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.  
  
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**.

Answer: At the first glance when we see the data after reading the csv, we can predict that both the columns are numerical values. However, since there are a lot of rows (2000), we cannot keep checking each and every ow if the value is numeric or not. There might be non-numeric values too in the data.

Columns with all numeric values can be treated as numeric(quantitative) and other as categorical(qualitative). To check that, we have couple of methods in R.

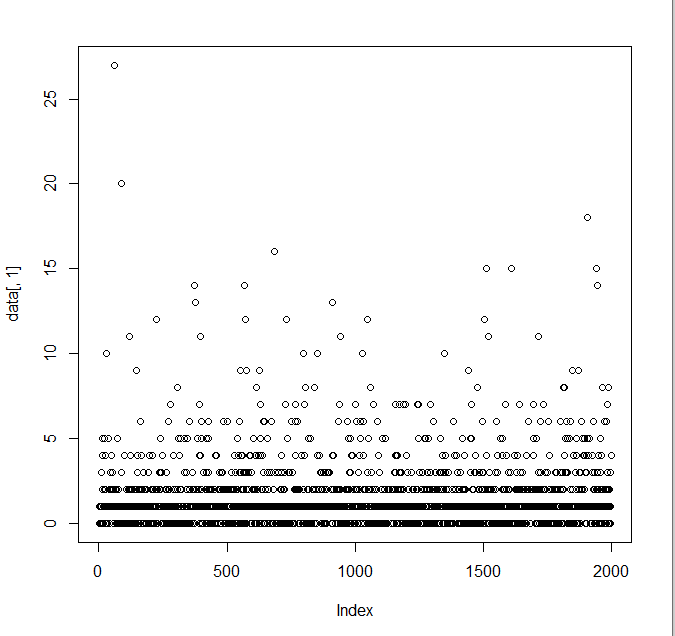
1. **str(data)** => This returns the number of observations, no of variables and data type of each column.
2. **sapply(data, class)** => This returns the class of each variable

Using sapply for the above read data, we get data type of V1 as integer and V2 as character, which confirms that V1 is numeric and V2 is categorical and has string type values in it.

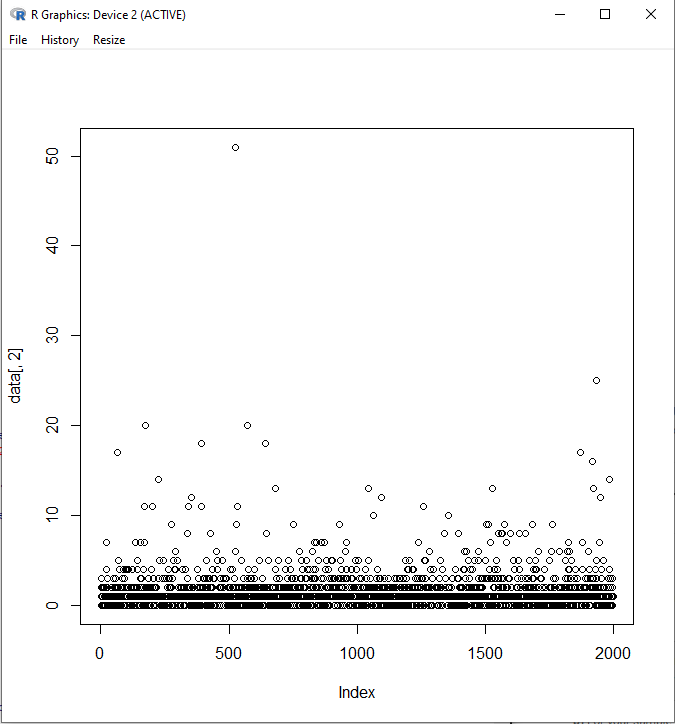
**b) What is the specific problem that causes one of these two attributes to be read in as qualitative (categorical) when it seems it should be quantitative (numeric)?**

Answer: When we use sapply function, it returns the data type of column as character, from which we can depict that there are some string values instead of numerical values in the 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.**  
Answer:

plot(data[,1]) gives below scattered plot. It shows values of V1 column on y axis and index values on x axis.



plot(data[,2]) also gives a scattered plot as below.



