MATH203 A1

created by Simon Hsu ID: 260610820

Q2.20

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
q1table = data_frame(goalkeeper= rep(c("Dive left","Stay middle","Dive right"),each=3),
match = rep(c("Team behind","Tied", "Team ahead"),3),
proportion = c(0.29,.48,0.51,0,.03,.01,0.71,0.49,0.48))
q1table
```

```
## # A tibble: 9 × 3
##    goalkeeper       match proportion
##         <chr>       <chr>      <dbl>
## 1   Dive left Team behind       0.29
## 2   Dive left        Tied       0.48
## 3   Dive left  Team ahead       0.51
## 4 Stay middle Team behind       0.00
## 5 Stay middle        Tied       0.03
## 6 Stay middle  Team ahead       0.01
## 7  Dive right Team behind       0.71
## 8  Dive right        Tied       0.49
## 9  Dive right  Team ahead       0.48
```

```
library(ggplot2)
colours <-c("red","green","blue")
barplot(as.matrix(q1table[3:3]), main="My side by side barchart", xlab="goalkeeper", ylab = "proportion
```

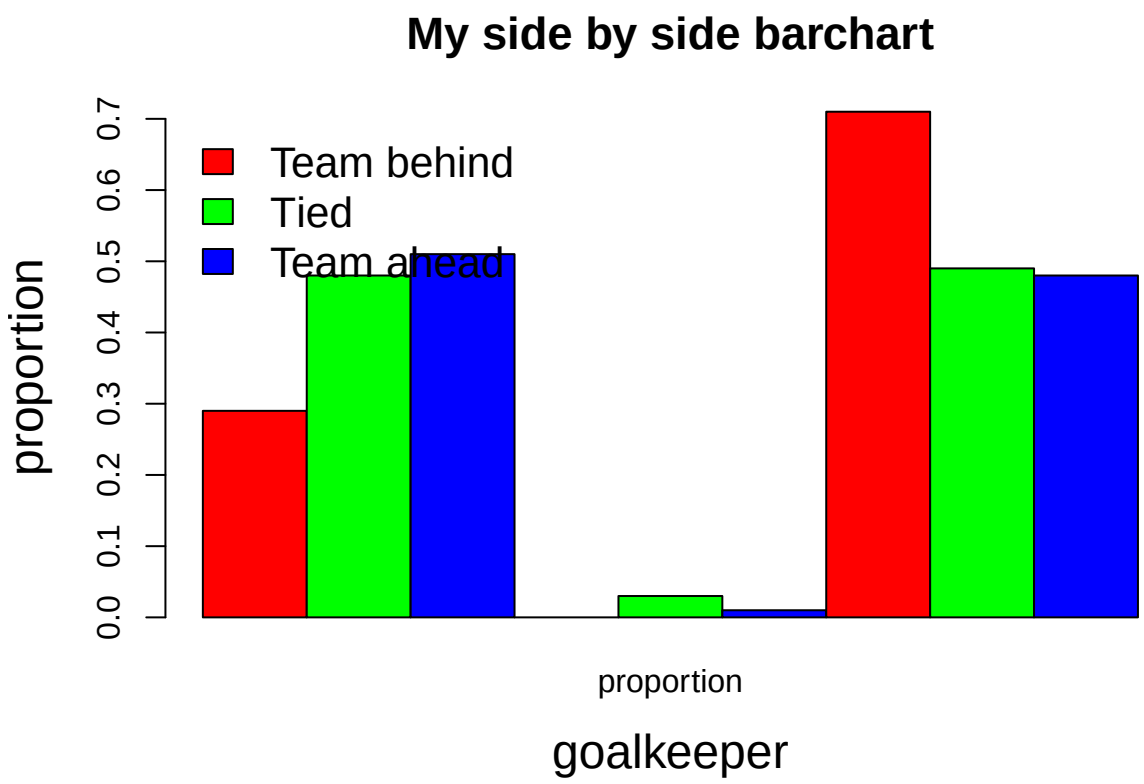
```
## Warning in plot.window(xlim, ylim, log = log, ...): "q1table.name" is not a
## graphical parameter
```

