Statistical Methods for Data Science

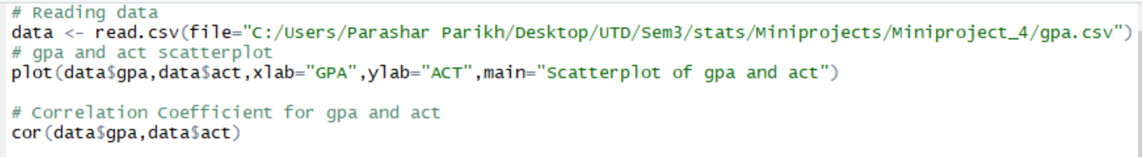
Mini Project 4

Mini Project Group No # 11
Names of group members: Parashar Parikh(pxp190016), Aakash Prajapati(axp190001)

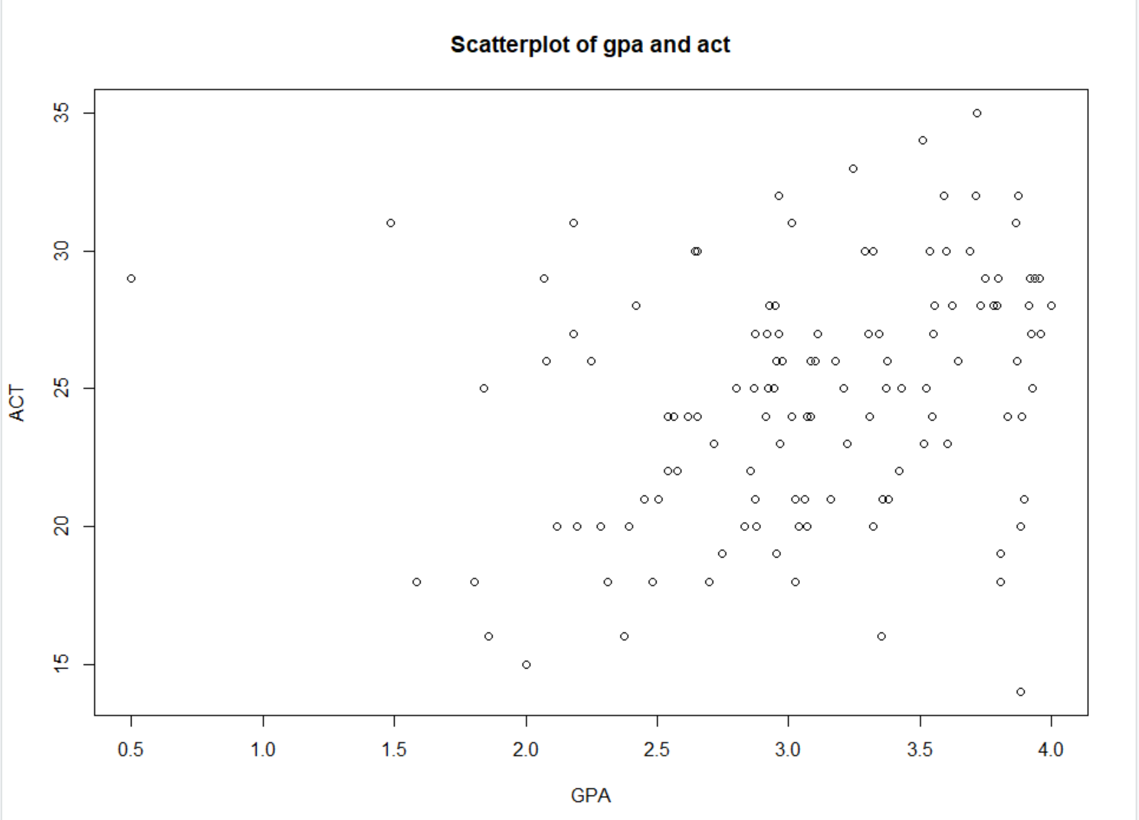
Contribution of each group member:
Both worked on all tasks and did documentation simultaneously.
Aakash
Q1 bootstrap calculation and plot graph, Q2 and Q3 confidence interval part
Parashar
Q1 scatterplot, Q2 and Q3 normal statistics of two data sets and plotted box and histogram plots.

