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.

Class Exercises #1

Find a different definition of data mining online. How does it compare to the one in the text on the previous slide?

Data Mining: The practice of analyzing large databases in order to generate new information. (Oxford dictionary)

Data mining: is the process of automatically discovering useful information in large data repositories.

Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. (Wikipedia)

Class Exercise #2

Give an example of something you did yesterday or today which resulted in data which could potentially be mined to discover useful information.

Cell phone

Laptop

Internet Surfing

Grocery Shopping

3) Do Chapter 2 textbook problem #2 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 more ability.

(a) Time in terms of AM or PM (binary, qualitative, ordinal)

(b) Brightness as measured by a light meter (continuous, quantitative, ratio)

(c) Brightness as measured by people's judgements (discrete, qualitative, ordinal)

(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)

(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)

(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)

(l) Density of substance in grams per cubic centimeter (discrete, quantitative, ratio)

(m) Cost check number. (When you attend an event, you can often give your cost to someone who, in turn, gives you a number that you can use to claim your cost 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/factor) or quantitative (numeric) using R. Explain how you can tell using R.

Solution:

```
> setwd("/Users/pemayangdon/Desktop/DS/Specialization/DataMining/data Mining
Assignments/DM Assignment2")
> data <- read.csv("myfirstdata.csv", header=FALSE)
>
```

```
>
> class(data[,1])
[1] "integer"
> class(data[,2])
[1] "character"
```

The first attribute is considered as quantitative because its type of attribute is integer. Whereas the second attribute is considered as qualitative as its attribute type is character.

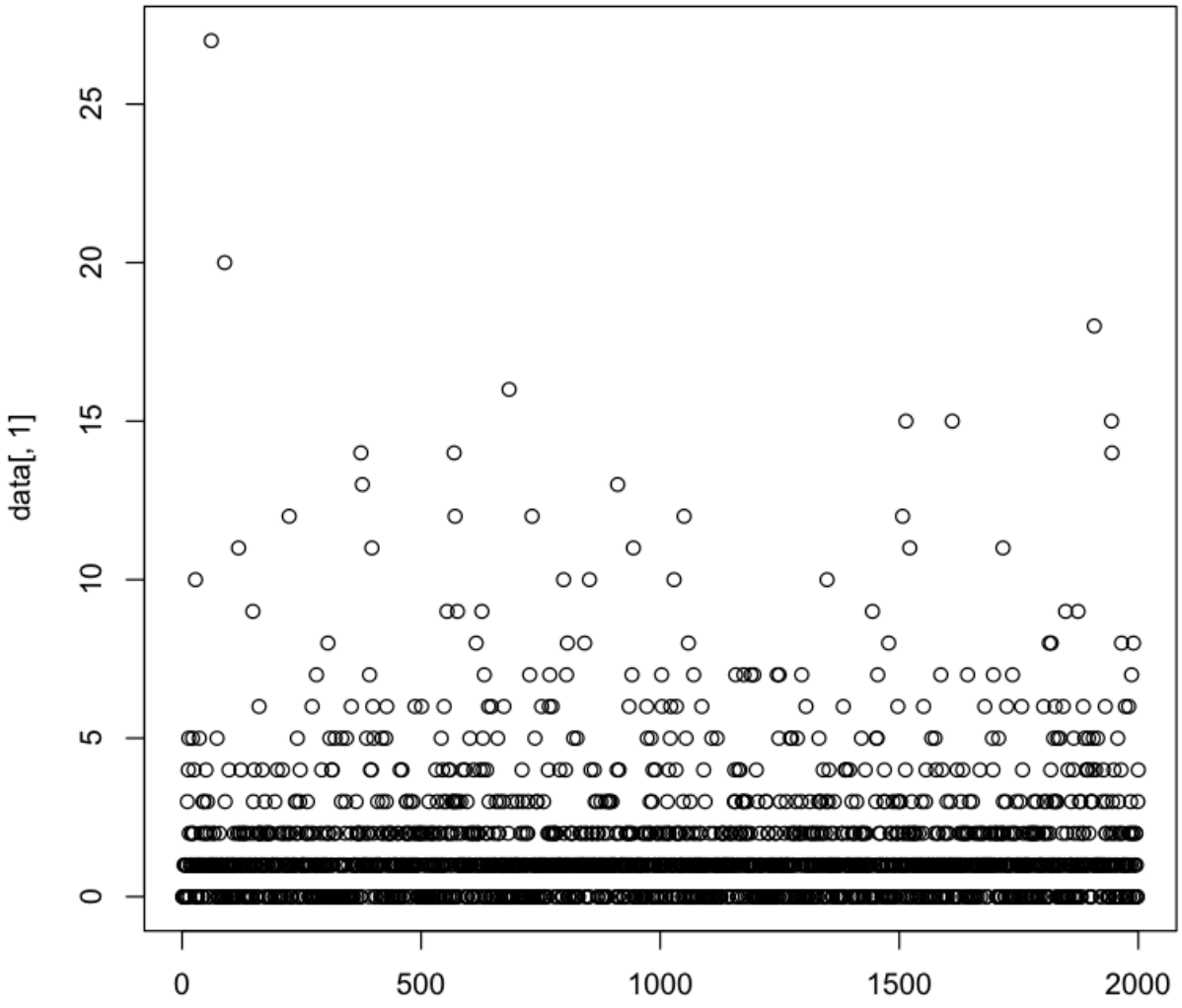
b) What is the specific problem that causes one of these two attributes to be read in as qualitative (categorical) when it seems it should be quantitative (numeric)?

