[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](#) on page 89.

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

(a) Time in terms of AM or PM.

Binary, qualitative, nominal (most people consider binary attributes to be nominal)

(b) Brightness as measured by a light meter.

Continuous, quantitative, ratio

(c) Brightness as measured by people's judgments.

Discrete, qualitative, ordinal (assuming we make them choose from a discrete set of ratings)

(d) Angles as measured in degrees between 0◦ and 360◦.

Continuous, quantitative, ratio

(e) Bronze, Silver, and Gold medals as awarded at the Olympics.

Discrete, qualitative, ordinal

(f) Height above sea level.

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

(g) Number of patients in a hospital.

Discrete, quantitative, ratio

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

Discrete, qualitative, nominal (but ISBN numbers do have some order information so it could be ordinal if you use that information)

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

Discrete, qualitative, ordinal

(j) Military rank.

Discrete, qualitative, ordinal

(k) Distance from the center of campus.

Continuous, quantitative, interval/ratio (depends)

(l) Density of a substance in grams per cubic centimeter.

Continuous, 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 (or ordinal if you are using the order information)

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.

➢   After reading the data, I used the *R* commands *is.factor* and *is.numeric* to determine that **column 1 is quantitative and column 2 is qualitative.**

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)?

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.

➢ The plot for column 1 shows the row numbers on the x axis and the column 1 values on the y axis. A point is drawn for each row.
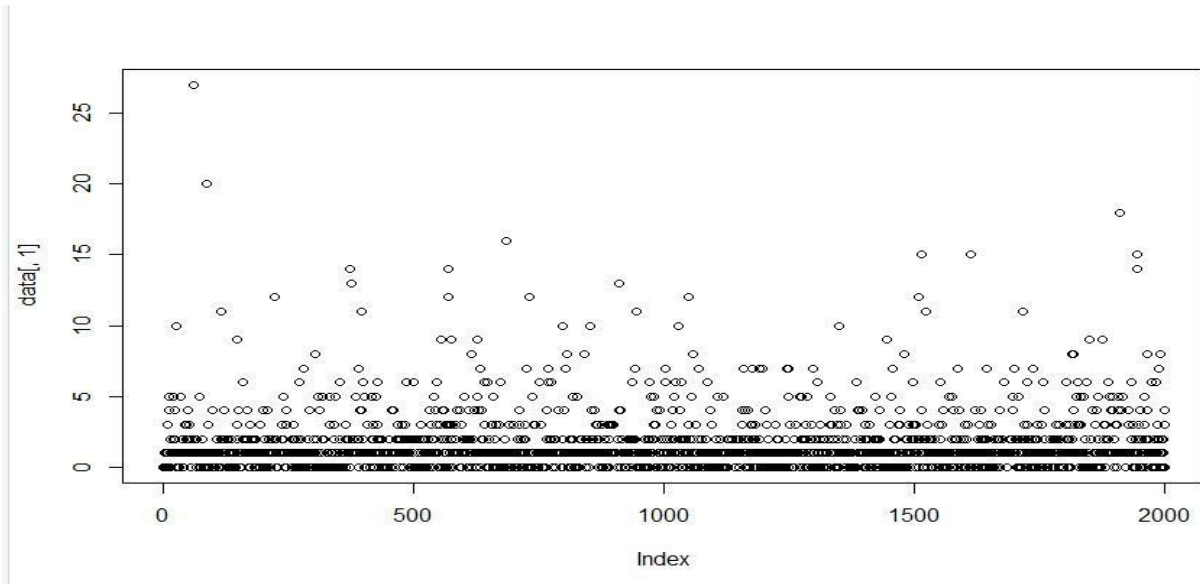
➢ The plot for column 2 shows logical data representation in the form of bar graph.
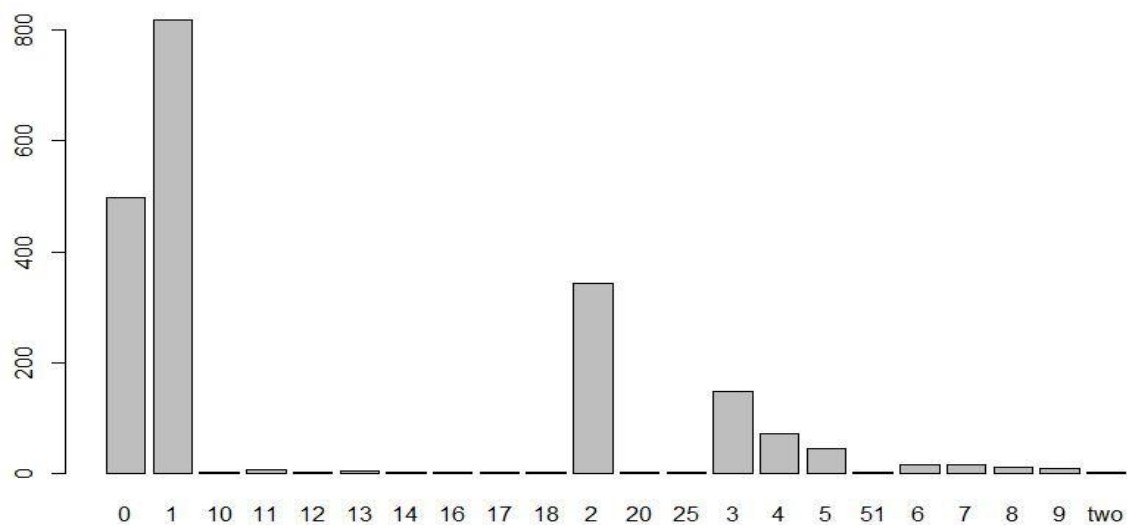
d) Read the data into Excel. Excel should have no problem opening the file directly since it is .csv. Create a new column that is equal to the second column plus 10. What is the result for the problem observations (rows) you identified in part b? What specific outcome does Excel display?

