ASS1

Clayton Glenn

February 8, 2018

Working Directory

```
getwd()
```

```
## [1] "C:/Users/cglen/Documents/Stat Methods/Assignments/ASS1"
```

Question 1

There are 4 assignments that equal 15% of my grade and all work must be shown to receive full credit. There are 16 Labs in the class that equal 10% of my grade. I cannot deposit my lab into the drop box after time is up. The 1 project I have is 10% of my total grade. The project is over Simple Linear Regression and needs to be submitted in the outlined format provided on canvas. Clickers are done in class and worth 10% of my total grade. Missing a couple class will most likely not change my grade. Chapter quizzes are online and equal 5% of my total grade. They are usually 10 questions and are graded automatically. I have 2 midterm exams worth a total of 20% of my grade. The midterms should not overlap in content, but the final exam will. My final in this class is worth 30% of my grade and is cumulative. The breakup of the final exam will be about 1/3 Exams 1 and 2, and 2/3 from chapters 8 and 10. The grading scale in this class is as follows: A(90-100), B(80-89), C(60-79), D(50-59), F(0-49) without the possibility of a curve of the total grade, so what you earn is what you get.

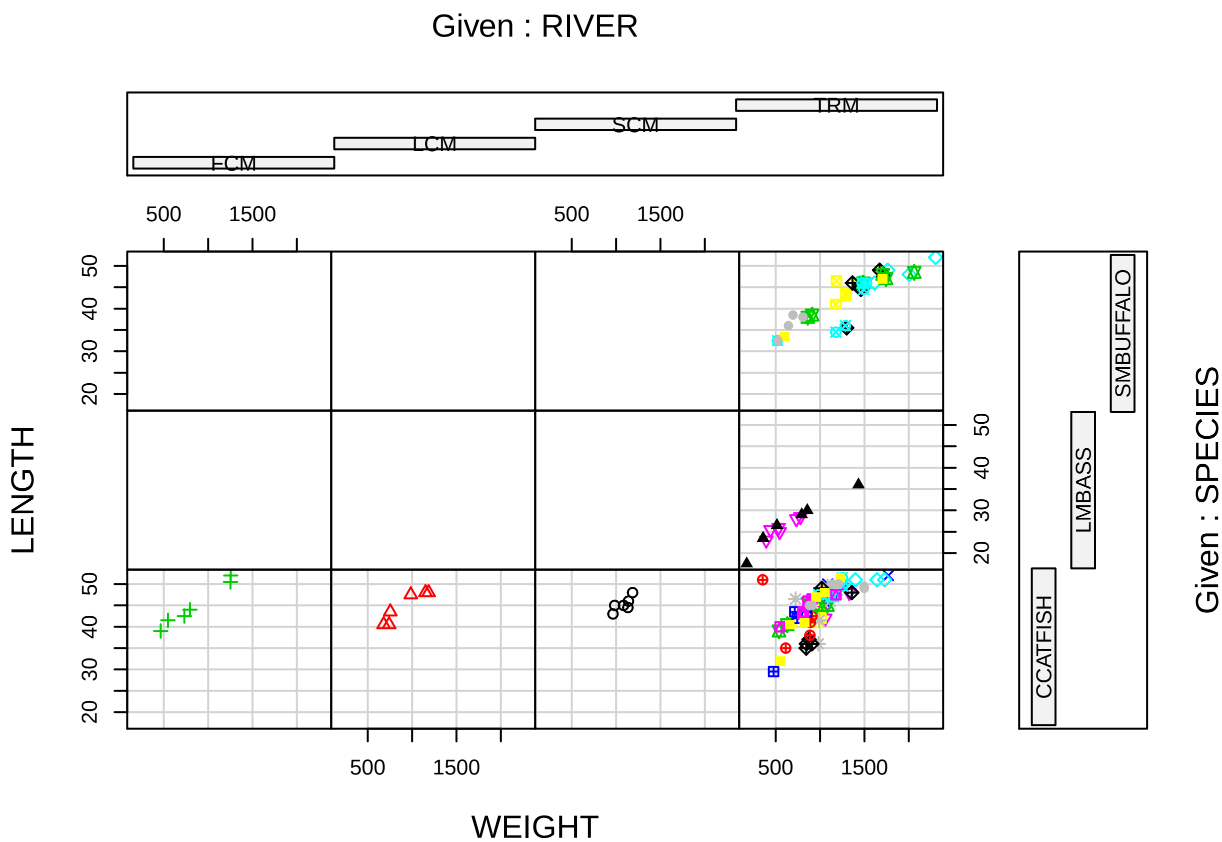
Question 2

Data Read

```
ddt = read.csv("DDT.csv")
```

Part A

```
m=with(ddt, as.numeric(levels(factor(MILE)))) # A
colm=c()
for(i in 1:length(ddt$MILE)){
colm[i]=which(ddt$MILE[i]==m) #B
}
coplot(LENGTH ~ WEIGHT | RIVER*SPECIES, ddt, col = colm, pch = colm)
```



Part B

This coplot shows the broken-up data of each river and species. The lower far left plot shows the number and size of catfish in the FCM River. The Lower Left Mid plot shows the number and size of catfish in the LCM River, and the Lower Third from the Left plot shows the number and size of catfish that are in the SCM River.