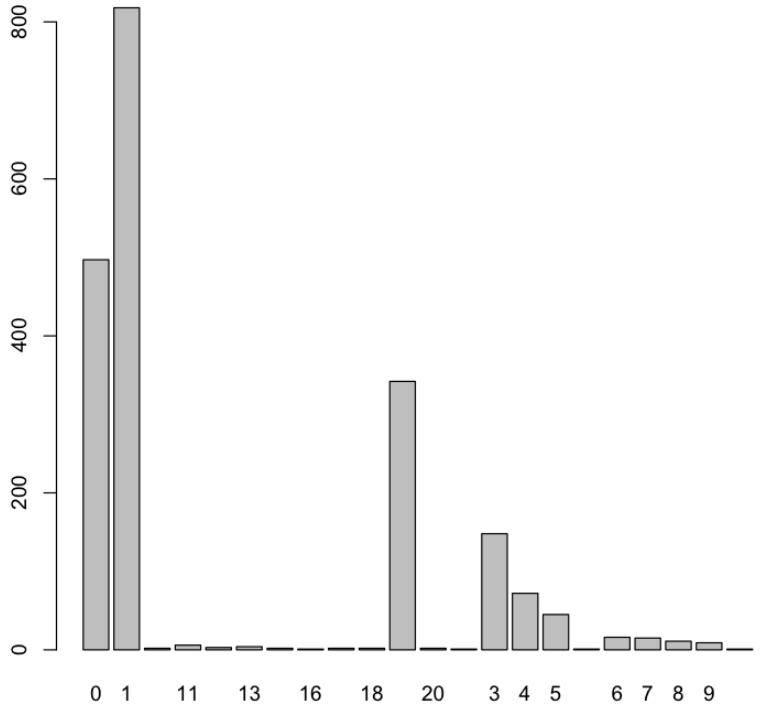
```
> data[,2] + 10
Error in data[, 2] + 10 : non-numeric argument to binary operator
```

The problem is that the second column have some non-numeric value, because of which it throws an error when numeric operations were applied on this type of data.

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 which shows the distribution of values of each row in 1st column. X axis represents the index number, and whereas the Y axis as the value of the row at that index.





The plot above depicts a histogram which represents the frequency of the values containing in 2nd column's rows. X axis is for the values, and the Y axis is the number of occurrences of that value in rows.

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?

1462	1	2	12
1463	0	two	#VALUE!
1464	1	1	11
1465	1	2	12

After adding a value 10 on the second column, one of the results is a value error, as the value is non-numeric and performing numeric operations on categorical is not meaningful.

