ASSIGNMENT-1

SAI TEJA KANNEGANTI

28 August 2016

DSA 5103: INTELLIGENT DATA ANALYTICS

Required Packages for this assignment :

library (moments) # Contains functions skewness and kurtosis  
 library (plyr) # To load dataframe baseball  
 library(datasets) # To load quakes data from datasets package  
 library(reshape) #contains replace function

##   
## Attaching package: 'reshape'

## The following objects are masked from 'package:plyr':  
##   
## rename, round\_any

**PROBLEM 1** : Vectors using R

**PROBLEM 1(a):**

x <- c(3,12,6,-5,0,8,15,1,-10,7) # To concatinate 10 numerical values in x.

**Problem 1(b)**:

y <- seq(min(x),max(x),length=10) #Finding minimum and maximum values of x and splitting interval into 10 numbers and assigning to y.  
y

## [1] -10.000000 -7.222222 -4.444444 -1.666667 1.111111 3.888889  
## [7] 6.666667 9.444444 12.222222 15.000000

**Problem 1(c)**

**For x**

sum(x)

## [1] 37

mean(x)

## [1] 3.7

sd(x)

## [1] 7.572611

var(x)

## [1] 57.34444

mad(x) # To find Mean Absolute Deviation.

## [1] 5.9304

quantile (x, probs=c(0.0,0.25,0.50,0.75,1.0)) # Displays value of random variable at these fractions.

## 0% 25% 50% 75% 100%   
## -10.00 0.25 4.50 7.75 15.00

quantile (x, probs=c(0,0.20,0.40,0.60,0.80,1.0))

## 0% 20% 40% 60% 80% 100%   
## -10.0 -1.0 2.2 6.4 8.8 15.0

**For y**

sum(y)

## [1] 25

mean(y)

## [1] 2.5

sd(y)

## [1] 8.41014

var(y)

## [1] 70.73045

mad(y)

## [1] 10.29583

quantile (y, probs=c(0.0,0.25,0.50,0.75,1.0))

## 0% 25% 50% 75% 100%   
## -10.00 -3.75 2.50 8.75 15.00

quantile (y, probs=c(0.0,0.20,0.40,0.60,0.80,1.0))

## 0% 20% 40% 60% 80%   
## -1.000000e+01 -5.000000e+00 -1.665335e-15 5.000000e+00 1.000000e+01   
## 100%   
## 1.500000e+01

Quartiles are values of random variable that divide into 4 groups.

Quintiles are values of random variable that divide into 5 groups.

**problem 1(d):**

z<- sample (x,7, replace = TRUE) # Sampling random variable x,to choose 7 elements with replacement.  
z

## [1] -5 -5 0 12 3 -5 0

**problem 1(e):**

skewness(x)

## [1] -0.3123905

kurtosis(x)# moments is package, skewness & kurtosis are funtions to determine skewness and kurtosis.

## [1] 2.355328

Skewness and kurtosis can also be determined in e1701 package, but this package is not available in R version 3.3.1

**problem 1(f)**

t.test(x,y) # t-test of 2 independent vectors with some default parameters.

##   
## Welch Two Sample t-test  
##   
## data: x and y  
## t = 0.33531, df = 17.805, p-value = 0.7413  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -6.324578 8.724578  
## sample estimates:  
## mean of x mean of y   
## 3.7 2.5

As p value is greater than 0.05 and 0 lies in the confidence interval, hypothesis is true,i.e. true difference between means is 0. So the difference of means is insignificant.

**problem 1(g)**

sort(x)

## [1] -10 -5 0 1 3 6 7 8 12 15

t.test(x, sort (x),paired=TRUE) # paired t-test between x and sort(x).

