STATISTICAL METHODS FOR DATA SCIENCE - MINI PROJECT 4

Names of group members: 1. Manneyaa Jayasanker Net ID: MXJ180040
2. Venkatesh Sankar Net ID: VXS200014

Contribution of each group member:

Manneyaa Jayasanker:

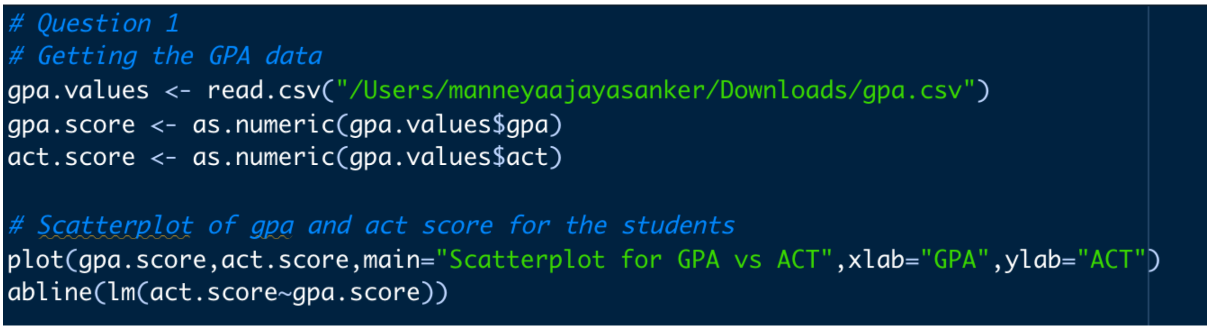
- Worked on R code for 1 and 2.
- Worked on conclusion for 1 and 2.
- Wrote documentation for 1 and 2.

Venkatesh Sankar:

- Worked on R code for 3.
- Worked on conclusion for 3.
- Wrote documentation for 3.

