Homework#1

Azadeh Gilanpour

August 30, 2018

Questions

Question# 1: Using R vector

(a)

create a vector with 10 specefic number and assiging to X

```
x<- c(3,12,6,-5,0,8,15,1,-10,7)
x
```

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

(b)

create a vector y with 10 elements of x from min to max

```
y<- y <-seq(min(x),max(x),length=10)
y
```

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

(c)

- summation of elements “x” and “y”

```
sum(x); sum(y)
```

```
## [1] 37
```

```
## [1] 25
```

- mean value of “x” and “y”

```
mean(x);mean(y)
```

```
## [1] 3.7
```

```
## [1] 2.5
```

- standard deviation values for “x” and “y”

```
sd (x); sd (y)
```

```
## [1] 7.572611
```

```
## [1] 8.41014
```

- Variances for “x” and “y”

```
var(x); var(y)

## [1] 57.34444

## [1] 70.73045
```

- Mean absolute deviation values for “x” and “y”

```
mad(x); mad(y)

## [1] 5.9304

## [1] 10.29583
```

- Quartiles for “x” and “y”

```
quantile(x); quantile(y)

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

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

- Quintiles for “x” and “y”

```
quantile(x, probs = seq(0, 1, 0.25)); quantile(y, probs = seq(0, 1, 0.25))

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

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

(d)

create vector z consist 7 elemnts randomly sample from x

```
z<- sample(x, 7,replace=TRUE)
z

## [1]   7   7  12 -10  12   0 -10
```

(e)

package for calculaiton of skewness and kurtosis

```
library(e1071)
```

- Skewness of vector “x” from package e1071

```
skewness(x)

## [1] -0.2667237
```

- kurtosis of vector “x” from package e1071

```
kurtosis(x)
```

```
## [1] -1.092184
```

(f)

t-test to compare the mean of “x” and “y”

```
t.test(x,y)
```

```
##
##  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
```

- The reults shows the difference between two sets are not significant. The p-values for the test is %74 which is very higher than %5. The confidence interval is also (-6.32, 8.72) which contains zero.

(g)

t-test to compare the mean of sorted “x” and “y”

```
t.test(sort(x),y,paired = T)
```

```
##
##  Paired t-test
##
## data:  sort(x) and y
## t = 2.164, df = 9, p-value = 0.05868
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05440584  2.45440584
## sample estimates:
## mean of the differences
##                     1.2
```

(h)

Logical vector to show the negative number of x

```
as.logical(x<0)
```

```
##  [1] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE
```

(i)

remove negative number from x

```
x<- x[!x<0]
x

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

Question# 2: Using R Introductory data exploration

(a)

Read data from file "college.csv"

```
college<-read.csv("college.csv")
```

(b)

```
# Use the first column of data as row's name
rownames(college) <- college[,1]

# Display the content of the data frame
View (college )

# Remove the first column's of data
college <- college[,-1]
```

(c)

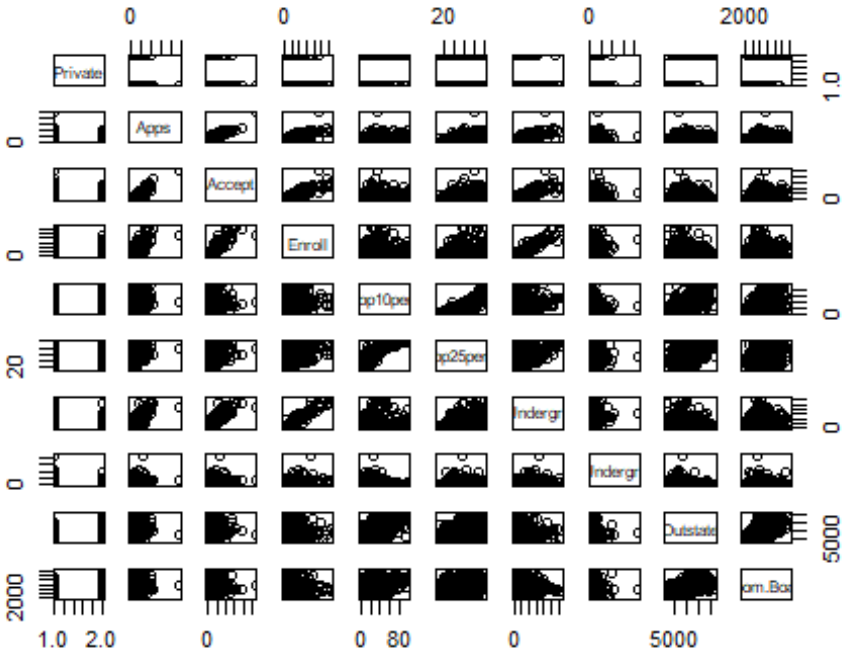
- i Shows the numerical summaries of "college" dataset

