Data Mining Assignment 2

1) Read Chapter 1 (all) and Chapter 2 (only sections 2.1, 2.2 and 2.3).  
  
2) Redo In Class Exercises #1 and #2, but use different examples from those which we used in class.  
  
3) Do Chapter 2 textbook [problem #2](http://www.cob.sjsu.edu/mease_d/bus297D/ch2textbookquestion.doc) on page 89.

Classify the following attributes as binary, discrete, or continuous. Also

classify them as qualitative (nominal or ordinal) or quantitative (interval or

ratio). Some cases may have more than one interpretation, so briefly indicate

your reasoning if you think there may be some ambiguity.

**Example:** Age in years. **Answer:** Discrete, quantitative, ratio

1. Time in terms of AM or PM.

Binary, qualitative, ordinal

1. Brightness as measured by a light meter.

Continuous, quantitative, ratio

1. Brightness as measured by people’s judgments.

Discrete, qualitative, ordinal

1. Angles as measured in degrees between 0*◦* and 360*◦*.

Continuous, quantitative, ratio

1. Bronze, Silver, and Gold medals as awarded at the Olympics.

Discrete, qualitative, ordinal

1. Height above sea level.

Continuous, quantitative, interval/ratio (depends on whether sea level is regarded as an arbitrary origin)

1. Number of patients in a hospital.

Discrete, quantitative, ratio

1. ISBN numbers for books. (Look up the format on the Web.)

Discrete, qualitative, nominal (ISBN numbers do have order information, though)

(i) Ability to pass light in terms of the following values: opaque, translucent,

transparent.

Discrete, qualitative, ordinal

1. Military rank.

Discrete, qualitative, ordinal

1. Distance from the center of campus.

Continuous, quantitative, interval/ratio (depends)

1. Density of a substance in grams per cubic centimetre.

Discrete, quantitative, ratio

(m) Coat check number. (When you attend an event, you can often give

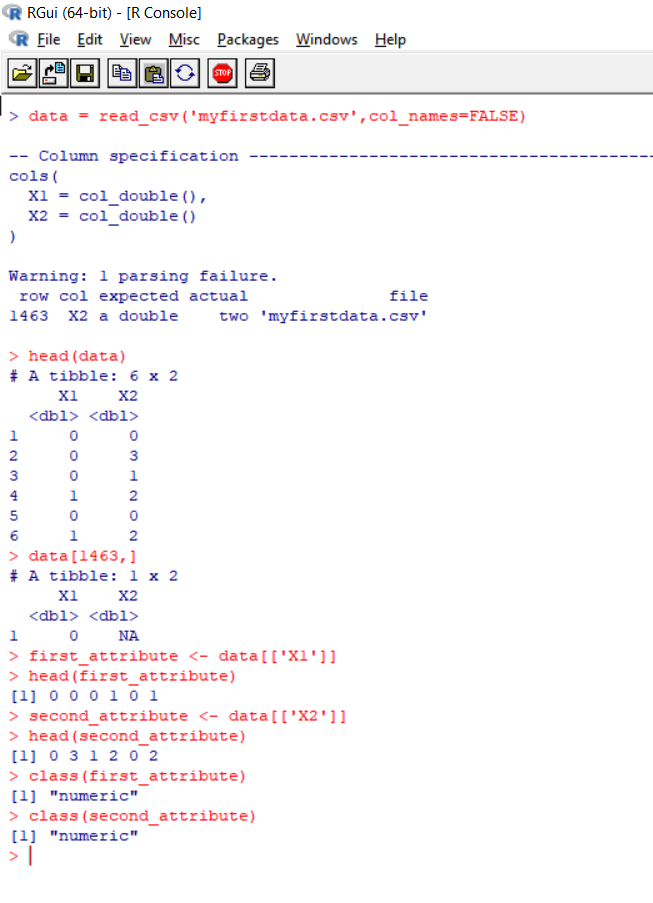
your coat to someone who, in turn, gives you a number that you can

use to claim your coat when you leave.)