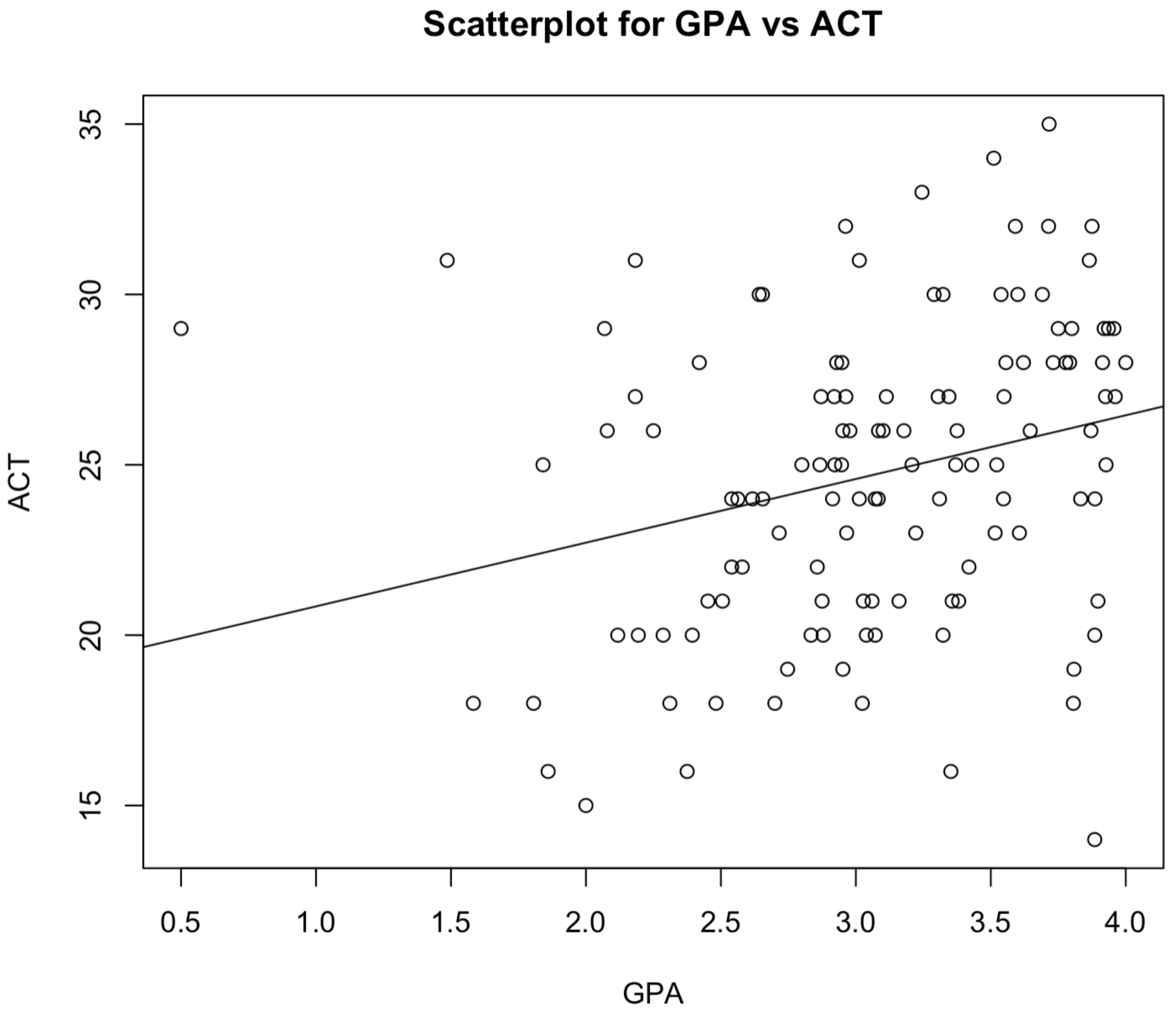
==

1)



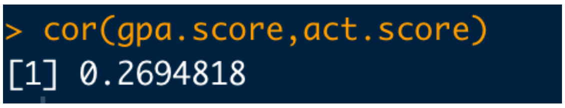
```
# Question 1
# Getting the GPA data
gpa.values <- read.csv("/Users/manneyaajayasanker/Downloads/gpa.csv")
gpa.score <- as.numeric(gpa.values$gpa)
act.score <- as.numeric(gpa.values$act)

# Scatterplot of gpa and act score for the students
plot(gpa.score,act.score,main="Scatterplot for GPA vs ACT",xlab="GPA",ylab="ACT")
abline(lm(act.score~gpa.score))
```

Output :



Inference :
From the scatter plot we know that GPA against ACT has positive slope.

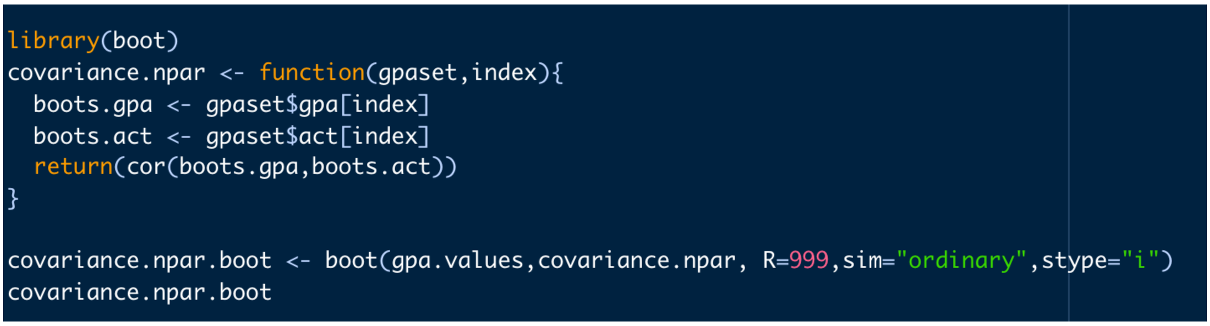
Correlation :



```
> cor(gpa.score,act.score)
[1] 0.2694818
```

