Homework_1

Sai Saketh Boyanapalli

August 29, 2017

Packages required

```
library(moments)
library(plyr)
library(datasets)
```

Question 1

1 Using R: Vectors

(a) Create a vector with 10 numbers (3, 12, 6, -5, 0, 8, 15, 1, -10, 7) and assign it 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) Using the commands seq, min, and max with one line of code create a new vector y with 10 elements ranging from the minimum value of x to the maximum value of x.

```
y <- seq(min(x),max(x),length.out = 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) Compute the sum, mean, standard deviation, variance, mean absolute deviation, quartiles, and quintiles for x and y.

```
#sum of x and y
sum(x)
```

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

```
sum(y)
```

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

```
#mean of x and y
mean(x)
```

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

```
mean(y)
```

```
## [1] 2.5

#standard deviation of x and y
sd(x)

## [1] 7.572611

sd(y)

## [1] 8.41014

#variance of x and y
var(x)

## [1] 57.34444

var(y)

## [1] 70.73045

#mean absolute deviation of x and y
mad(x)

## [1] 5.9304

mad(y)

## [1] 10.29583

#quartiles of x and y
quantile(x)

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

quantile(y)

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

#quintiles of x and y
quantile(x,probs = seq(0,1,0.2))

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

quantile(y,probs = seq(0,1,0.2))

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

(d) Create a new 7 element vector z by using R to randomly sample from x with replacement.

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

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

(e) Find a package (or packages) that provide the statistical measures skewness and kurtosis. Use the appropriate functions from the package to calculate the skewness and kurtosis of x.

The Skewness and Kurtosis functions are available in moments package.

```
skewness(x)

## [1] -0.3123905

kurtosis(x)

## [1] 2.355328
```

(f) Use t.test() to compute a statistical test for differences in means between the vectors x and y. Are the differences in means signifcant?

```
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
```

(g) Sort the vector x and re-run the t-test as a paired t-test.

```
sorted_x <- sort(x)
t.test(sorted_x,y,paired = TRUE)

##
##  Paired t-test
##
## data:  sorted_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) Create a logical vector that identifies which numbers in x are negative.

```
neq_x <- x[x<0]
```

(i) Use this logical vector to remove all entries with negative numbers from x. (Make sure to overwrite the vector x so that the new vector x has 8 elements!)

```
x <- x[!x %in% neq_x]
x

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

Question 2

Using R: Introductory data exploration

This exercise relates to the College data set, which can be found in the file .csv" in D2L. The file contains a number of variables for 777 different universities and colleges in the US.

(a)

Use the read.csv() function to read the data into a data frame in R. Call the data frame college. Make sure that you have the directory set to the correct location for the data (or that the data is in the same directory as the RStudio project).

```
college <- read.csv("college.csv",header = TRUE)
head(college)

##                              X Private Apps Accept Enroll Top10perc
## 1 Abilene Christian University     Yes 1660   1232    721        23
## 2           Adelphi University     Yes 2186   1924    512        16
## 3               Adrian College     Yes 1428   1097    336        22
## 4          Agnes Scott College     Yes  417    349    137        60
## 5    Alaska Pacific University     Yes  193    146     55        16
## 6            Albertson College     Yes  587    479    158        38
##   Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD
## 1        52        2885         537     7440       3300   450     2200  70
## 2        29        2683        1227    12280       6450   750     1500  29
## 3        50        1036          99    11250       3750   400     1165  53
## 4        89         510          63    12960       5450   450      875  92
## 5        44         249         869     7560       4120   800     1500  76
## 6        62         678          41    13500       3335   500      675  67
##   Terminal S.F.Ratio perc.alumni Expend Grad.Rate
## 1       78      18.1          12   7041        60
## 2       30      12.2          16  10527        56
## 3       66      12.9          30   8735        54
## 4       97       7.7          37  19016        59
```

```
## 5       72      11.9           2  10922        15
## 6       73       9.4          11   9727        55
```

