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.

Q. 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

a) Time in terms of AM or PM.
   Binary, qualitative, nominal
b) Brightness as measured by a light meter.
   Continuous, quantitative, ratio
c) Brightness as measured by people's judgments.
   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 centre of campus.
   Continuous, quantitative, ratio
l) Density of a substance in grams per cubic centimetre.
   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

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.

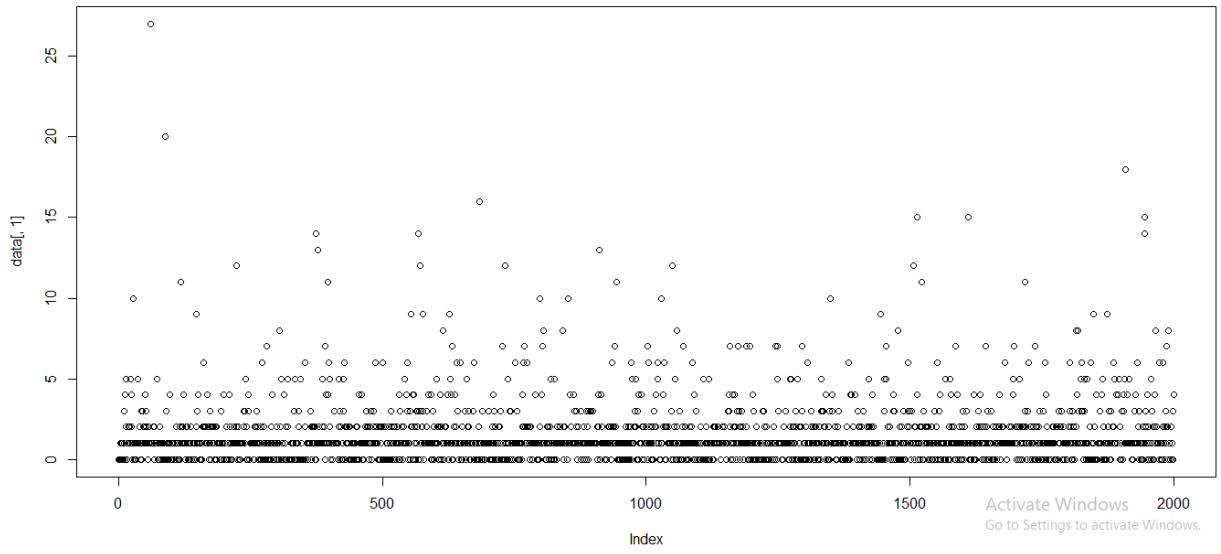
```
> setwd("F:\\Data Science\\DataScience_2019501111\\Data Mining\\DM Assignment2\\")
> data<-read.csv("myfirstdata.csv",header=FALSE)
> head(data)
  V1 V2
1  0  0
2  0  3
3  0  1
4  1  2
5  0  0
6  1  2
> c(is.factor(data[, 1]), is.numeric(data[, 1]))
[1] FALSE  TRUE
> c(is.factor(data[, 2]), is.numeric(data[, 2]))
[1] FALSE FALSE
>
```

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)?

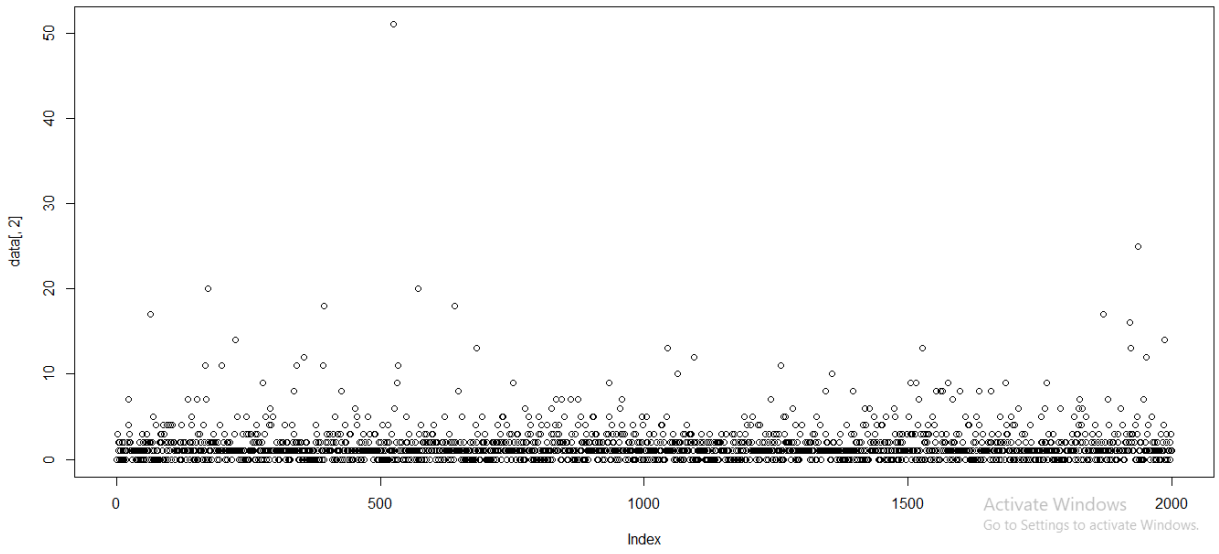
```
> data <- read.csv("myfirstdata.csv", header = F)
> which.nonnumeric <- function (column) {
+ which(is.na(suppressWarnings(as.numeric(as.character(column)))))
+ }
> for (name in names(data)) {
+  c <- data[[name]]
+  r <- which.nonnumeric(c)
+  v <- c[r]
+ msg <- ''
+ if (length(v)) {
+ msg <- sprintf("data$%s is qualitative (%s[%d] == '%s')", name, name, r, as.character(v))
+ } else {
+ msg <- sprintf("data$%s is quantitive (all rows are numeric)", name)
+ }
+ print(msg)
+ }
[1] "data$V1 is quantitive (all rows are numeric)"
[1] "data$V2 is qualitative (V2[1463] == 'two')"
>
```