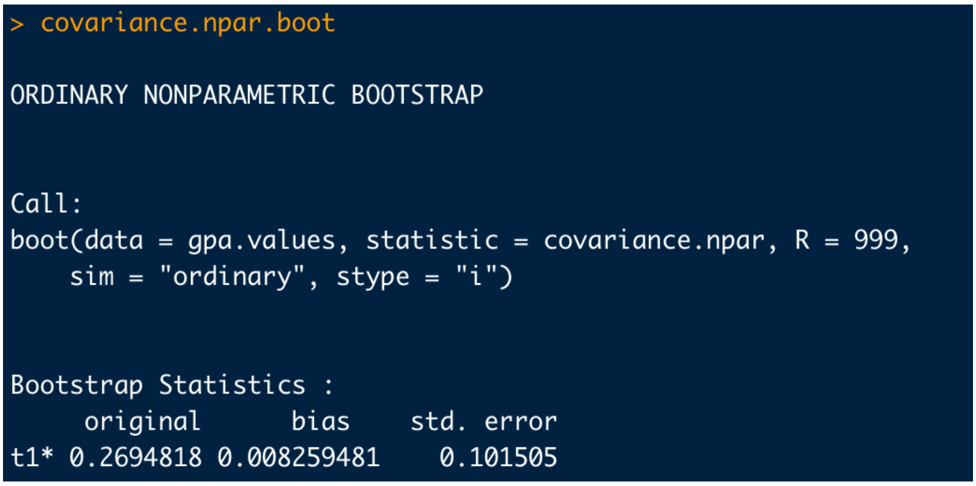
The value of 0.2694818 shows there is a positive correlation between two variables, but it is weak and likely unimportant linear relationship.

Estimates :



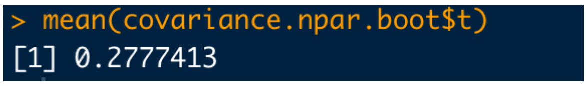
```
library(boot)
covariance.npar <- function(gpaset,index){
  boots.gpa <- gpaset$gpa[index]
  boots.act <- gpaset$act[index]
  return(cor(boots.gpa,boots.act))
}

covariance.npar.boot <- boot(gpa.values,covariance.npar, R=999,sim="ordinary",stype="i")
covariance.npar.boot
```

Output :



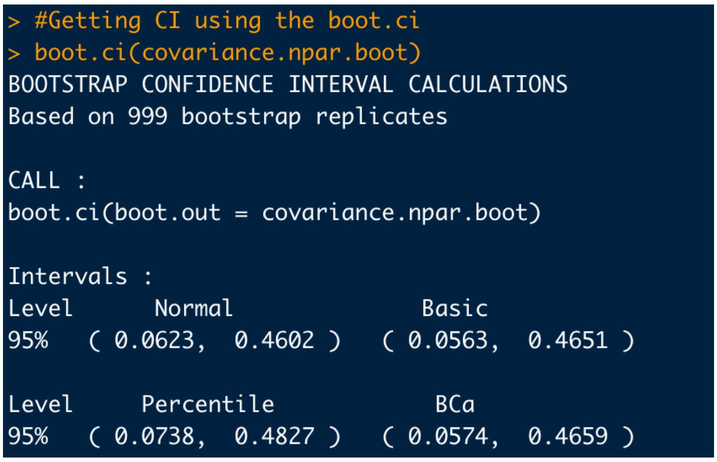
```
> covariance.npar.boot

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = gpa.values, statistic = covariance.npar, R = 999,
    sim = "ordinary", stype = "i")


Bootstrap Statistics :
     original       bias    std. error
t1* 0.2694818 0.008259481    0.101505
```

point estimate of bootstrap:



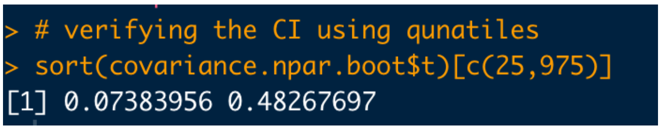
```
> mean(covariance.npar.boot$t)
[1] 0.2777413
```

Confidence intervals using boot.ci:



```
> #Getting CI using the boot.ci
> boot.ci(covariance.npar.boot)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

CALL :
boot.ci(boot.out = covariance.npar.boot)

Intervals :
Level      Normal              Basic
95%   ( 0.0623,  0.4602 )   ( 0.0563,  0.4651 )

Level     Percentile            BCa
95%   ( 0.0738,  0.4827 )   ( 0.0574,  0.4659 )
```