```
## Warning in axis(if (horiz) 2 else 1, at = at.l, labels = names.arg, lty =
## axis.lty, : "q1table.name" is not a graphical parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "q1table.name" is not a graphical parameter

## Warning in axis(if (horiz) 1 else 2, cex.axis = cex.axis, ...):
## "q1table.name" is not a graphical parameter
```

```
legend("topleft", c("Team behind","Tied", "Team ahead"), cex=1.3, bty="n", fill=colours)
```



The first 3 bars represents `dive left`, the middle three represents `stay middle`, and the last three represents `dive right`. (PS. i tried putting the goalkeeper position names on the chart and getting rid of the word 'proportion' on x axis but they don't seem to work)

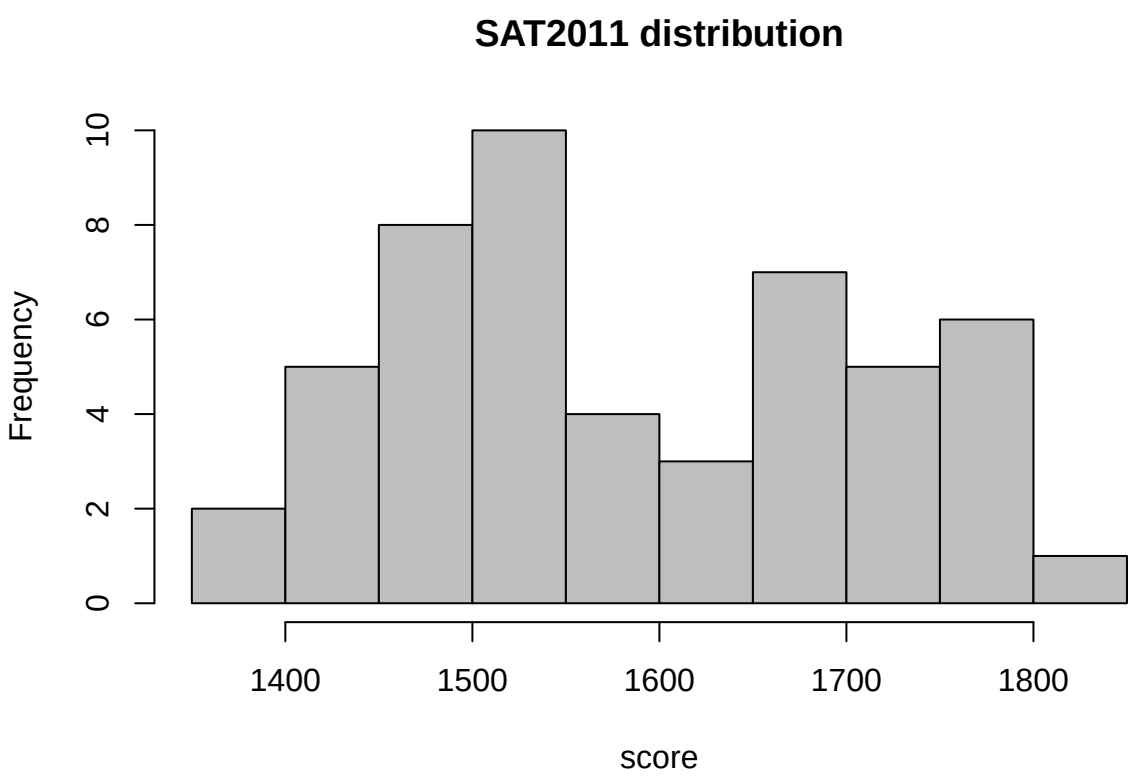