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.

```
> data <- read.csv("twomillion.csv",header=FALSE)
>
> sample_data <- sample(data[,1],10000,replace=TRUE)
> head(sample_data)
[1]  8.789100 12.076136 10.657138  7.513992  8.644192  9.539759
>
```

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.

```
> mean(sample_data)
[1] 9.431632
> max(sample_data)
[1] 16.65478
> var(sample_data)
[1] 4.042824
> quantile(sample_data, .25)
     25%
8.084657
>
```

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?

```
> mean(data$V1)
[1] 9.453041
> max(data$V1)
[1] 18.67771
> var(data$V1)
[1] 4.002815
> quantile(data$V1)
      0%       25%       50%       75%      100%
-0.111507  8.105759  9.453331 10.803445 18.677712
>
```

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.

```
write.csv(sample_data, file = "sample_data.csv")
```

e) Exactly what happens if you try to open the full data set with Excel?

```
> str(data)
'data.frame':   1048576 obs. of  1 variable:
 $ V1: num  9.03 11.22 7.39 7.22 10.42 ...
>
```

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 dollar

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").

Shireesha's Graph
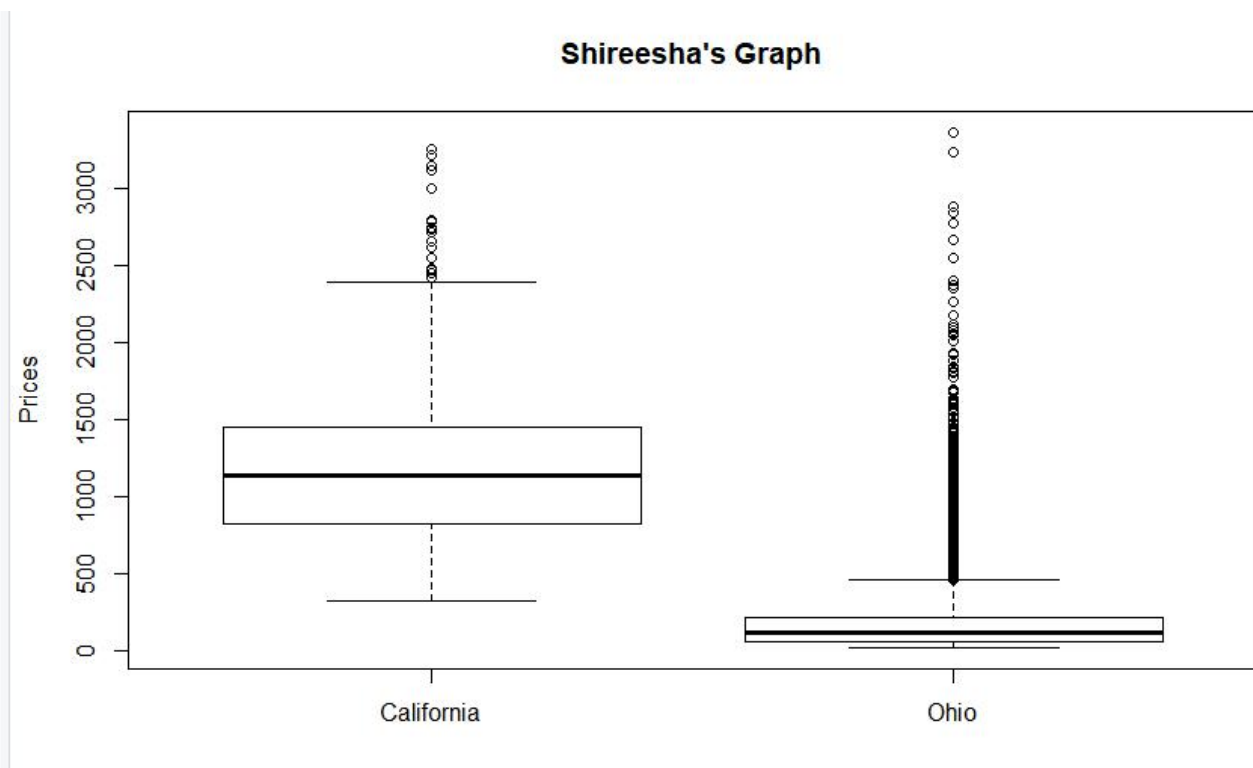
b) Use R to produce a frequency histogram for only the California house prices. Use intervals of width $500,000 beginning at 0 and ending at $3.5 million. Include the R commands and the plot. Put your name in the title of the plot
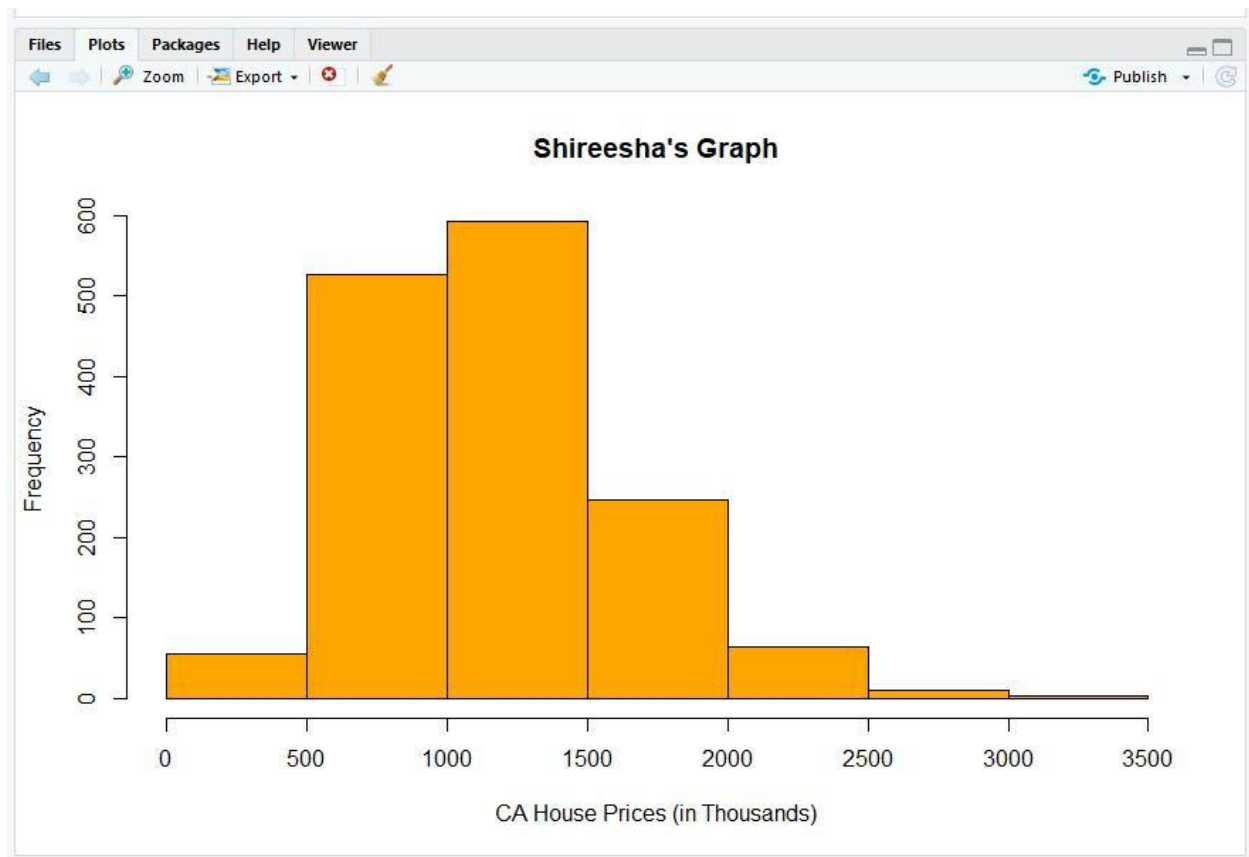
c) Use R to plot the ECDF of the California houses and Ohio houses on the same graph (as in ICE #11). Include a legend. Include the R commands and the plot. Put your name in the title of the plot.
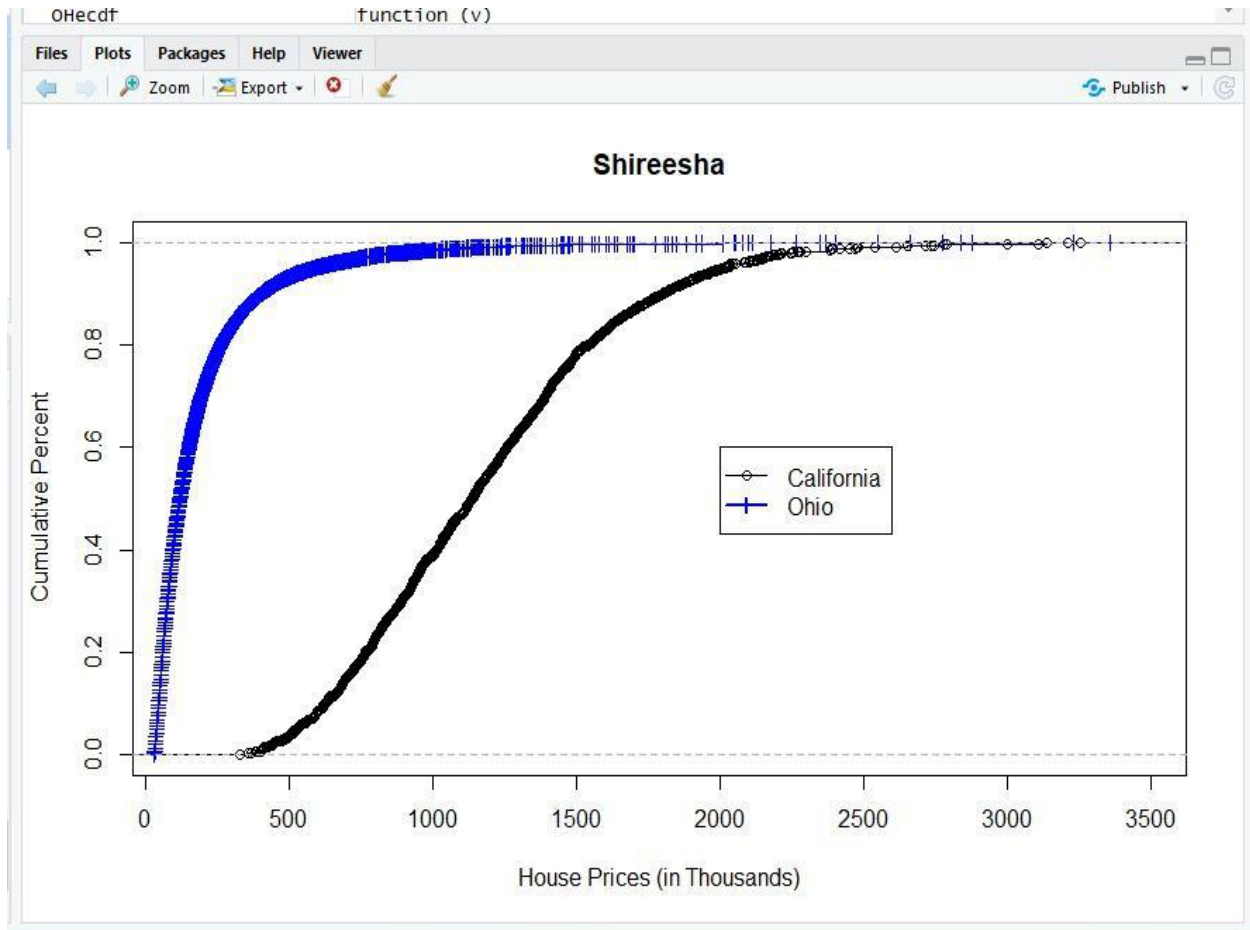
8) This question uses the data at http://www-stat.wharton.upenn.edu/~dmease/football.csv. Download it to your computer. This data set gives the total number of wins for each of the 117 Division 1A college football teams for the 2003 and 2004 seasons.