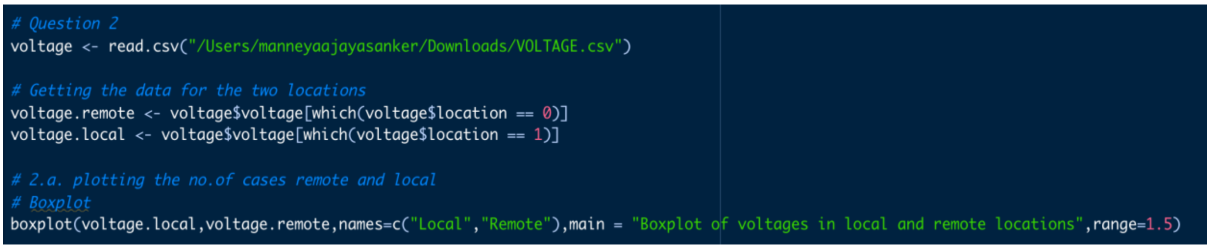
Verifying confidence intervals:



```
> # verifying the CI using qunatiles
> sort(covariance.npar.boot$t)[c(25,975)]
[1] 0.07383956 0.48267697
```

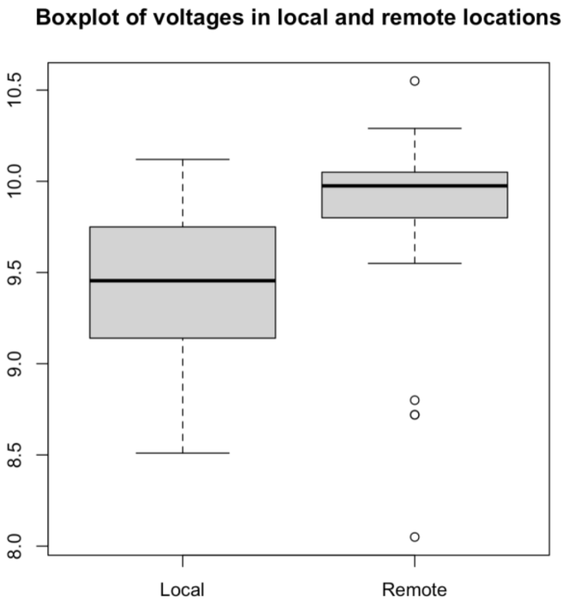
Inference:
From the observations we can say that point estimate of the correlation from the bootstrap is close to the correlation value from the samples and also boot.ci is close to the Quantiles values.

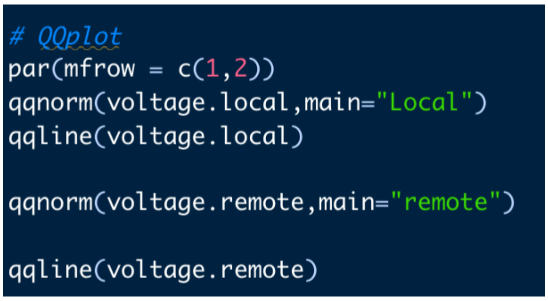
2)
a)



```
# Question 2
voltage <- read.csv("/Users/manneyaajayasanker/Downloads/VOLTAGE.csv")

# Getting the data for the two locations
voltage.remote <- voltage$voltage[which(voltage$location == 0)]
voltage.local <- voltage$voltage[which(voltage$location == 1)]

# 2.a. plotting the no.of cases remote and local
# Boxplot
boxplot(voltage.local,voltage.remote,names=c("Local","Remote"),main = "Boxplot of voltages in local and remote locations",range=1.5)
```

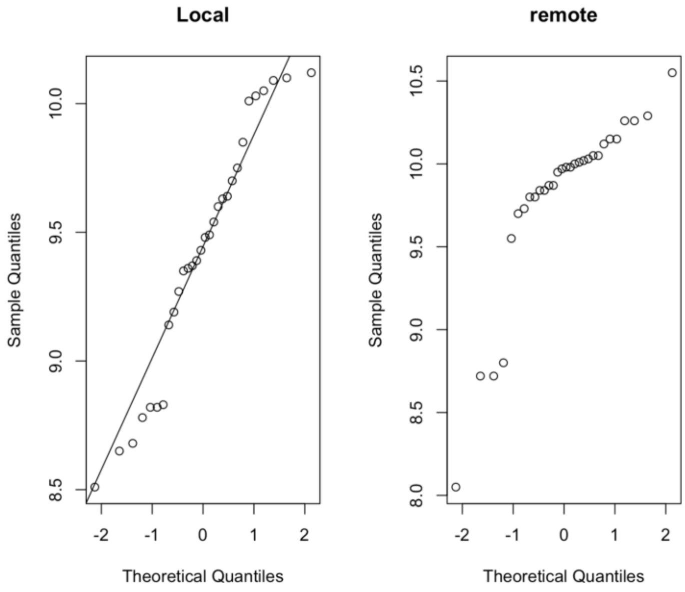
Output :