My inference: According to the side by side barchart, diving right while the team is behind is the trend here, while no matter in what situations, it is rare for the goal keeper to stay middle, and lastly, while the team is ahead or tied the goalkeeper ten to dive left.

Q2.46 [do not do (d)]

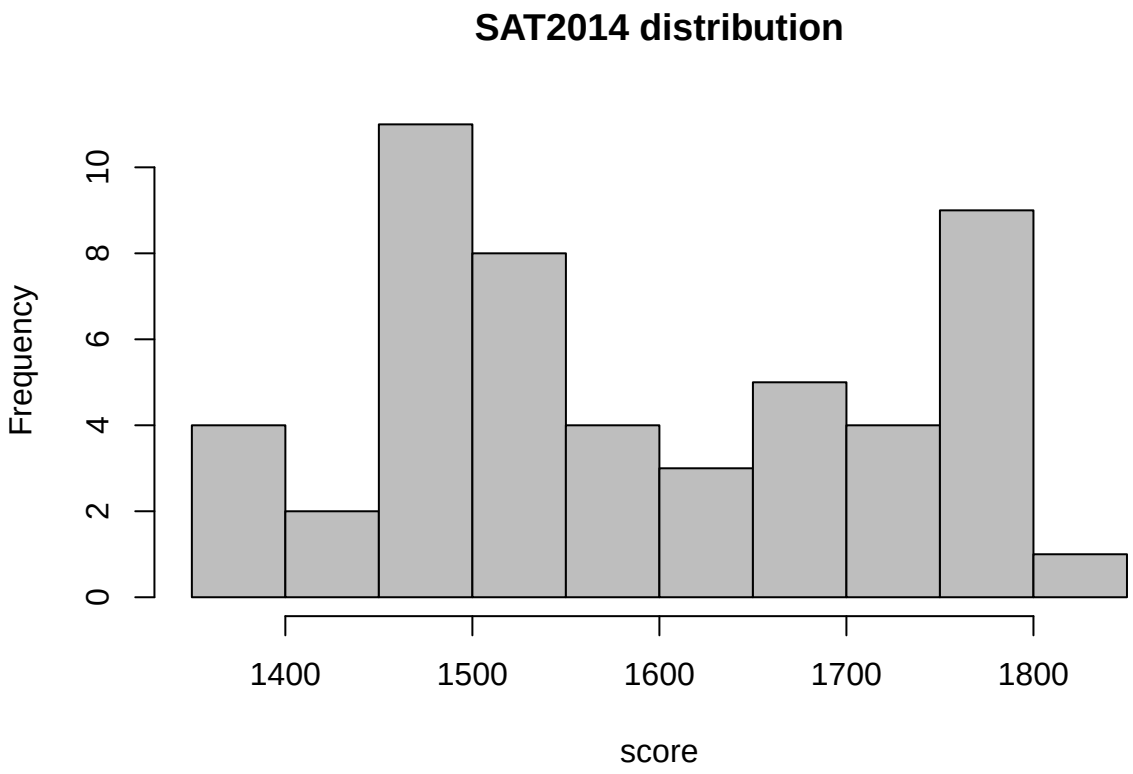
Replace part (d) with the the following: d) What is the mean, median, standard deviation and range of the SAT scores in 2011? for 2014?

This problem uses the "SAT" dataset

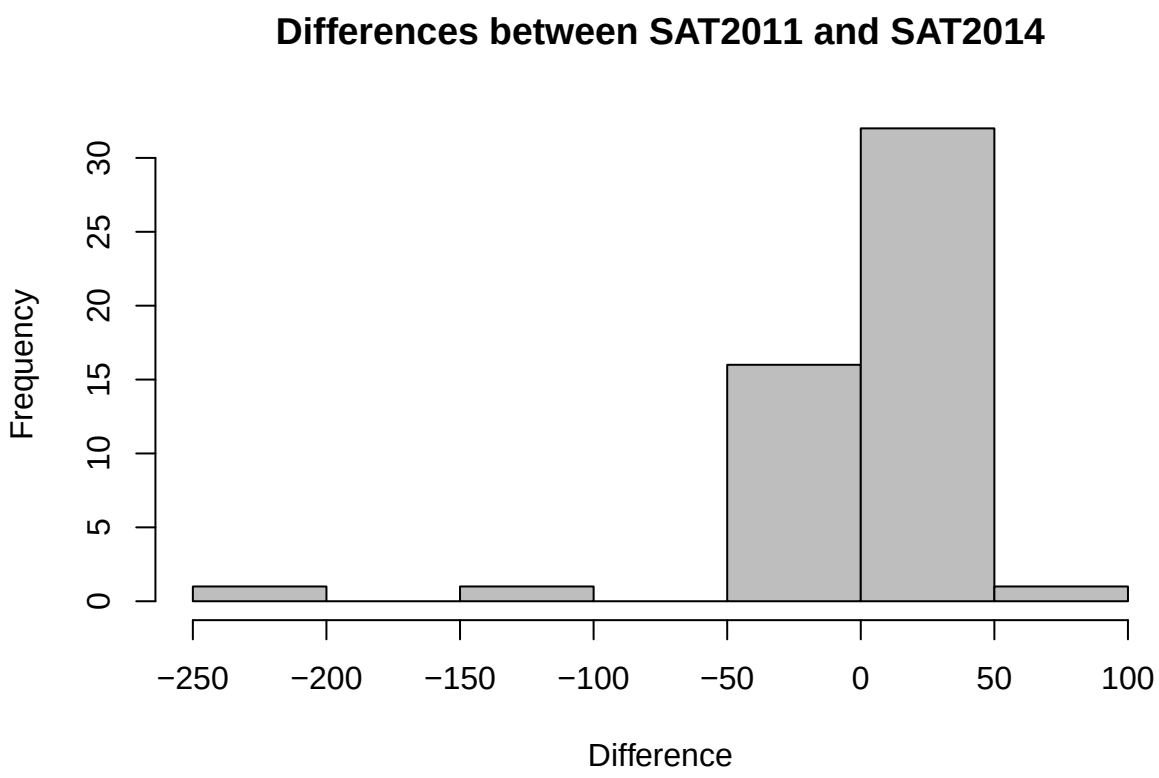
```
sat<-read.csv("SAT.csv")
#part a: The differences in thedistributions of state score from 2011 to 2014 are there are more higher
hist(sat$SAT2011, col="gray", main="SAT2011 distribution", xlab="score")
```