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.

```
> plot(data[, 1])
>
```



```
> plot(data[, 2])
Warning message:
In xy.coords(x, y, xlabel, ylabel, log) : NAs introduced by coercion
>
```



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 has the same interpretation.

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?

- We get an error on the row 1463 because we are trying to add string 'two' which is in second column and number 10 which is not possible.

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("F:\\Data Science\\DataScience_2019501111\\Data Mining\\DM Assignment2\\")
> data<-read.csv("twomillion.csv",header=FALSE)
> sam<-sample(seq(1,length(data[,1])), 10000, replace=T)
>
```

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.

```
> my_sample<-data[sam,1]
> mean(my_sample)
[1] 9.429238
> max(my_sample)
[1] 18.28539
> var(my_sample)
[1] 4.006404
> quantile(my_sample,.25)
     25%
8.087002
>
```

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
> #Differ....
> abs(mean(my_sample)-mean(data[,1]))
[1] 0.02380335
> abs(max(my_sample)-max(data[,1]))
[1] 0.3923254
> abs(var(my_sample)-var(data[,1]))
[1] 0.003589476
> abs(quantile(my_sample,.25)-quantile((data[,1]),.25))
       25%
0.01875719
>
```

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(my_sample,"my_sample.csv")
>
```

AVG	9.429238	
MAX	18.28539	
VAR	4.006404	
QUARTILE	8.087002	

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

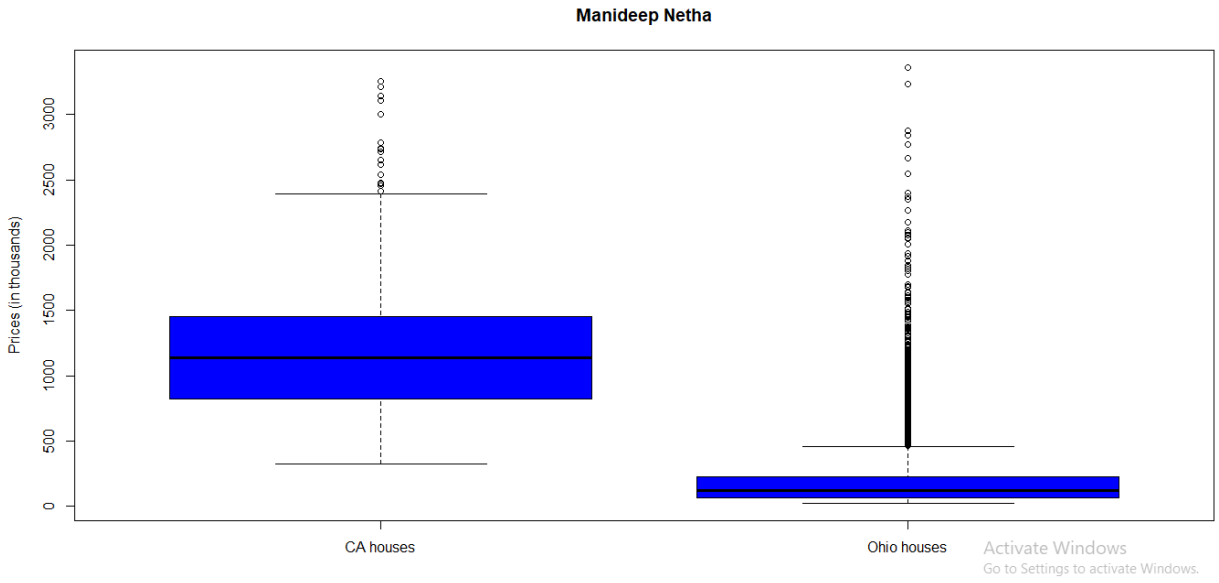
A) It displays total of 1048576 rows

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.

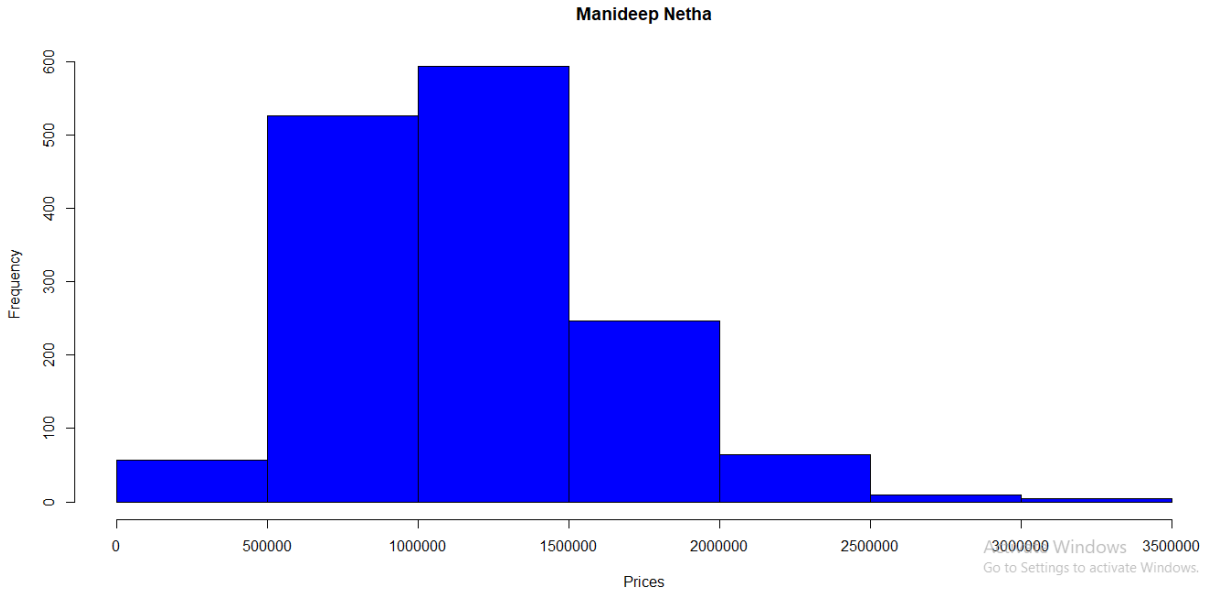
a) Use R to produce a single graph displaying a boxplot for each set (as in ICE #16). Include the R commands and the plot. Put your name in the title of the plot (for example, main="Britney Spears' Boxplots").

```
> 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="blue",main="Manideep Netha",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.

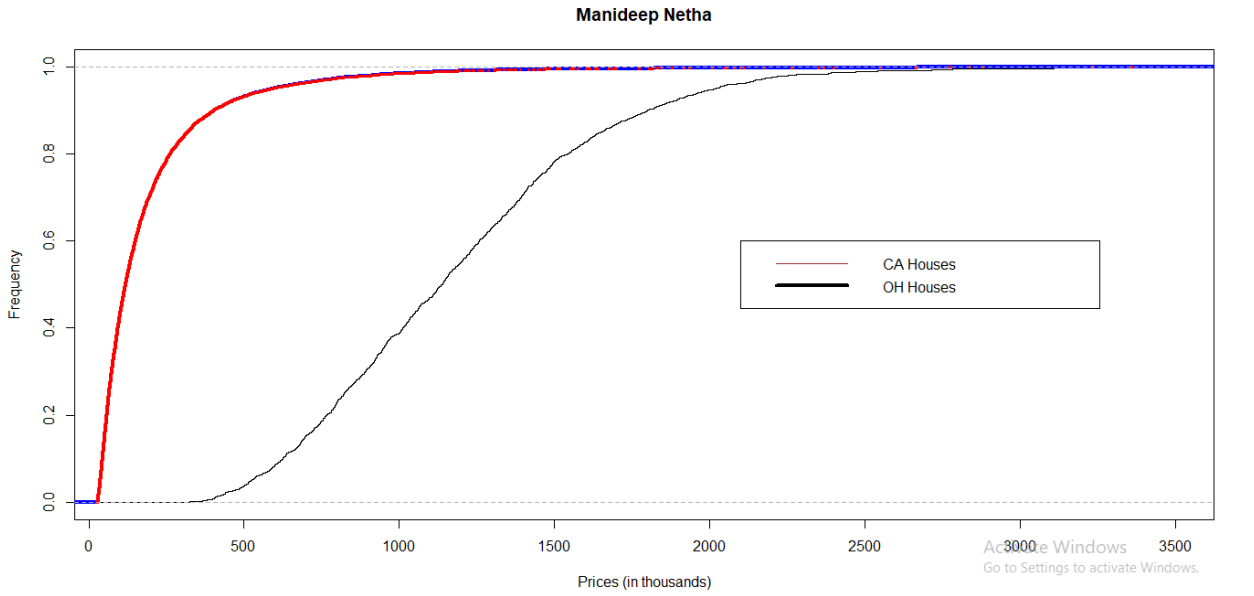
```
> hist(ca_data[,1]*1000,breaks=seq(0,3500000,by=500000),col="blue",xlab="Prices",ylab="Frequency",main="Manideep Netha")
>
```