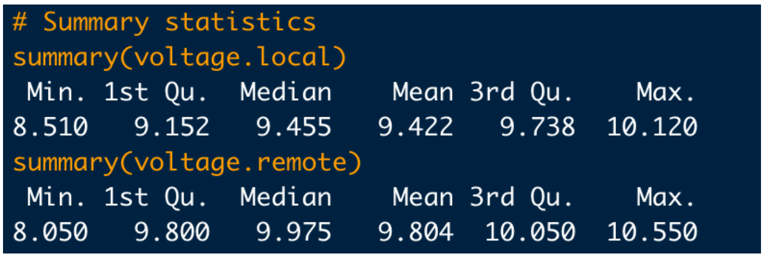


```
# QQplot
par(mfrow = c(1,2))
qqnorm(voltage.local,main="Local")
qqline(voltage.local)

qqnorm(voltage.remote,main="remote")

qqline(voltage.remote)
```

Output :



Summary Statistics:



```
# Summary statistics
summary(voltage.local)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
8.510   9.152   9.455   9.422   9.738  10.120
summary(voltage.remote)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
8.050   9.800   9.975   9.804  10.050  10.550
```

Inference:

From the boxplot we can say that the values of the remote locations are greater that those at the local. From the summary we can say the distribution is left skewed as the median is greater than the mean value.

The distribution of the voltage produced remotely and locally varies a lot. We can say that both distributions are approximately normally as seen by the QQ-plot.

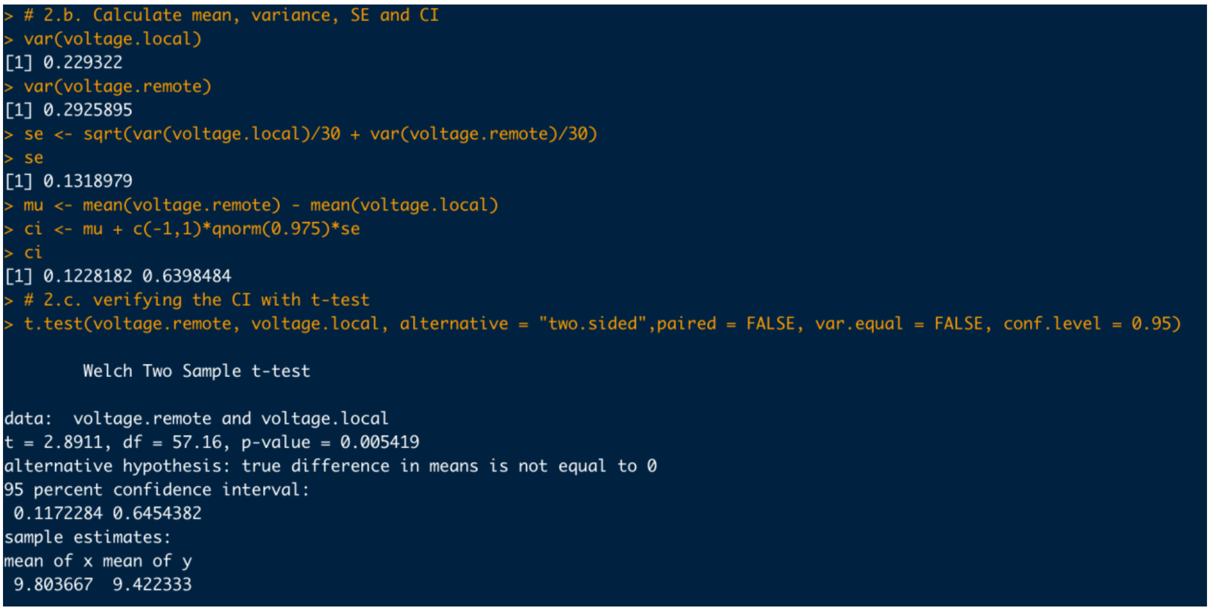
b)

Here hypothesis is:
Null Hypothesis - The difference of the mean between the two voltages of the population is equal to zero.

