Chapter 2 HW

Nehemya McCarter-Ribakoff

14 February 2017

Conceptual Questions

Exercise 2: Identify each as classification or regression, prediction or inference. Provide n and p

```
(a) We collect a set of data on the top 500 firms in the US. For each
firm we record profit, number of employees, industry and the
CEO salary. We are interested in understanding which factors
affect CEO salary.

Since we are looking for factors, or categories, this is classification.
We are not building a model for prediction, only studying patterns, so
this is inference.
n = 500, p = {profit, number of employees, industry, and CEO salary}

(b) We are considering launching a new product and wish to know
whether it will be a success or a 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.

Since products are labeled either success or failure, this is classification.
We are determining if our product will be a success, so this is prediction.
n = 20, p = {success or failure, price of product, marketing budget,
             competition budget, 10 other...}

(c) We are interested in predicting the % change in the US dollar 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 dollar, the % change in the US market,
the % change in the British market, and the % change in the
German market.

We are looking for a percent change. This is regression.
We are making a prediction model.
n = 52, p = {Δdollar, ΔUS, ΔBritish, ΔGerman}
```

Exercise 4: You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

1. Mobile phone application to determine the species of a plant in a given photo based off a training dataset.

   predictors = {plant.shape, plant.color, plant.structure, ...}
   Prediction model
   The model looks for a photo or species in its database that matches the given photo as closely as possible. This is an image classifier.

2. Virtual doctor providing diagnoses based on symptoms.

   predictors = {body.temp, inflammation(if any), reported.symptoms, ...}
   Goal: prediction
   The output of our model is an element from the set of all diseases.

3. Determine if decreasing size of fish population is correlated with region, species, diet, etc.

   predictors = {region, species, diet, etc.}
   Inference model.
   Our outcomes are discrete -- either there is a correlation or there isn't. Additionally, we are only checking if there is a pattern. No output is generated, so this is inference.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

1. Find the probability a tornado will occur based on climate data.
   predictors = {atmospheric.pressure, temperature, wind.data, ...}
   Goal: prediction
   Because a probability is a continuous value, this is a regression problem

2. Estimate the property value of a given plot of land
   predictors = {location, size, structures, ...}
   Goal: prediction
   A property value is continuous, but because the price at which a plot will sell is, to a degree, arbitrary, this problem is prediction, not inference.

3. Determine how much each factor impacts on property sales.
   predictors = {location, size, structures, ...}
   Goal: inference
   Our scenario is similar to the last example, but we are only looking for a better understanding of the relationship between our predictors and outcome. Since our end goal is quantitative ("how much"), this is also an inference problem.

Applied Questions

Exercise 8:

a. import data (assumes College.csv is in working directory)

```
college = read.csv(file = "College.csv", header = TRUE, sep = ",")
```

b. print csv table

```
fix(college)
  # remove uni names from list
rownames(college) = college[,1]
fix(college)
  # isolate college names (sets as private)
college = college[,-1]
  #let's look again
fix(college)
```

c. See parts below

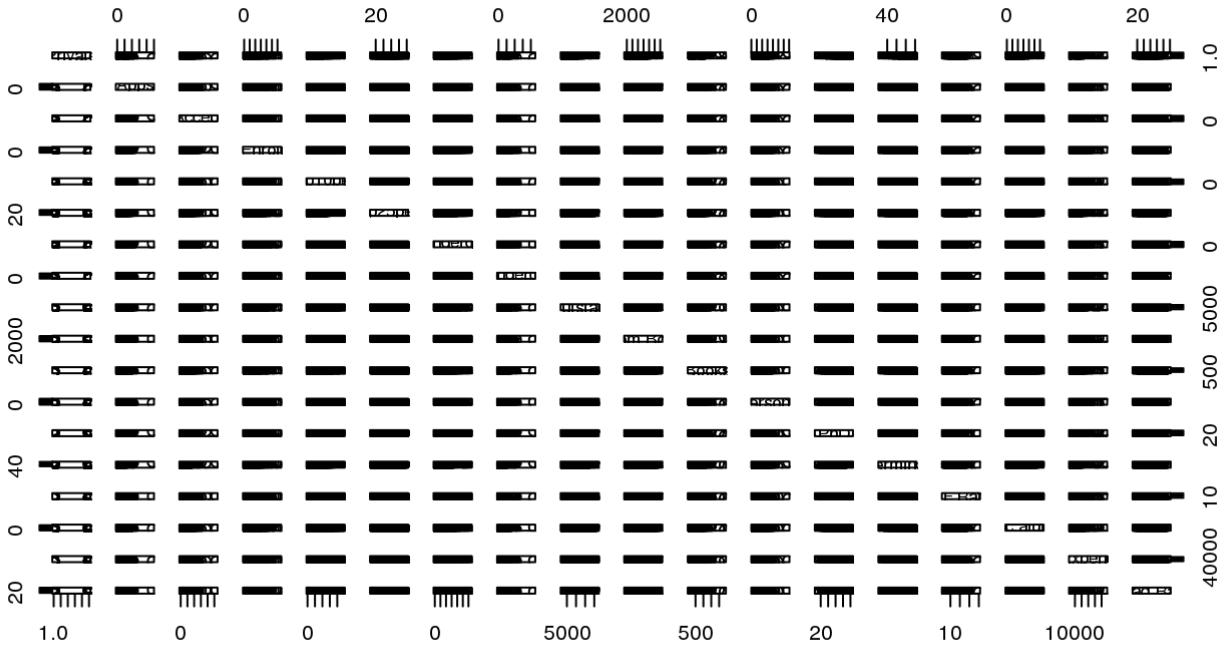
d. use summary() to produce numerical summary of dataset

```
summary(college)
```