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?

Answer: In column tow, row 1463, there is a string value “two”. when adding 10 to that string value, excel gives a “error in value” error and displays “#VALUE!” instead of adding them, because both values are of different data types.

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.**

COMMAND: **s = 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(s) = 9.426801

max(s) = 17.26488

quantile(s,.25) = 25% 8.055174  
  
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) = 9.4530

max(s) = 18.6777

quantile (s, .25) = 8.1058

Comparing both the values from sample and actual data, all the values (mean, max and quantile) are approximately same.

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.**

Mean => =AVERAGE(A:A) => 9.426801

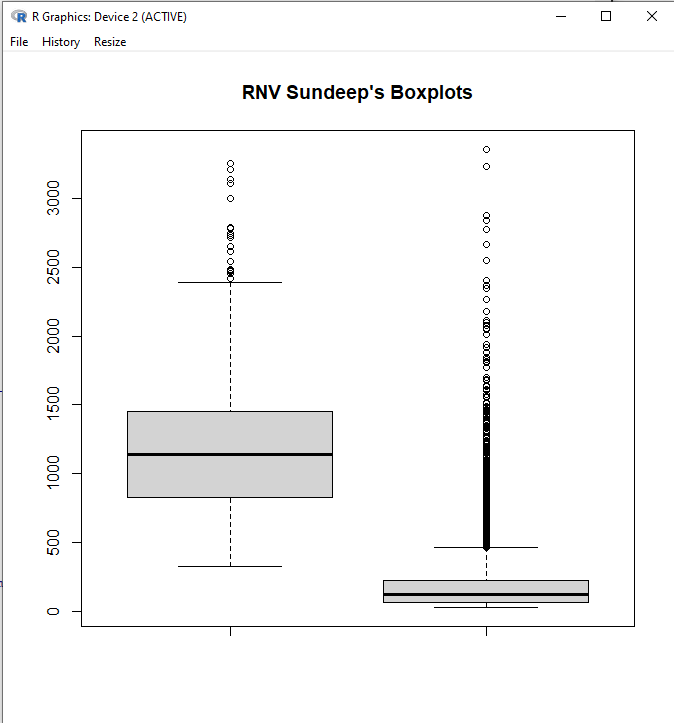
Maximum => =MAX(A:A) => 17.26488

Variance => =VAR.P(A:A) => 4.056299

1st quartile => =QUARTILE(A:A,1) => 8.055174

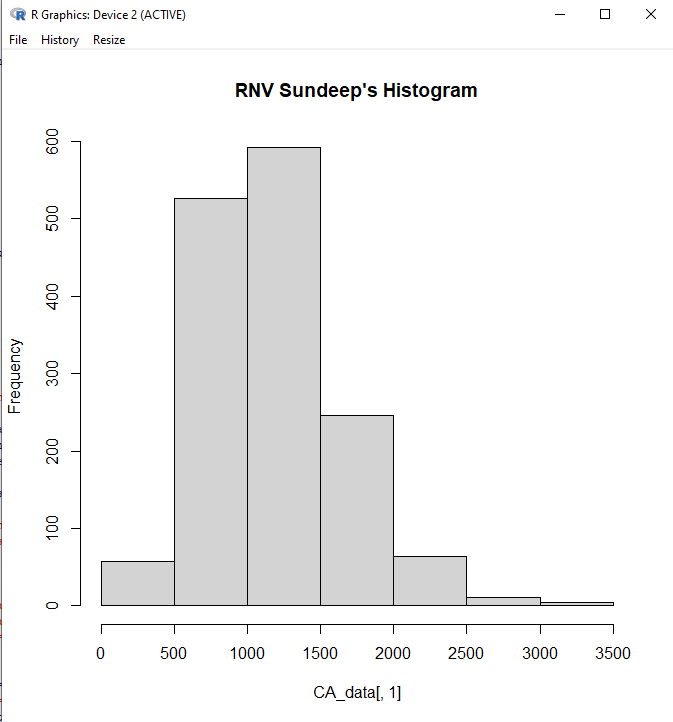
**e) Exactly what happens if you try to open the full data set with Excel?**  
  