Alternative Hypothesis - The difference of the mean between the two voltages of the population is not equal to zero.

We can use t.test to verify the hypothesis as the data is available to us. As the variances are not equal we know from the above plots we will be using var.equal = False which uses Welch's or Satterthwaite's approximation internally.

So we are going to calculate the standard error and from that we can construct our confidence interval :



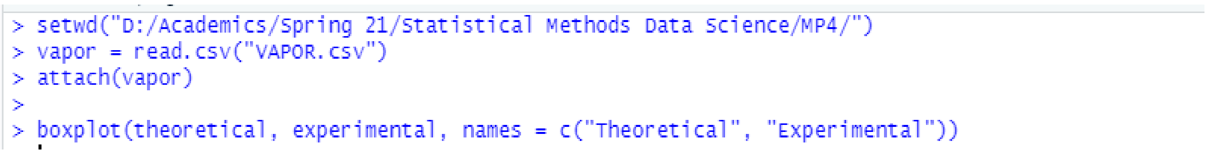
```
> # 2.b. Calculate mean, variance, SE and CI
> var(voltage.local)
[1] 0.229322
> var(voltage.remote)
[1] 0.2925895
> se <- sqrt(var(voltage.local)/30 + var(voltage.remote)/30)
> se
[1] 0.1318979
> mu <- mean(voltage.remote) - mean(voltage.local)
> ci <- mu + c(-1,1)*qnorm(0.975)*se
> ci
[1] 0.1228182 0.6398484
> # 2.c. verifying the CI with t-test
> t.test(voltage.remote, voltage.local, alternative = "two.sided",paired = FALSE, var.equal = FALSE, conf.level = 0.95)

        Welch Two Sample t-test

data:  voltage.remote and voltage.local
t = 2.8911, df = 57.16, p-value = 0.005419
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1172284 0.6454382
sample estimates:
mean of x mean of y
 9.803667  9.422333
```

So the interval that we have obtained is (0.11,0.64) which does not contain 0 so we can reject the null hypothesis. This implies that the difference between the means is not zero and hence the manufacturing process cannot be established locally. This can also be confirmed by performing t-test on the mean values.

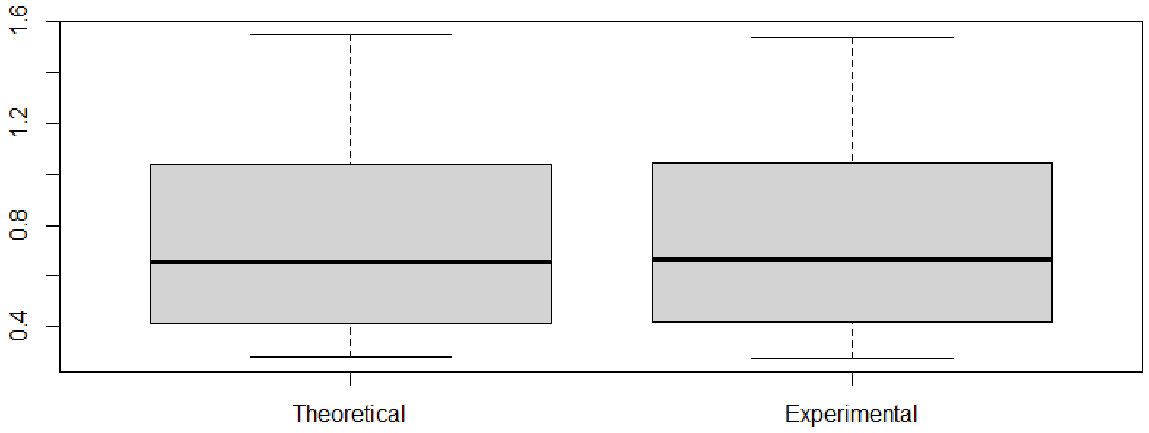
C)
From part A, we can conclude that the voltage reading at remote locations is more than the readings in local locations. For a manufacturing process, a high voltage is required to fuel the heavy equipment. So based on the data we collected in (A) and (B), it is clear that the manufacturing process can be done only in remote locations.

3.