```
hist(sat$SAT2014, col="gray", main="SAT2014 distribution", xlab="score")
```




```
#part b
Difference = sat$SAT2014 - sat$SAT2011
hist(Difference, col="gray", main="Differences between SAT2011 and SAT2014", xlab="Difference")
```




```
#part c
#from the graph in part b we can conclude that the most of the scores have either gone up by 50 or gone
#part d
#in 2011
mean(sat$SAT2011)
```

```
## [1] 1587.667
```

```
median(sat$SAT2011)
```

```
## [1] 1559
```

```
sd(sat$SAT2011)
```

```
## [1] 121.48
```

```
range(sat$SAT2011)
```

```
## [1] 1385 1807
```

```
#in 2014
mean(sat$SAT2014)
```

```
## [1] 1586.647
```

```
median(sat$SAT2014)
```

```
## [1] 1551
```

```
sd(sat$SAT2014)
```

```
## [1] 134.0434
```

```
range(sat$SAT2014)
```

```
## [1] 1351 1807
```

that concludes question 2, i dont know where should i put my answers at so i just print everything out, i hope the marker doesnt miss anything and deduct my point! :P

Q2.92

This problem uses the "NUC" dataset (see attached CSV file). In addition to finding the range, variance and standard deviation, find the mean, median and mode for each part a), b), c).