```
summary(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
```

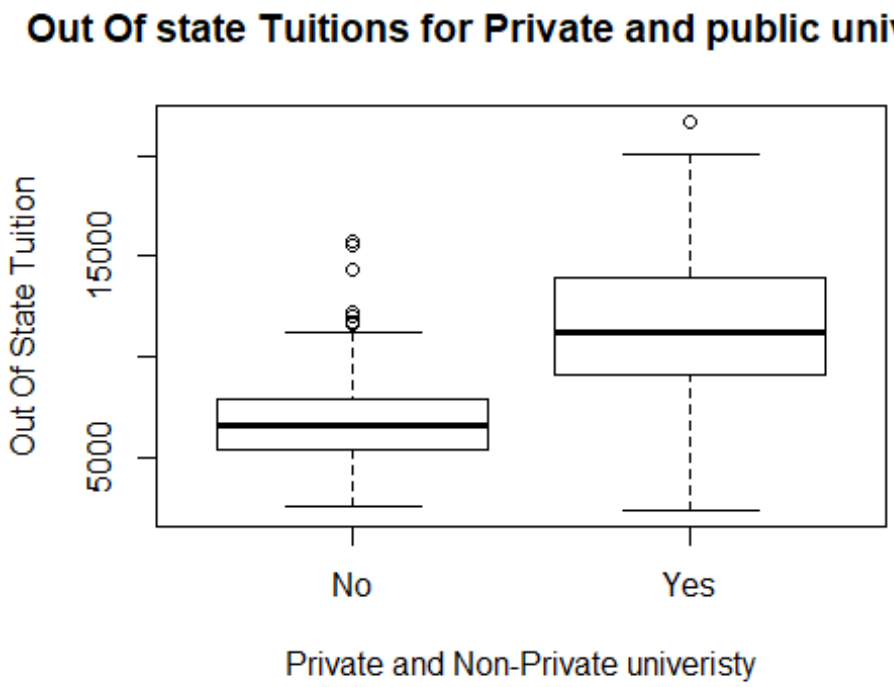
- ii

```
#Shows the description for "pairs" function
help("pairs")

## starting httpd help server ... done

# Used pairs fucntion to produce a scatterplot matrix for the first 10
columns
pairs(college[,1:10])
```



- iii Side-by-side boxplot for Outstate vs. Private

