IDA HW1 - Vignesh Murugan

2024-08-28

1st question

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

#1b
y <- seq(from=min(x), to=max(x), length.out=10)

# 1c
sum(x)
```

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

```
sum(y)
```

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

```
mean(x)
```

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

```
mean(y)
```

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

```
sd(x)
```

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

```
sd(y)
```

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

```r
var(x)
```

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

```r
var(y)
```

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

```r
mad(x)
```

```
## [1] 5.9304
```

```r
mad(y)
```

```
## [1] 10.29583
```

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

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

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

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

```r
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
```

```r
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
```

```r
# 1d
z <- sample(x, size=7, replace=TRUE )
z
```

```
## [1]  -5 -10   1  12  12   6   6
```

```r
# 1e
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
```

Differences in means are not significant

```r
# 1f
x <- x[order(x)]
x
```

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

```r
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 mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.05440584  2.45440584
## sample estimates:
## mean difference
##             1.2
```

```r
# 1g
negative_logical_vector <- x < 0
negative_logical_vector
```

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

```r
# 1h
non_negative_logical_vector <- x >= 0
x <- x[non_negative_logical_vector]
x
```

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

2nd question

```
# 2a
col1 <- c(1,2,3,NA,5)
col2 <- c(4,5,6,89,101)
col3 <- c(45,NA,66,121,201)
col4 <- c(14,NA,13,NA,27)
X <- rbind (col1,col2,col3,col4)

is.na(X)
```

```
##       [,1]  [,2]  [,3]  [,4]  [,5]
## col1 FALSE FALSE FALSE  TRUE FALSE
## col2 FALSE FALSE FALSE FALSE FALSE
## col3 FALSE  TRUE FALSE FALSE FALSE
## col4 FALSE  TRUE FALSE  TRUE FALSE
```

```
X[!complete.cases(X),]
```

```
##      [,1] [,2] [,3] [,4] [,5]
## col1    1    2    3   NA    5
## col3   45   NA   66  121  201
## col4   14   NA   13   NA   27
```

```
# 2b
y <- c(3,12,99,99,7,99,21)
y[y == 99] <- NA
y
```

```
## [1]  3 12 NA NA  7 NA 21
```

```
sum(is.na(y))
```

```
## [1] 3
```

3rd question

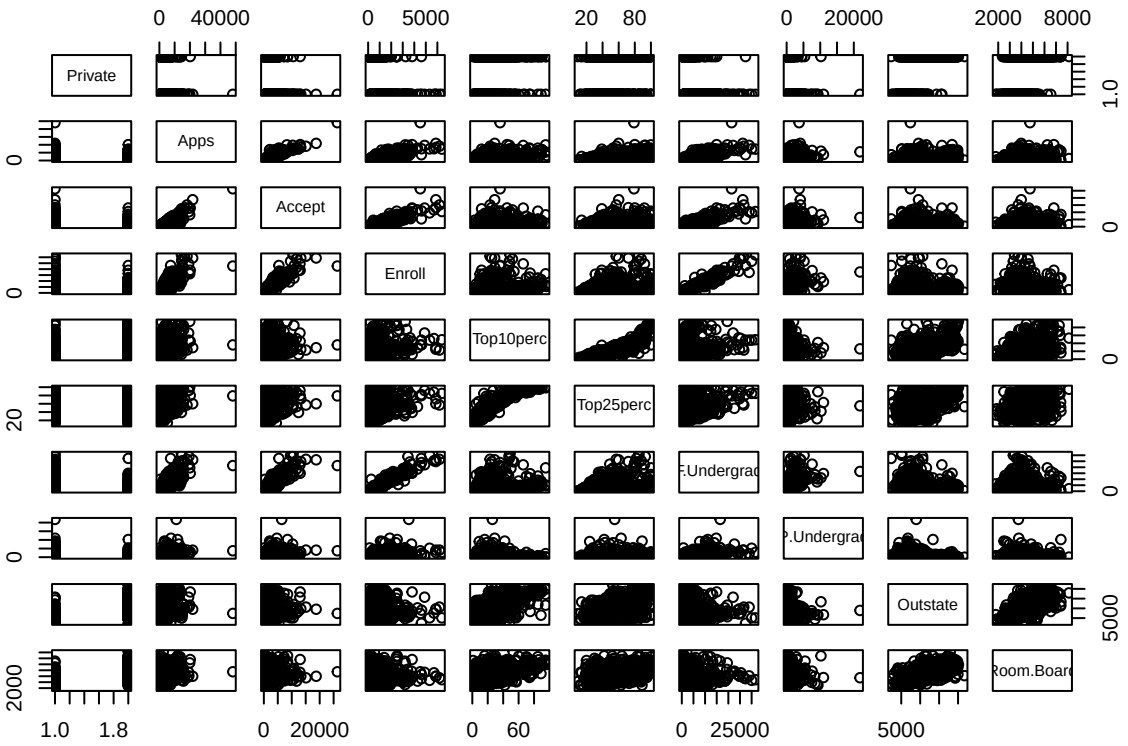
```r
# 3a
college <- read.csv("college.csv")
college <- as.data.frame(college)