```
nuc<-as_data_frame(read.csv("NUC.csv"))
nuc
```

```
## # A tibble: 20 × 2
##         STATE PLANTS
##        <fctr>  <int>
## 1     Alabama      5
## 2     Arizona      3
## 3  California      4
## 4     Florida      5
## 5     Georgia      4
## 6    Illinois     11
## 7      Kansas      1
## 8   Louisiana      2
## 9        Mass      1
## 10       Miss      1
## 11    NewHamp      1
## 12    NewYork      6
## 13  NCarolina      5
## 14       Ohio      3
## 15       Penn      9
## 16  SCarolina      7
## 17  Tennessee      3
## 18      Texas      4
## 19    Vermont      1
## 20  Wisconsin      3
```

```
#The following code creates a dataset eliminating the largest value from the dataset (by arranging the
nuc_partb = nuc %>% arrange(desc(PLANTS)) %>% slice(2:n())
nuc_partb
```

```
## # A tibble: 19 × 2
##         STATE PLANTS
##        <fctr>  <int>
## 1        Penn      9
## 2   SCarolina      7
## 3     NewYork      6
## 4     Alabama      5
## 5     Florida      5
## 6   NCarolina      5
## 7  California      4
## 8     Georgia      4
## 9       Texas      4
## 10    Arizona      3
## 11       Ohio      3
## 12  Tennessee      3
## 13  Wisconsin      3
## 14  Louisiana      2
## 15     Kansas      1
## 16       Mass      1
## 17       Miss      1
## 18    NewHamp      1
## 19    Vermont      1
```

```
#The following bit of code creates a dataset eliminating both the largest (first) and smallest (last) v
nuc_partc = nuc %>% arrange(desc(PLANTS)) %>% slice(2:(n()-1))
nuc_partc
```

```
## # A tibble: 18 × 2
##         STATE PLANTS
##        <fctr>  <int>
## 1        Penn      9
## 2   SCarolina      7
## 3     NewYork      6
## 4     Alabama      5
## 5     Florida      5
## 6   NCarolina      5
## 7  California      4
## 8     Georgia      4
## 9       Texas      4
## 10    Arizona      3
## 11       Ohio      3
## 12  Tennessee      3
## 13  Wisconsin      3
## 14  Louisiana      2
## 15     Kansas      1
## 16       Mass      1
## 17       Miss      1
## 18    NewHamp      1
```

```r
#part a
mean(nuc$PLANTS)
```

```
## [1] 3.95
```

```r
median(nuc$PLANTS)
```

```
## [1] 3.5
```

```r
var(nuc$PLANTS)
```

```
## [1] 7.523684
```

```r
sd(nuc$PLANTS)
```

```
## [1] 2.742934
```

```r
range(nuc$PLANTS)
```

```
## [1]  1 11
```

```r
a <-c(nuc$PLANTS)
tempa <- table(as.vector(a))
names(tempa)[tempa==max(tempa)]
```

```
## [1] "1"
```

```r
#part b
#the differences between a and b is every value went down, and the maxrange went from 11 to 9
mean(nuc_partb$PLANTS)
```

```
## [1] 3.578947
```

```r
median(nuc_partb$PLANTS)
```

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

```r
var(nuc_partb$PLANTS)
```

```
## [1] 5.035088
```

```r
sd(nuc_partb$PLANTS)
```

```
## [1] 2.2439
```

```r
range(nuc_partb$PLANTS)
```

```
## [1] 1 9
```

```r
b <-c(nuc_partb$PLANTS)
tempb <- table(as.vector(b))
names(tempb)[tempb==max(tempb)]
```

```
## [1] "1"
```

```r
#part c
#the differences between a and b is every value went up, and the minrange went from 1 to 3 and so is th
mean(nuc_partc$PLANTS)
```

```
## [1] 3.722222
```

```r
median(nuc_partc$PLANTS)
```

```
## [1] 3.5
```

```r
var(nuc_partc$PLANTS)
```

```
## [1] 4.918301
```

```r
sd(nuc_partc$PLANTS)
```

```
## [1] 2.217724
```

```r
range(nuc_partc$PLANTS)
```

```
## [1] 1 9
```

```r
c <-c(nuc_partc$PLANTS)
tempc <- table(as.vector(c))
names(tempc)[tempc==max(tempc)]
```

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

Q2.110

This problem uses the "SAND" dataset (see attached CSV file).

