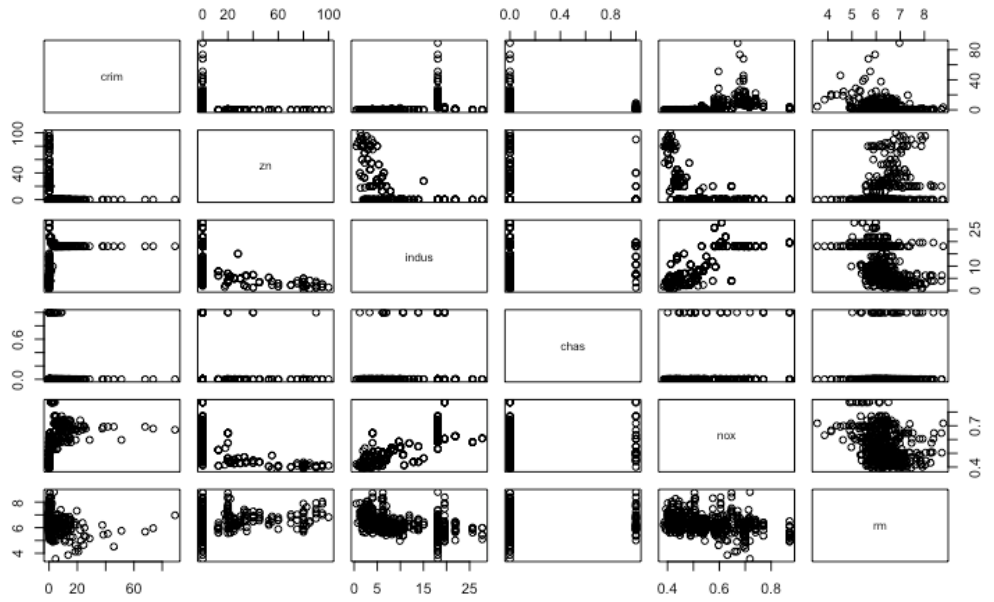
- No = 699, Yes = 78

❖ **Problem 6: This exercise involves the Boston housing data set.**

➢ **(a) To begin, load in the Boston data set. The Boston data set is part of the ISLR2 library. You may have**

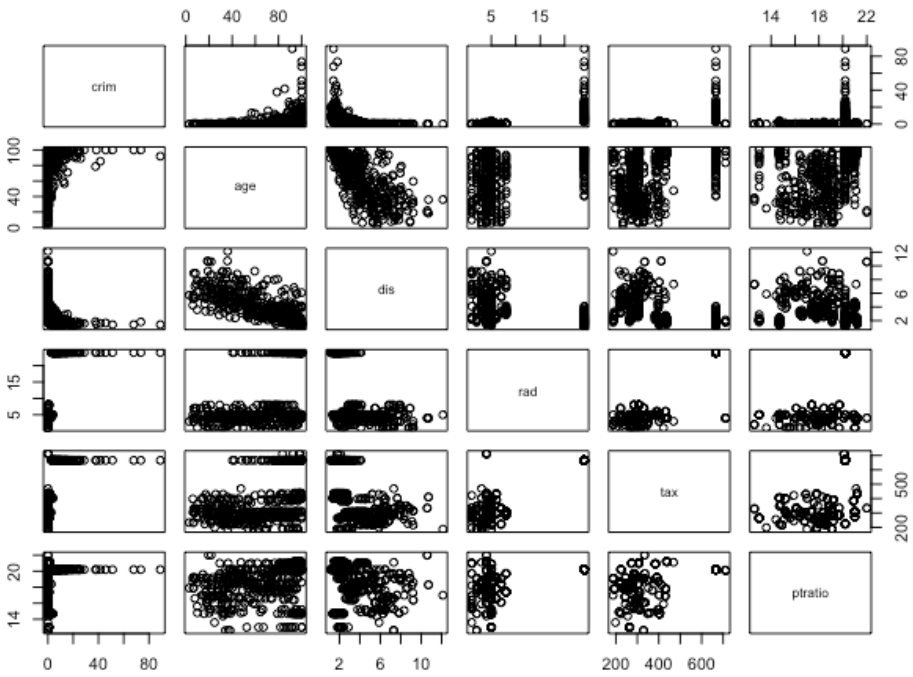
➢ **How many rows are in this data set? How many columns? What do the rows and columns represent?**

- There are 506 rows and 13 columns. The rows represent the number of observation numbers of the suburbs while the columns represent the total number of variables.
- **(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.**



-
- The data is rather unorganized for the most part however, Crime is generally High as Nox increases.