# 3b
rownames(college) <- college[,1]

if (interactive()) {
  View(college)
}

college <- college[,-1]

# 3c
#i
summary(college)
```

```
##    Private                Apps           Accept          Enroll
##  Length:777         Min.   :   81   Min.   :   72   Min.   :  35
##  Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
##  Mode  :character   Median : 1558   Median : 1110   Median : 434
##                     Mean   : 3002   Mean   : 2019   Mean   : 780
##                     3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902
##                     Max.   :48094   Max.   :26330   Max.   :6392
##    Top10perc       Top25perc      F.Undergrad     P.Undergrad
##  Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0
##  1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0
##  Median :23.00   Median : 54.0   Median : 1707   Median :  353.0
##  Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3
##  3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0
##  Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0
##     Outstate       Room.Board       Books           Personal
##  Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250
##  1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
##  Median : 9990   Median :4200   Median : 500.0   Median :1200
##  Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341
##  3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700
##  Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800
##       PhD            Terminal       S.F.Ratio      perc.alumni
##  Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00
##  1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00
##  Median : 75.00   Median : 82.0   Median :13.60   Median :21.00
##  Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74
##  3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00
##  Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00
##      Expend        Grad.Rate
##  Min.   : 3186   Min.   : 10.00
##  1st Qu.: 6751   1st Qu.: 53.00
##  Median : 8377   Median : 65.00
##  Mean   : 9660   Mean   : 65.46
##  3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :56233   Max.   :118.00
```

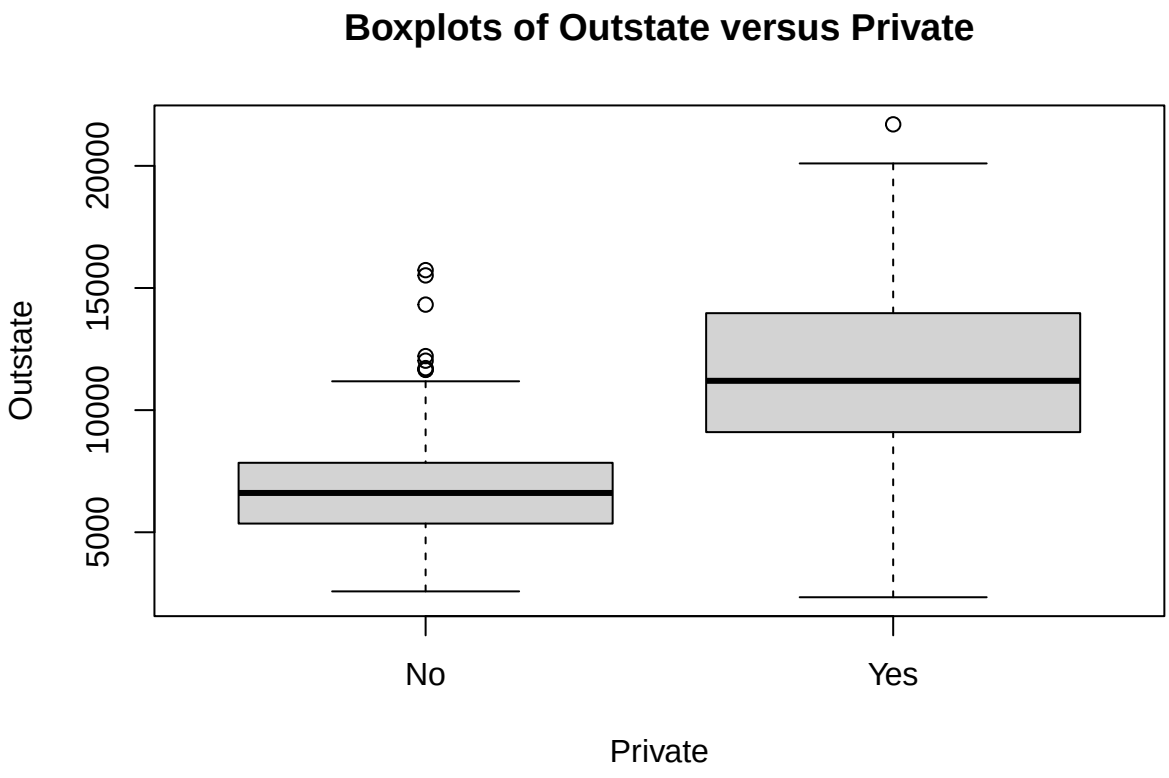
```r
college$Private <- as.factor(college$Private)