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.

```
> setwd("/Users/pemayangdon/Desktop/DS/Specialization/DataMining/data Mining
Assignments/DM Assignment2")
> data <- read.csv("twomillion.csv", header=FALSE)
>

>
> samp <- sample(data[,1], 10000, replace=TRUE)
>
```

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(samp)
[1] 9.432448
> max(samp)
[1] 17.03243
> var(samp)
[1] 3.971426
> quantile(samp,.25)
     25%
8.090964
```

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[,1])
[1] 9.453041
> max(data[,1])
[1] 18.67771
> var(data[,1])
[1] 4.002815
> quantile(data[,1],.25)
     25%
8.105759
```

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(samp, "sample.csv")
```

Average	9.43244773
max	17.0324334
var	3.97142618
quartile	8.09096438

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

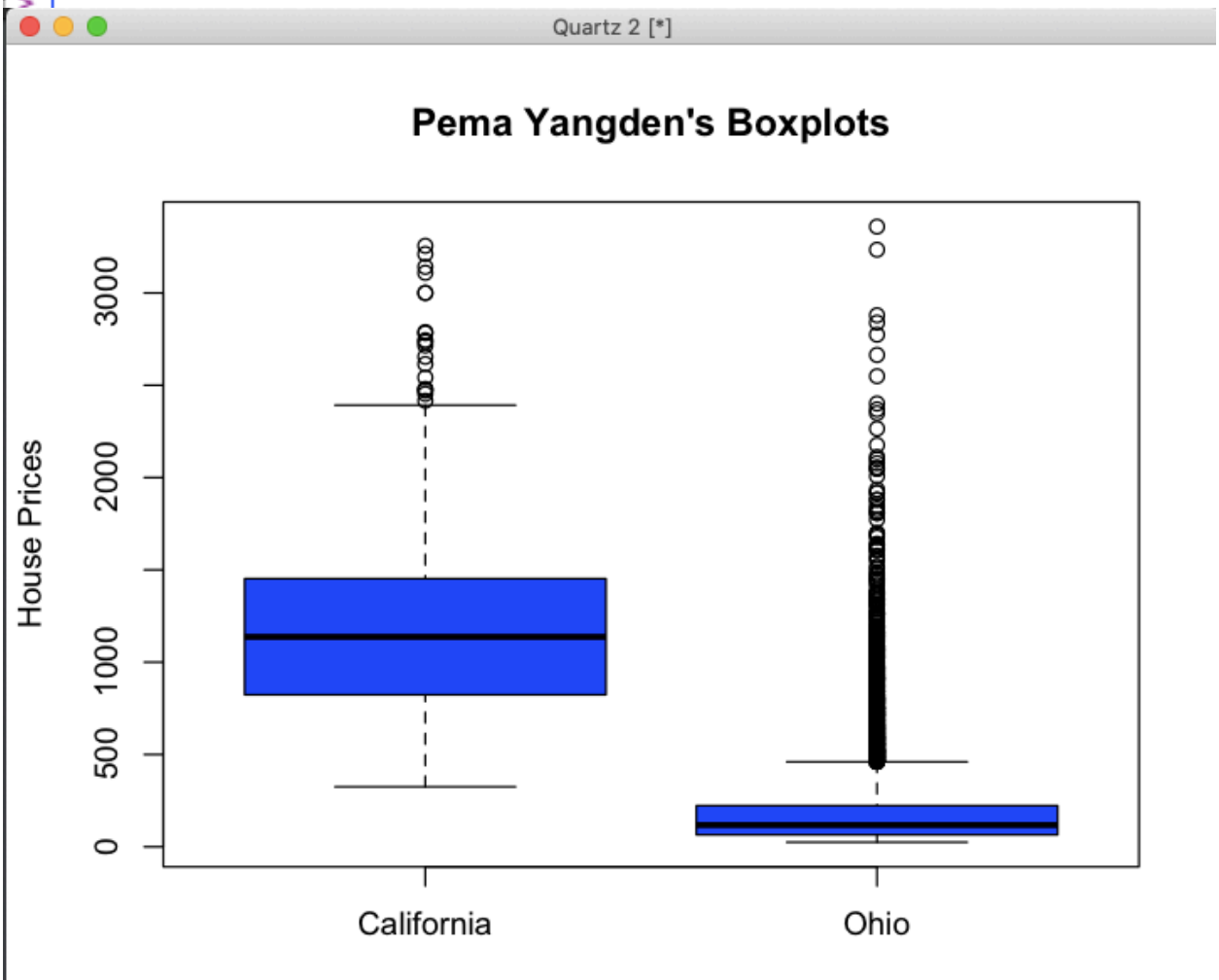
When I try to open the complete data set with excel, it shows the entire data i.e 1048576 rows, whereas in R, it displays up to 99999 rows, as it has some limits on the display of data.

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").

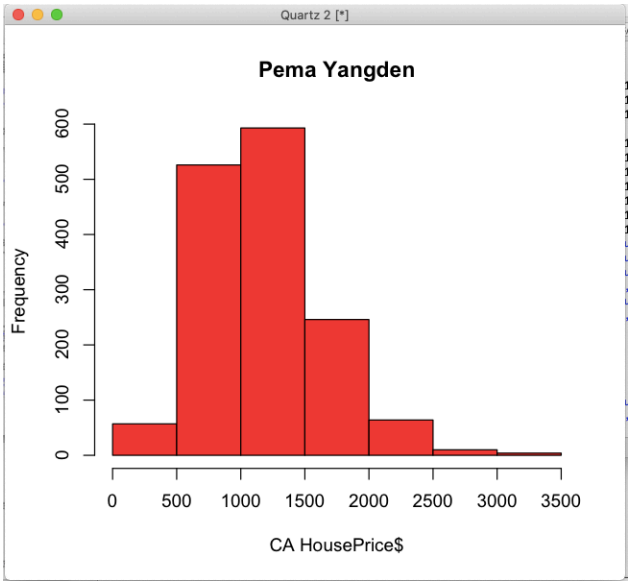
```
> boxplot(CA_house_prices[,1],OH_house_price[,1],col="blue",
+ main="Pema Yangden's Boxplots",
+ names=c("California","Ohio"),ylab="House Prices")
> |
```



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.

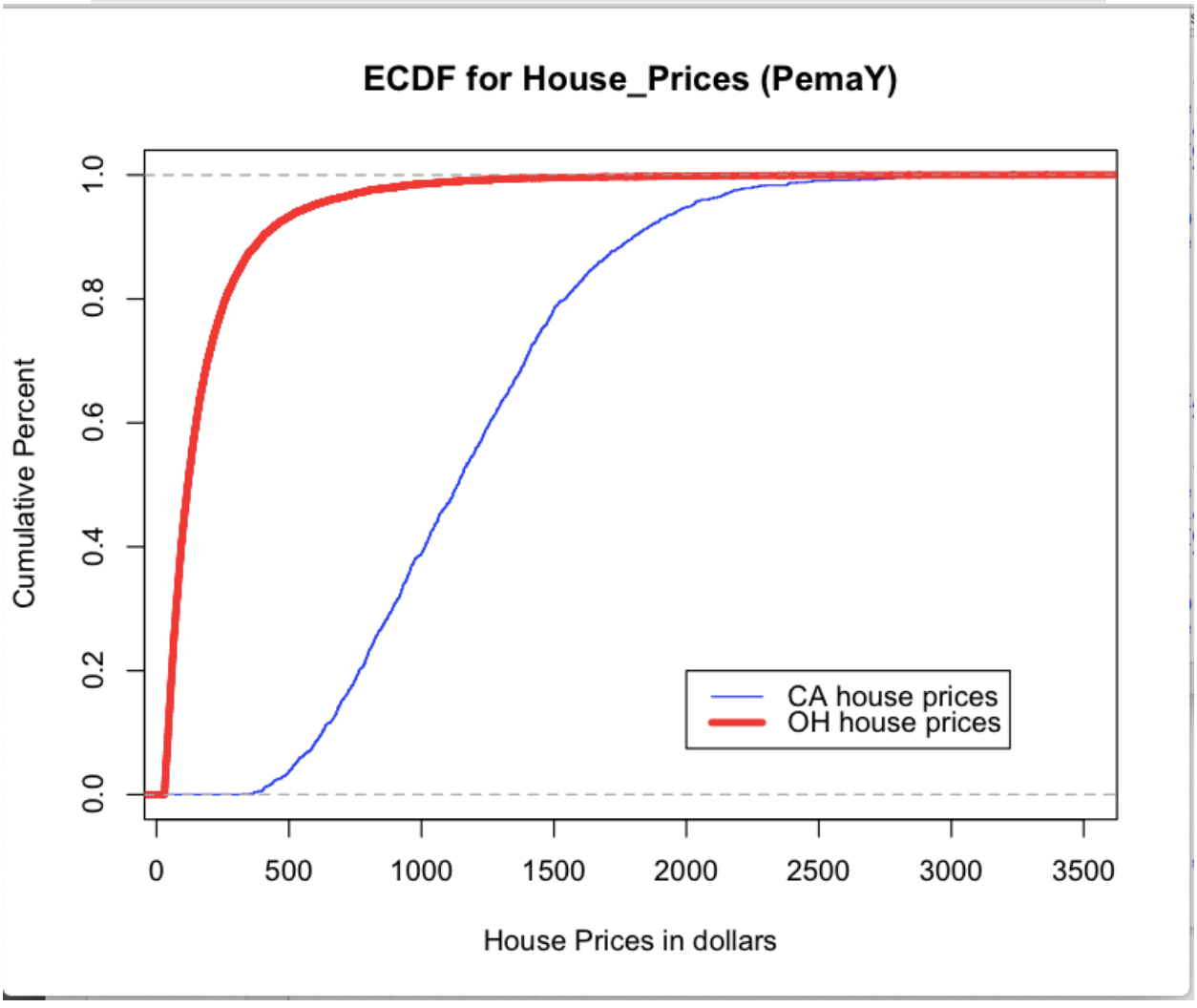
```
> setwd("/Users/pemayangdon/Desktop/DS/Specialization/2. DataMining/data Mining
Assignments/DM Assignment2")
>
> CA_house_prices <- read.csv("CA_house_prices.csv", header=FALSE)
>