```r
sand<-as_data_frame(read.csv("SAND.csv"))
sand
```

```
## # A tibble: 100 × 3
##    PermA PermB PermC
##    <dbl> <dbl> <dbl>
## 1   55.4 150.0  52.2
## 2   57.2 148.3  53.8
## 3   59.7 147.9  58.8
## 4   57.9 145.7  58.0
## 5   59.9 146.5  55.4
## 6   59.3 146.7  55.0
## 7   59.9 146.8  58.2
## 8   58.3 145.2  61.7
## 9   56.2 148.7  63.3
## 10  57.4 147.1  62.3
## # ... with 90 more rows
```

```
#i dont really understand where are the rules for me to imply in order to complete these questions, i w
#part a: group A sandstone slices will fall at minimum area around 55,20
mean(sand$PermA)
```

```
## [1] 73.623
```

```
sd(sand$PermA)
```

```
## [1] 14.47502
```

```
#part b: group B sandstone slices will fall at maximum area around 150.00
mean(sand$PermB)
```

```
## [1] 128.537
```

```
sd(sand$PermB)
```

```
## [1] 21.97164
```

```
#part c: group C sandstone slices will fall at minimum area around 52.20
mean(sand$PermC)
```

```
## [1] 83.07
```

```
sd(sand$PermC)
```

```
## [1] 20.04847
```

```
#par d
#type B appears to decay faster.
```

Q2.131

part a) z scores:2.0,-1.0,0.5,-2.5 respectively

2=(x-2.7)/.5 (2)(.5) = x-2.7 1=x-2.7 x=3.7

-1.0=(x-2.7)/.5 (-1.0)(.5)=x-2.7 -.5=x-2.7 x=-.5+2.7=2.2

.5=(x-2.7)/.5 (.5)(.5)=x-2.7 .25=x-2.7 x=.25+2.7=2.95

-2.5=(x-2.7)/.5 (-2.5)(.5) = x-2.7 -1.25 =x-2.7 x=2.7-1.25 = 1.45

par b) student is on probation for z-scores below -1.6 which can be converted in GPA to -1.6=(x-2.7)/.5 (-1.6)(.5)=x-2.7 -.08=x-2.7 x=1.9 (GPA)