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




```
#iii
plot(college$Private, college$Outstate, xlab= "Private", ylab= "Outstate",
     main="Boxplots of Outstate versus Private")
```

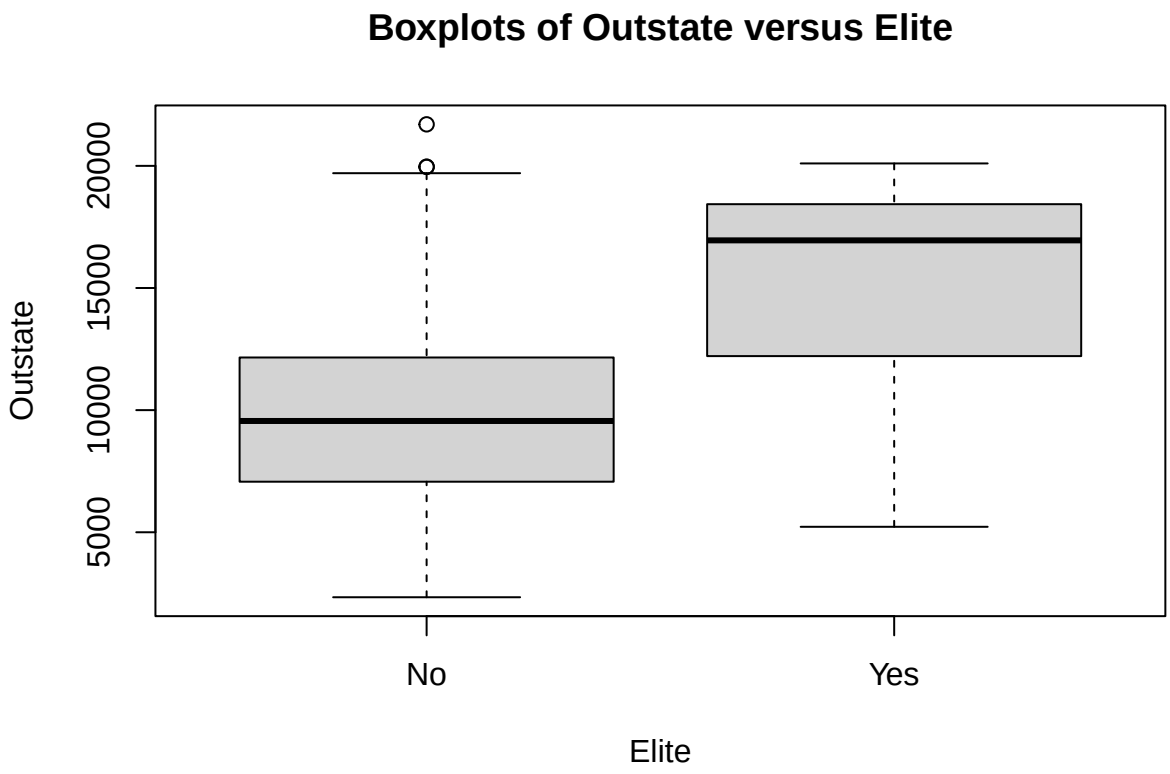



```
# iv
Elite <- rep("No", nrow(college))
# Create a character vector of size 777 with all elements being "No" and assigns it to variable called E

Elite[college$Top10perc >50] <- "Yes"
# If the values in "Top10Perc" variable are greater than 50 then those indexes in Elite are turned to y

Elite <- as.factor(Elite)
# we convert the character vector to factor vector

college <- data.frame(college ,Elite)
# creates a new variable in college called Elite

# v
summary(college)
```

