Mini Project 4

Names of group members: Karan Risbud(KSR190005), Shubham Vartak(SXV200115)

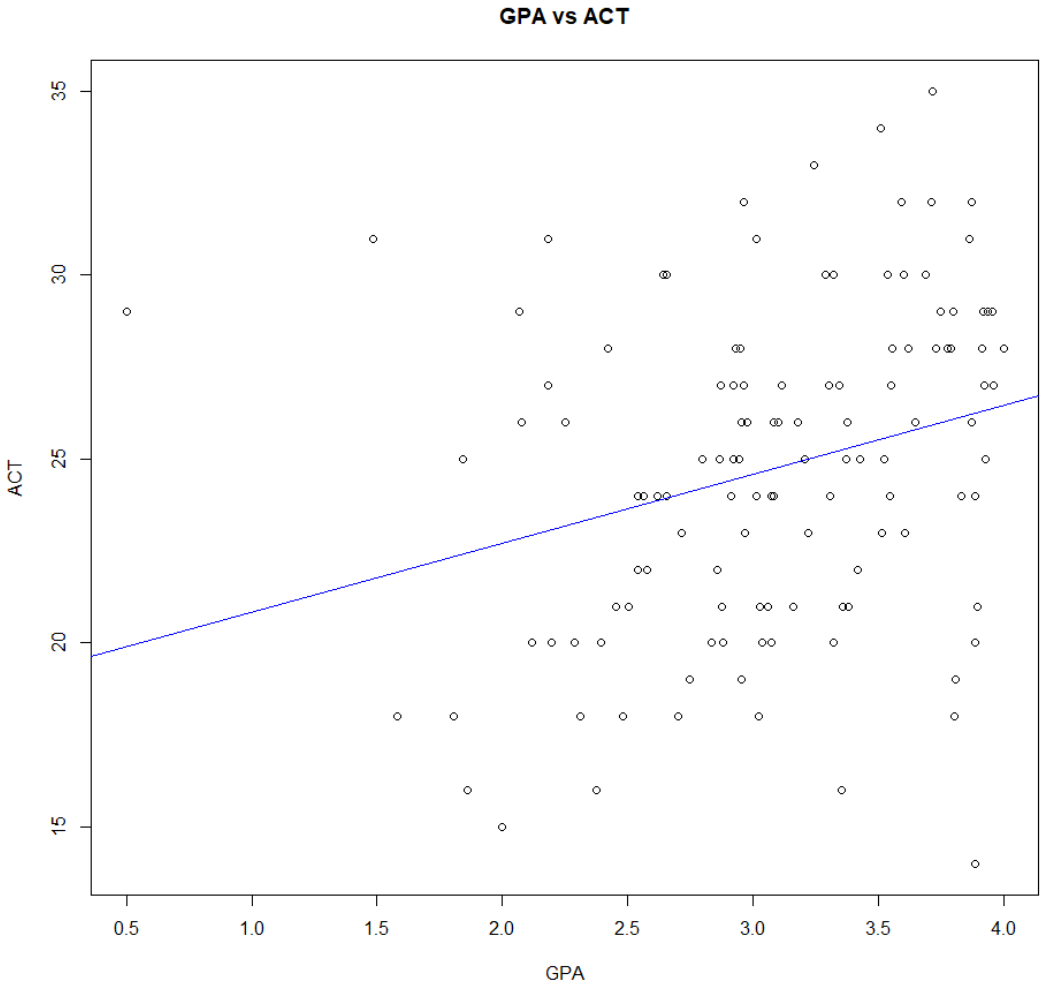
Contribution of each group member: Both the Project group members worked together on the project. Collaborated to solve the problem and implementation of R programming.