(b)

```
# this will assisgn the row names of the data frame to college names
rownames (college) <- college [,1]
View (college )
head(college)
```

```
##                                                 X Private Apps
## Abilene Christian University Abilene Christian University     Yes 1660
## Adelphi University                     Adelphi University     Yes 2186
## Adrian College                             Adrian College     Yes 1428
## Agnes Scott College                   Agnes Scott College     Yes  417
## Alaska Pacific University       Alaska Pacific University     Yes  193
## Albertson College                       Albertson College     Yes  587
##                              Accept Enroll Top10perc Top25perc F.Undergrad
## Abilene Christian University   1232    721        23        52        2885
## Adelphi University             1924    512        16        29        2683
## Adrian College                 1097    336        22        50        1036
## Agnes Scott College             349    137        60        89         510
## Alaska Pacific University       146     55        16        44         249
## Albertson College               479    158        38        62         678
##                              P.Undergrad Outstate Room.Board Books
## Abilene Christian University         537     7440       3300   450
## Adelphi University                  1227    12280       6450   750
## Adrian College                        99    11250       3750   400
## Agnes Scott College                   63    12960       5450   450
## Alaska Pacific University            869     7560       4120   800
## Albertson College                     41    13500       3335   500
##                              Personal PhD Terminal S.F.Ratio perc.alumni
## Abilene Christian University     2200  70       78      18.1          12
## Adelphi University               1500  29       30      12.2          16
## Adrian College                   1165  53       66      12.9          30
## Agnes Scott College               875  92       97       7.7          37
## Alaska Pacific University        1500  76       72      11.9           2
## Albertson College                 675  67       73       9.4          11
##                              Expend Grad.Rate
## Abilene Christian University   7041        60
## Adelphi University            10527        56
## Adrian College                 8735        54
## Agnes Scott College           19016        59
## Alaska Pacific University     10922        15
## Albertson College              9727        55
```

```
# Now that we have assigned each row to the appropriate college name we can
remove the column with college names
college <- college [,-1]
head(college)
```

```
##                               Private Apps Accept Enroll Top10perc
## Abilene Christian University     Yes 1660   1232    721        23
## Adelphi University               Yes 2186   1924    512        16
## Adrian College                   Yes 1428   1097    336        22
## Agnes Scott College              Yes  417    349    137        60
## Alaska Pacific University        Yes  193    146     55        16
## Albertson College                Yes  587    479    158        38
##                               Top25perc F.Undergrad P.Undergrad Outstate
## Abilene Christian University        52        2885         537     7440
## Adelphi University                  29        2683        1227    12280
## Adrian College                      50        1036          99    11250
## Agnes Scott College                 89         510          63    12960
## Alaska Pacific University           44         249         869     7560
## Albertson College                   62         678          41    13500
##                               Room.Board Books Personal PhD Terminal
## Abilene Christian University       3300   450     2200  70       78
## Adelphi University                 6450   750     1500  29       30
## Adrian College                     3750   400     1165  53       66
## Agnes Scott College                5450   450      875  92       97
## Alaska Pacific University          4120   800     1500  76       72
## Albertson College                  3335   500      675  67       73
##                               S.F.Ratio perc.alumni Expend Grad.Rate
## Abilene Christian University      18.1          12   7041        60
## Adelphi University                12.2          16  10527        56
## Adrian College                    12.9          30   8735        54
## Agnes Scott College                7.7          37  19016        59
## Alaska Pacific University         11.9           2  10922        15
## Albertson College                  9.4          11   9727        55
```

(c)

(i) summary() function will give us the summary of the data

```
summary(college)
```