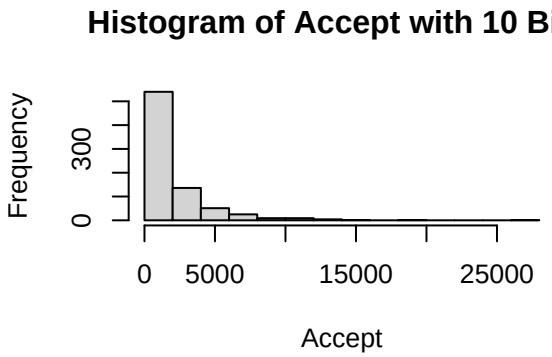
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00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

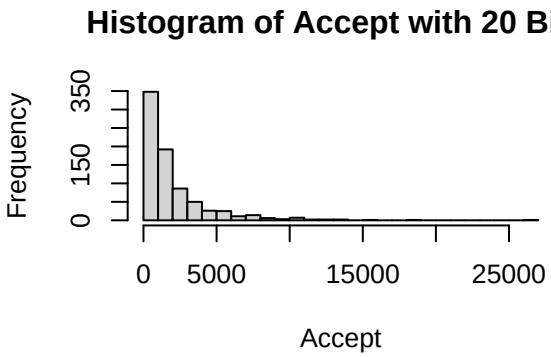
```
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           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
```

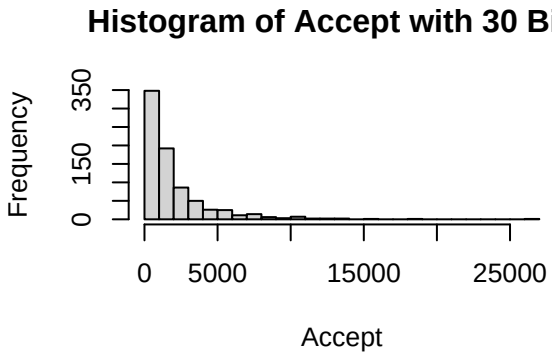
```r
# vi
plot(college$Elite, college$Outstate, xlab= "Elite", ylab= "Outstate",
     main="Boxplots of Outstate versus Elite")
```

