Homework 1

Pramod Aravind Byakod

OU ID - 113436879

Question 1

```
#Create a vector "x" and assign values
x = c(3, 12, 6, -5, 0, 8, 15, 1, -10, 7)
#display value of vector "x"
x
```

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

```
#Create a vector "y" and assign values using seq command
y = seq(min(x),max(x),length=10)
#Compute sum of x
sum(x)
```

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

```
#Compute mean of x
mean(x)
```

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

```
#Compute standard deviation of x
sd(x)
```

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

```
#Load the lsr package
library(lsr)
#Compute mean absolute deviation of x
aad(x)
```

```
## [1] 5.9
```

```
#Compute variance of x
var(x)
```

```
## [1] 57.34444
```

```
#Compute sum of y
sum(y)
```

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

```
#Compute mean of y
mean(y)
```

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

```
#Compute standard deviation of y
sd(y)
```

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

```r
#Compute mean absolute deviation of y
aad(y)
```

```
## [1] 6.944444
```

```r
#Compute variance of y
var(y)
```

```
## [1] 70.73045
```

```r
#load the package "moments". Used to compute skewness and kurtosis
library(moments)
#find skewness of x
skewness(x)
```

```
## [1] -0.3123905
```

```r
#find kurtosis of x
kurtosis(x)
```

```
## [1] 2.355328
```

```r
#compute a statistical test for differences in means between the vectors 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
```

```r
#Sort the vector x
x = sort(x)
#Paired test on x and y after x has been sorted
t.test(x,y,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  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
```

```r
#Differences in mean are not significant
#Create a logical vector to identify the negetive values in x
x_neg = x<0
#Display the previously created logical vector
```

```
x_neg
```

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

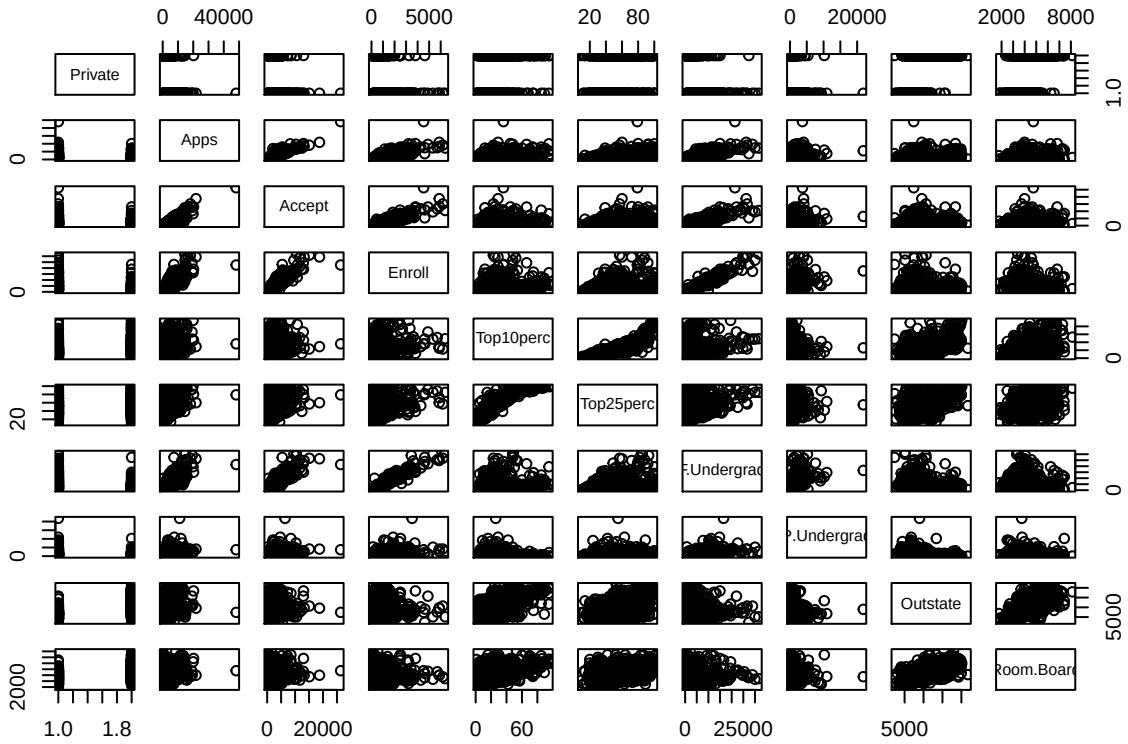
```
#Remove all the negetive values from x
x = x[!x_neg]
```