Part C

```
with(ddt, as.numeric(levels(factor(MILE))))
```

```
##  [1]   1   3   5 275 280 285 290 295 300 305 310 315 320 325 330 340 345
```

Line A shows the Each independent mile number in an array of numbers.

Part D

```
which(ddt$MILE[i]==m)
```

```
## [1] 17
```

Line B shows the Length of the array of independent mile numbers.

Part E

The top six plots are empty due to the fact that FCM, LCM, and SCM Rivers do not contain SMBUFFALO or LMBASS, so no data is shown for the coplots.

Part F

$$Mean = \frac{\sum_{i=1}^{n} X_i}{N}$$

```r
with(ddt[ddt$RIVER=="FCM" & ddt$SPECIES=="CCATFISH",],{
  mean(DDT)
})
```

```
## [1] 45
```

The mean value of DDT in CCATFISH Caught in the FCM river is 45.

Question 3

Part A

Length of Maximum Span(Feet) = Quantitative

Part B

Number of Vehicle Lanes = Quantitative

Part C

Toll Bridge = Qualitative

Part D

Average Daily Traffic = Quantitative

Part E

Condition of Deck(good, fair, or poor) = Qualitative

Part F

Bypass or Detour Length(Miles) = Quantitative

Part G

Route Type(interstate, U.S., state country, or city) = Qualitative

Question 4

Simple Random Sampling(Simple)

Randomly chose units out of a population.

Stratified Random Sampling(Complex)

Sampling used when units can be separated into strata or groups by characteristics.

Cluster Sampling(Complex)

Sampling used to break down large samples into clusters and then compare them.

Systematic Sampling(Complex)

Sampling used by selecting every Kth element in a population for a random sample.

Question 5

Data Read

```
mtbe = read.csv("MTBE.csv")
mtbeo=na.omit(mtbe)
```

Part A

```
i=sample(1:223,5,replace=FALSE)
mtbe[i,]
```

```
##       pH SpConduct DissOxy RoadsPct IndPct UrbanPct DevPct WellClass
## 18  7.36    2932.0    0.41     2.30   1.91    46.07  46.07   Private
## 117 7.90     516.6    0.40     3.42   1.91    66.43  71.67    Public
## 153 7.00     392.3    0.42     3.71   3.09    41.40  50.07    Public
## 191 8.12     234.8    0.53     1.20   0.00    18.59  29.22    Public
## 17  8.14     209.3    0.59     2.93   0.00    49.47  49.47   Private
##     Aquifier   Depth   SafeYld Distance MTBE.Detect MTBE.Level HouseDen
## 18   Bedrock      NA        NA  2079.43 Below Limit        0.2   278.95
## 117  Bedrock  68.580 325.51098  1478.05 Below Limit        0.2   278.95
## 153  Bedrock 153.924  94.62528   330.46 Below Limit        0.2   221.41
## 191  Bedrock  91.440  22.71007   716.44 Below Limit        0.2    31.78
## 17   Bedrock  92.964        NA  2652.00 Below Limit        0.2   159.78
##     PopDen
## 18   84.24
## 117  84.24
## 153  92.23
## 191  13.84
## 17    0.00
```

Part B

Standard Deviation of depth of Bedrock wells

$$StandardDeviation = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{N-1}$$

```
sd(mtbeo[mtbeo$Aquifier=="Bedrock",]$Depth)
```

```
## [1] 56.45357
```