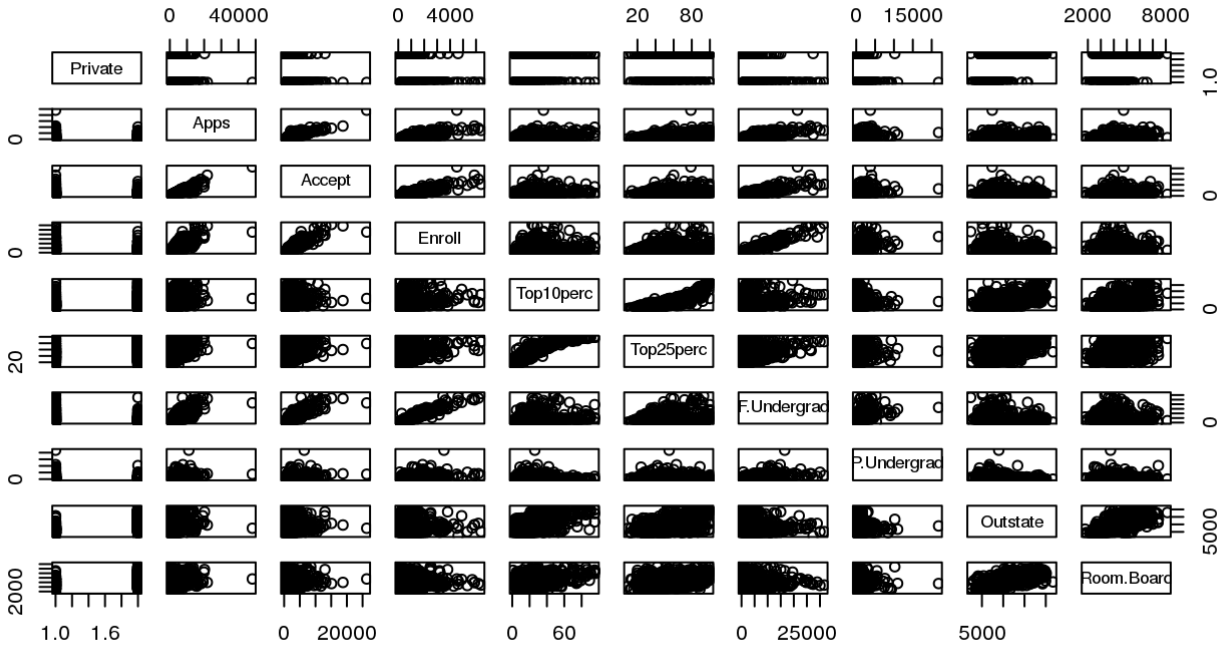
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

ii. use pairs() to produce a scatterplot matrix

```
pairs(college)        # references whole set
```

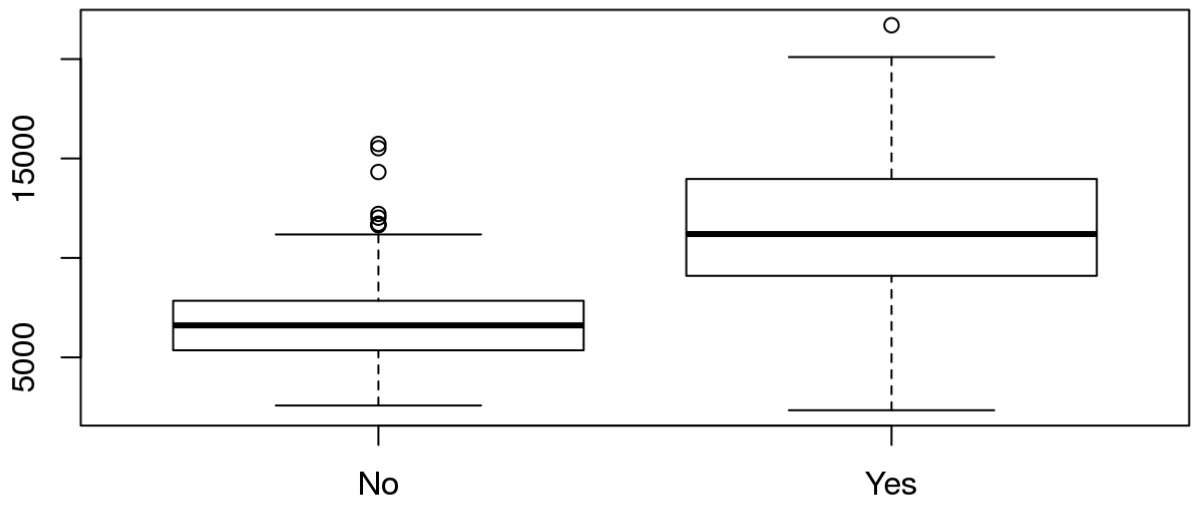


```
pairs(college[,1:10]) # references first 10 columns
```



iii. Use plot() to produce side-by-side boxplots of OutState vs Private

```
plot(college$Private, college$Outstate, main="Private vs. Out of State Universities")
```