##   
## Paired t-test  
##   
## data: x and sort(x)  
## t = 0, df = 9, p-value = 1  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -8.273986 8.273986  
## sample estimates:  
## mean of the differences   
## 0

Since we are performing paired t-test between x and sort(x),mean of the differences is 0. **problem 1(h)**

n <- x[x<0] # n is a vector having negative valus of x.  
n

## [1] -5 -10

**problem 1(I)**

x <- x[!x %in% n] # !x indicates not x,!x includes n,so x doesnot include n.  
x

## [1] 3 12 6 0 8 15 1 7

**PROBLEM 2**

**problem 2 (a):**

college <- read.csv("E:/Masters/IDA/Assignments/Assignment 1/college.csv", header = TRUE, sep = ",") # read.csv() function to read tablular data & create a data frame.

**problem 2(b):**

rownames (college) <- college [,1] # Assigning first column(university name) as row names of dataframe.  
  
View (college) # To View dataframe.  
  
college <- college [,-1] # Removing first column of the dataframe.

**problem 2(c)**

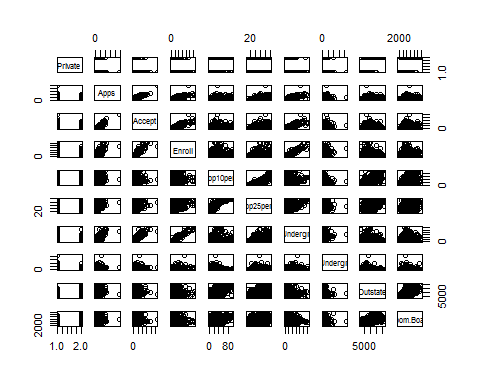
**1**.

summary(college) #Provides numerical summary of datavariables in college.

## Private Apps Accept Enroll Top10perc   
## No :212 Min. : 81 Min. : 72 Min. : 35 Min. : 1.00   
## Yes:565 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242 1st Qu.:15.00   
## Median : 1558 Median : 1110 Median : 434 Median :23.00   
## Mean : 3002 Mean : 2019 Mean : 780 Mean :27.56   
## 3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902 3rd Qu.:35.00   
## Max. :48094 Max. :26330 Max. :6392 Max. :96.00   
## Top25perc F.Undergrad P.Undergrad Outstate   
## Min. : 9.0 Min. : 139 Min. : 1.0 Min. : 2340   
## 1st Qu.: 41.0 1st Qu.: 992 1st Qu.: 95.0 1st Qu.: 7320   
## Median : 54.0 Median : 1707 Median : 353.0 Median : 9990   
## Mean : 55.8 Mean : 3700 Mean : 855.3 Mean :10441   
## 3rd Qu.: 69.0 3rd Qu.: 4005 3rd Qu.: 967.0 3rd Qu.:12925   
## Max. :100.0 Max. :31643 Max. :21836.0 Max. :21700   
## Room.Board Books Personal PhD   
## Min. :1780 Min. : 96.0 Min. : 250 Min. : 8.00   
## 1st Qu.:3597 1st Qu.: 470.0 1st Qu.: 850 1st Qu.: 62.00   
## Median :4200 Median : 500.0 Median :1200 Median : 75.00   
## Mean :4358 Mean : 549.4 Mean :1341 Mean : 72.66   
## 3rd Qu.:5050 3rd Qu.: 600.0 3rd Qu.:1700 3rd Qu.: 85.00   
## Max. :8124 Max. :2340.0 Max. :6800 Max. :103.00   
## Terminal S.F.Ratio perc.alumni Expend   
## Min. : 24.0 Min. : 2.50 Min. : 0.00 Min. : 3186   
## 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00 1st Qu.: 6751   
## Median : 82.0 Median :13.60 Median :21.00 Median : 8377   
## Mean : 79.7 Mean :14.09 Mean :22.74 Mean : 9660   
## 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00 3rd Qu.:10830   
## Max. :100.0 Max. :39.80 Max. :64.00 Max. :56233   
## Grad.Rate   
## Min. : 10.00   
## 1st Qu.: 53.00   
## Median : 65.00   
## Mean : 65.46   
## 3rd Qu.: 78.00   
## Max. :118.00