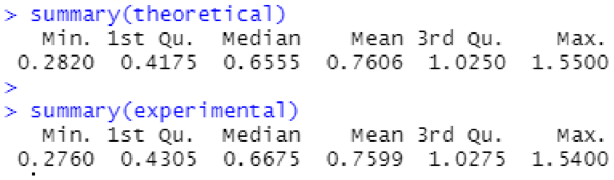
Performing analysis on the given VAPOR dataset. First, we import the dataset and construct boxplot analysis of experimental and theoretical values.



```
> setwd("D:/Academics/Spring 21/Statistical Methods Data Science/MP4/")
> vapor = read.csv("VAPOR.csv")
> attach(vapor)
>
> boxplot(theoretical, experimental, names = c("Theoretical", "Experimental"))
```



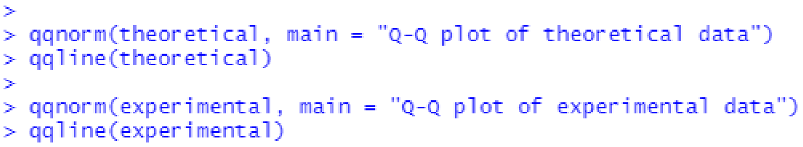
Then, we perform a summary analysis of three quartiles along with min and max values of both the theoretical and experimental values.



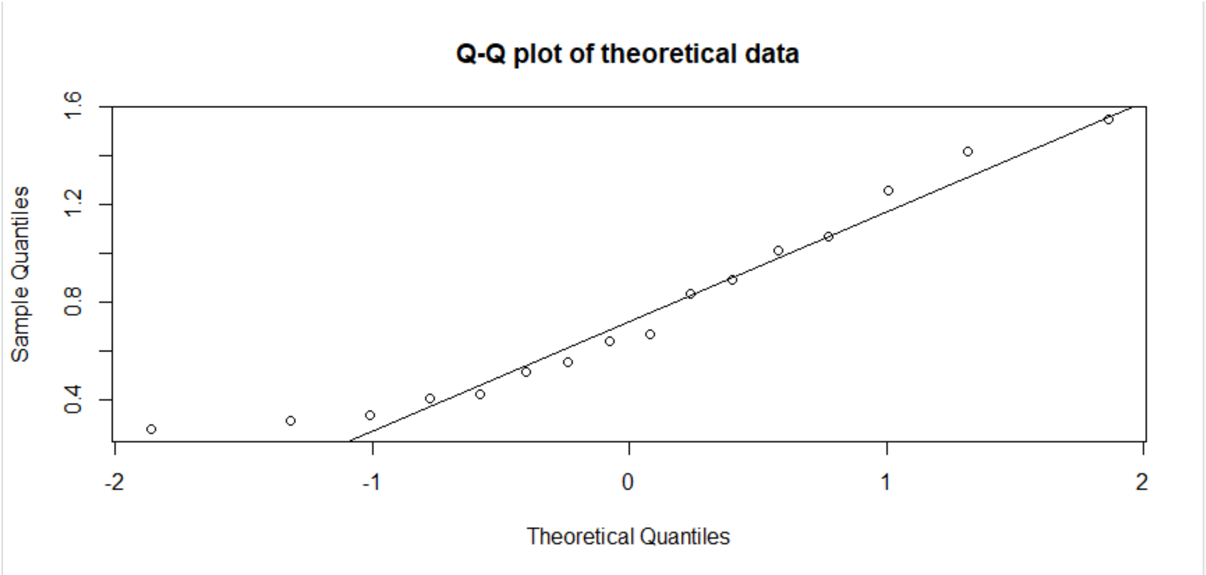
```
> summary(theoretical)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2820  0.4175  0.6555  0.7606  1.0250  1.5500 
>
> summary(experimental)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2760  0.4305  0.6675  0.7599  1.0275  1.5400 
```