Question 1.

1.

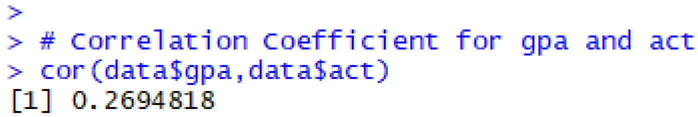


```
# Reading data
data <- read.csv(file="C:/Users/Parashar Parikh/Desktop/UTD/Sem3/stats/Miniprojects/Miniproject_4/gpa.csv")
# gpa and act scatterplot
plot(data$gpa,data$act,xlab="GPA",ylab="ACT",main="Scatterplot of gpa and act")

# Correlation Coefficient for gpa and act
cor(data$gpa,data$act)
```

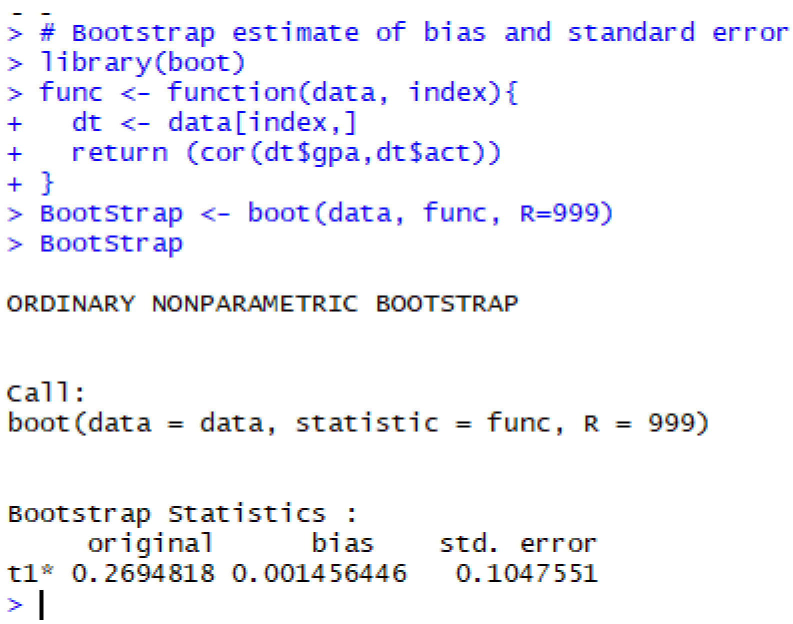


From the above scatterplot, we can observe the linear relation between student's ACT score and GPA which is not very strong. We can see from the plot that dependency between ACT score on GPA isn't strong because the plot is scattered. There is one more observation, increase in GPA corresponds to increase in ACT score.

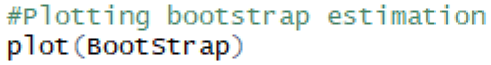
2.