Q1. (6 points) In the class, we talked about bootstrap in the context of one-sample prob lems. But the idea of nonparametric bootstrap is easily generalized to more general situations. For example, suppose there are two dependent variables X1 and X2 and we have i.i.d. data on (X1, X2) from n independent subjects. In particular, the data consist of (Xi1, Xi2), i = 1, . . . , n, where the observations Xi1 and Xi2 come from the ith subject. Let θ be a parameter of interest — it's a feature of the distribution of (X1, X2). We have an estimator ˆθ of θ that we know how to compute from the data. To obtain a draw from the bootstrap distribution of ˆθ, all we need to do is the following: randomly select n subject IDs with replacement from the original subject IDs, extract the observations for the selected IDs (yielding a resample of the original sample), and compute the estimate from the resampled data. This process can be 1 repeated in the usual manner to get the bootstrap distribution of ˆθ and obtain the desired inference. Now, consider the gpa data stored in the gpa.txt file available on eLearning. The data consist of GPA at the end of freshman year (gpa) and ACT test score (act) for randomly selected 120 students from a new freshman class. Make a scatterplot of gpa against act and comment on the strength of linear relationship between the two variables. Let ρ denote the population correlation between gpa and act. Provide a point estimate of ρ, bootstrap estimates of bias and standard error of the point estimate, and 95% confidence interval computed using percentile bootstrap. Interpret the results. (To review population and sample correlations, look at Sections 3.3.5 and 11.1.4 of the textbook. The sample correlation provides an estimate of the population correlation and can be computed using cor function in R.)

R Code:

```
library(boot)
data = read.csv(file="D:/Fall'21/STATS/mini_project_4/gpa.csv")
gpa = data$gpa
act = data$act
plot(gpa,act, xlab = "GPA",ylab = "ACT",main = "GPA vs ACT")
abline(lm(act~gpa),col="blue")
```

Plot:



Observations:

1. The line has a positive slope which indicates that there is positive relation between gpa and act.
2. By viewing the line in scatter plot and seeing the correlation coeffecient we can say that their linear relationship is not strong.(very weakly related).

R Code:

```
# correlation between gpa and act
cor(gpa,act)
```

```
Console   Terminal ×   Jobs ×
R 4.1.1 · D:/Fall'21/STATS/mini_project_4/
> cor(gpa,act)
[1] 0.2694818
```

R Code:

```
# Function calculates correlation between GPA and ACT
corr.npar <- function(data,indices) {
  result <- cor(data$gpa[indices],data$act[indices])
  return(result)
}
```

```
(corr.npar.boot <- boot(data,corr.npar,R=999,
                        sim="ordinary", stype="i"))


ORDINARY NONPARAMETRIC BOOTSTRAP


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


Bootstrap Statistics :
     original      bias    std. error
t1* 0.2694818 0.004793941   0.1044929
```

R Code:

```
mean(corr.npar.boot$t)
```

```
> mean(corr.npar.boot$t)
[1] 0.2742757
```

R Code:

```
boot.ci(corr.npar.boot)
```

```
> boot.ci(corr.npar.boot)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

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

Intervals :
Level      Normal              Basic
95%   ( 0.0599,  0.4695 )   ( 0.0547,  0.4553 )

Level     Percentile            BCa
95%   ( 0.0837,  0.4843 )   ( 0.0668,  0.4676 )
Calculations and Intervals on Original Scale
```

R Code:

```
sort(corr.npar.boot$t)[c(25, 975)]
```

```
> sort(corr.npar.boot$t)[c(25, 975)]
[1] 0.08368123 0.48425994
>
```