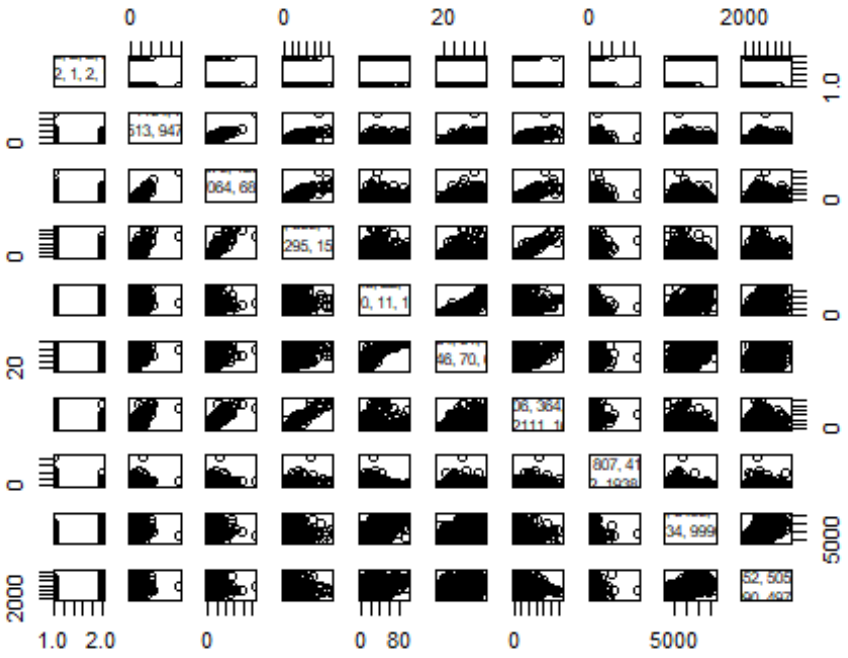
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00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

```
?pairs # using ? before a function shows us the documentation for it.
```

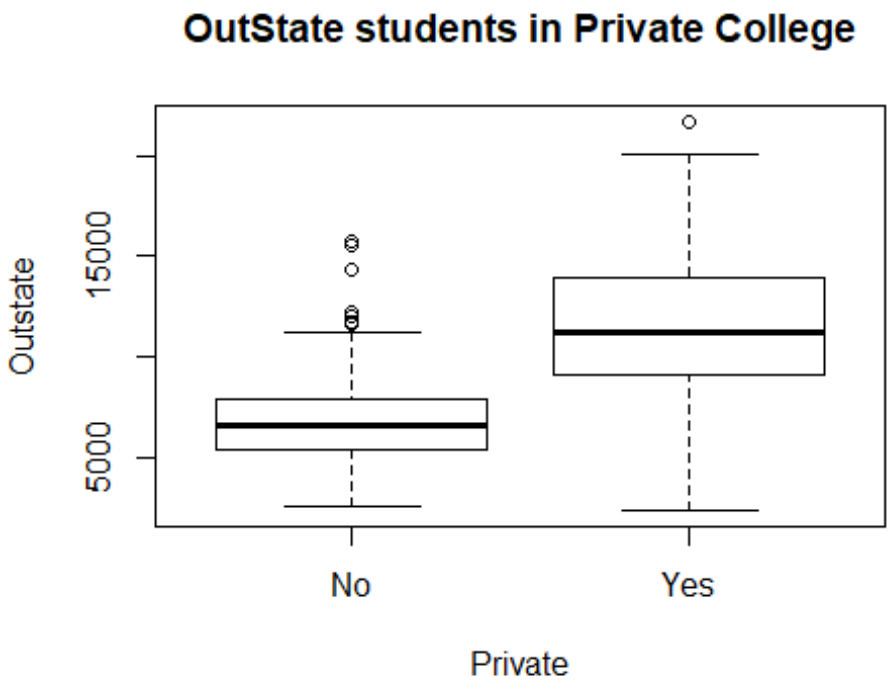
```
## starting httpd help server ... done
```

```
pairs(college[,1:10],college)
```



(iii)

```
# This function creates a boxplot for no of OutState students in private
colleges
plot(college$Private,college$Outstate,main = "OutState students in Private
College", xlab = "Private",ylab = "Outstate" )
```



iv. Using the following bit of code you will create a new qualitative variable, called Elite by binning the Top10perc variable. That is, Elite will classify the universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%. Add comments to each line below explaining what the corresponding code is doing and then run the code.

```
Elite <- rep ("No", nrow(college )) # this line creates a list with value
'NO' with the length set to no of rows in college. using rep function.
Elite [college$Top10perc >50] <- "Yes" # In this line the college with
top10percent greater than 50, the elite value is set to "Yes"
Elite <- as.factor (Elite) #The values in Elite are factored to two levels
college <- data.frame(college ,Elite) # Elite is addes as one of the
variables to college data frame.
```