part c) the limits of z scores are z = 1.0 and 2.0; in terms of GPA are GPA = 3.2 and 3.7;the assumption i make about the distribution is mound-shaped, symmetric.

Q2.146

(a) before treatment: The approximate 25th percentile PASI score is 10. The approximate median is 15. The approximate 75th percentile is 28.
(b) after treatment: The approximate 25th percentile PASI score is 2. The approximate median is 4. The approximate 75th percentile is 7.
(c) comment on the effectiveness of treament: From looking at the 75th percentile after treatment is lower than the 25th percentile before treatment, it means the treament is effective.

Q2.152

z score = (x-mean)/sd

part a) z = (4-7)/1 = -3

part b) from the answer in part a i believe the librarian's claim is incorrect.

part c) the standard normal distributions states that between 1.96 and -1.96 lie 95% of observation. So a z-score of -3 means you have a less than 5% chance of finding a sample with so large a deviation from mean.

part d) if the standard deviation were 2, my answers to parts b and c will change, due to the increased chance of finding a sample with the deviation from mean.

Q2.188

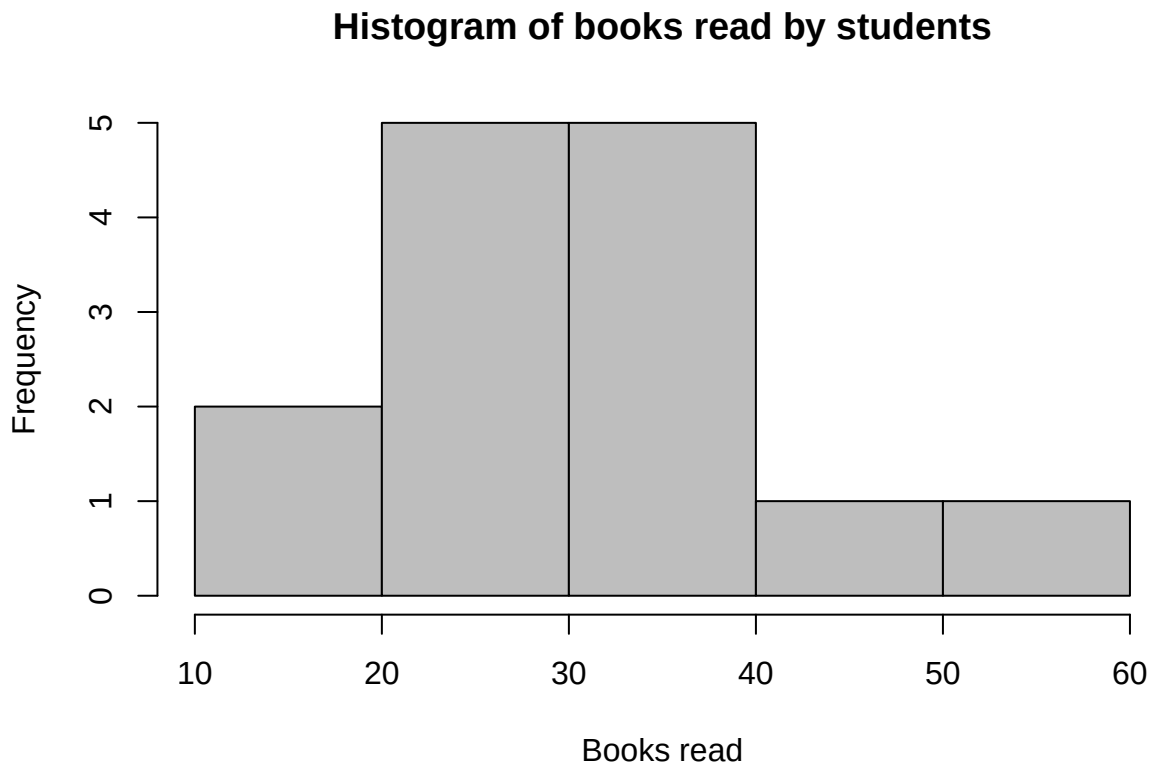
This problem uses the "JREAD" dataset (see attached CSV file). Solve only parts a), c) and d). In part a), rather than constructing a stem-and-leaf plot, construct a histogram.