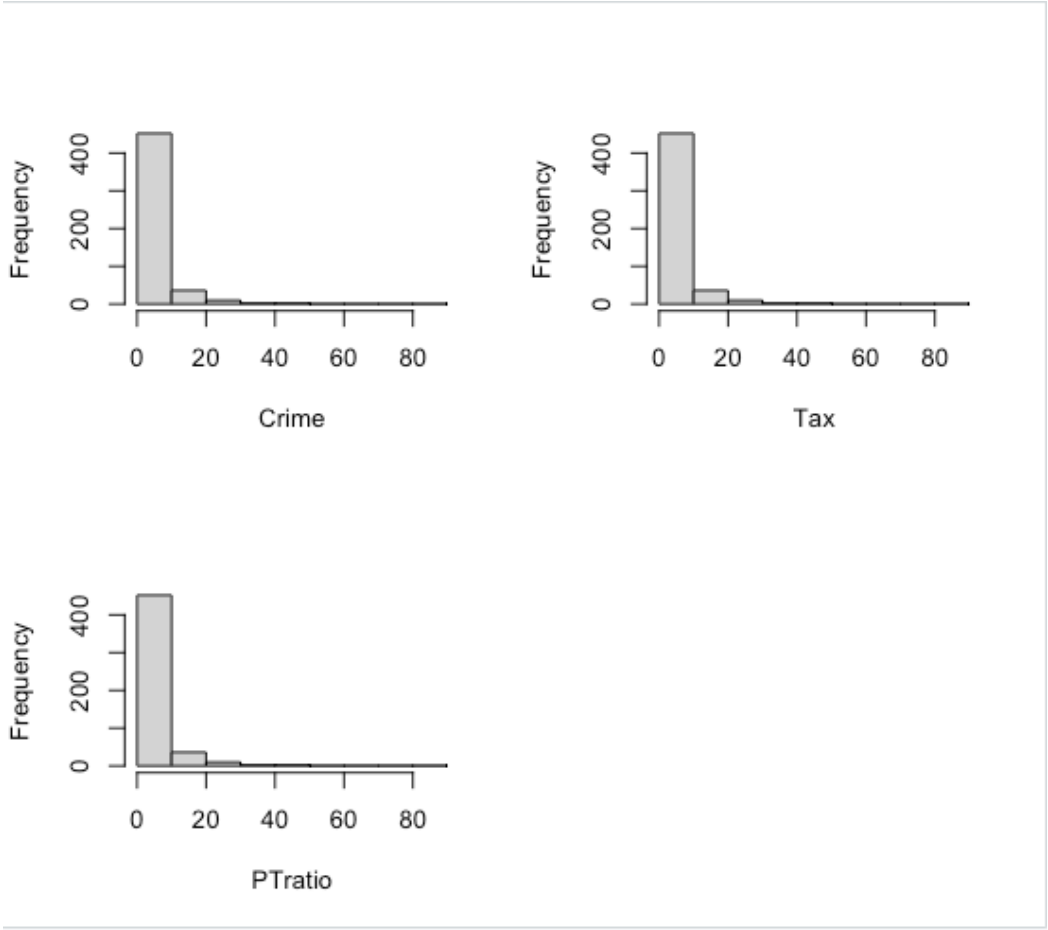
- **(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.**



- The strongest correlation regarding crime rates is with rad and tax. As both increase, so does the crime rate.

```
             zn      indus        chas       nox         rm       age        dis
[1,] -0.2004692 0.4065834 -0.05589158 0.4209717 -0.2192467 0.3527343 -0.3796701
           rad       tax   ptratio
[1,] 0.6255051 0.5827643 0.2899456
```

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



- The crime rate is skewed to the right, when it comes to the tax graph we can see a massive jump from 500(almost non-existent) to 700. While the parent to teacher ratio graph is skewed to the left.

- **(e) How many of the census tracts in this data set bound the Charles river? 3**
  - There are 35 suburbs that are bound to the Charles River.
- **(f) What is the median pupil-teacher ratio among the towns in this data set?**
  - The median pupil-teacher ratio is 19.05%
- **(g) Which census tract of Boston has the lowest median value of owner occupied homes? What are the values of the other predictors for that**

census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

- The lowest median value of owner occupied homes is 399
-

- **(h) In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.**
  - The Number of suburbs with an average of seven rooms or more is 64.
  - The number of suburbs with an average of eight rooms or more is 13.

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

-