Question 2

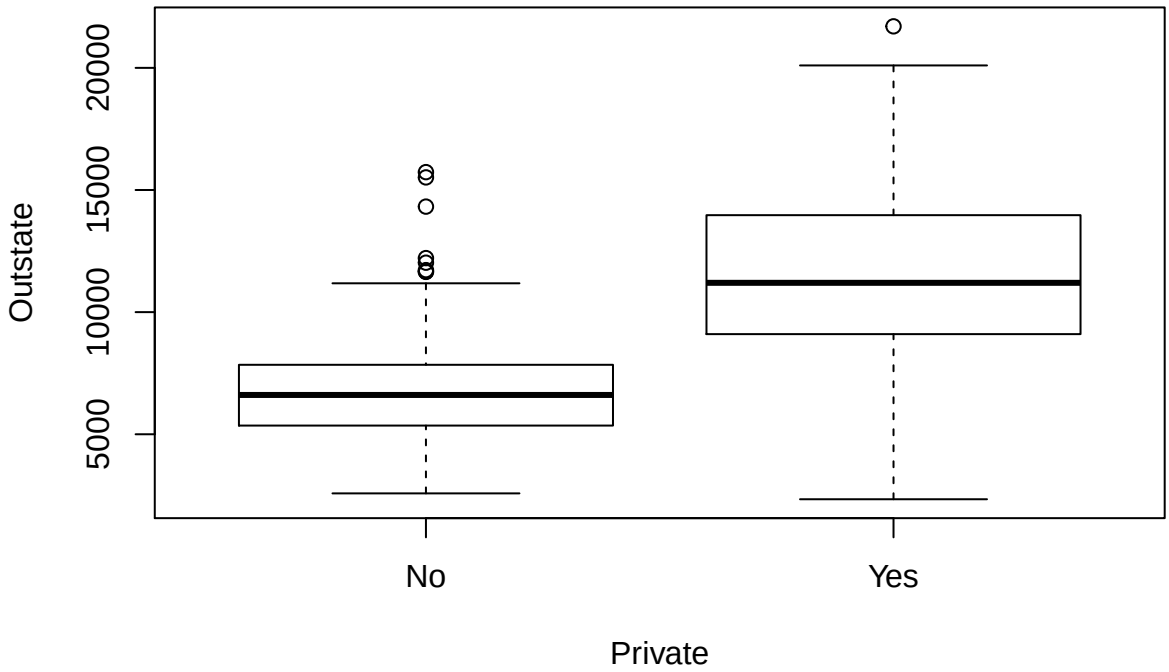
```
#Read the .csv file and store data into a dataframe
college = read.csv("college.csv")
#Remove the first coloumn from college
college <- college [,-1]
#Display summary of every variable in the data frame college
summary(college)
```

```
##  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
```

```r
#Get descreption for pairs function
?pairs
#Produce a scatterplot matrix of the first ten columns
pairs(college[,1:10])
```




```r
#Boxplots of Outstate versus Private
plot(college$Outstate ~ college$Private,main = "Outstate vs Private", ylab = "Outstate", xlab = "Private
```