Discrete, qualitative, nominal

4) This question uses the data at <http://www.cob.sjsu.edu/mease_d/bus297D/myfirstdata.csv>. Download it to your computer.  
  
a) Read in the data in R using data←read.csv("myfirstdata.csv",header=FALSE).

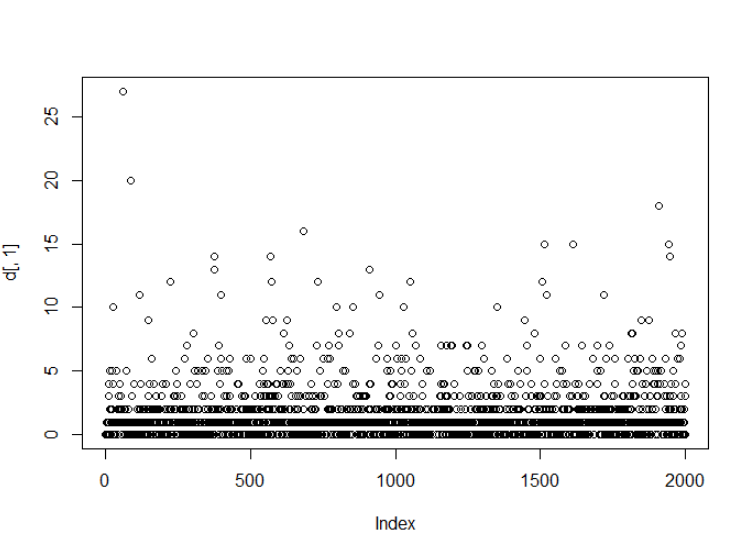
Note, you first need to specify your working directory using the setwd() command. Determine whether each of the two attributes (columns) is treated as qualitative (categorical) or quantitative (numeric) using R. Explain how you can tell using R.



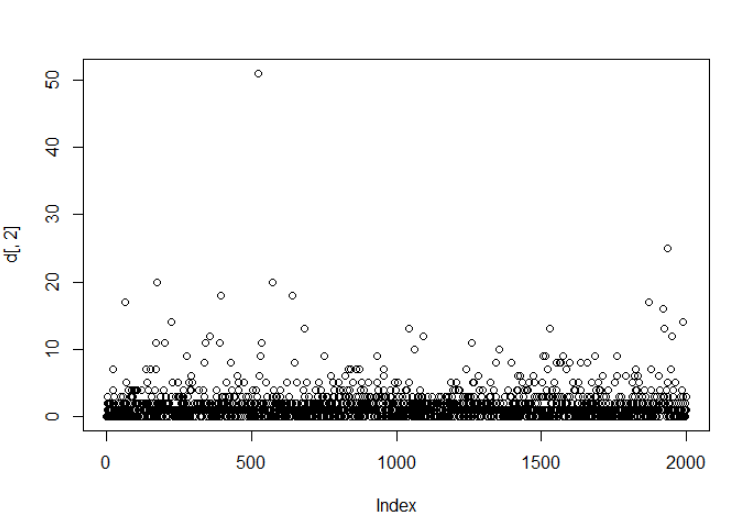
It treats both attributes as numeric as it is automatically parsing wrong type value in second column ‘two’ at 1463 row as NA.  
  
b) What is the specific problem that causes one of these two attributes to be read in as qualitative (categorical) when it seems it should be quantitative (numeric)?

It shows that both of the attributes are quantitative.  
  
c) Use the command plot() in R to make a plot for each column by entering plot(data[,1]) and plot(data[,2]). Because one variable is read in as quantitative (numeric) and the other as qualitative (categorical) these two plots are showing completely different things by default. Explain exactly what is being plotted in each of the two cases. Include these two plots in your homework.