```
plot(Outstate~Private,data=college,
     main="Out Of state Tuitions for Private and public university",
     xlab="Private and Non-Private univeristy",
     ylab="Out Of State Tuition")
```



- iv

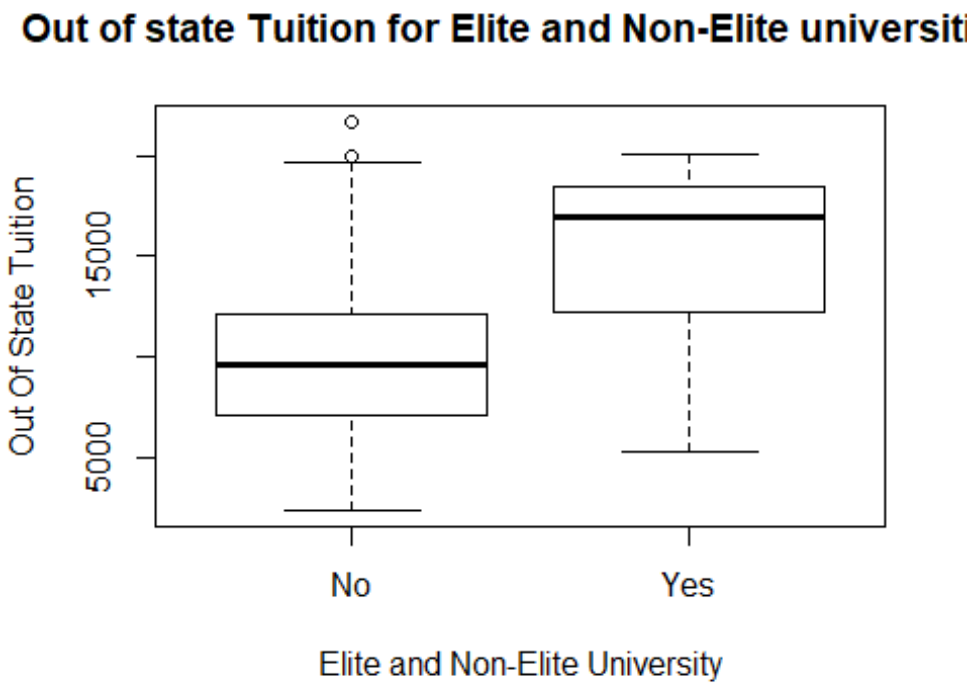
```
# Creates a vector in size of rows of the college filling with No
Elite <- rep("No", nrow(college))

# Find the Top10prec greater than 50 and replace the corresponding elements
of "Elite" with "Yes"
Elite [college$Top10perc>50] <- "Yes"

# the "Elite" vector into a factor vector with 2 levels
Elite <- as.factor(Elite)

# Adds "Elite" as a new column to the "College" data frame
college <- data.frame(college, Elite)
```

- v Shows the number of the elite and non-elite colleges

```
summary(Elite)

##  No Yes
## 699  78
```

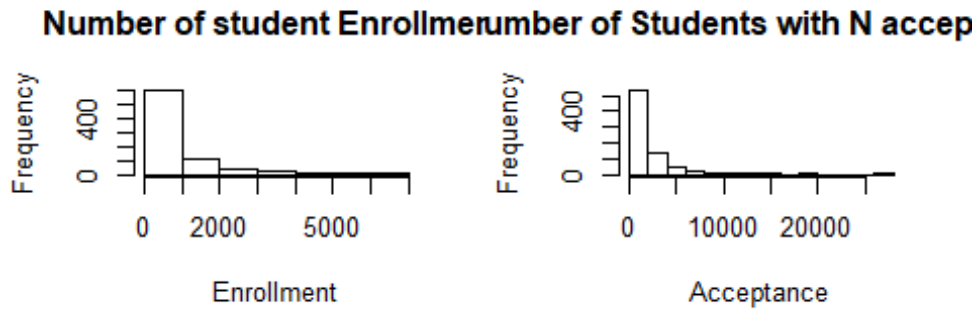
- vi # Side-by-side boxplot for Outstate vs. Elite

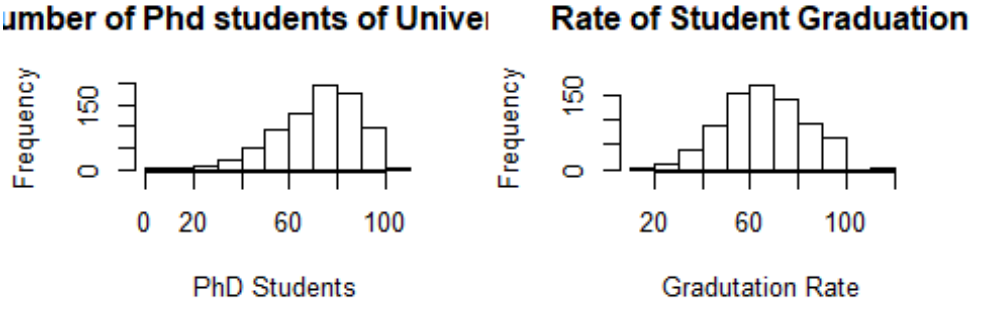
```
plot(Outstate~Elite,data=college,
     main="Out of state Tuition for Elite and Non-Elite universities",
     xlab="Elite and Non-Elite University",
     ylab="Out Of State Tuition")
```



- vii