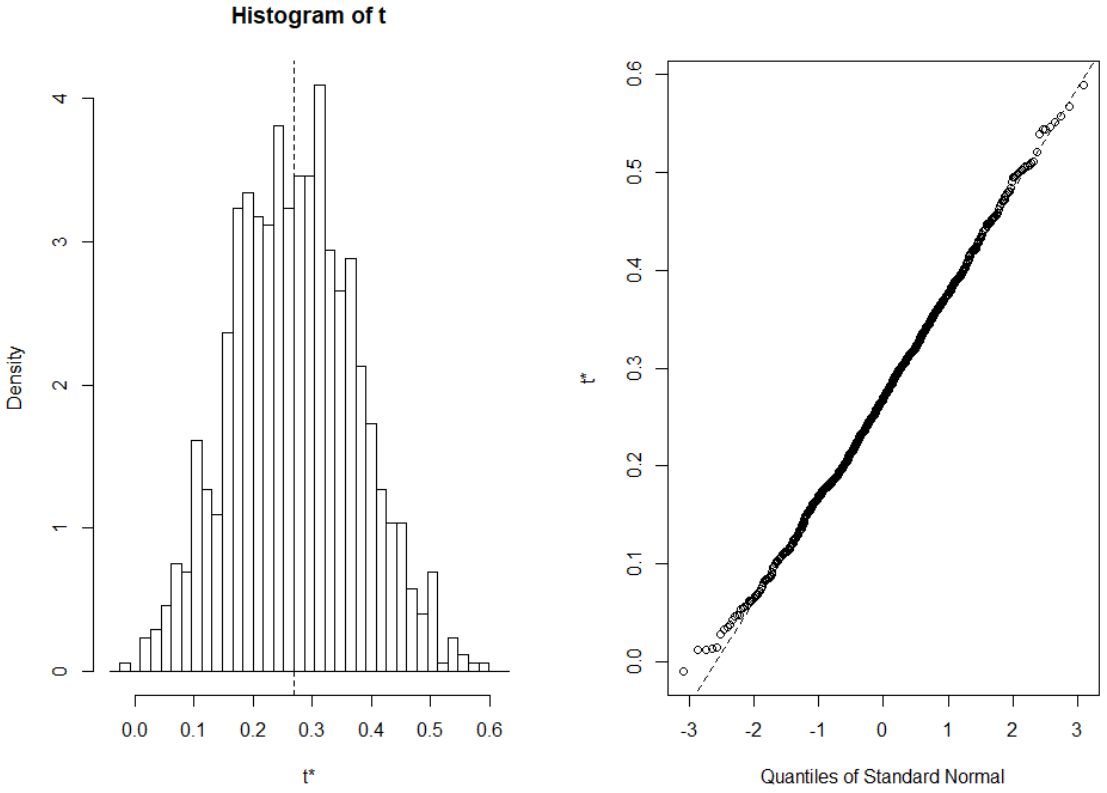
```
>
> # Correlation Coefficient for gpa and act
> cor(data$gpa,data$act)
[1] 0.2694818
```

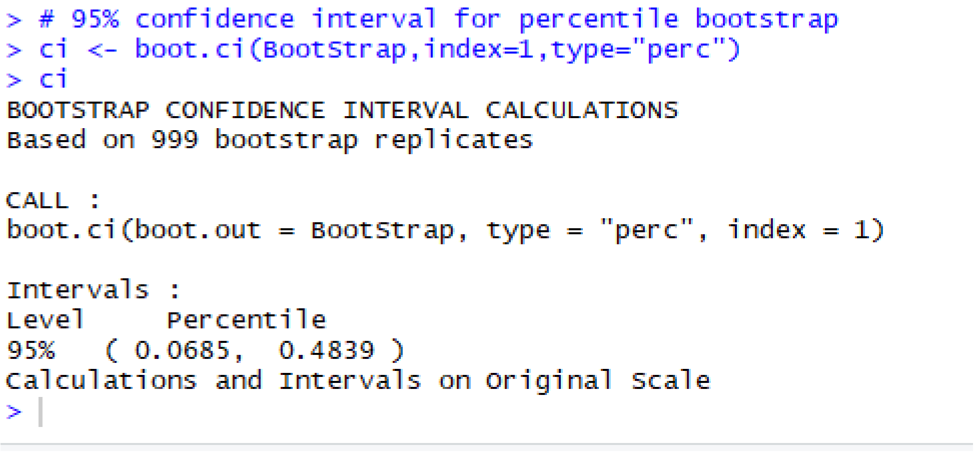


```
> # Bootstrap estimate of bias and standard error
> library(boot)
> func <- function(data, index){
+   dt <- data[index,]
+   return (cor(dt$gpa,dt$act))
+ }
> BootStrap <- boot(data, func, R=999)
> BootStrap

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = data, statistic = func, R = 999)


Bootstrap Statistics :
     original       bias    std. error
t1* 0.2694818 0.001456446   0.1047551
> 
```



```
#Plotting bootstrap estimation
plot(BootStrap)
```