For first column,

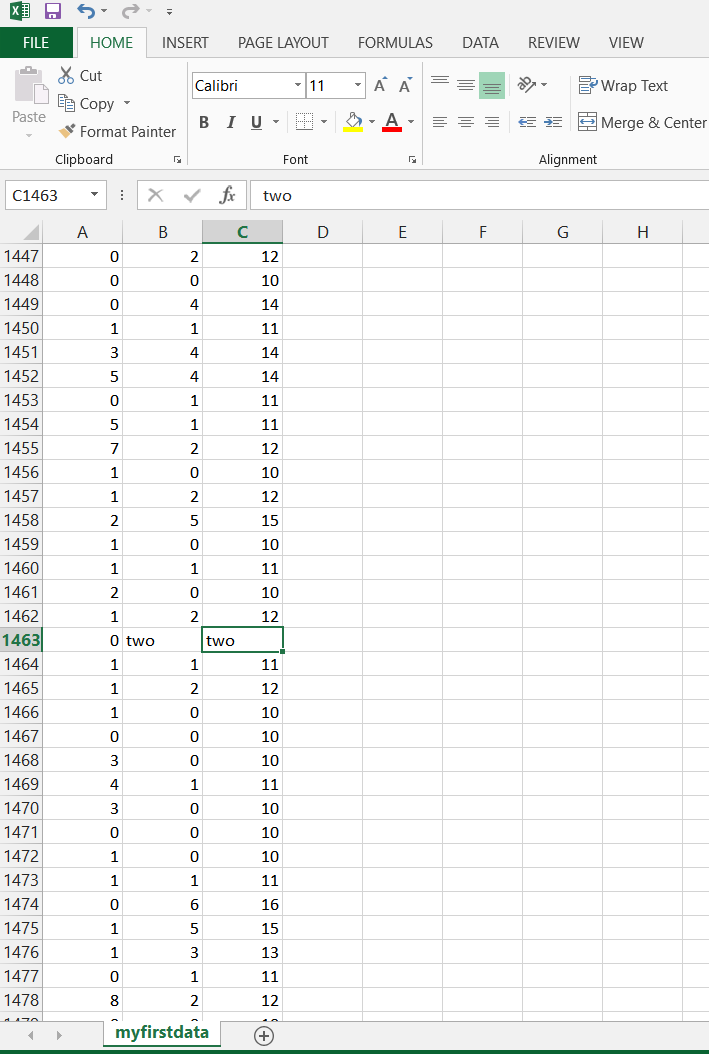
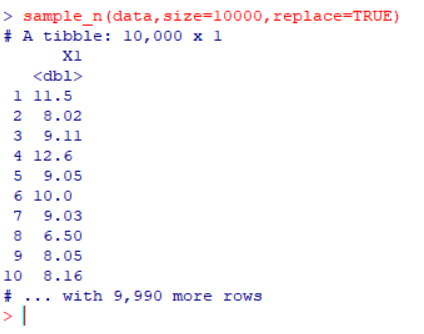


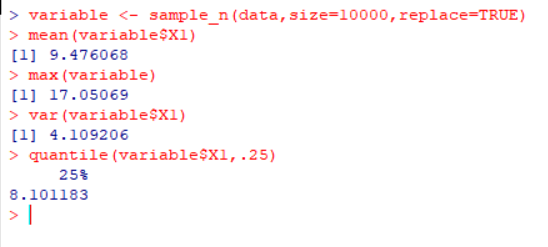
For Second Column,

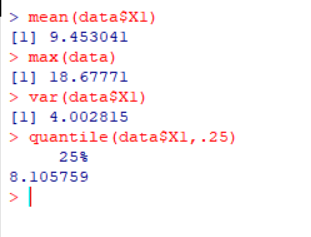
  
  
d) Read the data into Excel. Excel should have no problem opening the file directly since it is .csv. Create a new column that is equal to the second column plus 10. What is the result for the problem observations (rows) you identified in part b? What specific outcome does Excel display?

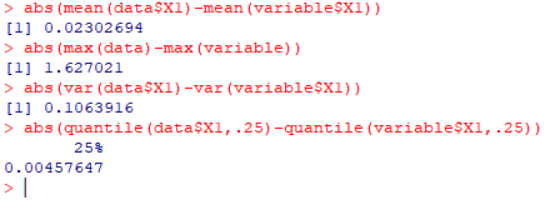
I have created a new column by adding 10 adding to the second column.