```
# Divides the screen into 4 windows
par(mfrow=c(2,2))
# Draw histograms for Number of new students enrolled, number of applicants
accepted,Percent of faculty with Ph.D.s , and Graduation rate.
hist(college$Enroll,xlab = "Enrollment",main = "Number of student
Enrollment",breaks = 5)

hist(college$Accept,xlab = "Acceptance",main = "Number of Students with N
acceptance",breaks=10)

hist(college$PhD,xlab="PhD Students",main="Number of Phd students of
Universities", breaks = 10)

hist(college$Grad.Rate,xlab="Gradutation Rate",main = "Rate of Student
Graduation",breaks = 10)
```





Question#3: Using R Mlating data in data frames

(a)

```
# package include of baseball dataset
library(plyr)
#Load a baseball data set
data("baseball")
# Shows the help and description of basegall dataset from plyr package
?baseball
```

(b)

Set 0 “sacrifies flies” for players beore 1954

```
baseball$sf[baseball$year<1954]=0
```

- 1 Set 0 missing values for “Hit by pitch”

```
baseball$hbp[is.na(baseball$hbp)]<- 0
```

- 2 Excludes all player records at bats with fewer than 50

```
baseball<- subset(baseball,baseball$ab>=50)
```

(c)

```
#Compute on base percentage
obp<- with(baseball, (h + bb + hbp) / (ab + bb + hbp + sf))
```

```r
 #Adds "obp" as data in a new column
baseball<- data.frame(baseball,obp)
```

(d)

```r
# Sorts the sata set based on "obp"
baseball<- baseball[order(obp),]
# Displays the top 5 on base percentage
head(baseball[,c("id", "year", "obp")], n =5)
```

```
##              id year        obp
## 41939 aguirha01 1962 0.03947368
## 44890 simmocu01 1965 0.04687500
## 46933 cardwdo01 1968 0.04918033
## 83686 leiteal01 2003 0.05454545
## 25361 johnssi01 1933 0.05479452
```

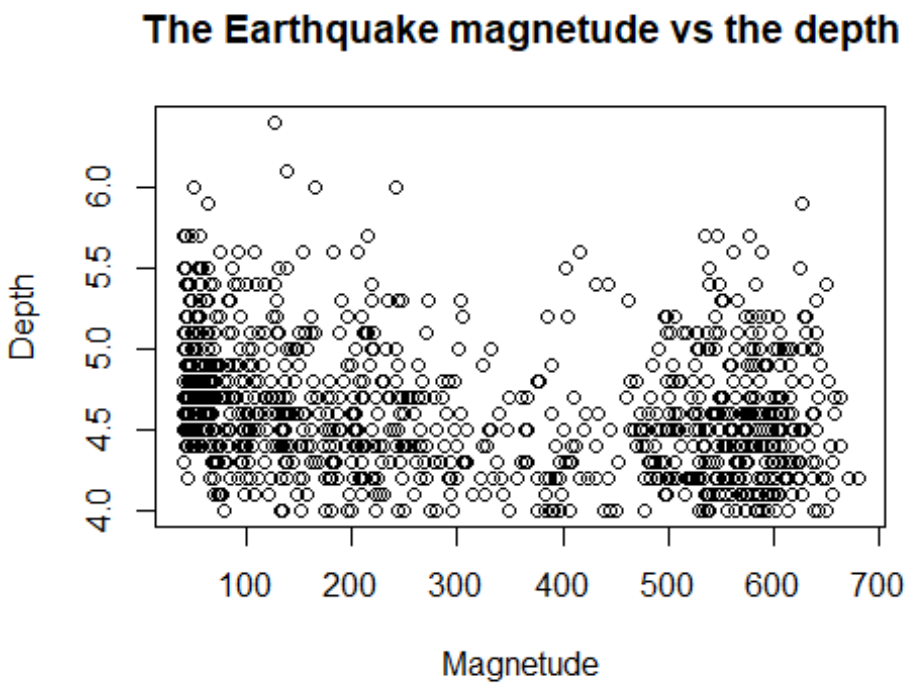
Question# 4 :Using R “aggregate()” function

(a)