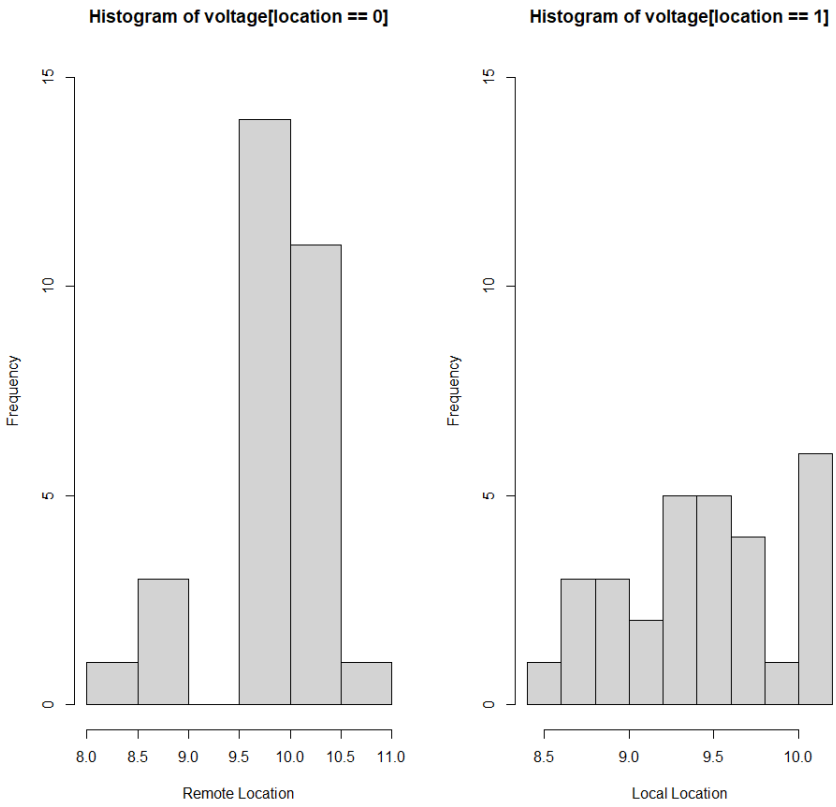
Observations:

1. We can see that the correlation estimate is 0.269
2. The strength of the relationship varies in degree based on the value of the correlation coefficient.
3. As the CI of correlation coeffecient lies between 0.08 to 0.48, we can say there is a positive correlation between two variables, but it is weak.

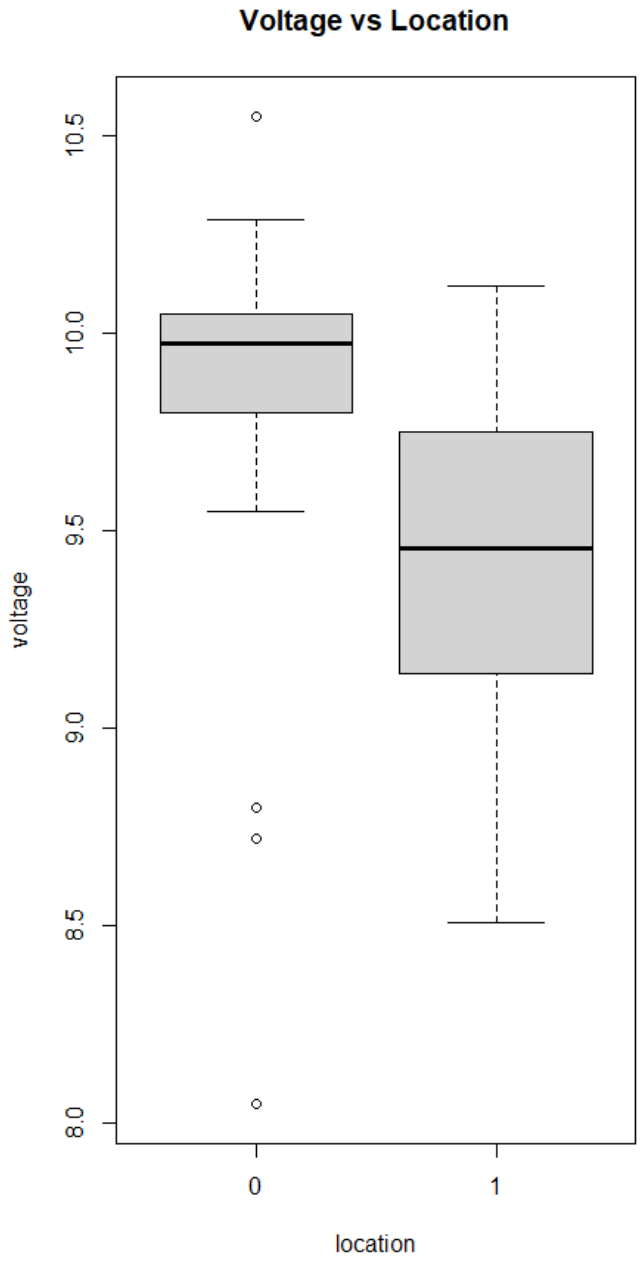
Q2. (7 points) Consider the data stored in the file VOLTAGE.DAT on eLearning. These data come from a Harris Corporation/University of Florida study to determine whether a manufacturing process performed at a remote location can be established locally. Test devices (pilots) were set up at both the remote and the local locations and voltage readings on 30 separate production runs at each location were obtained. In the dataset, the remote and local locations are indicated as 0 and 1, respectively.