It successfully added 10 to all values of 2nd column except for the non-numeric value “two” in 1463 row.

  
  
5) This question uses the data at <http://www.cob.sjsu.edu/mease_d/bus297D/twomillion.csv>. Download it to your computer.  
  
a) Read the data into R using data<-read.csv("twomillion.csv",header=FALSE). Note, you first need to specify your working directory using the setwd() command. Extract a simple random sample with replacement of 10,000 observations (rows). Show your R commands for doing this.  
  
b) For your sample, use the functions mean(), max(), var() and quantile(,.25) to compute the mean, maximum, variance and 1st quartile respectively. Show your R code and the resulting values.

  
  
c) Compute the same quantities in part b on the entire data set and show your answers. How much do they differ from your answers in part b?



  
  
d) Save your sample from R to a csv file using the command write.csv(). Then open this file with Excel and compute the mean, maximum, variance and 1st quartile. Provide the values and name the Excel functions you used to compute these.

AVERAGE(A2:A10001) = 9.476068

MAX(A2:A10001) = 17.05069

VAR(A2:A10001) = 4.109206

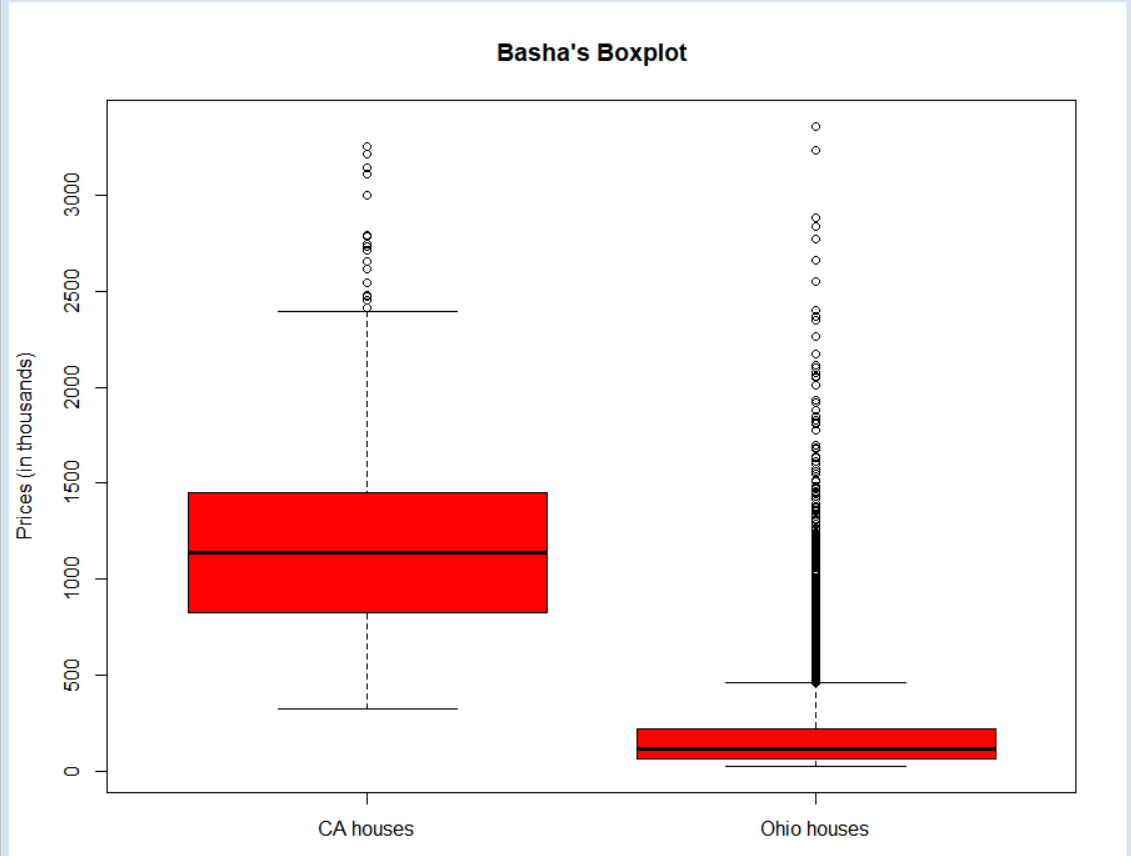
QUARTILE(A2:A10001,1) = 8.101183  
  
e) Exactly what happens if you try to open the full data set with Excel?

