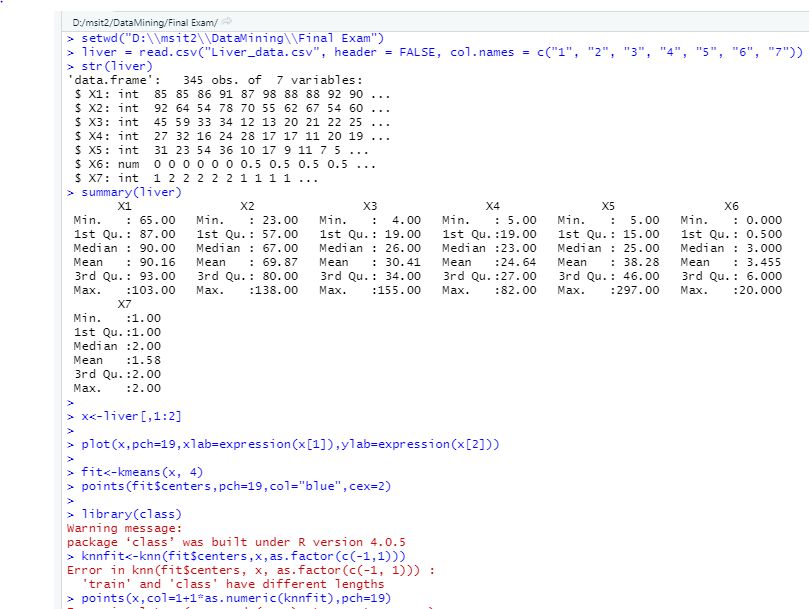
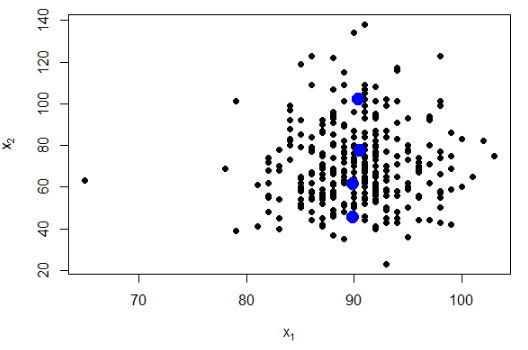
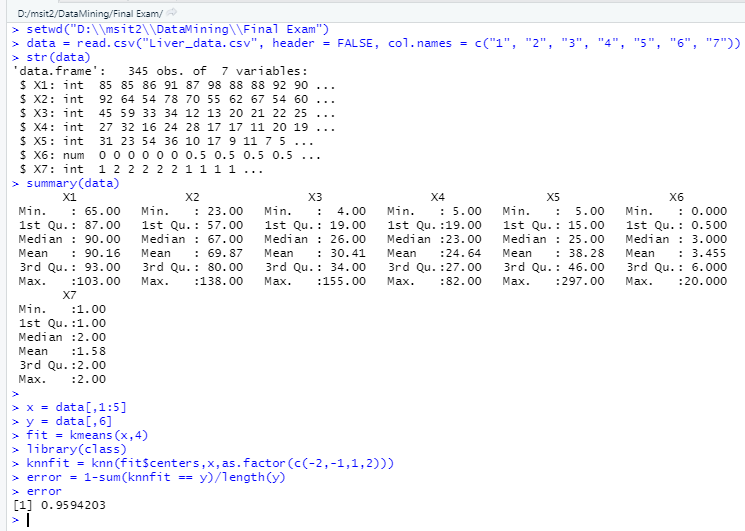
1. Create k-means clusters for k=4 for the Liver Disorders Data Set at <http://archive.ics.uci.edu/ml/datasets/Liver+Disorders> . Explain the input parameters you provided for the clustering algorithm. Plot the fitted cluster centers using a different color. Finally assign the cluster membership for the points to the nearest cluster center. Color the points according to their cluster membership. (10+10=20M)





1. Compute the misclassification error that would result if you used your clustering rule to classify the data by assigning the majority class of the cluster. (10M)



1. a) For the dataset BSE\_Sensex\_Index.csv, create an extra column of successive differences for each column of numeric values in this data file. Extract two simple random samples with replacement of 1000 and 3000 observations (rows). Show your R commands for doing this.

Do the same thing by using Excel. Show your Excel commands.