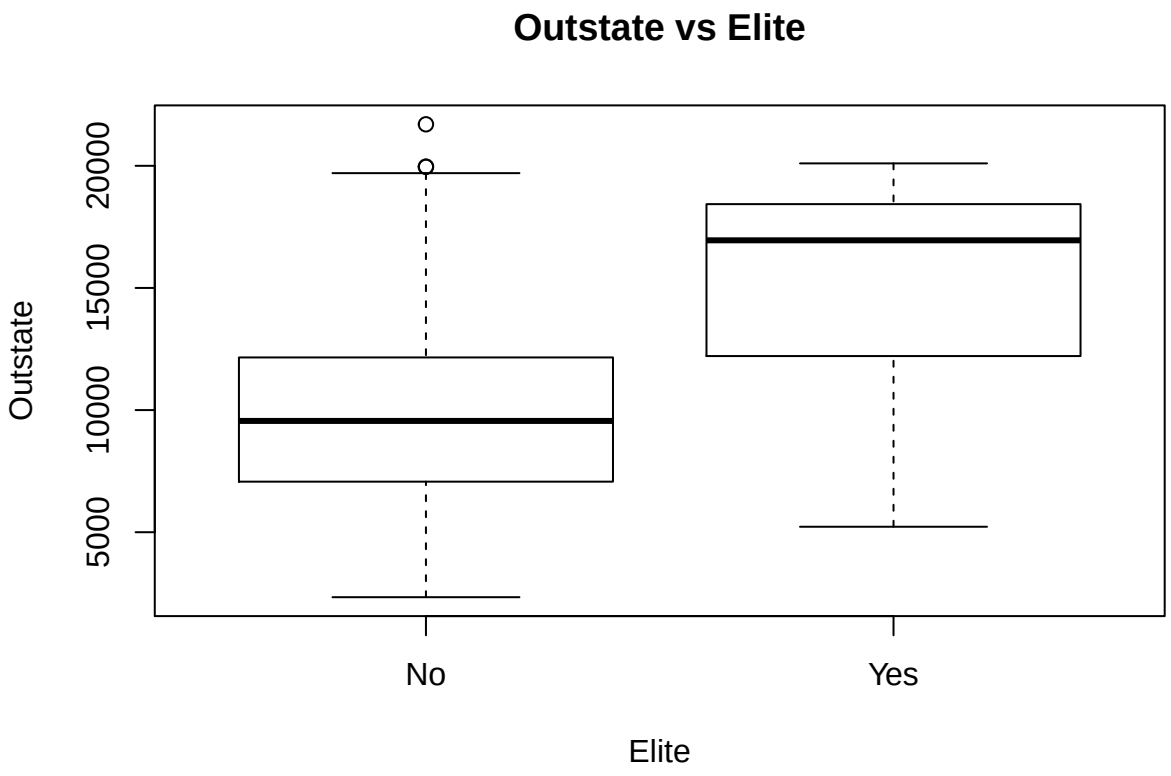



```
#Create a character vector Elite with a string "No" repeated for number of rows in college data frame
Elite <- rep ("No", nrow(college ))
#Assign string "Yes" to the Elite vector for each Top10perc variable of college is greater than 50
Elite [college$Top10perc >50] <- "Yes"
#Encode the Elite vector as a factor
Elite <- as.factor (Elite)
#Add Elite vector to data frame college. Elite will now appear as a column in college
college <- data.frame(college ,Elite)
#To check how many Elite universities are present
summary(college)
```

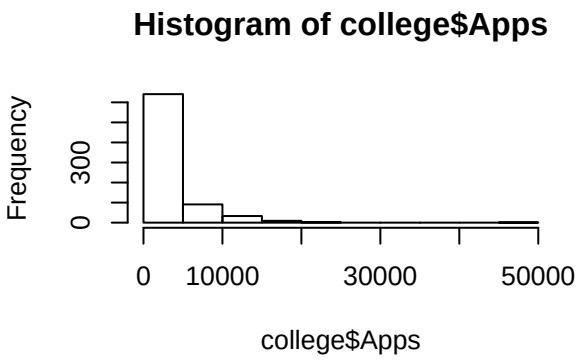
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            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
```

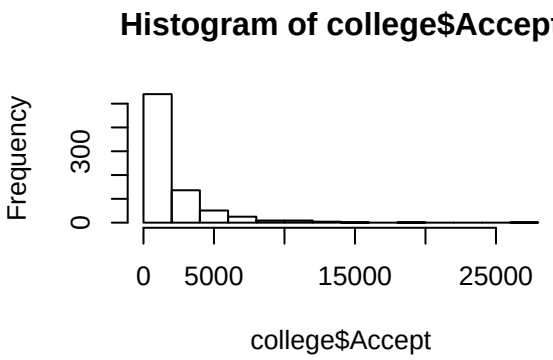
```
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   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      Elite
##  Min.   : 10.00   No :699
##  1st Qu.: 53.00   Yes: 78
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

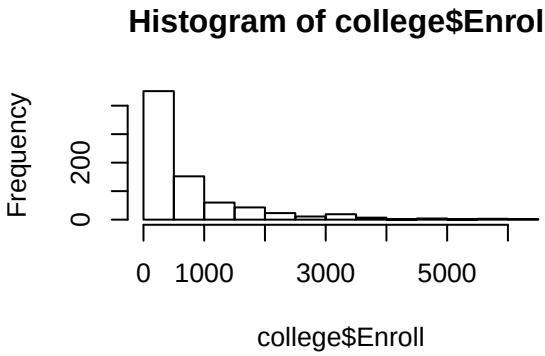
```r
#Plot Outstate vs Elite
plot(college$Outstate ~ college$Elite,main = "Outstate vs Elite", ylab = "Outstate", xlab = "Elite")
```