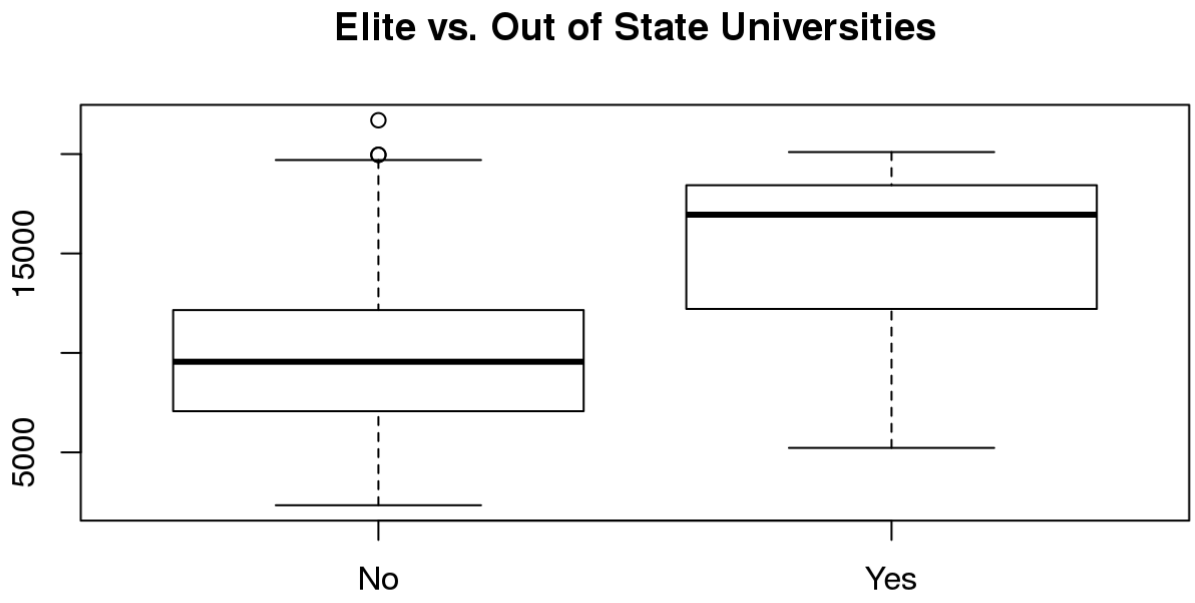


iv. Create qualitative variable Elite by binning Top10perc

```
Elite = rep("No",nrow(college))
Elite[college$Top10perc>50]="Yes"
Elite=as.factor(Elite)
college=data.frame(college,Elite)
# summary() to see how many elite universities there are
summary(Elite)
```

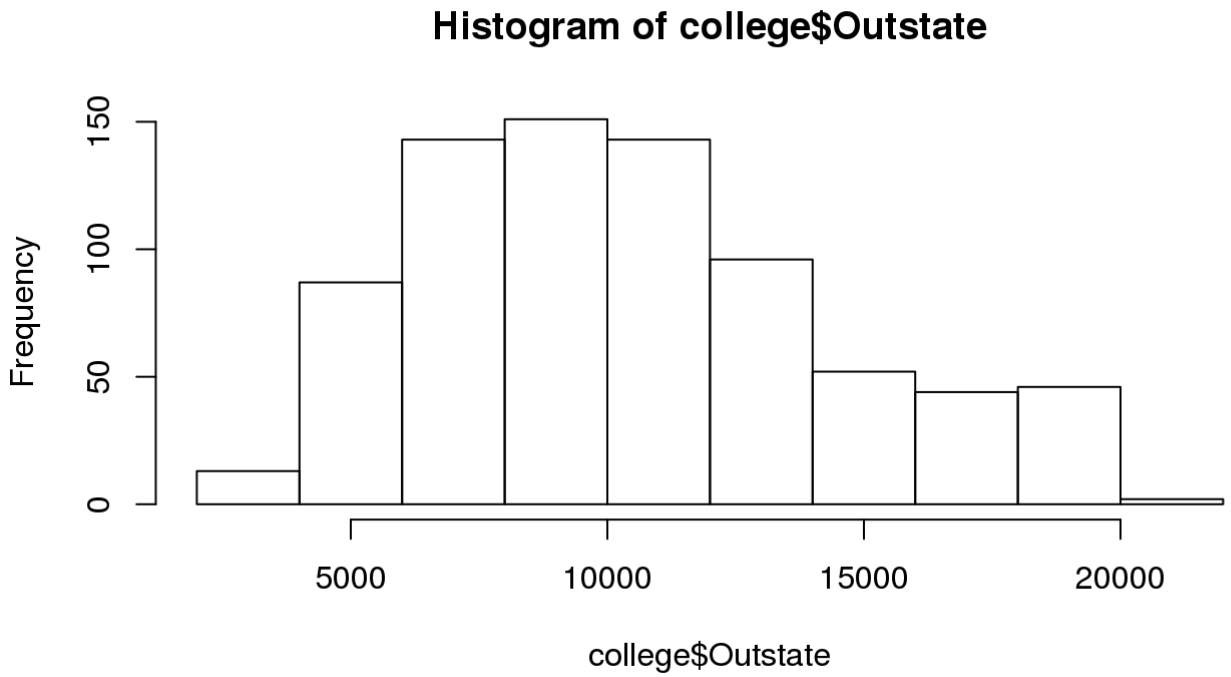
```
##  No Yes
## 699  78
```

```
# plot() for side-by-side boxplots of OutState vs Elite
plot(college$Elite, college$Outstate, main="Elite vs. Out of State Universities")
```