Question 6

Data Read

```
eq = read.csv("EARTHQUAKE.csv")
```

Random Sample

```
eqsam=sample(1:2929,30,replace=FALSE)
eq[eqsam,]
```

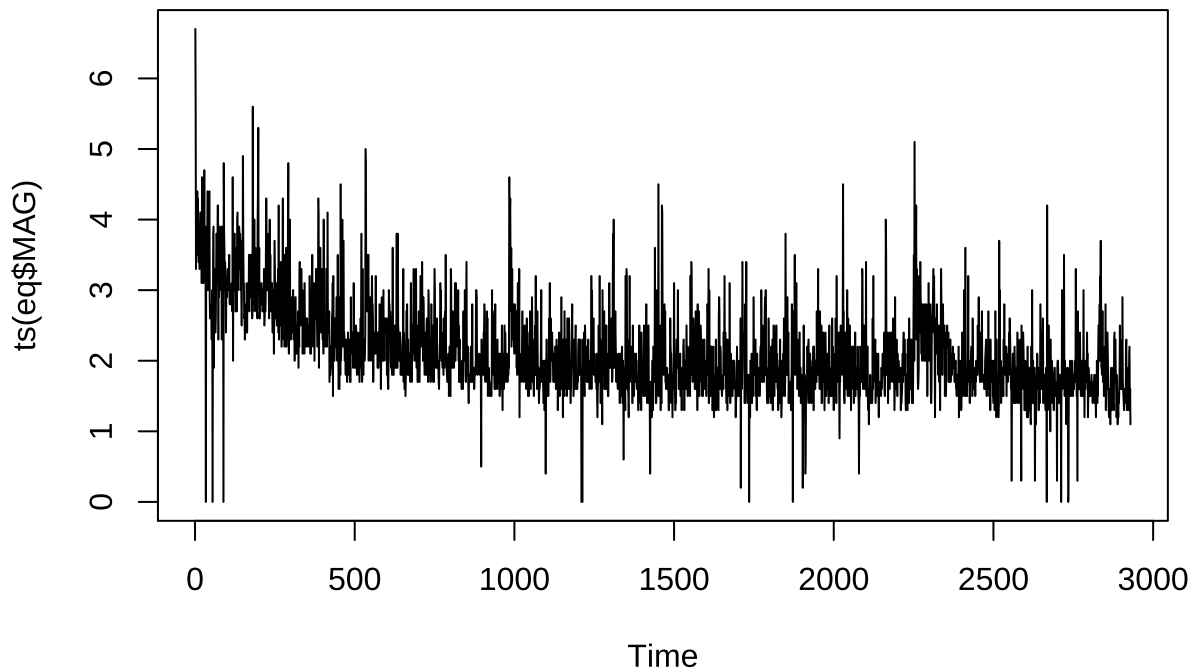
```
##      YEAR MONTH DAY HOUR MINUTE MAGNITUDE
## 872  1994     1  21    8      8       1.9
## 1169 1994     1  22   15     35       2.1
## 118  1994     1  17   17     56       4.6
## 464  1994     1  19   15      2       2.2
## 2390 1994     2   1    1     20       1.7
## 2521 1994     2   2   12     12       1.7
## 758  1994     1  20   19     50       1.9
## 1376 1994     1  23   18     17       2.1
## 1598 1994     1  25    0     53       1.7
## 1961 1994     1  27    6     26       2.5
## 2804 1994     2   5    0     55       1.5
## 2666 1994     2   3   16     22       1.5
## 1499 1994     1  24   10     47       2.0
## 1649 1994     1  25    9     17       1.6
## 475  1994     1  19   16      5       2.2
## 2103 1994     1  28    8     34       1.5
## 177  1994     1  17   22     57       3.5
## 2591 1994     2   3    3      6       1.8
## 658  1994     1  20    9     36       1.8
## 2687 1994     2   3   21     50       1.4
## 1017 1994     1  21   21     45       1.9
## 2388 1994     2   1    1      0       1.7
## 2647 1994     2   3   12     32       2.8
## 331  1994     1  18   19     28       3.1
## 1618 1994     1  25    4     16       2.2
## 5    1994     1  17   12     36       3.8
## 1464 1994     1  24    5     59       2.7
## 2794 1994     2   4   23      9       1.9
```

```
## 274  1994     1  18   13     23        3.0
## 2044 1994     1  27   21     17        1.5
```

Part A

Section i

```r
plot(ts(eq$MAG))
```



Section ii

$$Median = X_{N/2}$$

```r
median(eq$MAGNITUDE)
```

```
## [1] 2
```

The median of the whole Earthquake data file based on magnitude is 2

Question 7

Part A

The scientists used a stratified sample.

Part B

The Population was all fish in the Tennessee River and its tributaries.

Part C

The qualitative variables in DDT file is River and Species.

Question 8

Part A

A bar graph describes the data.

Part B

Number of Robots

Part C

According to the graph, the most used robot design is Legs Only.

Part D

$$RelativeFrequency = \frac{Class}{N}$$

```r
15/106
```

```
## [1] 0.1415094
```

```r
8/106
```

```
## [1] 0.0754717
```

```r
63/106
```

```
## [1] 0.5943396
```

```r
20/106
```

```
## [1] 0.1886792
```

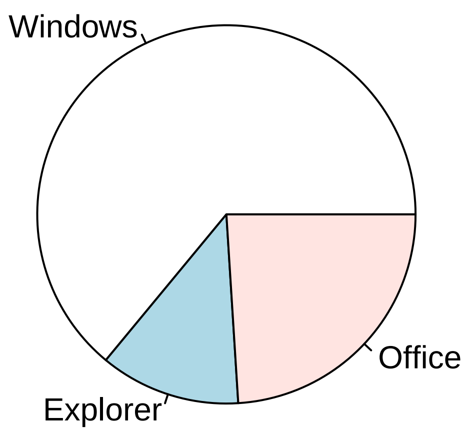
Part E - Wrong

```
freq=c(15,8,63,20)
RL=c("None","Both","Legs0","Wheels0")
x=rep(RL,freq)
```

Question 9

Part A

```
mpfreq=c(32,6,12)
mpt=c("Windows","Explorer","Office")
pie(mpfreq, mpt)
```



Based on the pie chart, Explorer has the lowest proportion of security issues.