> hist(CA_house_prices[,1],breaks=seq(0,3500,by=500), col="red", xlab="CA
HousePrice$", ylab="Frequency",main="Pema Yangden")
>
```



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.

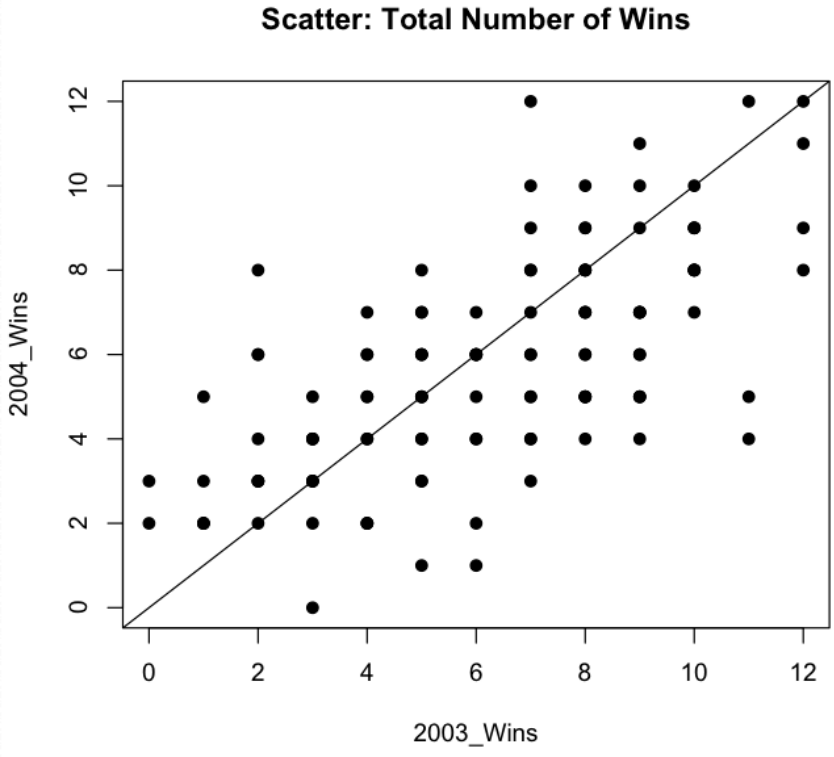
```
> plot(ecdf(CA_house_prices[,1]),
+ verticals= TRUE,
+ do.p=FALSE,
+ main="ECDF for House_Prices (PemaY)",
+ xlab="House Prices in dollars",
+ ylab="Cumulative Percent", col="blue")
> lines(ecdf(OH_house_price[,1]),
+ verticals= TRUE,do.p = FALSE,
+ col.h="red",col.v="red",lwd=4)
> legend(2000,.2,c("CA house prices","OH house prices"),
+ col=c("blue","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.

```
> plot(football$X2003.Wins,football$X2004.Wins, main="Scatter: Total Number of
Wins", xlim=c(0,12), ylim=c(0,12),pch=19, xlab="2003_Wins", ylab="2004_Wins")
> plot(football$X2003.Wins,football$X2004.Wins, main="Scatter: Total Number of
Wins", xlim=c(0,12), ylim=c(0,12),pch=19, xlab="2003_Wins", ylab="2004_Wins")
> 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).

Because when both variables are discrete, many points in a scatter plot may be plotted over top of one another, which tends to skew the relationship. A solution is to add a small amount of random noise to the points so that they are jittered a little bit. Or when there are too many points to display on a scatter plot, sampling may also be efficient.

c) Compute the correlation in R using the function cor().