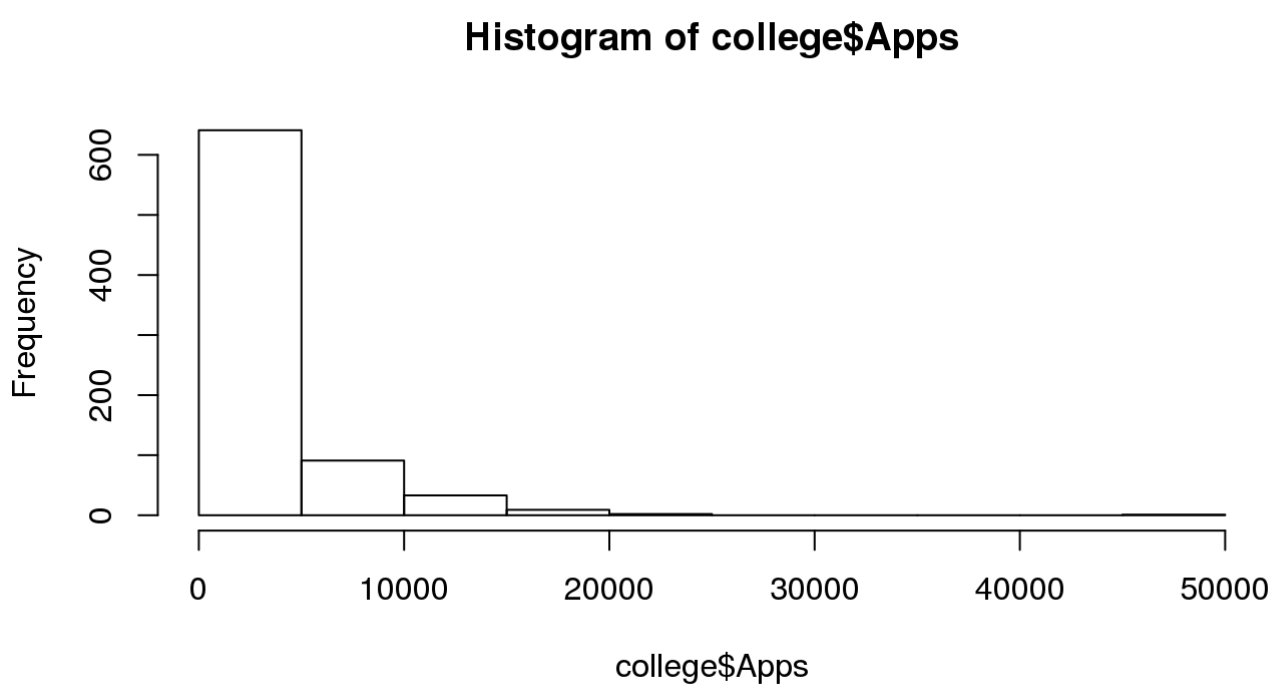


v. Use hist() to produce some histograms

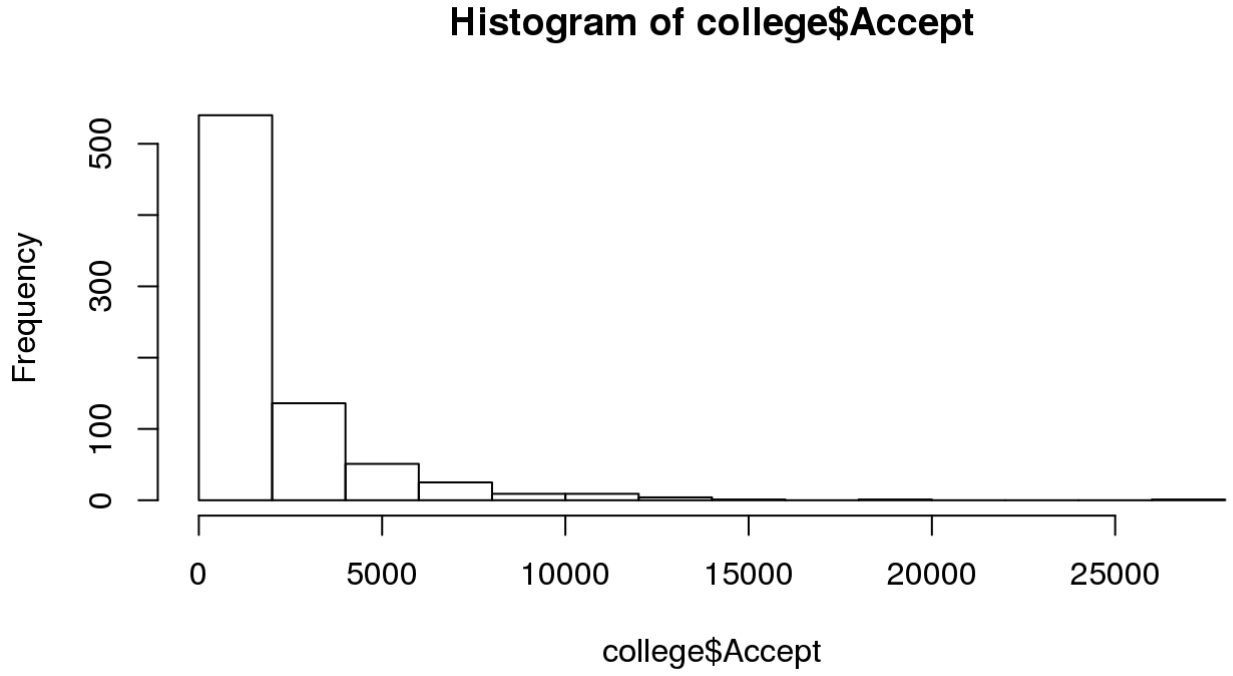
```
# output format
par(mfrow=c(1,1))
# some histograms
hist(college$Outstate)
```