Q2a) (1 points) Perform an exploratory analysis of the data by examining the distri butions of the voltage readings at the two locations. Comment on what you see. Do the two distributions seem similar? Justify your answer.

```
vlt = read.csv(file="D:/Fall'21/STATS/mini_project_4/voltage.csv")
location = vlt$location
voltage = vlt$voltage

par(mfrow=c(1,2))
hist(voltage[location==0],xlab = "Remote Location",ylim = c(0,15))
hist(voltage[location==1],xlab = "Local Location",ylim = c(0,15))
```



boxplot(voltage~location,main = "Voltage vs Location")

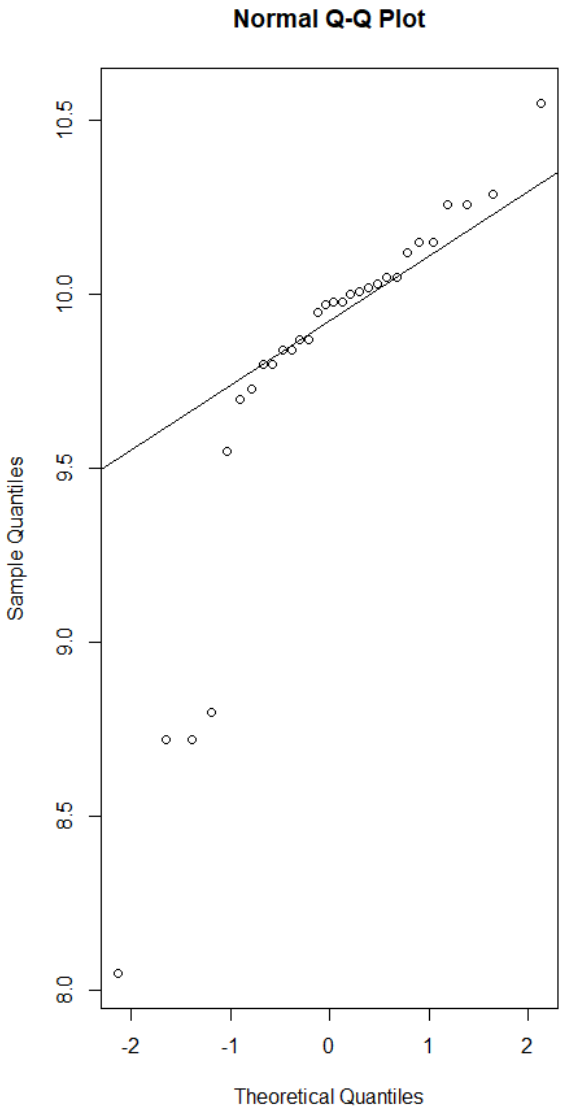


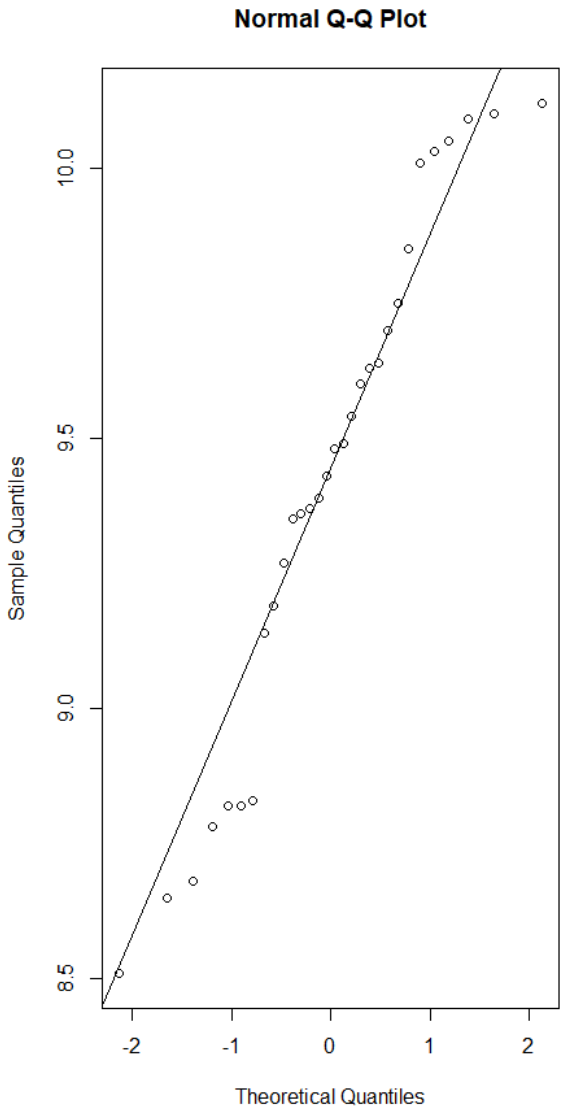
summary((voltage[location==0]))
summary((voltage[location==1]))

```
> summary((voltage[location==0]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.050   9.800   9.975   9.804  10.050  10.550
> summary((voltage[location==1]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.510   9.152   9.455   9.422   9.738  10.120
```

par(mfrow=c(1,2))

```
qqnorm((voltage[location==0]))
qqline((voltage[location==0]))

qqnorm((voltage[location==1]))
qqline((voltage[location==1]))
```





Observations:

1. From the box plot,histogram and summary of the data we can conclude that at local locations voltage used is less.
2. Also the distribution is slightly left skewed as mean<median and there are outliers in the case of remote location.