It opened the full data set normally but took some time and showed all values.  
  
6) Read Chapter 3 (only sections 3.1, 3.2 and 3.3).  
  
7) This question uses a sample of 1500 California house prices at <http://www-stat.wharton.upenn.edu/~dmease/CA_house_prices.csv> and a sample of 10,000 Ohio house prices at <http://www-stat.wharton.upenn.edu/~dmease/OH_house_prices.csv>. Download both data sets to your computer. Note that the house prices are in thousands of dollars.  
  
a) Use R to produce a single graph displaying a boxplot for each set (as in ICE #16). Include the R commands and the plot. Put your name in the title of the plot (for example, main="Britney Spears' Boxplots").

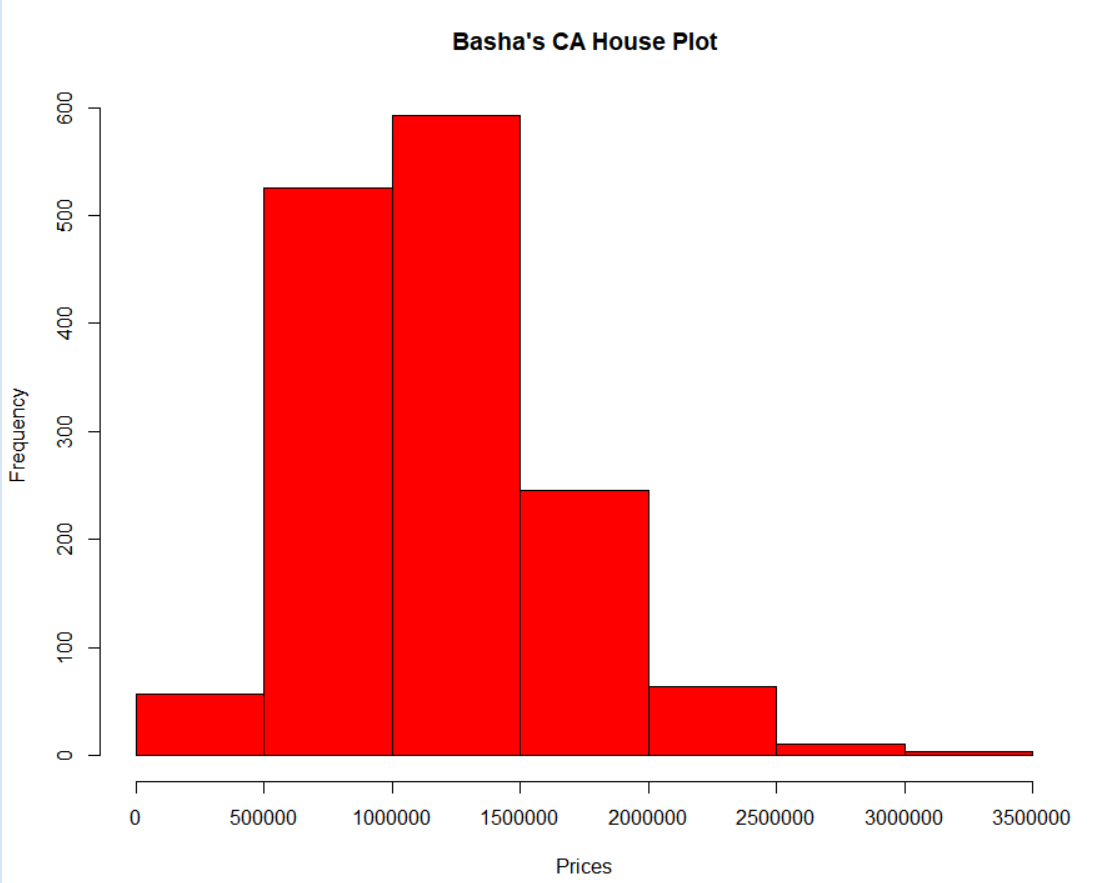
> ca\_data<-read.csv("CA\_house\_prices.csv",header=FALSE)