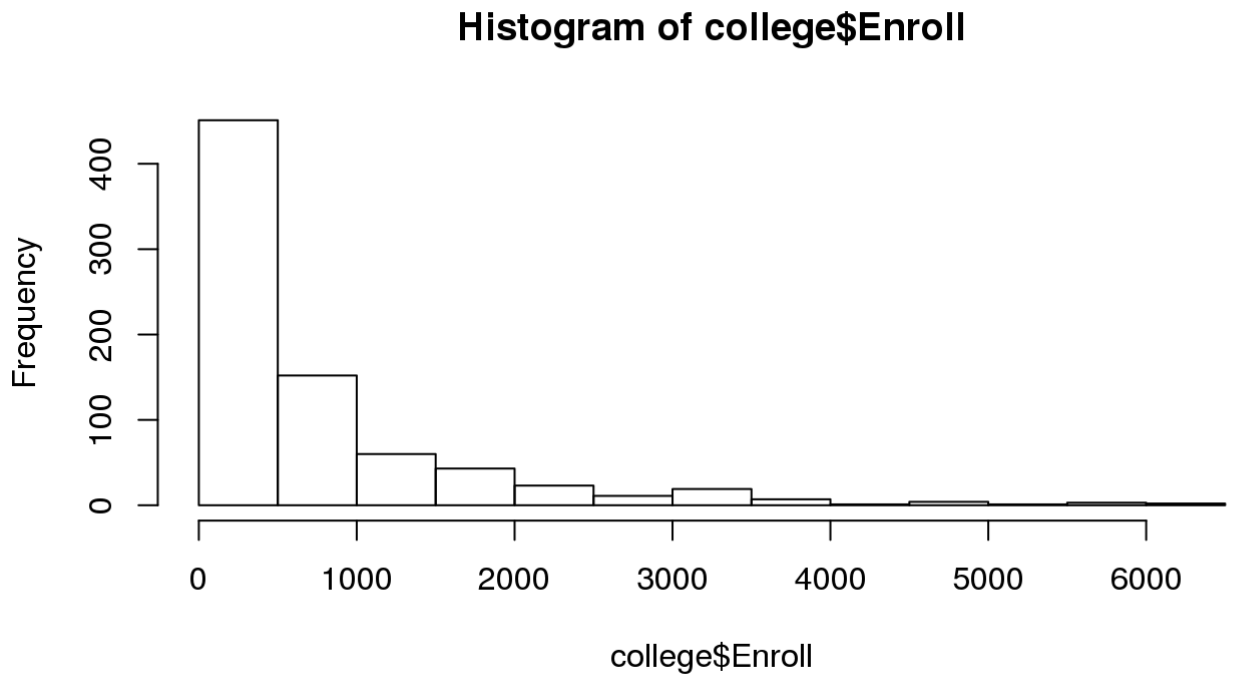

```
hist(college$Apps)
```



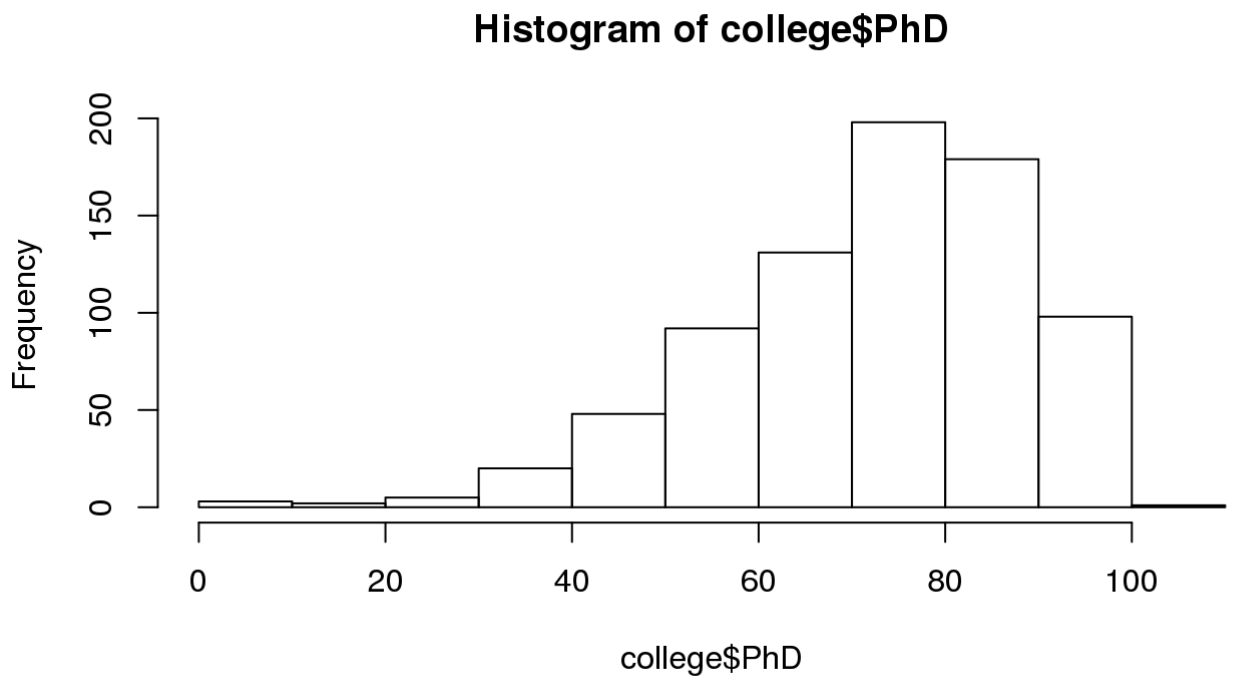

```
hist(college$Accept)
```




```
hist(college$Enroll)
```




```
hist(college$PhD)
```



Exercise 10:

```
# a) load data
library(MASS)
#Boston
#Command is commented out to save paper
# info on set
?Boston
```