3. From QQ plot it shows approximate normality.

Q2b) (5 points) The manufacturing process can be established locally if there is no difference in the population means of voltage readings at the two locations. Does it appear that the manufacturing process can be established locally? Answer this question by constructing an appropriate confidence interval. Clearly state the assumptions, if any, you may be making and be sure to verify the assumptions.

1. We use 2 independent sample to calculate the CI using large samples technique.
2. Here null hypothesis H0: mean()-mean(local)=0
3. Alternate hypothesis H1: mean(remote)-mean(local)!=0

R code:

```
x1=mean(voltage[location==0])
x2=mean(voltage[location==1])
s1=var(voltage[location==0])
s2=var(voltage[location==1])
ci = (x1-x2) + c(-1,1)*qnorm(0.975)*sqrt((s1/30)+(s2/30))
ci
```

```
> ci
[1] 0.1228182 0.6398484
>
```

R code:

```
t.test(voltage[location==0],voltage[location==1], alternative = "two.sided", conf.level = 0.95, var.equal = FALSE)
```

```
> t.test(voltage[location==0],voltage[location==1], alternative = "two.sided", conf.level = 0.95, var.equal = FALSE)

        Welch Two Sample t-test

data:  voltage[location == 0] and voltage[location == 1]
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
```

Observations:

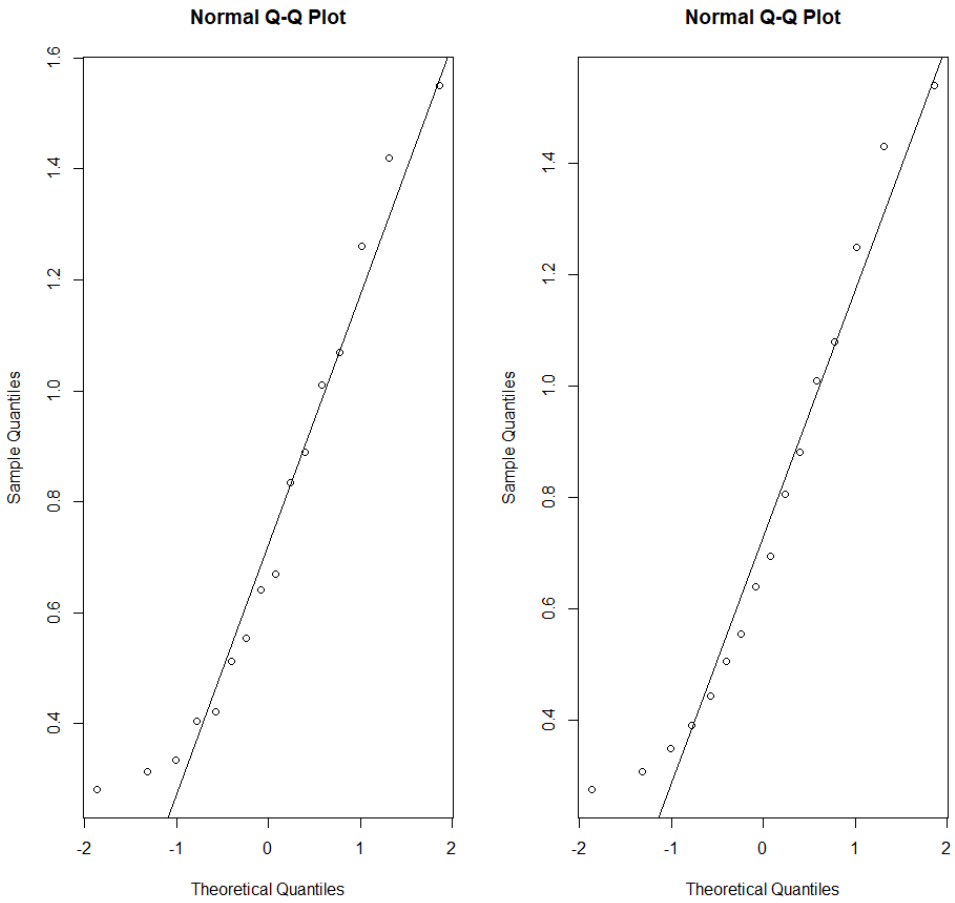
1. By using large sample size we get CI same as that of t test so the normality assumption is correct.
2. We observe that 0 does not lie in the confidence interval, we can say that the difference in population means of voltages at two locations will not be 0.
3. This means that we reject null hypothesis thus manufacturing process cannot be established at local locations.

Q2c) (1 point) How does your conclusion in (b) compare with what you expected from the exploratory analysis in (a)?

1. From part A using plots we observed that remote locations require a higher voltage compared to local locations
2. From part B we calculated the CI and observed that since 0 does not lie in CI and the CI is positive, the mean voltage required at remote location is more compared to that at local location
3. Thus, the manufacturing process cannot be done locally as the heavy equipment may require higher voltage.
4. Also, from qq plot in part A we could observe normality in data which could be observed in part B when constructing CI.

Q3. (7 points) The file VAPOR.DAT on eLearning provide data on theoretical (calculated) and experimental values of the vapor pressure for dibenzothiophene, a heterocycloaro matic compound similar to those found in coal tar, at given values of temperature. If the theoretical model for vapor pressure is a good model of reality, the true mean difference between the experimental and calculated values of vapor pressure will be zero. Perform an appropriate analysis of these data to see whether or not this is the case. Be sure to justify all the steps in the analysis.