Part B - Wrong

```
pareto<-function(mpfreq,mn="Microsoft Security",...){
  mpfreq.tab<-table(mpfreq)
  xx.tab<-sort(mpfreq.tab, decreasing=TRUE,index.return=FALSE)
  cs<-cumsum(as.vector(xx.tab))
```

```
  lenx<-length(mpfreq.tab)
  bp<-barplot(xx.tab,ylim=c(0,max(cs)),las=2)
  lb<-seq(0,cs[lenx],l=11)
  axis(side=4,at=lb,labels=paste(seq(0,100,length=11),"%",
  sep =""),las=1,line=-1,col="Blue",col.axis="Red")
  for(i in 1:(lenx-1)){
    segments(bp[i],cs[i],bp[i+1],cs[i+1],col=i,lwd=2)
  }
  title(main=mn,...)
}
#plot(pareto)
```

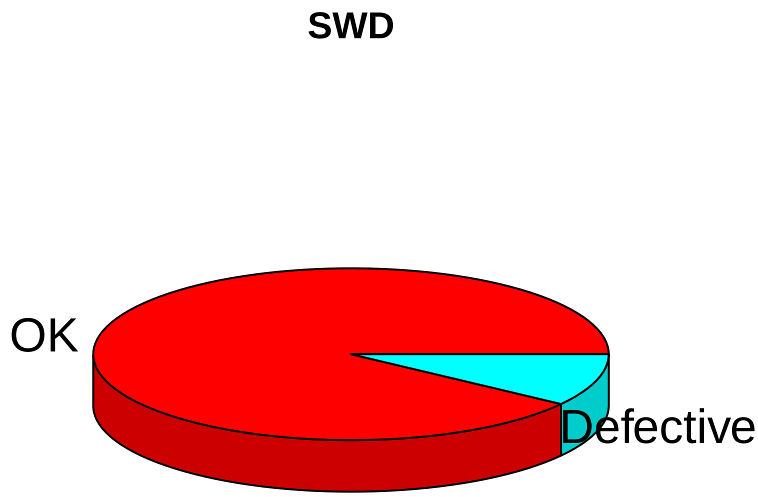
Based on the pareto graph, Windows should be the most focused on by Microsoft.

Question 10

```
swd=read.csv("SWDEFECTS.csv", header=TRUE)
#head(swd)
library(plotrix)
tab=table(swd$defect)
rtab=tab/sum(tab)
round(rtab,2)
```

```
##
## FALSE  TRUE
##   0.9   0.1
```

```
pie3D(rtab,labels=list("OK","Defective"),main="SWD")
```



The likelihood of software code being defective is 10%. The probability of OK software is 10:1.