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.

Shireesha

2003 Wins
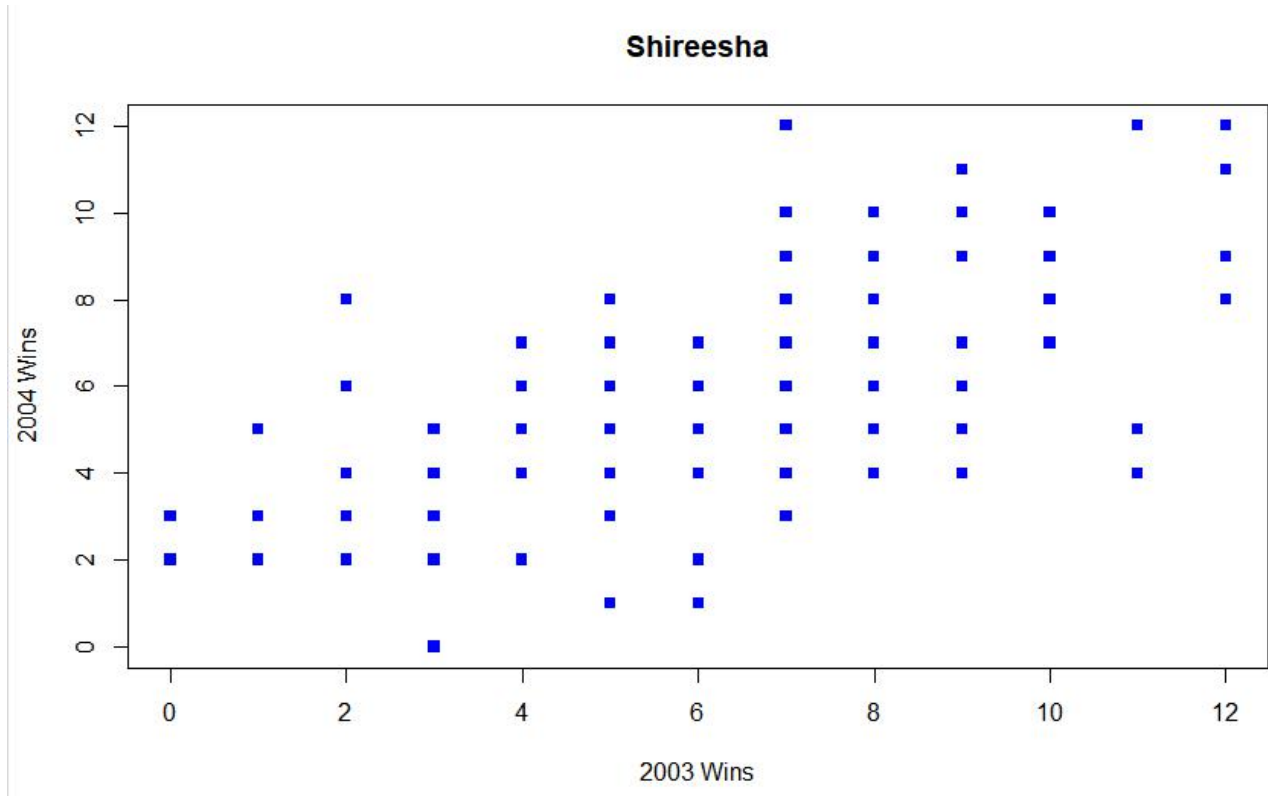
b) Why are there fewer than 117 points visible on your graph in part a? Describe the solution we discussed in class to deal with this problem (but don't actually do it)

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
>
```

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
>
```

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?

```
> data<-read.csv("OH_house_prices.csv", header=FALSE)
>
> median(data[,1])
[1] 118
>
> mean(data[,1])
[1] 190.3176
`
```

Mean is larger than median.

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

c) How does the median change if you add 10 (thousand dollars) to all the values?

```
> median(data[,1]+10)
[1] 128
>
```

d) How does the median change if you multiply all the values by 2?

```
> median(2*data[,1])
[1] 236
>
```

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.

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
>
```

d) Using R, how does the value in part a change if you multiply all the values by 100?

```
> sd(100*ages)
[1] 831.5218
>
```