```
jread<-as_data_frame(read.csv("JREAD.csv"))
jread
```

```
## # A tibble: 14 × 2
##    BOOKS  GRADE
##    <int> <fctr>
## 1     53      A
## 2     42      A
## 3     40      A
## 4     40      B
```

```
## 5     39      A
## 6     34      A
## 7     34      A
## 8     30      A
## 9     28      B
## 10    24      A
## 11    22      C
## 12    21      B
## 13    20      B
## 14    16      B
```

```
#part a
hist(jread$BOOKS, col="gray", main="Histogram of books read by students", xlab="Books read")
```




```
#part c
mean(jread$BOOKS)
```

```
## [1] 31.64286
```

```
median(jread$BOOKS)
```

```
## [1] 32
```

```
m <-c(jread$BOOKS)
tempm <- table(as.vector(m))
names(tempm)[tempm==max(tempm)]
```

```
## [1] "34" "40"
```

```
#part d
#the mean and median indicate the skewness of the distribution of the data is symmetrical.
```

Q2.192

This problem uses the "TILL" dataset (see attached CSV file). To find the minimum use the min() command, to find the maximum use the max() command.

```
till<-as_data_frame(read.csv("TILL.csv"))
till
```

```
## # A tibble: 26 × 2
##    BOREHOLE RATIO
##      <fctr> <dbl>
## 1    UMRB-1  3.75
## 2    UMRB-1  4.05
## 3    UMRB-1  3.81
## 4    UMRB-1  3.23
## 5    UMRB-1  3.13
## 6    UMRB-1  3.30
## 7    UMRB-1  3.21
## 8    UMRB-2  3.32
## 9    UMRB-2  4.09
## 10   UMRB-2  3.90
## # ... with 16 more rows
```

```
#part a
max(till$RATIO)
```

```
## [1] 5.06
```

```
min(till$RATIO)
```

```
## [1] 2.25
```

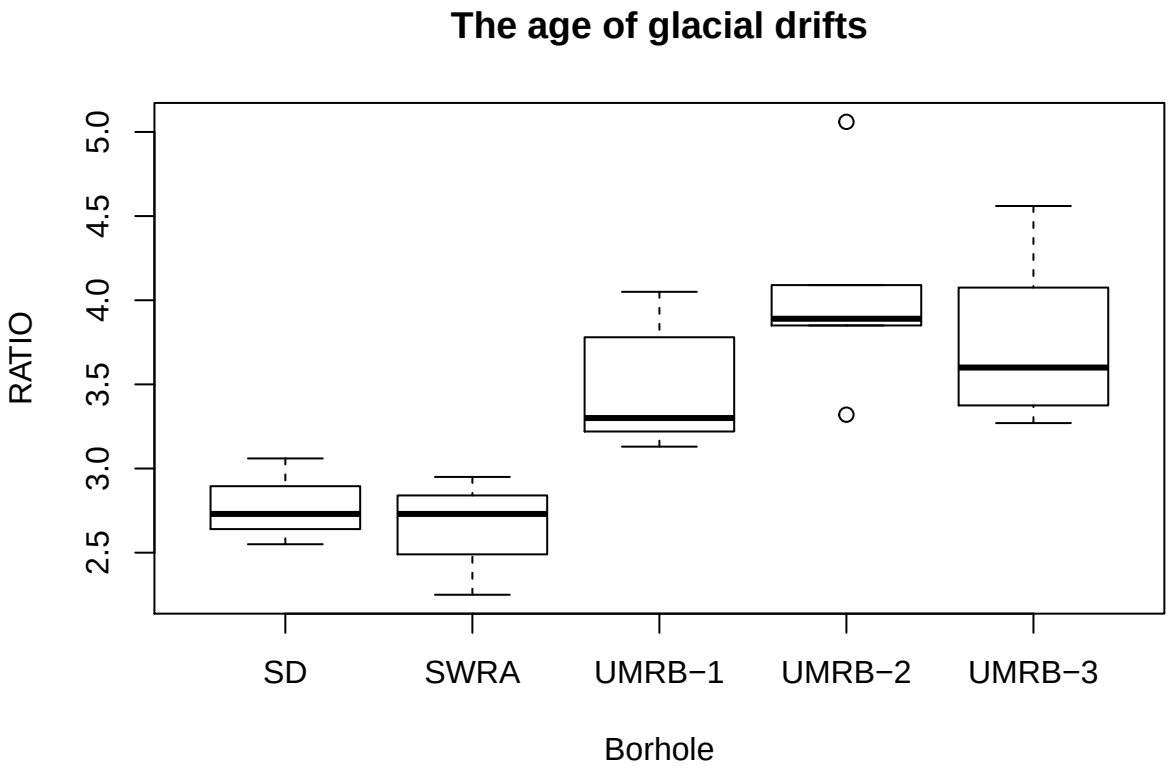
```
mean(till$RATIO)
```

```
## [1] 3.506923
```

```
#part b
#yes i consider the largest ratio to be unsusally large due to the difference between mean and max rati
#part c
till<-read.csv("TILL.csv")
head(till)
```

```
##   BOREHOLE RATIO
## 1   UMRB-1  3.75
## 2   UMRB-1  4.05
## 3   UMRB-1  3.81
## 4   UMRB-1  3.23
## 5   UMRB-1  3.13
## 6   UMRB-1  3.30
```

```r
boxplot(till$RATIO~till$BOREHOLE, ylab="RATIO", main="The age of glacial drifts", xlab="Borhole")
```



Q2.210

Same average but different standard deviation If i have to choose i would definately choose the professor's class who had a smaller standard deviation. (first professor's) a small standard deviation indicates that the data points are clustered closely around the mean