**Note:** Successive difference for date d1= (date d1 value-immediate available previous date of d1 value)/immediate available previous date of d1. For the last row fill up values with mean of its immediate three previous row values.

b) For your samples, use the functions mean(), max(), var() and quartile(,.25) to compute the mean, maximum, variance and 1st quartile respectively for each column which has successive differences. Show your R code and the resulting values.

Do the same thing by using Excel. Show your Excel commands.

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? Do you find any significant difference between two sample values like mean in comparison with entire data? If so what explanation you can give for that?

Do the same thing by using Excel. Show your Excel commands.

d) Use R to produce a single graph displaying a boxplot for open, close, high and low. Include the R commands and the plot.

Do the same thing by using Excel. Show your Excel commands

e) Use R to produce a frequency histogram for Close values. Use intervals of width 2000 beginning at 0. Include the R commands and the plot.

Do the same thing by using Excel. Show your Excel commands. (10+10=20M)

**Code :**

setwd("D:\\msit2\\DataMining\\Late")

getwd()

stock = read.csv("BSE\_Sensex\_Index.csv",header = FALSE)

str(stock)

summary(stock)

stock$diffV2 <- c(0, diff(stock$V2))

stock$diffV3 <- c(0, diff(stock$V3))

stock$diffV4 <- c(0, diff(stock$V4))

stock$diffV5 <- c(0, diff(stock$V5))

stock$diffV6 <- c(0, diff(stock$V6))

stock$diffV7 <- c(0, diff(stock$V7))

summary(stock)

str(stock)

sample1 = sample(seq(1,length(stock$V2)),1000,replace = T)

sample2 = sample(seq(1,length(stock$V2)),3000,replace = T)

sampleV2\_1 = stock[sample1,8]

mean(sampleV2\_1)

max(sampleV2\_1)

var(sampleV2\_1)

quantile(sampleV2\_1,.25)

sampleV2\_2 = stock[sample2,8]

mean(sampleV2\_2)

max(sampleV2\_2)

var(sampleV2\_2)

quantile(sampleV2\_2,.25)

sampleV3\_1 = stock[sample1,9]

mean(sampleV3\_1)

max(sampleV3\_1)

var(sampleV3\_1)

quantile(sampleV3\_1,.25)

sampleV3\_2 = stock[sample2,9]

mean(sampleV3\_2)

max(sampleV3\_2)

var(sampleV3\_2)

quantile(sampleV3\_2,.25)

sampleV4\_1 = stock[sample1,10]

mean(sampleV4\_1)

max(sampleV4\_1)

var(sampleV4\_1)

quantile(sampleV4\_1,.25)

sampleV4\_2 = stock[sample2,10]

mean(sampleV4\_2)

max(sampleV4\_2)

var(sampleV4\_2)

quantile(sampleV4\_2,.25)

sampleV5\_1 = stock[sample1,11]

mean(sampleV5\_1)

max(sampleV5\_1)

var(sampleV5\_1)

quantile(sampleV5\_1,.25)

sampleV5\_2 = stock[sample2,11]

mean(sampleV5\_2)

max(sampleV5\_2)

var(sampleV5\_2)

quantile(sampleV5\_2,.25)

sampleV6\_1 = stock[sample1,12]

mean(sampleV6\_1)

max(sampleV6\_1)

var(sampleV6\_1)

quantile(sampleV6\_1,.25)

sampleV6\_2 = stock[sample2,12]

mean(sampleV6\_2)

max(sampleV6\_2)

var(sampleV6\_2)

quantile(sampleV6\_2,.25)

sampleV7\_1 = stock[sample1,13]

mean(sampleV7\_1)

max(sampleV7\_1)

var(sampleV7\_1)

quantile(sampleV7\_1,.25)

sampleV7\_2 = stock[sample2,13]

mean(sampleV7\_2)

max(sampleV7\_2)

var(sampleV7\_2)

quantile(sampleV7\_2,.25)

mean(stock$diffV2)

max(stock$diffV2)

var(stock$diffV2)

quantile(stock$diffV2,.25)

abs(mean(sampleV2\_1)-mean(stock$diffV2))

abs(max(sampleV2\_1)-max(stock$diffV2))

abs(var(sampleV2\_1)-var(stock$diffV2))

abs(quantile(sampleV2\_1,.25)-quantile(stock$diffV2,.25))