**6) Read Chapter 3 (only sections 3.1, 3.2 and 3.3).  
  
7) This question uses a sample of 1500 California house prices at** [**http://www-stat.wharton.upenn.edu/~dmease/CA\_house\_prices.csv**](http://www-stat.wharton.upenn.edu/~dmease/CA_house_prices.csv) **and a sample of 10,000 Ohio house prices at** [**http://www-stat.wharton.upenn.edu/~dmease/OH\_house\_prices.csv**](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").**

**COMMAND**: boxplot(CA\_data[,1], OH\_data[,1], main="RNV Sundeep's Boxplots")

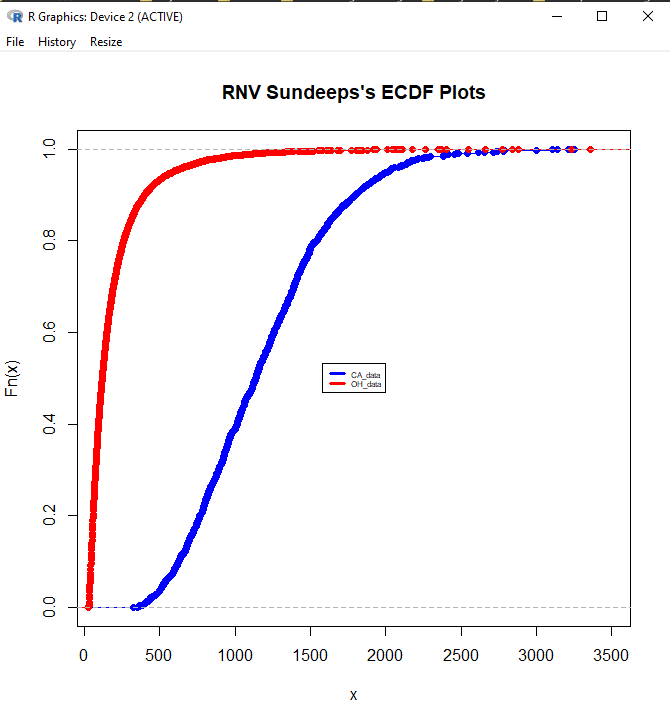


**b) Use R to produce a frequency histogram for only the California house prices. Use intervals of width $500,000 beginning at 0 and ending at $3.5 million. Include the R commands and the plot. Put your name in the title of the plot.**