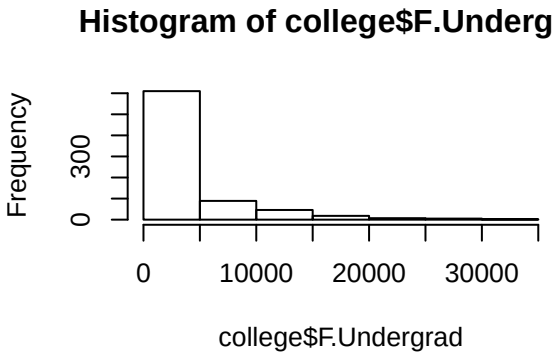



```r
#Divide the print window into 4 regions
par(mfrow=c(2,2))
#Histogram for number of pplications variable of college data frame
hist(college$Apps)
#Histogram for accepted applications variable of college data frame
hist(college$Accept)
#Histogram for number of enrolled students variable of college data frame
hist(college$Enroll)
#Histogram for F.Undergrad variable of college data frame
hist(college$F.Undergrad)
```










```
#Reset par function previously executed
par(mfrow=c(1,1))
```

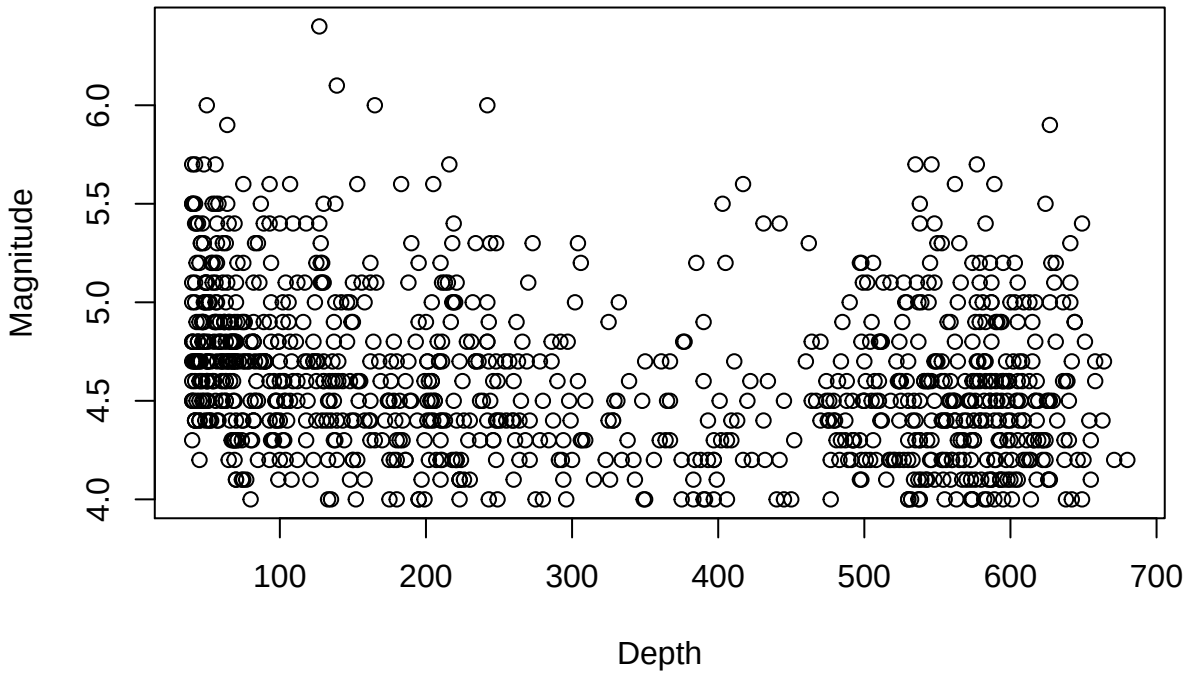
Question 3

```
#load plyr package
library(plyr)
#set sf variable to 0 for each of the corresponding year variable lesser than 1954 in baseball data fra
baseball$sf[baseball$year < 1954] = 0
#Replace "NA" entries of hbp variable by 0
baseball$hbp[is.na(baseball$hbp)] = 0
#Keep the only rows where ab variable is greater than 50 in baseball data frame
baseball = baseball[baseball$ab > 50,]
#Add a column obp to baseball data frame and calculate the respective value
baseball$obp=(baseball$h+baseball$bb+baseball$hbp)/(baseball$ab+baseball$bb+baseball$hbp+baseball$sf)
#Sort the baseball data frame according to obp variable
baseball = baseball[with(baseball, order(obp)),]
#Print top 5 rows with columns id,year and obp from baseball data frame
head(baseball[,c("id","year","obp")],5)
```