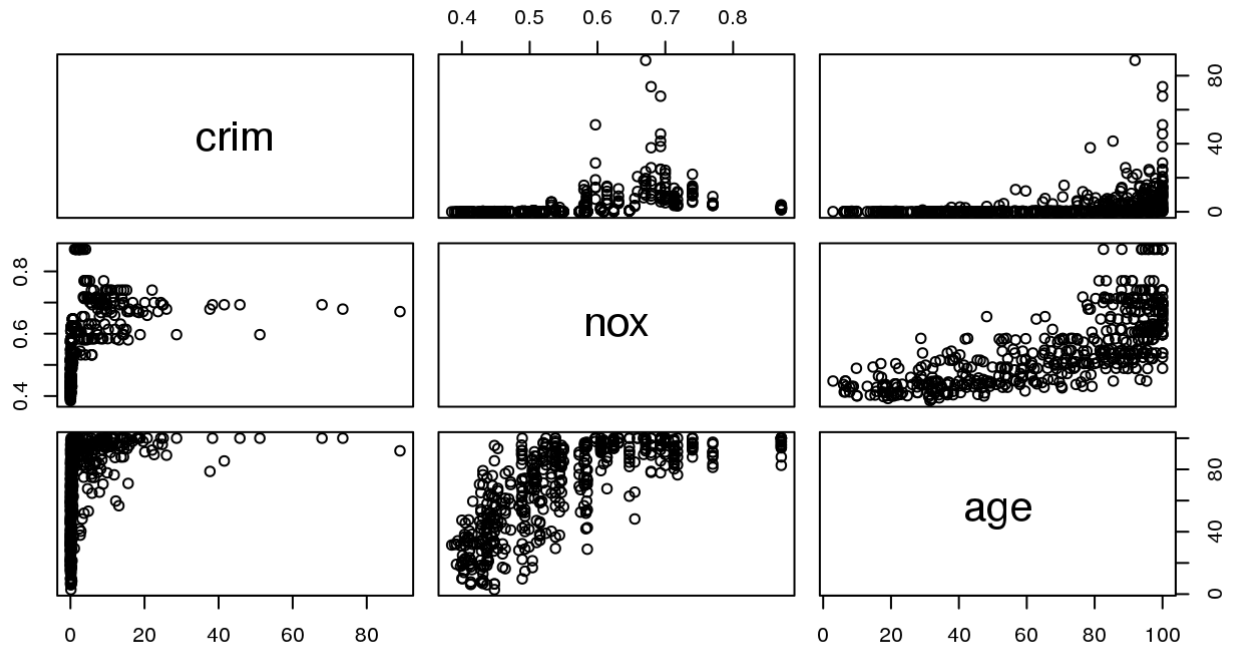
How many rows? How many columns? What do they represent?

506 rows, 14 columns. Columns represent crime rate, proportional land zoned, proportional non-retail business acres per town, Charles river tract, ntriogen oxides concentration, avg rooms per dwelling, proportion of owner-occupied units built prior to 1940, weighted mean of distances to five Boston employment centres, index of accessibility to radial highways, full-value property-tax rate, pupil-teacher ration by town, proportaion of blacks by town, percent lower status of the population, and median value of owner-occupied homes. Rows represent distinct suburbs.

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

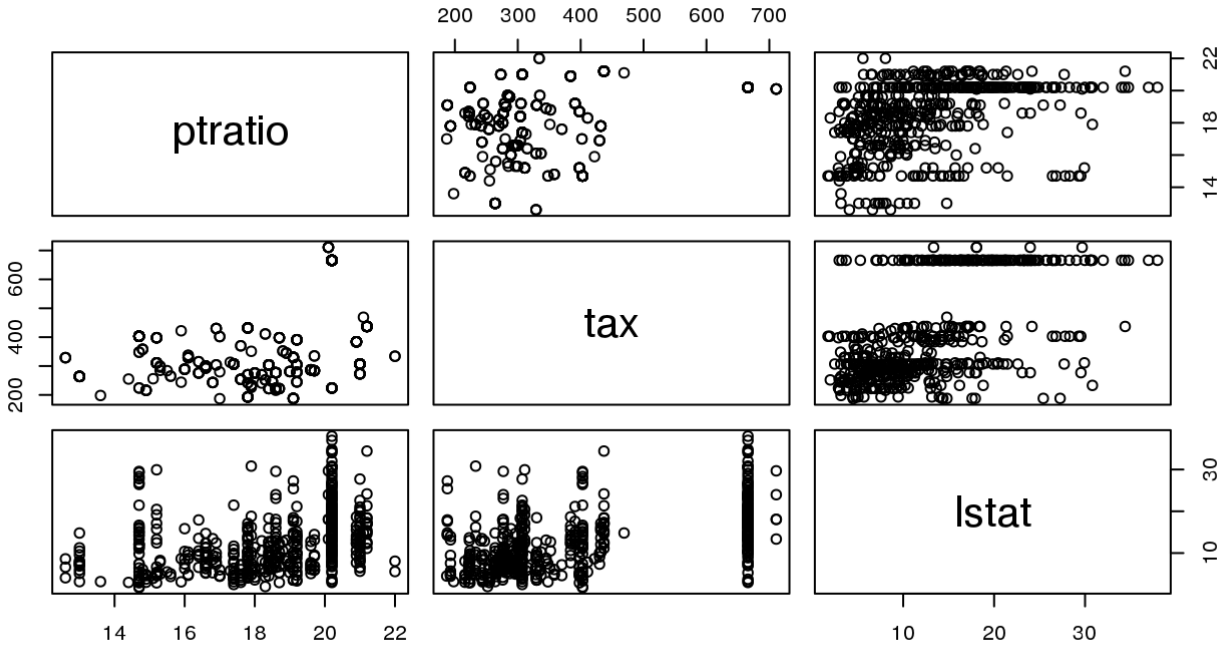
   Let's take a look at how crime, nitrogen oxides, and home age relate

```
attach(Boston) # so we may access column names as variables
pairs(~ crim + nox + age)
```



There appears to be a weak relationship between nox and areas with a high concentration of old homes.. Crime appears to experience a spike in areas with a high concentration of old homes, despite being fairly low and steady prior to the spike.

```
pairs(~ ptratio + tax + lstat)
```