```r
# package including "quakes" data set for problem #4
library(datasets)
#Load a quakes data set
data("quakes")
```

(b)

```r
# Reset display windows
par(mfrow=c(1,1))

# Plots the earthquake magnetude agianst the depth
plot(mag~depth,data=quakes,
     main= "The Earthquake magnetude vs the depth",
     xlab= "Magnetude",
     ylab = "Depth")
```



(c)

Use aggregate to compute the average earthquake depth for each magnitude level. Store these results in a quakeAvgDepth

```
quakeAvgDepth<- aggregate(quakes$depth,list("Magnitude"=quakes$mag),mean)
```

(d)

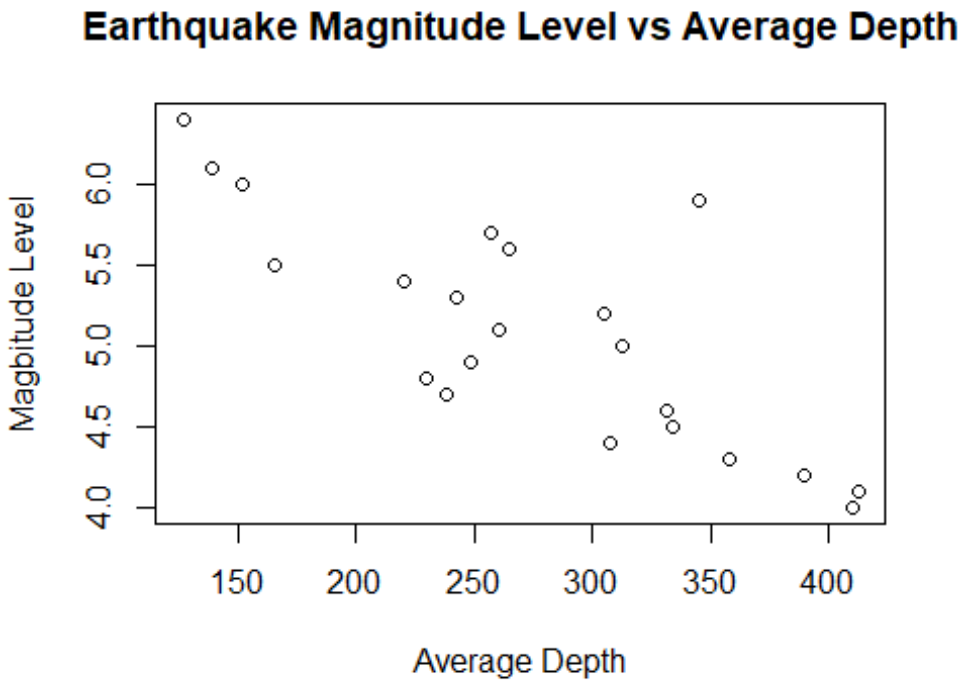
Rename the quakeAvgDepth

```
names(quakeAvgDepth) <- c("Magnitude Level","Average Depth")
```

(e)

Plot the magnitude vs. the average depth.

```
plot( quakeAvgDepth$`Magnitude Level`~ quakeAvgDepth$`Average Depth`,
      main= "Earthquake Magnitude Level vs Average Depth",
      xlab="Average Depth",
      ylab="Magbitude Level")
```



(f)

The mean trend showes as the average depth increases the earthquack gets weaker,