**2.**

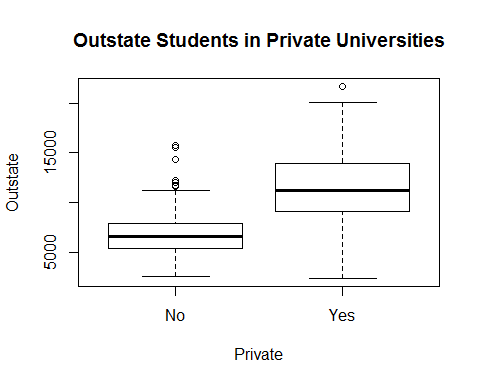
pairs(college[,1:10]) # pairs is used to produce a matrix of scatter plots.



Scatterplot of 1 to 10 columns of dataframe college.

**3**.

plot(college$Private, college$Outstate,main="Outstate Students in Private Universities",xlab = "Private", ylab = "Outstate")



Side by side boxplots of outstate students Vs private universities.

**4.**

Elite <- rep ("NO", nrow(college)) # Replicates NO to all rows in college dataframe.  
Elite [college$Top10perc >50] <- "Yes" # Elite contains universities having Top10perc greater than 50  
Elite <- as.factor (Elite) # as.factor is used to encode a vector as a factor(1 for no,2 for yes)  
college <- data.frame(college ,Elite) # To create Elite data variable in college dataframe

**5.**

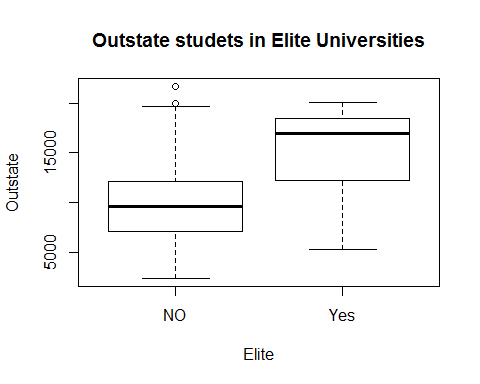
summary (Elite)

## NO Yes   
## 699 78

There are 78 Elite universities in given college list.

**6**.

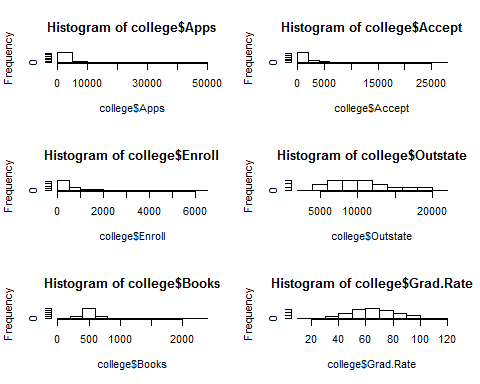
plot(college$Elite,college$Outstate, main="Outstate studets in Elite Universities",ylab="Outstate",xlab="Elite")



Side by side boxplots of outstate students in Elite universities.

**7.**

par(mfrow=c(3,2)) # used for multiple number of plots simultaniously (r\*c).  
hist(college$Apps,breaks=9) # Histogram for no of applications(Apps) with 9 breaks  
hist(college$Accept,breaks=11)  
hist(college$Enroll,breaks=14)  
hist(college$Outstate,breaks=9)  
hist(college$Books,breaks=9)  
hist(college$Grad.Rate,breaks=14)



Some histograms with differing numbers of bins for few of the quantitative variables.

**Problem 3 :** Manipulating data into data frames

**Problem 3(a):**

data(baseball) # data loads specified data sets.

**problem 3(b)**

baseball$sf[baseball$year< 1954] <- 0 # equating value of sf to 0,when year in baseball dataframe is less than 1954.  
baseball$hbp[is.na(baseball$hbp)] <- 0 # Equating all missing values(NA) in hbp to 0.  
baseball <- baseball[c(baseball$ab>=50), ] #Baseball contains rows of players whose ab >= 50(deletes rows of players with ab < 50) & all columns remains unchanged.