abs(mean(sampleV2\_2)-mean(stock$diffV2))

abs(max(sampleV2\_2)-max(stock$diffV2))

abs(var(sampleV2\_2)-var(stock$diffV2))

abs(quantile(sampleV2\_2,.25)-quantile(stock$diffV2,.25))

mean(stock$diffV3)

max(stock$diffV3)

var(stock$diffV3)

quantile(stock$diffV3,.25)

abs(mean(sampleV3\_1)-mean(stock$diffV3))

abs(max(sampleV3\_1)-max(stock$diffV3))

abs(var(sampleV3\_1)-var(stock$diffV3))

abs(quantile(sampleV3\_1,.25)-quantile(stock$diffV3,.25))

abs(mean(sampleV3\_2)-mean(stock$diffV3))

abs(max(sampleV3\_2)-max(stock$diffV3))

abs(var(sampleV3\_2)-var(stock$diffV3))

abs(quantile(sampleV3\_2,.25)-quantile(stock$diffV3,.25))

mean(stock$diffV4)

max(stock$diffV4)

var(stock$diffV4)

quantile(stock$diffV4,.25)

abs(mean(sampleV4\_1)-mean(stock$diffV4))

abs(max(sampleV4\_1)-max(stock$diffV4))

abs(var(sampleV4\_1)-var(stock$diffV4))

abs(quantile(sampleV4\_1,.25)-quantile(stock$diffV4,.25))

abs(mean(sampleV4\_2)-mean(stock$diffV4))

abs(max(sampleV4\_2)-max(stock$diffV4))

abs(var(sampleV4\_2)-var(stock$diffV4))

abs(quantile(sampleV4\_2,.25)-quantile(stock$diffV4,.25))

mean(stock$diffV5)

max(stock$diffV5)

var(stock$diffV5)

quantile(stock$diffV5,.25)

abs(mean(sampleV5\_1)-mean(stock$diffV5))

abs(max(sampleV5\_1)-max(stock$diffV5))

abs(var(sampleV5\_1)-var(stock$diffV5))

abs(quantile(sampleV5\_1,.25)-quantile(stock$diffV5,.25))

abs(mean(sampleV5\_2)-mean(stock$diffV5))

abs(max(sampleV5\_2)-max(stock$diffV5))

abs(var(sampleV5\_2)-var(stock$diffV5))

abs(quantile(sampleV5\_2,.25)-quantile(stock$diffV5,.25))

mean(stock$diffV6)

max(stock$diffV6)

var(stock$diffV6)

quantile(stock$diffV6,.25)

abs(mean(sampleV6\_1)-mean(stock$diffV6))

abs(max(sampleV6\_1)-max(stock$diffV6))

abs(var(sampleV6\_1)-var(stock$diffV6))

abs(quantile(sampleV6\_1,.25)-quantile(stock$diffV6,.25))

abs(mean(sampleV6\_2)-mean(stock$diffV6))

abs(max(sampleV6\_2)-max(stock$diffV6))

abs(var(sampleV6\_2)-var(stock$diffV6))

abs(quantile(sampleV6\_2,.25)-quantile(stock$diffV6,.25))

mean(stock$diffV7)

max(stock$diffV7)

var(stock$diffV7)

quantile(stock$diffV7,.25)

abs(mean(sampleV7\_1)-mean(stock$diffV7))

abs(max(sampleV7\_1)-max(stock$diffV7))

abs(var(sampleV7\_1)-var(stock$diffV7))

abs(quantile(sampleV7\_1,.25)-quantile(stock$diffV7,.25))

abs(mean(sampleV7\_2)-mean(stock$diffV7))

abs(max(sampleV7\_2)-max(stock$diffV7))

abs(var(sampleV7\_2)-var(stock$diffV7))

abs(quantile(sampleV7\_2,.25)-quantile(stock$diffV7,.25))

boxplot(stock$diffV2,

stock$diffV3,

stock$diffV4,

stock$diffV5,

stock$diffV6,

stock$diffV7 ,col = 'blue', main = 'Boxplot', names=c("Open","High", "Low", "Close",

"volume", "adj"))

stock$c = as.numeric(stock$V4)

hist(stock$c,breaks=seq(0,20000,by=2000),col='blue',xlab = "Close",ylab = "Frequency",main = "Histogram Plot")