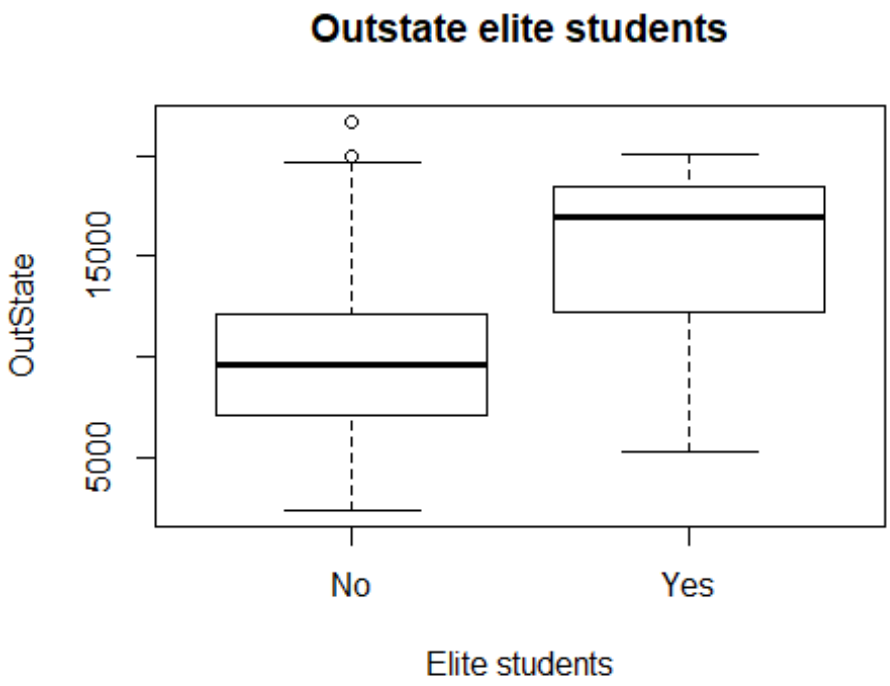
v. Use the summary() function to see how many elite universities there are.

```
summary(college$Elite)

##  No Yes
## 699  78
```

we can see there are 78 elite students in total. vi.

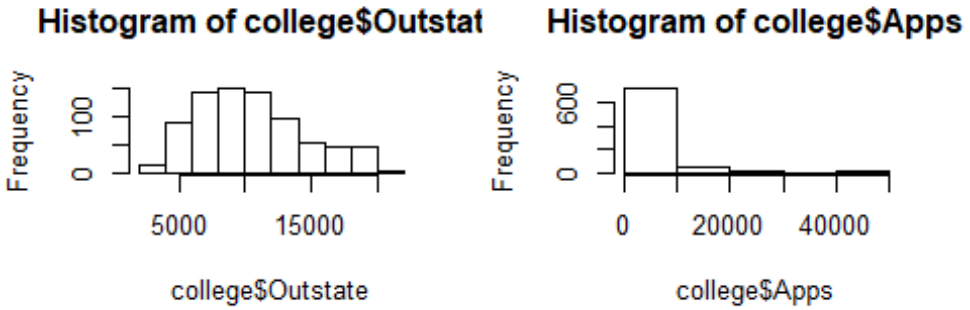
```
plot(college$Elite,college$Outstate,main = "Outstate elite students", xlab =
"Elite students", ylab = "OutState")
```

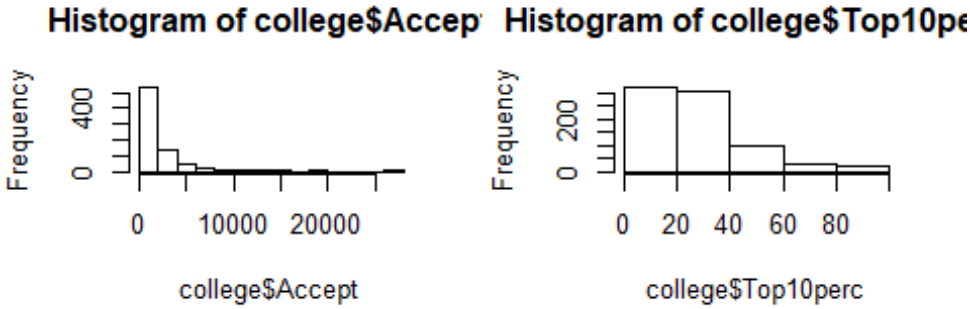


we can clearly see that there are more outstate elite students.