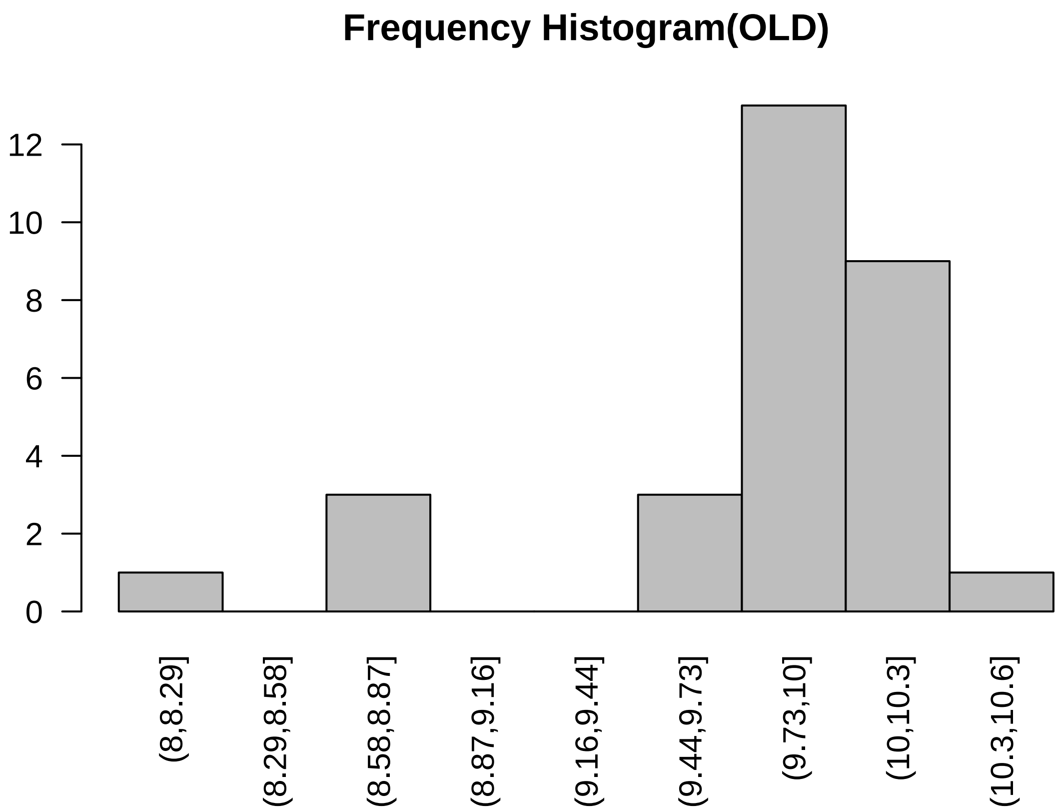
Question 11

Data Read

```
voltage.df<-read.csv("VOLTAGE.csv", header=TRUE)
old<-subset(voltage.df,subset=LOCATION=="OLD")
new<-subset(voltage.df,subset=LOCATION=="NEW")
```

Part A

```
old$VOLTAGE->vtn
lept<-min(vtn)-0.05
rept<-max(vtn)+0.05
rnge<-rept-lept
inc<-rnge/9
seq(lept, rept,by=inc)->cl
cvtn<-cut(vtn,breaks=cl)
new.tab=table(cvtn)
barplot(new.tab,space=0,main="Frequency Histogram(OLD)",las=2)
```



Part B

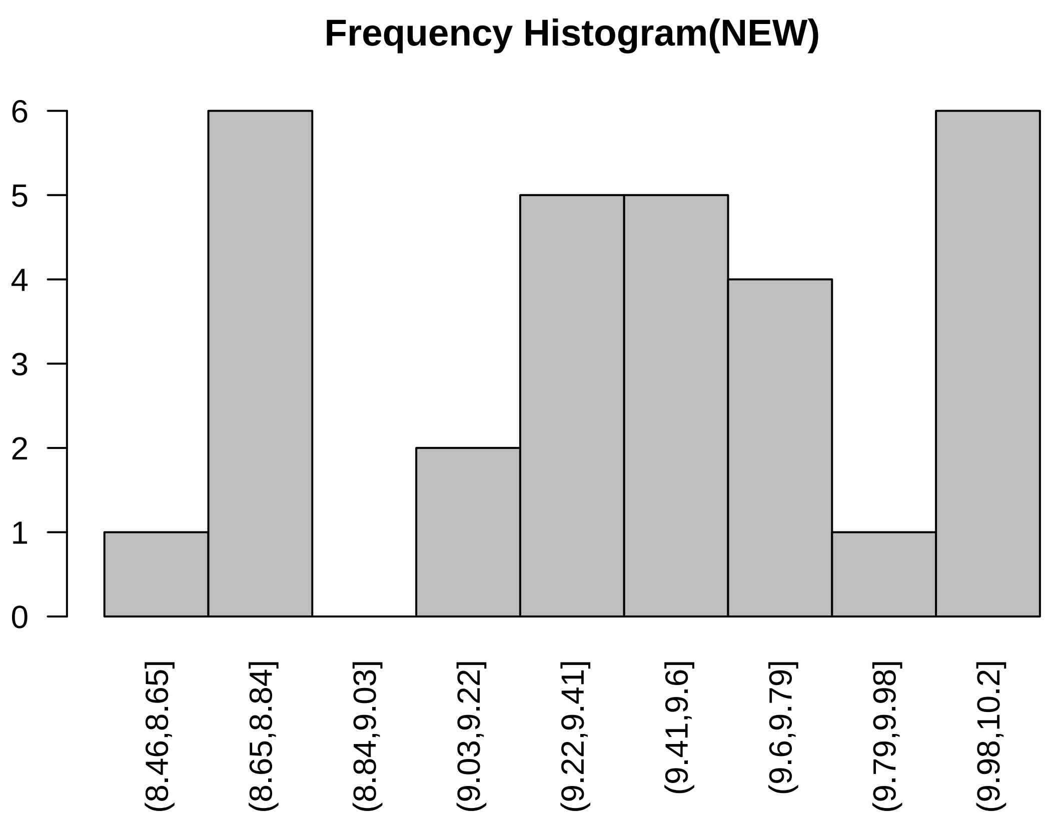
```
stem(vtn)
```

```
##
##   The decimal point is at the |
##
##    8 | 1
##    8 | 778
##    9 |
##    9 | 6778888999
##   10 | 000000011122333
##   10 | 6
```

Part C

```
new$VOLTAGE->vtn
lept<-min(vtn)-0.05
rept<-max(vtn)+0.05
rnge<-rept-lept
inc<-rnge/9
seq(lept, rept,by=inc)->cl
cvtn<-cut(vtn,breaks=cl)
```

```r
new.tab=table(cvtn)
barplot(new.tab,space=0,main="Frequency Histogram(NEW)",las=2)
```



Part D

The New Process is better than the Old process due to less outliers of voltage.