**COMMAND**: hist(CA\_data[,1], breaks = c(500 \* (0:7)), main="RNV Sundeep's Histogram")

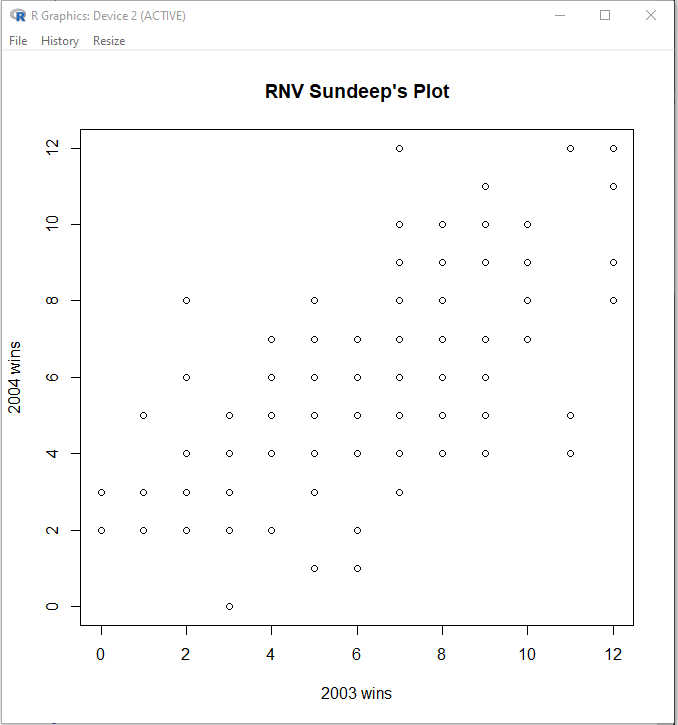


**c) Use R to plot the ECDF of the California houses and Ohio houses on the same graph (as in ICE #11). Include a legend. Include the R commands and the plot. Put your name in the title of the plot.**



**8) This question uses the data at** [**http://www-stat.wharton.upenn.edu/~dmease/football.csv**](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.**

**COMMAND**: plot(data[,2], data[,3], xlim = c(0,12), ylim = c(0,12), xlab = "2003 wins", ylab = "2004 wins", main = "RNV Sundeep's Plot")



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(data[,2], data[,3]) = 0.6537691

**d) How does the value in part c change if you add 10 to all the values for 2004?**  
data[,3] <- data[,3] + 10

cor(data[,2], data[,3]) = 0.6537691

After adding 10 to 2004 column data, cor remains same

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

data[,3] <- data[,3] + 2

cor(data[,2], data[,3]) = 0.6537691

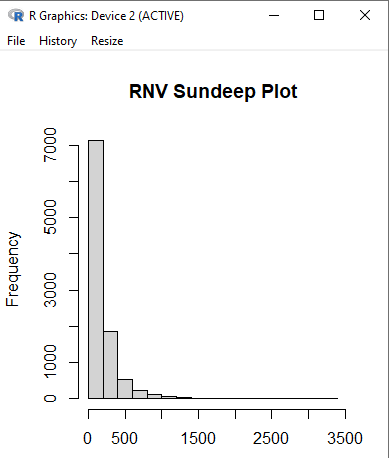
After adding 2 to 2004 column data, cor remains same

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

data[,3] <- data[,3] -2

cor(data[,2], data[,3]) = 0.6537691

After subtracting 2 from 2004 column data, cor remains same

**9) This question uses the sample of 10,000 Ohio house prices at** [**http://www-stat.wharton.upenn.edu/~dmease/OH\_house\_prices.csv**](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 = 118, mean = 190.3132. Median is less than mean  
**b) What does your answer to part a suggest about the shape of the distribution (right-skewed or left-skewed)?**  


Plotting the given data shows that the data is left-skewed.

**c) How does the median change if you add 10 (thousand dollars) to all the values?**  
data[,1] <- data[,1] + 10

median(data[,1]) = 128

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

data[,1] <- data[,1] \* 2

median(data[,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.**

sd(ages) = 8.315218  
  
**b) Compute the same value by hand and show all the steps.**

sd = sqrt((sum(x - mean) ^ 2) / n)  
mean = 19 + 23 + 30 + 30 + 45 + 25 + 24 + 20 / 8 = 27

sqrt( (19 - 27)^2 + (23 - 27)^2 + (30 - 27)^2 + (30 - 27)^2 + (45 - 27)^2 + (25 - 27)^2 + (24 - 27)^2 + (20 - 27)^2 / 8) )

sqrt((64 + 16 + 9 + 9 + 324 + 4 + 9 + 49) / 8)

sqrt(69.05)

**8.30**

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

sd(ages) = 8.315213

Value remains same even if we add 10 to data

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

ages <- ages \* 100

sd(ages) = 831.5213

After multiplying ages with 100, we get sd as 831.