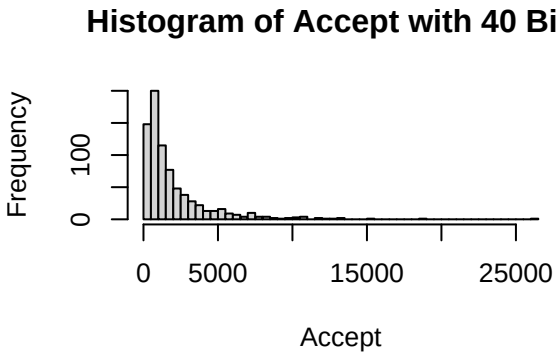



```
# vii
par(mfrow=c(2,2))

hist(college$Accept, breaks=10,
     main="Histogram of Accept with 10 Bins", xlab="Accept")

hist(college$Accept, breaks=20,
     main="Histogram of Accept with 20 Bins", xlab="Accept")

hist(college$Accept, breaks=30,
     main="Histogram of Accept with 30 Bins", xlab="Accept")

hist(college$Accept, breaks=40,
     main="Histogram of Accept with 40 Bins", xlab="Accept")
```









4th question

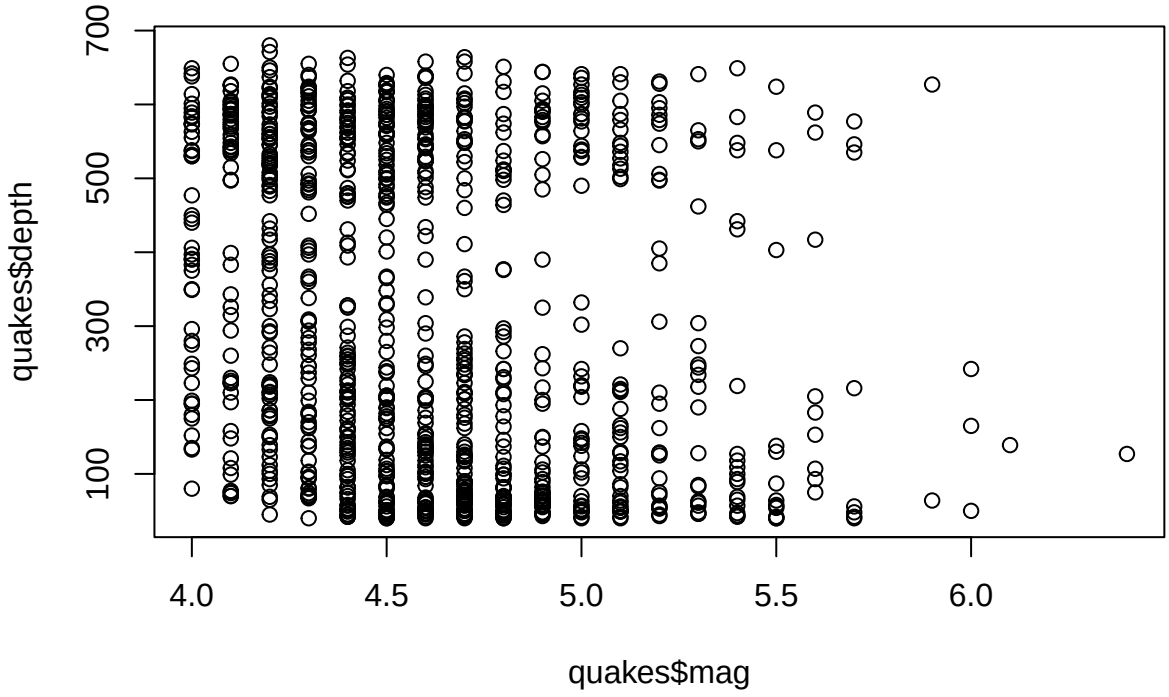
```
library(plyr)

# 4a
baseball <- baseball

# 4b
baseball$sf[baseball$year < 1954] <- 0

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

baseball <- baseball[baseball$ab >= 50,]

# 4c
baseball$obp <- with(baseball, (h + bb + hbp) / (ab + bb + hbp + sf))
```

```r
# 4d
top_players <- baseball[order(-baseball$obp),
                        c("year", "id", "obp")]
head(top_players, 5)
```