Part E

$$Mean = \frac{\sum_{i=1}^{n} X_i}{N}$$

$$Median = X_{N/2}$$

$$Mode = ClassMostOften$$

```r
mean(old$VOLTAGE)
```

```
## [1] 9.803667
```

```r
median(old$VOLTAGE)
```

```
## [1] 9.975
```

```r
which.max(old$VOLTAGE)
```

```
## [1] 7
```

```r
mean(new$VOLTAGE)
```

```
## [1] 9.422333
```

```r
median(new$VOLTAGE)
```

```
## [1] 9.455
```

```r
which.max(new$VOLTAGE)
```

```
## [1] 8
```

We can use the median for the central tendency because the median is close enough to mean and is a better looking number.

Part F

$$\mathcal{Z}Score = \frac{10.5 - \bar{X}_i}{S}$$

```r
(10.5-mean(old$VOLTAGE))/sd(old$VOLTAGE)
```

```
## [1] 1.287324
```

Part G

```r
(10.5-mean(new$VOLTAGE))/sd(new$VOLTAGE)
```

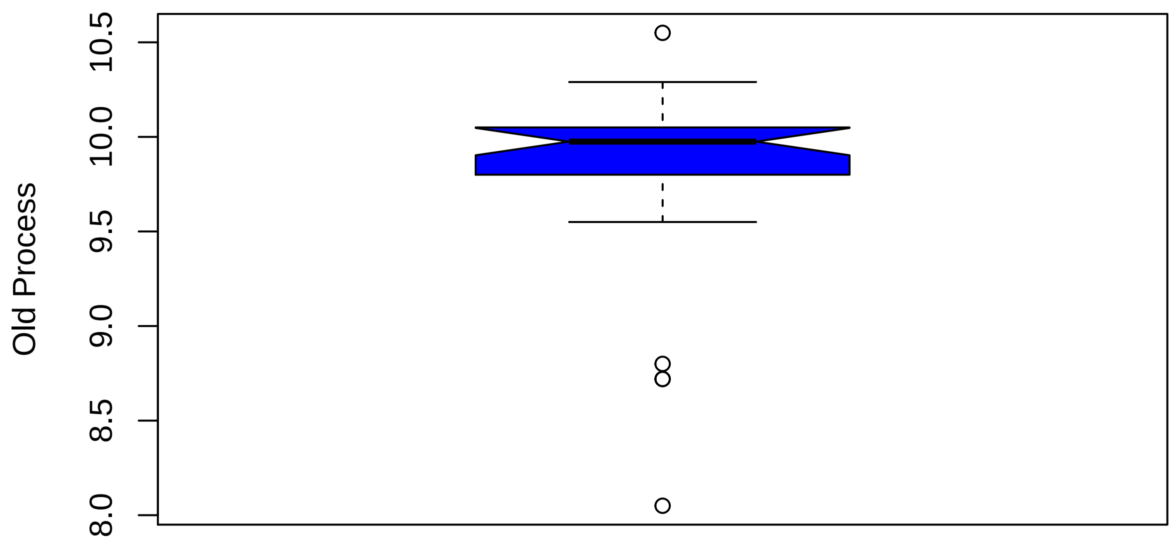
```
## [1] 2.25041
```

Part H

Based on parts F and G, 10.5 Voltage will more likely occur at the old process. This is due to 10.5 being closer to mean+sd

Part I

```r
with(old,boxplot(VOLTAGE,ylab="Old Process",col="Blue",notch=TRUE))
```



There are 4 outliers in the old process data.

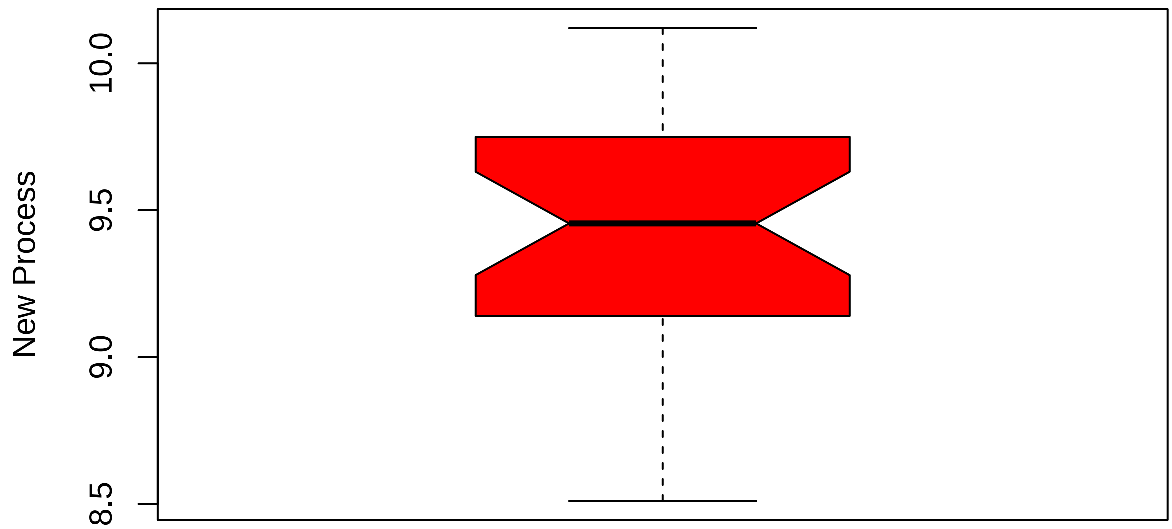
Part J

```
old[  (old$VOLTAGE-mean(old$VOLTAGE))/sd(old$VOLTAGE) <= -2 |
      (old$VOLTAGE-mean(old$VOLTAGE))/sd(old$VOLTAGE) >= 2,]
```

```
##    VOLTAGE LOCATION
## 6     8.05      OLD
## 20    8.72      OLD
## 28    8.72      OLD
```