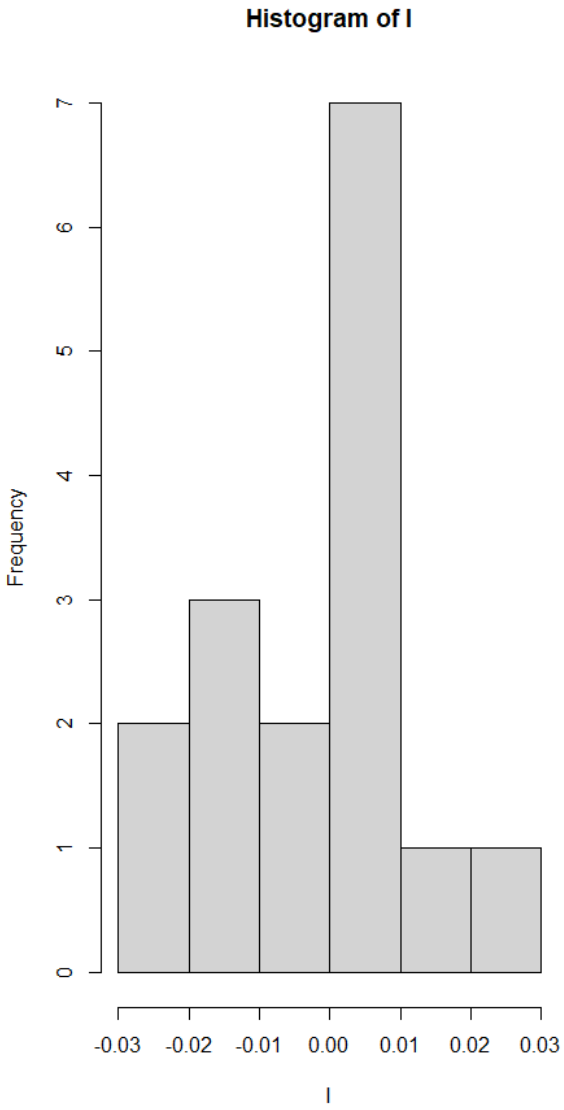
```
vapor = read.csv("D:/Fall'21/STATS/mini_project_4/VAPOR.csv")
temprature = vapor$temprature
theoretical = vapor$theoretical
experimental = vapor$experimental

par(mfrow=c(1,2))
qqnorm(theoretical)
qqline(theoretical)

qqnorm(experimental)
qqline(experimental)
```

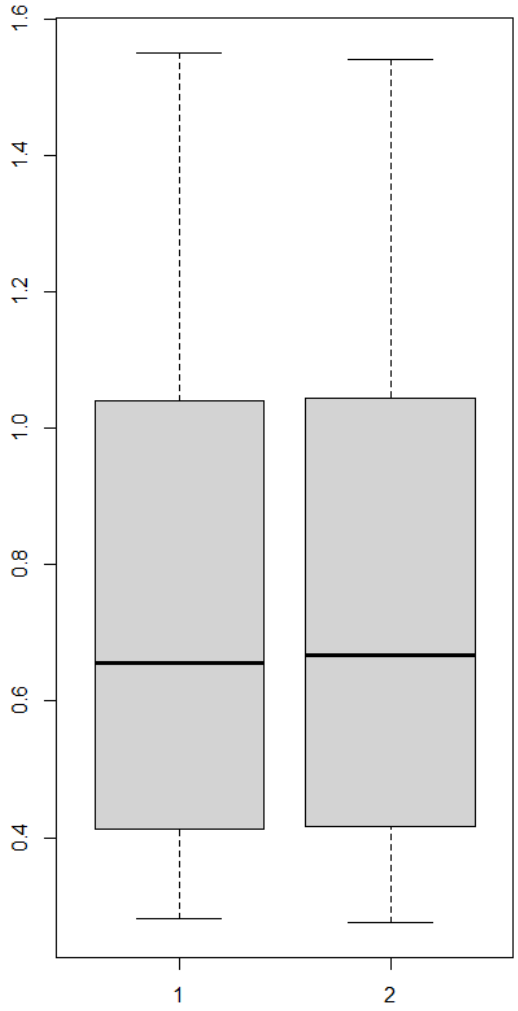


I = (theoretical - experimental)
summary(I)

```
>
> I = (theoretical - experimental)
> summary(I)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.0260000 -0.0100000  0.0040000  0.0006875  0.0085000  0.0290000
```

hist(I)
boxplot(theoretical,experimental)





Observations:

1. From the qq plot and the five point summary we can observe that the data is approximately normal.
2. From the boxplot we observe that the data is very similar and mean and median are nearly equal.
3. Data is slightly rightskewed as observe by the plot and the summary.
4. Now we have to test the mean difference between theoretical and experimental means.
5. Null hypothesis H0: mean(theoretical)-mean(experimental) = 0
6. Alternate hypothesis H1: mean(theoretical)-mean(experimental)!=0
7. we calculate the CI using the t distribution.

R code:

t.test(theoretical,experimental,paired = TRUE,conf.level = 0.95,var.equal = FALSE,alternative="two.sided")

```
        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

> |
```

Observations:

1. By observing the CI we can see that 0 lies in between the interval hence we accept the null hypothesis and reject the alternate hypothesis.
2. from t.test we can observe the mean difference very close to 0 which is also a strong evidence to accept H0.