> oh\_data<-read.csv("OH\_house\_prices.csv",header=FALSE)

> boxplot(ca\_data[,1],oh\_data[,1],col="red",main="Basha's Boxplot",names=c("CA houses","Ohio houses"),ylab="Prices (in thousands)")

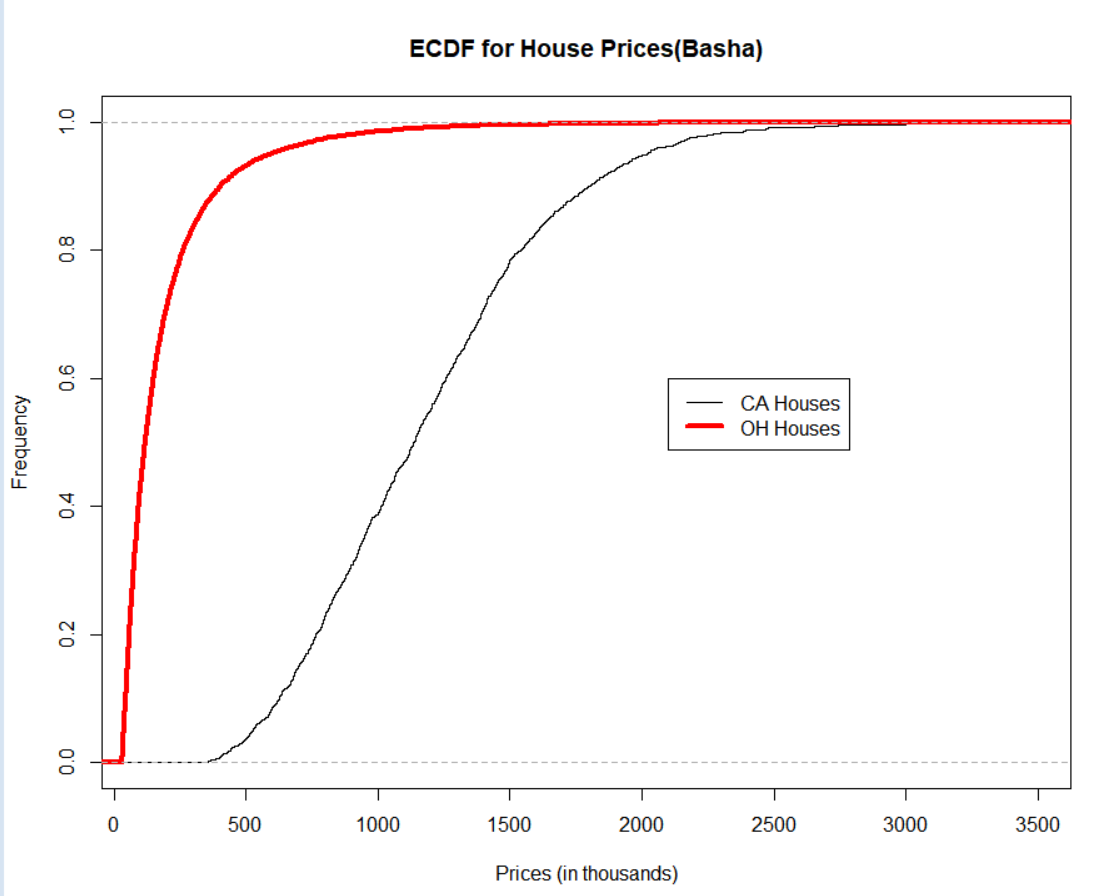
  
  
b) Use R to produce a frequency histogram for only the California house prices. Use intervals of width $500,000 beginning at 0 and ending at $3.5 million. Include the R commands and the plot. Put your name in the title of the plot.