Part K

```
with(new,boxplot(VOLTAGE,ylab="New Process",col="Red",notch=TRUE))
```



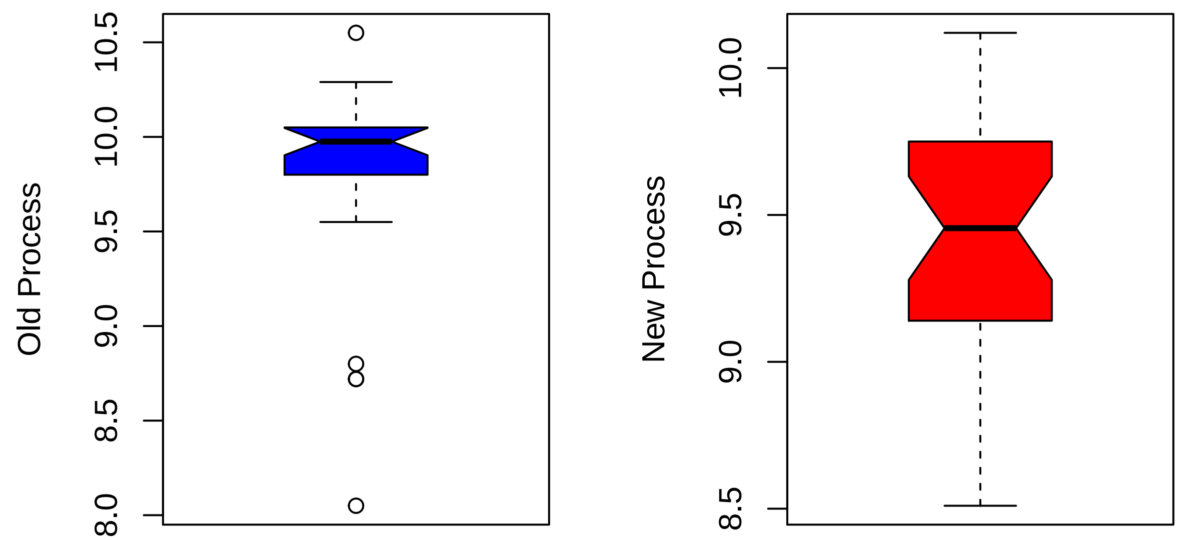
Part L

```
new[  (new$VOLTAGE-mean(new$VOLTAGE))/sd(new$VOLTAGE) <= -2 |
      (new$VOLTAGE-mean(new$VOLTAGE))/sd(new$VOLTAGE) >= 2,]
```

```
## [1] VOLTAGE  LOCATION
## <0 rows> (or 0-length row.names)
```

Part M

```
layout(matrix(c(1,2),nr=1,nc=2))
with(old,boxplot(VOLTAGE,ylab="Old Process",col="Blue",notch=TRUE))
with(new,boxplot(VOLTAGE,ylab="New Process",col="Red",notch=TRUE))
```



Question 12

$$\frac{\sum_{i=1}^{n} X_i}{N} - \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{N-1}$$

```
RP <- c(1.72,2.5,2.16,2.13,1.06,2.24,2.31,2.03,1.09,1.4,2.57,2.64,1.26,2.05,1.19,2.13,1.27,1.51,2.41,1.9
mean(RP) - sd(RP)*2
```

```
## [1] 0.8331772
```

```
mean(RP) + sd(RP)*2
```

```
## [1] 2.928823
```

Question 13

Data Read

```
gobi.df<-read.csv("GOBIANTS.CSV", header=TRUE)
dry.df = within(gobi.df, {
  reg <- ifelse(Region == "Gobi Desert", "GS","DS")
  reg<-factor(reg)
```

```
})
des.df = subset(gobi.df,subset=Region=="Gobi Desert")
```

Part A

$$Mean = \frac{\sum_{i=1}^{n} X_i}{N}$$

$$Median = X_{N/2}$$

$$Mode = ClassMostOften$$

```
mean(gobi.df$AntSpecies)
```

```
## [1] 12.81818
```

The Average Ant Species in all 11 sites.

```
median(gobi.df$AntSpecies)
```

```
## [1] 5
```

The Amount of Ant Species at the Site that has the exact middle amount of Ant Species.

```
which.max(gobi.df$AntSpecies)
```

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

The Most common number of Species at the sites.