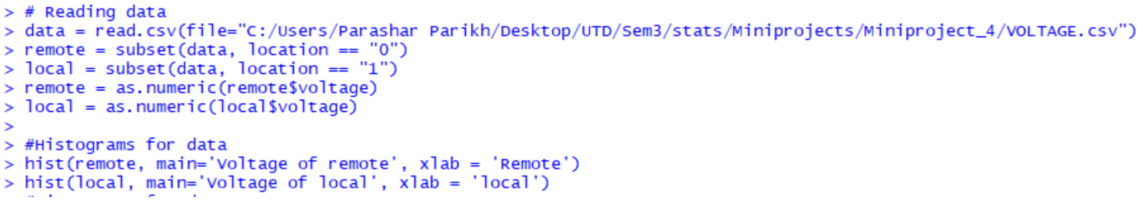
```
> # 95% confidence interval for percentile bootstrap
> ci <- boot.ci(BootStrap,index=1,type="perc")
> ci
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

CALL :
boot.ci(boot.out = BootStrap, type = "perc", index = 1)

Intervals :
Level     Percentile
95%   ( 0.0685,  0.4839 )
Calculations and Intervals on Original Scale
>
```

Observations:

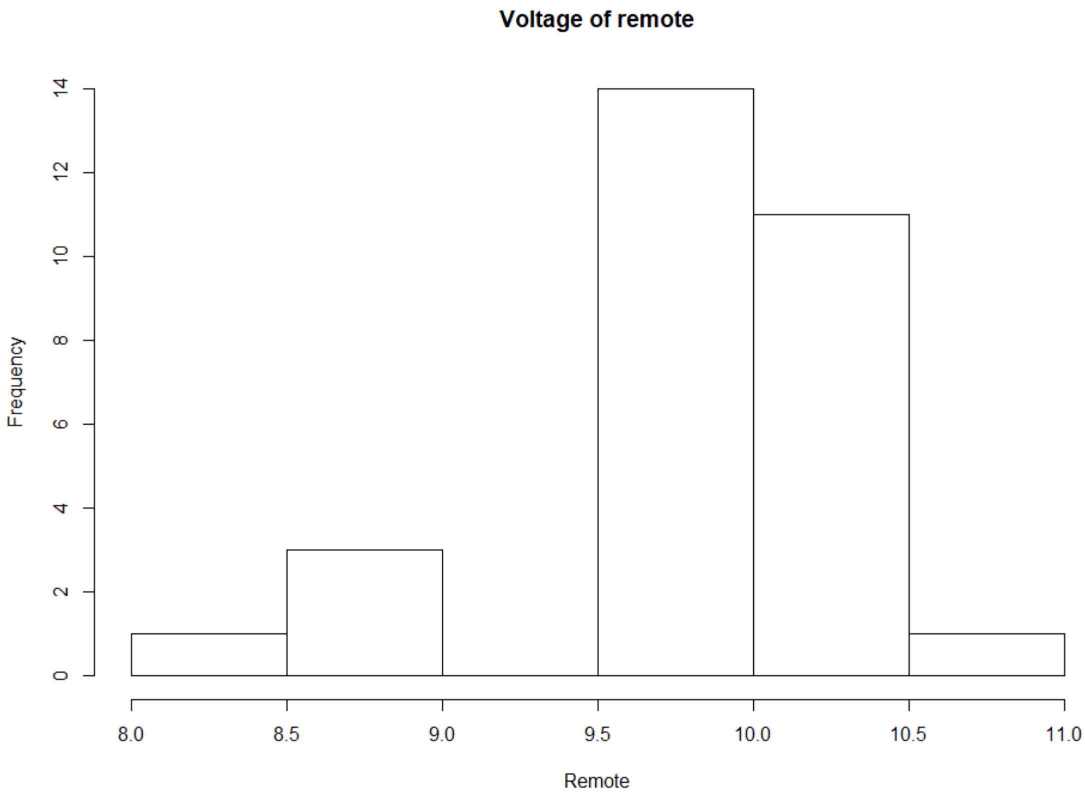
From the above bootstrap estimate of bias and standard error and confidence interval plot, we can observe that the data distribution becomes normal distribution after large number of bootstrap sampling. The correlation coefficient is positive that proves that proves the positive association in initial observation and also there are some outlines in the data. We can also say that 95% confidence interval shows, that the point estimation of correlation coefficient is within the range with 0.95 probability. We can also say that the strength of relation between GPA and ACT score is strong with positive association because bootstrap confidence interval in which the correlation coefficient is, is positive apart from the outliner data.