vii. Use the hist() function to produce some histograms with diering numbers of bins for a few of the quantitative variables.

```
par(mfrow=c(2,2)) # this command will divide plot window into 4 sections
hist(college$Outstate,breaks = 10) # this will create a Histogram.
hist(college$Apps,breaks = 5) # breaks is used to set no of bins.
hist(college$Accept,breaks = 15)
hist(college$Top10perc,breaks = 6)
```





##Question 3

#Using R: Manipulating data in data frames (a) Load the data frame baseball in the plyr package. Use ?baseball to get information about the data set and definitions for the variables.

```
data("baseball") # data is used to load a specific data set
?baseball
```

(b) You will calculate the on base percentage for each player, but first clean up the data:

- Before 1954, sacrifice flies were counted as part of sacrifice hits, so for players before 1954,sacrifice flies (i.e. the variable sf) should be set to 0.

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

- Hit by pitch (the variable hbp) is often missing { set these missings to 0.

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

## Warning in is.na(baseball$hpb): is.na() applied to non-(list or vector) of
## type 'NULL'
```

- Exclude all player records with fewer than 50 at bats (the variable ab).

```
baseball <- baseball[-c(baseball$ab < 50), ]
```

(c) Compute on base percentage in the variable obp according to the formula:

```
obp <-((baseball$h + baseball$bb + baseball$hbp)/(baseball$ab + baseball$bb +
baseball$hbp + baseball$sf))
baseball <- data.frame(baseball,obp)
```

(d) Sort the data based on the computed obp and print the year, player name, and on base percentage for the top five records based on this value.

```
Sorted_obp <- baseball[order(-obp) , ] # (-obp indicates decreasing order)
top_five <- Sorted_obp[1:5, ]
top_five[,c("year","id","obp")]

##       year        id obp
## 6074  1894 brownpe01   1
## 13924 1913 griffcl01   1
## 14537 1914 griffcl01   1
## 16076 1916 davisha01   1
## 17429 1918 haineje01   1
```

Question 4

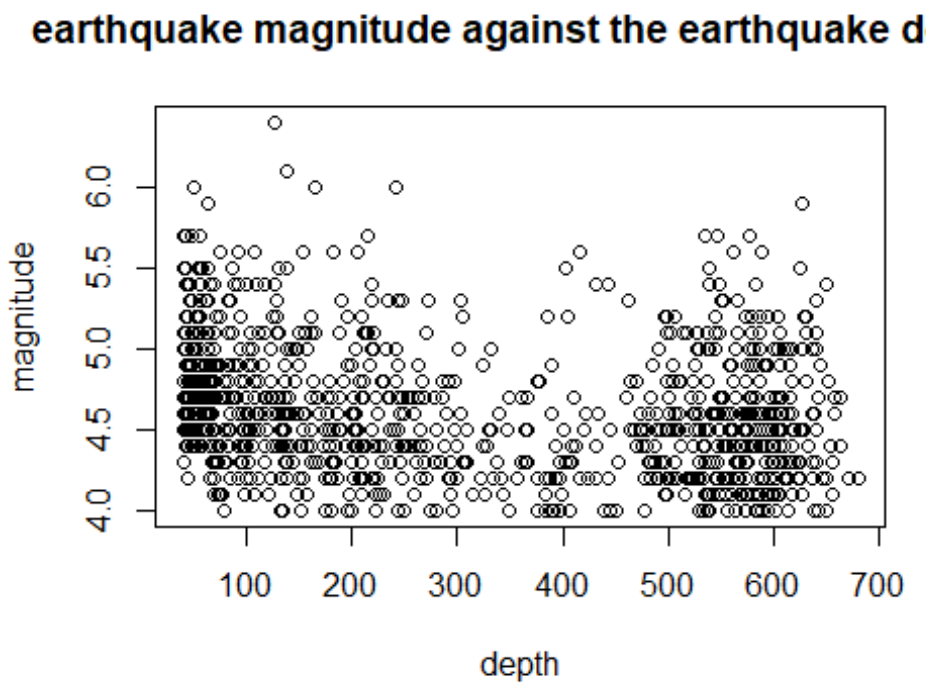
Using R: aggregate() function

(a) Load the quakes data from the datasets package.