Part B

The mean value best suits the data, due to the high volume of species at few sites.

Part C

$$Mean = \frac{\sum_{i=1}^{n} X_i}{N}$$

$$Median = X_{N/2}$$

$$Mode = ClassMostOften$$

```
mean(dry.df[dry.df$reg == "DS",]$PlantCov)
```

```
## [1] 40.4
```

```
median(dry.df[dry.df$reg == "DS",]$PlantCov)
```

```
## [1] 40
```

```
which.max(dry.df[dry.df$reg == "DS",]$PlantCov)
```

```
## [1] 2
```

Part D

$$Mean = \frac{\sum_{i=1}^{n} X_i}{N}$$

$$Median = X_{N/2}$$

$$Mode = ClassMostOften$$

```
mean(des.df$PlantCov)
```

```
## [1] 28
```

```
median(des.df$PlantCov)
```

```
## [1] 26
```

```
which.max(des.df$PlantCov)
```

```
## [1] 4
```

Part E

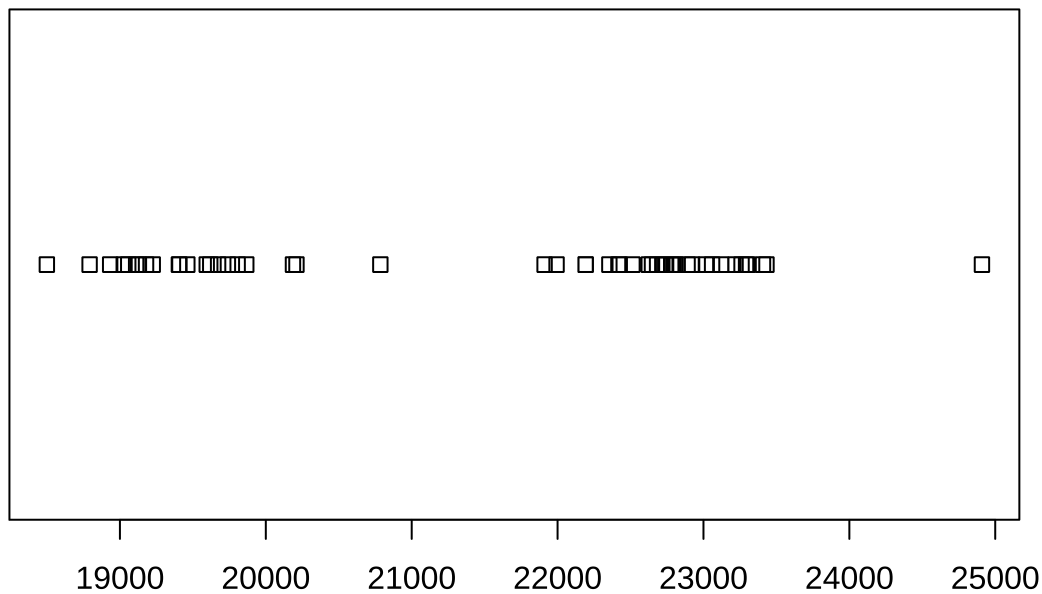
The ant species seems more abundant with less plant cover, so the Dry Steppe is more bountiful for Ants.

Question 14

Part A

```
gal.df<-read.csv("GALAXY2.CSV", header=TRUE)
low.df = gal.df[gal.df$VELOCITY < 21000,]
high.df = gal.df[gal.df$VELOCITY > 21000,]
```

```
plot(gal.df)
```



Part B

Yes, there are two clusters. One cluster is between 19000 and 20000, and the other cluster is between 22000 and 23000.

Part C

$$Mean = \frac{\sum_{i=1}^{n} X_i}{N}$$

$$StandardDeviation = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{N-1}$$

```
mean(low.df)
```

```
## [1] 19462.24
```

```
sd(low.df)
```

```
## [1] 532.2868
```

```
mean(high.df)
```

```
## [1] 22838.47
```

```
sd(high.df)
```

```
## [1] 560.9767
```

Part D

The galaxy Velocity of 20000 would fit within A1775A because the velocity is much closer to A1775A's Mean + SD than A1775B's Mean - SD.

Question 15

```r
library(ggplot2)
gg <- ggplot(ddt, aes(x=RIVER, y=LENGTH, color=SPECIES, fill = SPECIES))
gg <- gg + geom_boxplot(colour = "#1F3552")
gg <- gg + ggtitle("Clayton Glenn")
show(gg)
```