> hist(ca\_data[,1]\*1000,breaks=seq(0,3500000,by=500000),col="red",xlab="Prices",ylab="Frequency",main="Basha's CA House Plot")

  
  
c) Use R to plot the ECDF of the California houses and Ohio houses on the same graph (as in ICE #11). Include a legend. Include the R commands and the plot. Put your name in the title of the plot.  
> plot(ecdf(ca\_data[,1]),verticals= TRUE,do.p = FALSE,main ="ECDF for House Prices(Basha)",xlab="Prices (in thousands)",ylab="Frequency")

> lines(ecdf(oh\_data[,1]),verticals= TRUE,do.p = FALSE,col.h="red",col.v="red",lwd=4)

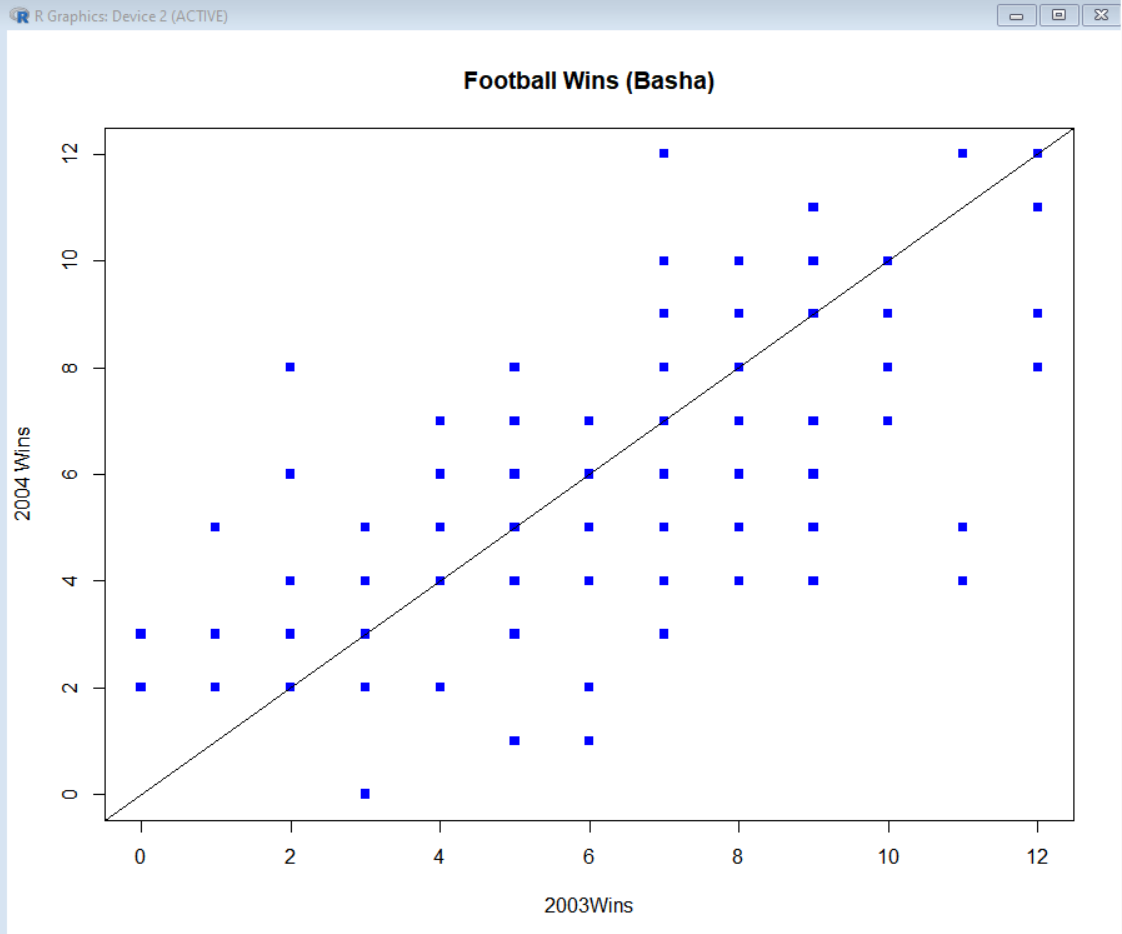
> legend(2100,.6,c("CA Houses","OH Houses"), col=c("black","red"),lwd=c(1,4))