**problem 3(c)**

obp <-((baseball$h+baseball$bb+baseball$hbp)/(baseball$ab+baseball$bb+baseball$hbp+baseball$sf))   
baseball <- data.frame(baseball,obp) # combining obp variable(column) to baseball dataframe.

**problem 3(d)**

Result <- baseball[order(-obp) , ] # Result is a dataframe containing rows sorted as per obp(-obp indicates decreasing order)  
Result <- Result[1:5, ] # Considering only 1 to 5 rows.  
Result[3:22] <- list(NULL) # Removing 3 to 22 columns.  
head(Result)

## id year obp  
## 84983 bondsba01 2004 0.6094003  
## 82594 bondsba01 2002 0.5816993  
## 29489 willite01 1941 0.5528053  
## 7772 mcgrajo01 1899 0.5474860  
## 19883 ruthba01 1923 0.5445402

Result contains player name, year and on base percentage of top 5 records.

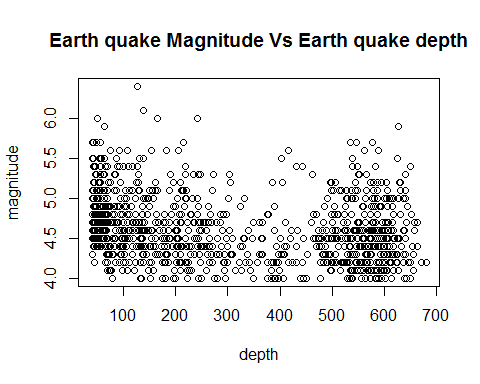
**problem 4**

**problem 4(a)**

data(quakes) # to load quakes from datasets package.

**problem 4 (b)**

plot(quakes$depth,quakes$mag,main= "Earth quake Magnitude Vs Earth quake depth",xlab= "depth",ylab= "magnitude") # To plot a graph, need to give x-axis,y-axis quantities, we use $ inorder to access particular column in dataframe, later main is used for title, xlab,ylab used for labels.



Plot between Earth quake magnitude and depth from surface

**problem 4 (c)**

quakeAvgDepth <- aggregate(quakes$depth,by = list (quakes$mag),mean) #we are using aggregate function to calculate average depth for given sequencence of magnitude,since we are calculating average depth, function written is mean.

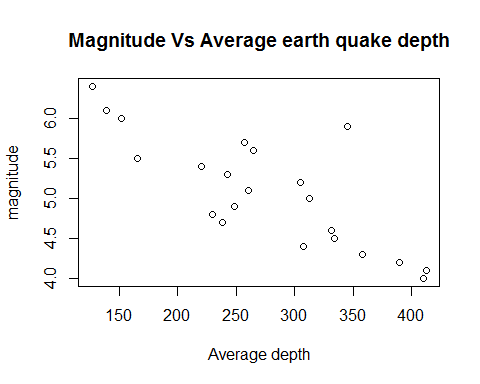
**problem 4(d)**

colnames(quakeAvgDepth) <- c("Magnitude","Avg\_depth") # To list column names of dataframe.(we can also use rename command).

Renamed variables to Magnitude, Avg depth.

**problem 4 (e)**

plot(quakeAvgDepth$Avg\_depth,quakeAvgDepth$Magnitude,main= "Magnitude Vs Average earth quake depth",xlab= "Average depth",ylab= "magnitude") # Used renamed variables of "quakeAvgDepth" data frame to plot graph.



Plot between Earth quake magnitude and Average depth from surface

**problem 4(f)**

From the graphs we can understand that:

1. Generally magnitude of earthquake increases with decrease in depth from earth surface.(From plot 4(e))

2. More number of earthquakes occur at lower magnitudes (from 4.0 to 5.0).(From plot 4(b)).