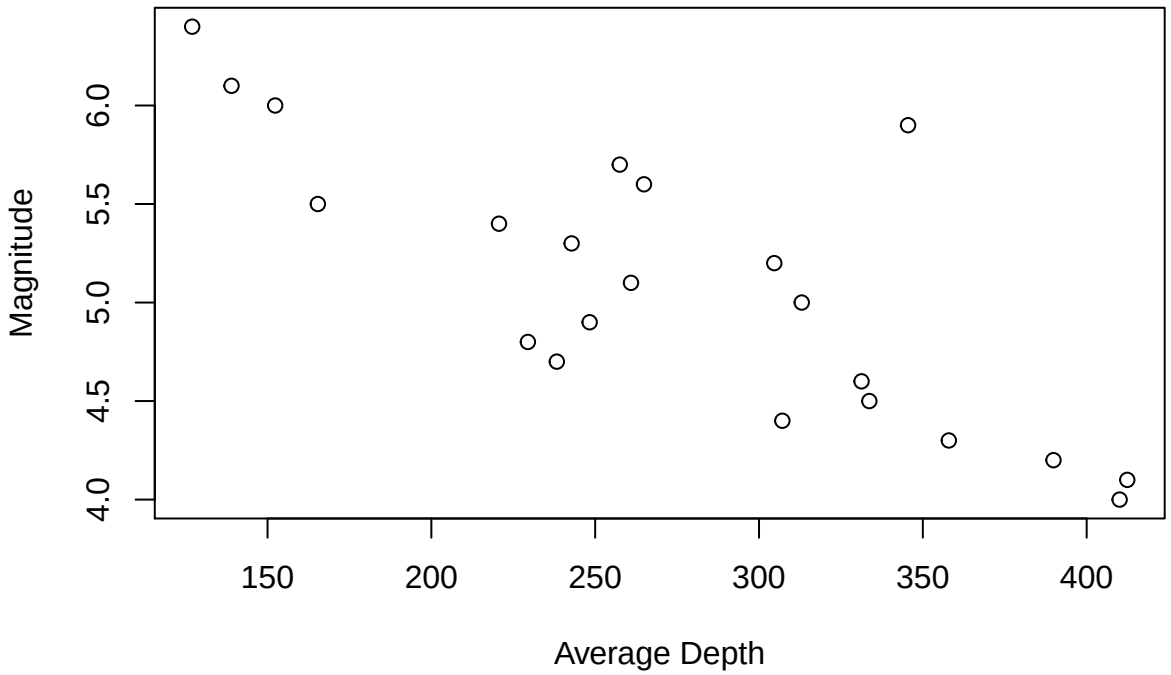
```
##              id year        obp
## 41939 aguirha01 1962 0.03947368
## 44890 simmocu01 1965 0.04687500
## 46933 cardwdo01 1968 0.04918033
## 83686 leiteal01 2003 0.05454545
## 25361 johnssi01 1933 0.05479452
```

Question 4

```
#Load datasets package
library("datasets")
#Plot Magnitude vs Depth from quake data frame
plot(quakes$mag ~ quakes$depth,xlab="Depth",ylab="Magnitude")
```




```
#Compute the average earthquake depth for each magnitude level
quakeAvgDepth=aggregate(quakes$depth ~ quakes$mag, data = quakes, mean)
#Rename the first column of quakeAvgDepth to "mag_level"
colnames(quakeAvgDepth)[1] = "mag_level"
#Rename the second column of quakeAvgDepth to "avg_eq_depth"
colnames(quakeAvgDepth)[2] = "avg_eq_depth"
#Plot mag_level vs avg_eq_depth of quakeAvgDepth dataframe
plot(quakeAvgDepth$mag_level ~ quakeAvgDepth$avg_eq_depth,xlab="Average Depth",ylab="Magnitude")
```




```
#There is a relation between earthquake depth and magnitude. In the seconf graph as the average depth i
```