```
data("quakes")
```

(b) Plot the recorded earthquake magnitude against the earthquake depth using the plot command.

```
plot(quakes$depth,quakes$mag,main ="earthquake magnitude against the
earthquake depth", xlab = "depth",ylab = "magnitude")
```



(c) Use aggregate to compute the average earthquake depth for each magnitude level. Store these results in a new data frame named quakeAvgDepth.

```
quakeAvgDepth <- aggregate(quakes$depth ~ quakes$mag,quakes, FUN = mean)
```

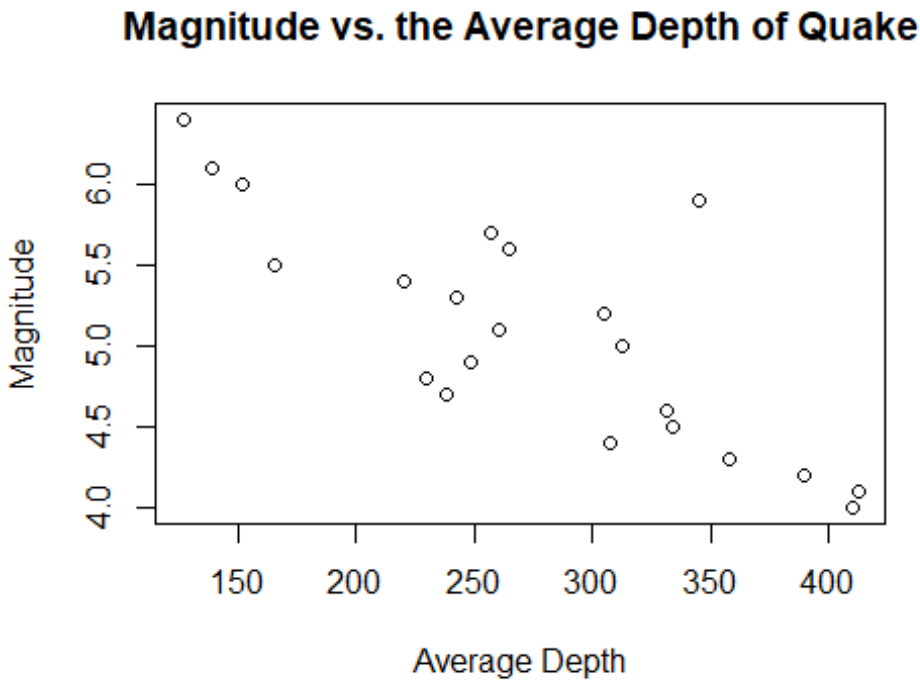
(d) Rename the variables in quakeAvgDepth to something meaningful.

```
colnames(quakeAvgDepth) <- c("Magnitude of
Earthquake","corresponding_Average_Depth")
head(quakeAvgDepth)
```

```
##   Magnitude of Earthquake corresponding_Average_Depth
## 1                     4.0                    410.0652
## 2                     4.1                    412.4000
## 3                     4.2                    389.8778
## 4                     4.3                    357.9294
## 5                     4.4                    307.1188
## 6                     4.5                    333.6729
```

(e) Plot the magnitude vs. the average depth.

```
plot(quakeAvgDepth$corresponding_Average_Depth,quakeAvgDepth$`Magnitude of
Earthquake`,main="Magnitude vs. the Average Depth of Quake",xlab="Average
Depth",ylab="Magnitude")
```



(f) From the two plots, do you think there is a relationship between earthquake depth and magnitude?

From the Two graphs we can see that the depth of the quake decreases with the increase in magnitude.