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

5th question

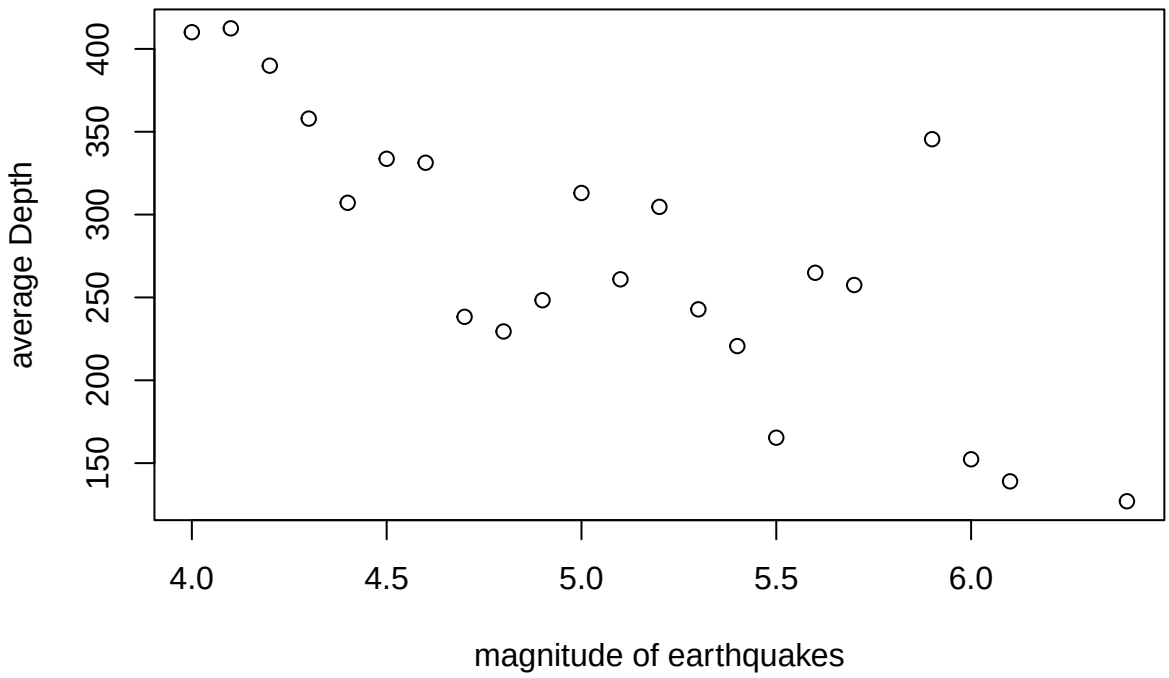
```r
# 5a
quakes <- quakes

# 5b
par(mfrow=c(1,1))
plot(quakes$mag, quakes$depth)
```




```
# 5c
quakeAvgDepth <- aggregate(depth ~ mag, data = quakes, mean)

# 5d
names(quakeAvgDepth)[2] <- "AvgDepth"
names(quakeAvgDepth)
```

```
## [1] "mag"      "AvgDepth"
```

```
# 5e
plot(quakeAvgDepth$mag,quakeAvgDepth$AvgDepth,
     xlab = "magnitude of earthquakes", ylab = "average Depth")
```



5f

From the plots we can say that as magnitude increases the depth decreases I believe there is a reason for this,so I spent sometime surfing the web.It seems that higher magnitude quakes on higher depths are less harmful towards earths surface.So You could say that the data only contains earthquakes which cause substantial damage to property on surface.