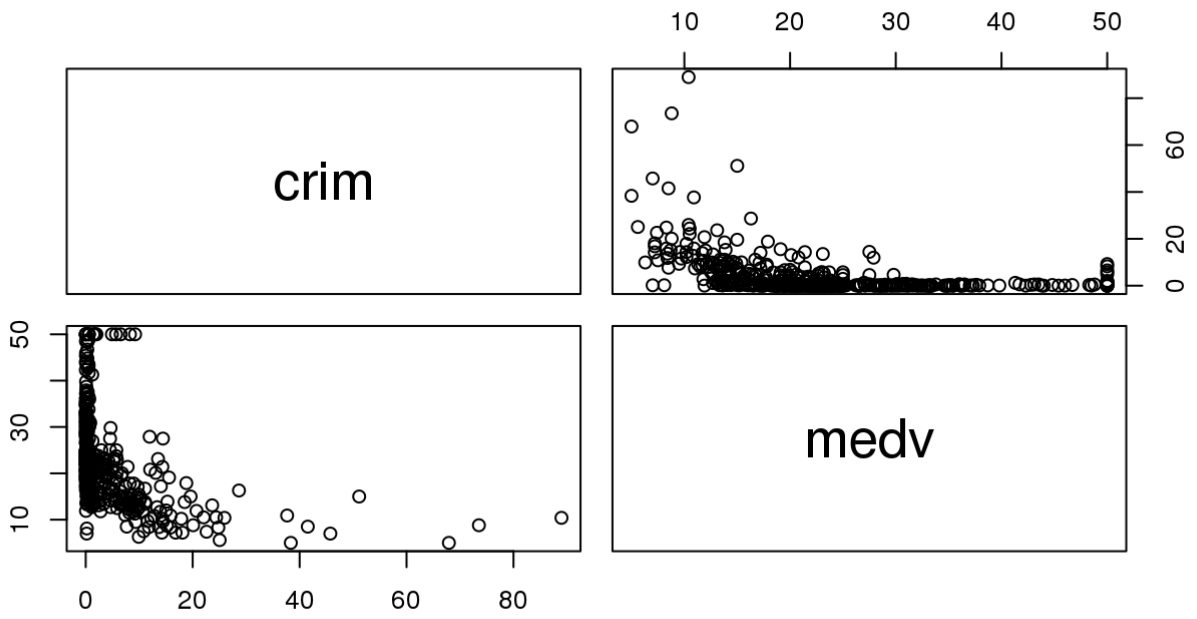


I see no distinct relationship between the pupil-teacher ratio and full-value property-tax rates. Pupil-teacher ratios seem to be varied for areas with areas with fewer lower-status populations. The ratios present for poorer regions are higher, but their are also far few data points, which may indicate fewer teachers and pupils altogether. Areas with the smallest lower-status populations appear to have a bubble of low tax rates, a gap in the middle, and a few high tax rates. Residents with hight tax rates seem to live in neighborhoods across the spectrum.

c. Are any of the predictors associated with per capita crime rate? If so, describe the realtionship.

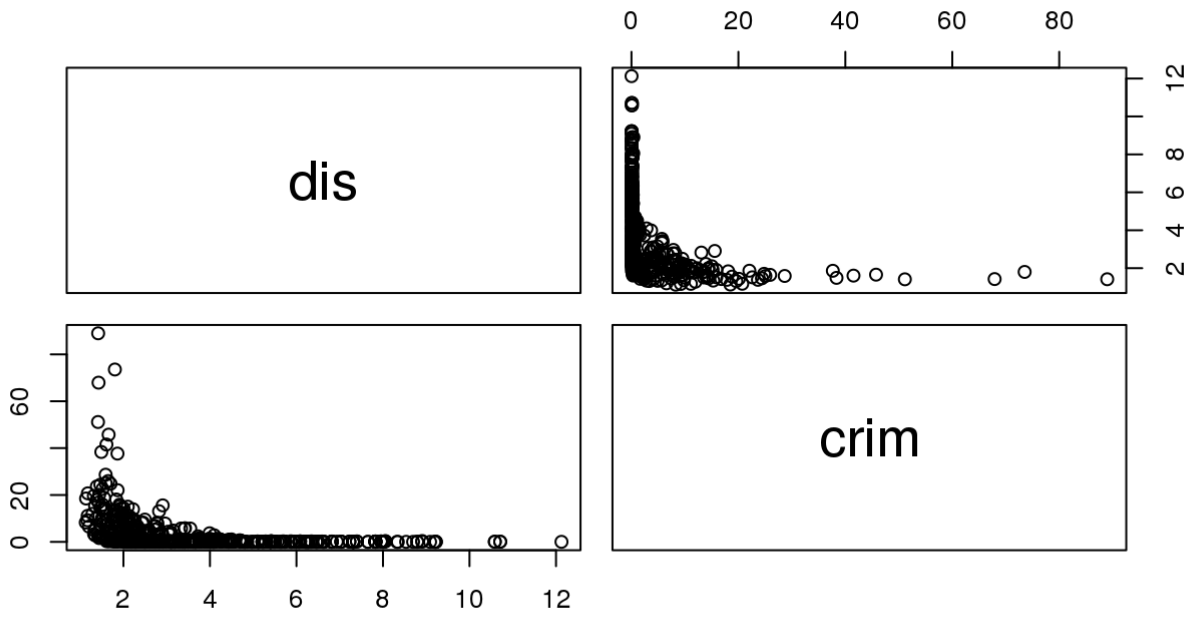
The followig predictors exhibit some correlation with per capita crim.

```
pairs(~ crim + medv)
```