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.

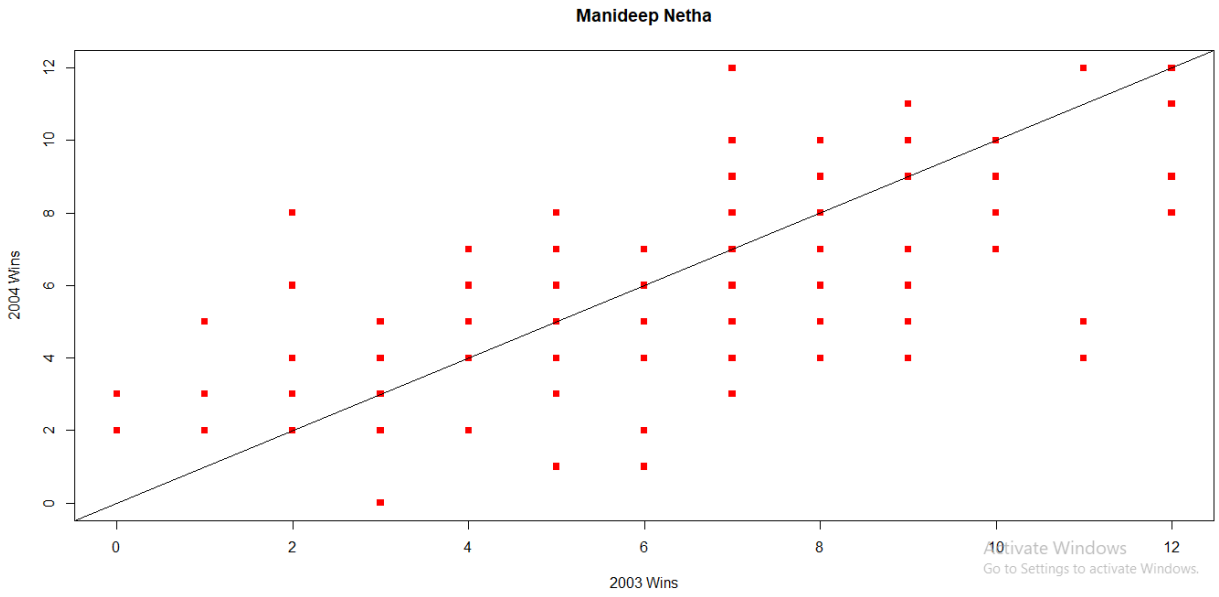
```
> plot(ecdf(ca_data[,1]),verticals= TRUE,do.p = FALSE,main ="Manideep Netha",xlab="Prices (in thousands)",ylab="Frequency")
> lines(ecdf(oh_data[,1]),verticals= TRUE,do.p = FALSE,col.h="blue",col.v="red",lwd=4)
> legend(2100,.6,c("CA Houses","OH Houses"), col=c("brown","black"),lwd=c(1,4))
>
```



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="red",xlab="2003 Wins",ylab="2004 Wins",main="Manideep Netha")
> 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).

A) The solution is to add a small amount of noise to the points because some information is plotted on the same set of axes and is not visible because they have been plotted on top of each other.

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
>
```

Differences in sign

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
```

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)?

Mean is greater than median so data is right skewed.

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

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

The new median is increased by 10

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

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

new median value is double of the old median value

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.

A) List of numbers: 19,23,30,30,45,25,24,20
B) Mean: (19+23+30+30+45+25+24+20) / 8 = 216/ 8 = 27
C) List of deviations: -8, -4, 3, 3, 18, -2,-3, -7
D) Squares of deviations: 64, 16, 9, 9, 324, 4, 9, 49
E) Sum of deviations: 64+16+9+9+324+4+9+49 = 484
F) Divided by one less than the number of items in the list: 484 / 7 = 69.14285
G) 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 changes

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
>
```

10 times the old median value