  
8) This question uses the data at <http://www-stat.wharton.upenn.edu/~dmease/football.csv>. Download it to your computer. This data set gives the total number of wins for each of the 117 Division 1A college football teams for the 2003 and 2004 seasons.   
  
a) Use plot() in R to make a scatter plot for this data with 2003 wins on the x-axis and 2004 wins on the y-axis. Use the range 0 to 12 for both the x-axis and y-axis. Include the R commands and the plot. Put your name in the title of the plot.

> football<-read.csv("football.csv", header=TRUE)

> plot(football[,2],football[,3],xlim=c(0,12),ylim=c(0,12),pch=15,col="blue",xlab="2003Wins",ylab="2004 Wins",main="Football Wins (Basha)")

> abline(c(0,1))

  
  
b) Why are there fewer than 117 points visible on your graph in part a? Describe the solution we discussed in class to deal with this problem (but don't actually do it).

Some data are overlapped on one another. Adding a small amount of noise will solve the problem.  
  
c) Compute the correlation in R using the function cor().

> cor(football[,2],football[,3])

[1] 0.6537691  
  
d) How does the value in part c change if you add 10 to all the values for 2004?

> cor(football[,2],football[,3]+10)

[1] 0.6537691

No change  
  
e) How does the value in part c change if you multiply all the 2004 values by 2?  
> cor(football[,2],football[,3]\*2)

[1] 0.6537691

No change

f) How does the value in part c change if you multiply all the 2004 values by -2?  
> cor(football[,2],football[,3]\*-2)

[1] -0.6537691

9) This question uses the sample of 10,000 Ohio house prices at <http://www-stat.wharton.upenn.edu/~dmease/OH_house_prices.csv>. Download the data set to your computer. Note that the house prices are in thousands of dollars.  
  
a) What is the median value? Is it larger or smaller than the mean?

> median(oh\_data[,1])

[1] 118

> mean(oh\_data[,1])

[1] 190.3176  
  
b) What does your answer to part a suggest about the shape of the distribution (right-skewed or left-skewed)?

Data is right-skewed, the mean is greater than the median  
  
c) How does the median change if you add 10 (thousand dollars) to all the values?  
> median(oh\_data[,1]+10)

[1] 128

It increases by 10

d) How does the median change if you multiply all the values by 2?  
> median(oh\_data[,1]\*2)

[1] 236

Doubled

10) This question uses the following people's ages: 19,23,30,30,45,25,24,20. Store them in R using the syntax ages<-c(19,23,30,30,45,25,24,20).  
  
a) Compute the standard deviation in R using the sd() function.  
> ages<-c(19,23,30,30,45,25,24,20)

> sd(ages)

[1] 8.315218

b) Compute the same value by hand and show all the steps.  
list of numbers: 19,23,30,30,45,25,24,20

mean: (19+23+30+30+45+25+24+20) / 8 = 216/ 8 = 27

list of deviations: -8, -4, 3, 3, 18, -2,-3, -7

squares of deviations: 64, 16, 9, 9, 324, 4, 9, 49

sum of deviations: 64+16+9+9+324+4+9+49 = 484

divided by one less than the number of items in the list: 484 / 7 = 69.14285

square root of this number: square root (69.14285) = **about 8.31521**

c) Using R, how does the value in part a change if you add 10 to all the values?

> sd(ages+10)

[1] 8.315218

No Change  
  
d) Using R, how does the value in part a change if you multiply all the values by 100?

> sd(ages\*10)

[1] 83.15218

Multiplied by 10