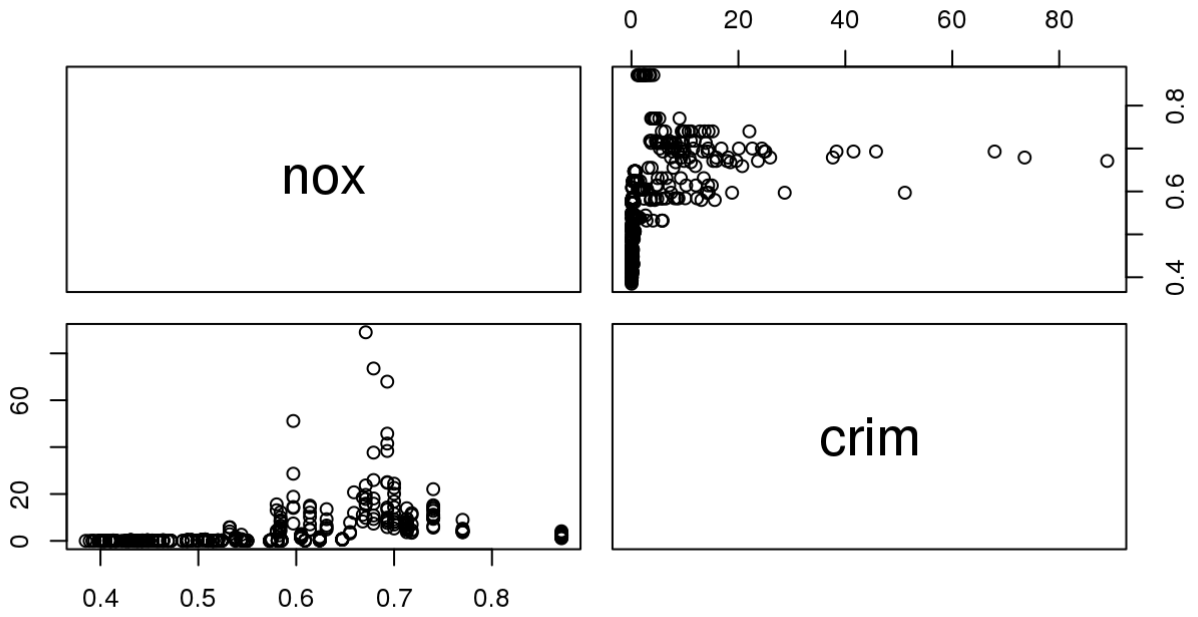
Higher crime rates appear to be associate with lower median value homes.

```
pairs(~ dis + crim)
```



Most data points occur at distances closer to employment centers. Further distances have lower per capita crime, but the data also tapers off. This could be a matter of sampling.

```
pairs(~ nox + crim)
```



Higher crime rates appear at areas with a rate between 0.6 and 0.8 nitrogen oxide parts per 10 million

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

Let's take a closer look at each

```
summary(crim)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00632  0.08204  0.25650  3.61400  3.67700 88.98000
```

Our range is 88.980 - 0.00632 = 88.97368. This is quite a large range. On top of that, our median and mean are quite different from one another, which suggests there are a lot of outliers. Here are the highest per capita crime rates

```
tail(sort(crim))
```

```
## [1] 41.5292 45.7461 51.1358 67.9208 73.5341 88.9762
```

Let's take a look at property tax rates.

```
summary(tax)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   187.0   279.0   330.0   408.2   666.0   711.0
```

Our range is 711 - 187 = 524. Important to keep in mind our numbers here are per $10,000, so that's a difference of 524 / 10,000 = 0.0524, or about a 5% difference from the highest to the lowest tax rate. This is a smaller range than I had expected, given the United State's progressive tax system. Here are the set's highest.

```
tail(sort(tax))
```

```
## [1] 666 711 711 711 711 711
```

Lastly, we have pupil-teacher ratios.

```
summary(ptratio)
```

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

The range is 22 - 12.6 = 9.4. Since we are talking about a number of students, we may say the highest and lowest sampled classrooms differ by about 10 students. Here, our mean and median are somewhat close, which tells us the data is fairly normal. The highest ratios are as follows:

```
tail(sort(ptratio))
```

```
## [1] 21.2 21.2 21.2 21.2 22.0 22.0
```

e. How many of the suburbs in this data set bound the Charles river?

```
sum(chas)
```

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

35 suburbs bound the Charles River

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

```
median(ptratio)
```

```
## [1] 19.05
```

Median = 19.05 pupils per teacher

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.

```
lowest.medv = which.min(medv)
Boston[lowest.medv,]
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio black
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.9
##     lstat medv
## 399 30.59    5
```

This suburb has a crime rate, industry zoning, and nox concentration above the 3rd quadrant. None of its land is zoned for lots over 25,000 sq.ft. It is important to note here that the median is also 0 for this predictor. Its population percentage in the lower status (30.59%) is near the highest recorded. All of its owner-occupied homes were build prior to 1940. Its owner-occupied dwellings average about 6 rooms, which is the same as the first quadrant, and sounds about right for a suburban area. These numbers suggest this is a poor, crime-ridden suburb.

The suburb has a weighted mean of 1.4896 (no units included) to an employment center. Residents are not too far from work. In fact, they are closer than most suburbs in Boston, sitting well below the mean and median, as well as the first quadrant. However, They are also the farthest suburb to any radial highways. It may be difficult for residents to get out to attain work in a different area.

This suburb also has a pupil-teacher ratio at the third quadrant of our data, about 1 pupil away from being the highest. There are relatively few teachers for the students in this area. This area also has the highest population of black individuals of all suburbs in Boston.

h. 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

```
sum(rm > 7)
```

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

64 suburbs average more than seven rooms per dwelling

```
sum(rm > 8)
```

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

13 suburbs average more than eight rooms per dwelling.

Teamwork report

Team member	Conceptual	Applied	Contribution %
Nehemya	Yes	Yes	100%
Total			100%