```
> cor(football$X2003.Wins, football$X2004.Wins)
[1] 0.6537691
```

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

```
> cor(football$X2003.Wins, football$X2004.Wins+10)
[1] 0.6537691
```

e) How does the value in part c change if you multiply all the 2004 values by 2?

```
> cor(football$X2003.Wins, football$X2004.Wins * 2)
[1] 0.6537691
```

f) How does the value in part c change if you multiply all the 2004 values by -2?

```
> cor(football$X2003.Wins, football$X2004.Wins * -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?

```
> mean(OH_house_price[,1])
[1] 190.3176
> median(OH_house_price[,1])
[1] 118
```

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

The distribution of data is right skewed since the mean is greater than the median value.

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

```
> mean(OH_house_price[,1]+10)
[1] 200.3176
> median(OH_house_price[,1]+10)
[1] 128
```

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

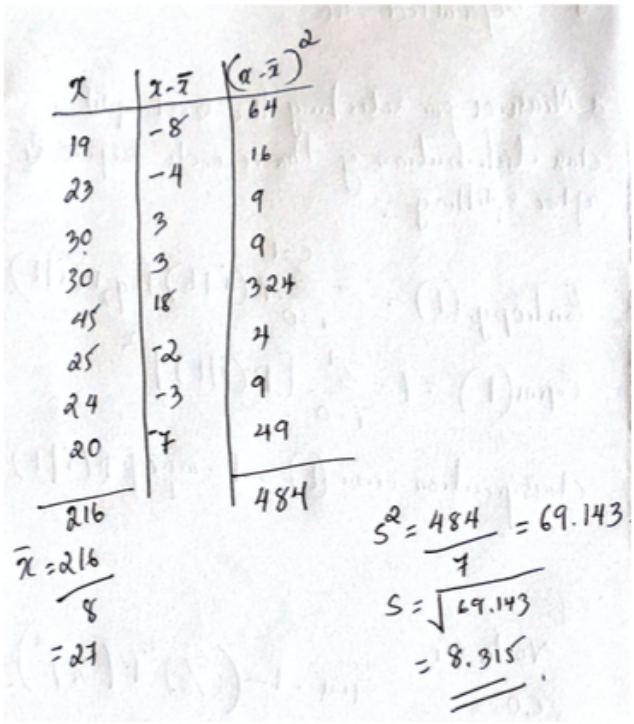
```
> mean(OH_house_price[,1] * 2)
[1] 380.6352
> median(OH_house_price[,1] * 2)
[1] 236
```

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.



x	$x-\bar{x}$	$(x-\bar{x})^2$
19	-8	64
23	-4	16
30	3	9
30	3	9
45	18	324
25	-2	4
24	-3	9
20	-7	49
216		484

$\bar{x} = \frac{216}{8} = 27$

$s^2 = \frac{484}{7} = 69.143$

$s = \sqrt{69.143} = 8.315$

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

```
> new_ages <- ages+10
> new_ages
[1] 29 33 40 40 55 35 34 30
> sd(new_ages)
[1] 8.315218
```

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

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