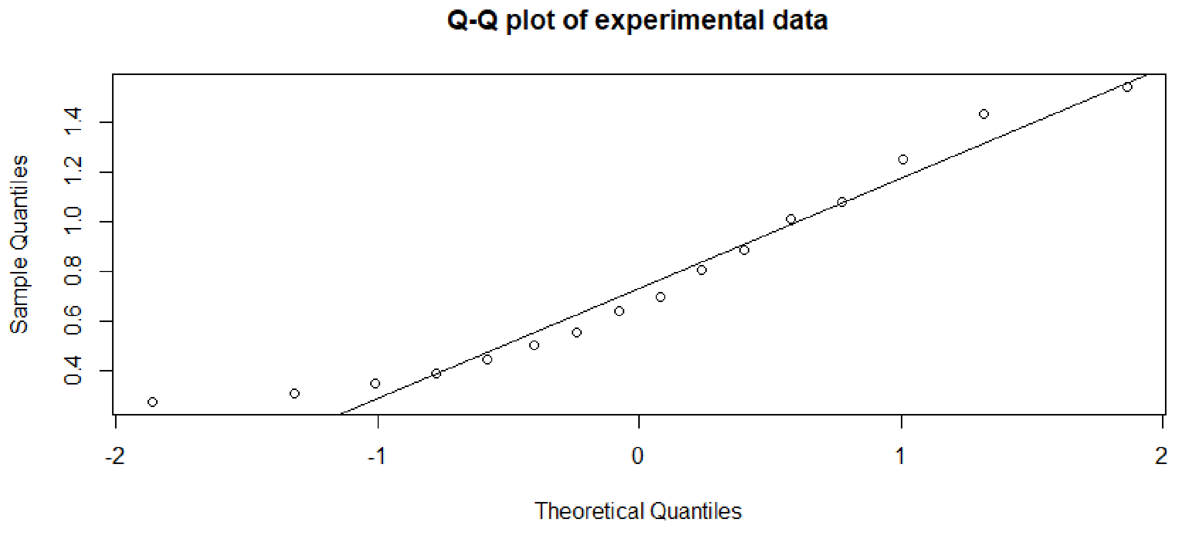
From the summary, we can infer that the theoretical and experimental pressure looks almost similar. We can also imply that the distributions are similar based on this.

Then, we perform Q-Q plot of both the experimental and theoretical pressure values.

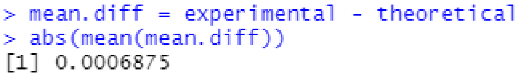


```
>
> qqnorm(theoretical, main = "Q-Q plot of theoretical data")
> qqline(theoretical)
>
> qqnorm(experimental, main = "Q-Q plot of experimental data")
> qqline(experimental)
```





From these Q-Q plot, we can infer that both the distributions are normal. We then calculate the mean difference of theoretical and experimental values.



```
> mean.diff = experimental - theoretical
> abs(mean(mean.diff))
[1] 0.0006875
```

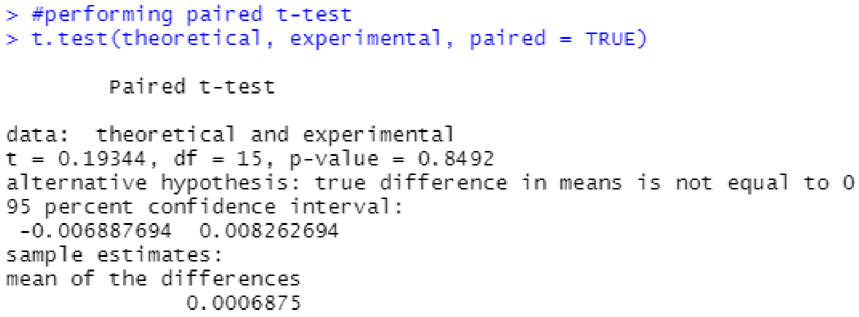
We can infer that the mean difference between experimental and theoretical values is zero based on the above calculation.

Next, we perform hypothesis testing inorder to further validate the results.

Null hypothesis (H_0) : The true mean difference between experimental and theoretical values is equal to zero.

Alternative hypothesis (H_1): The true mean difference between experimental and theoretical values is not equal to zero.

We then perform a paired t-test .



```
> #performing paired t-test
> t.test(theoretical, experimental, paired = TRUE)

        Paired t-test

data:  theoretical and experimental
t = 0.19344, df = 15, p-value = 0.8492
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.006887694  0.008262694
sample estimates:
mean of the differences
              0.0006875
```

From the paired t-test results, we can infer that the 95% confidence interval is [-0.0069, 0.0083] which contains zero in it. Hence, the difference can be equal to zero which accepts the Null Hypothesis. Therefore, we can conclude that the theoretical model for vapor pressure is a good model of reality.