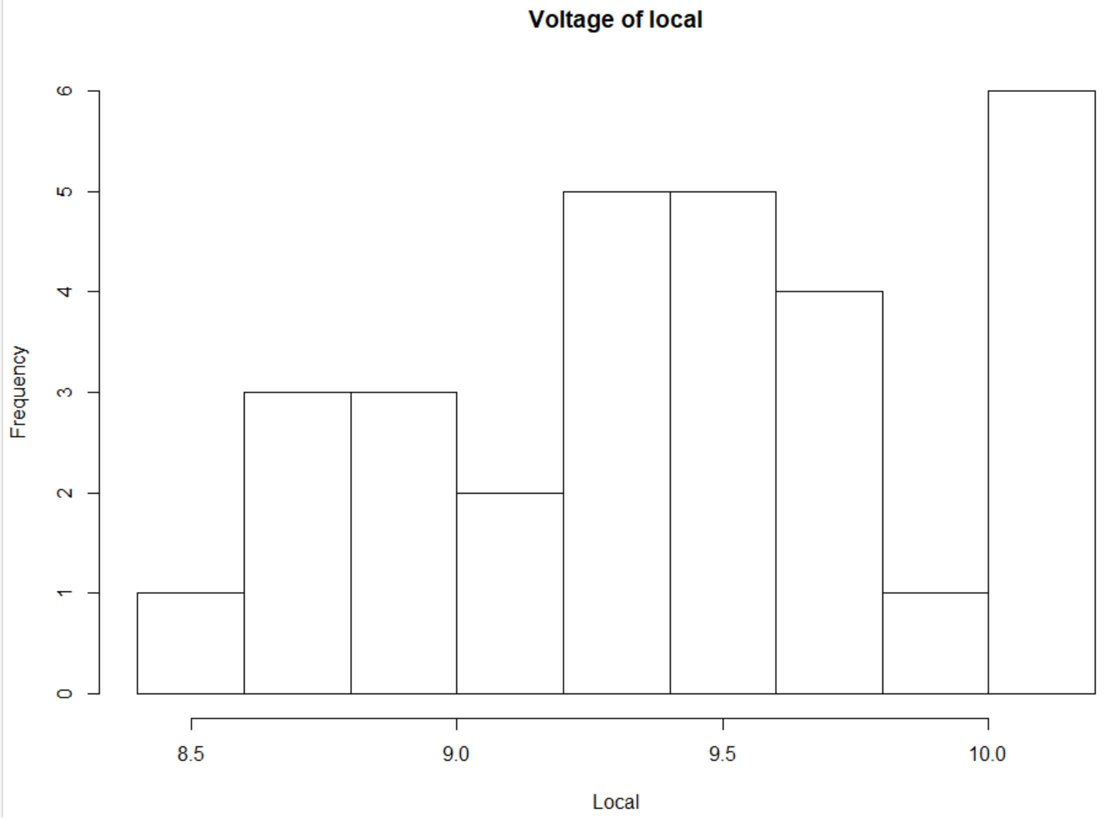
Question 2.

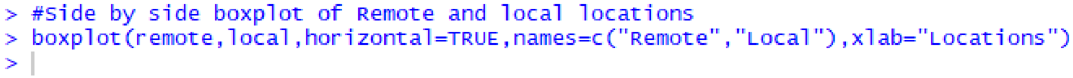
a.



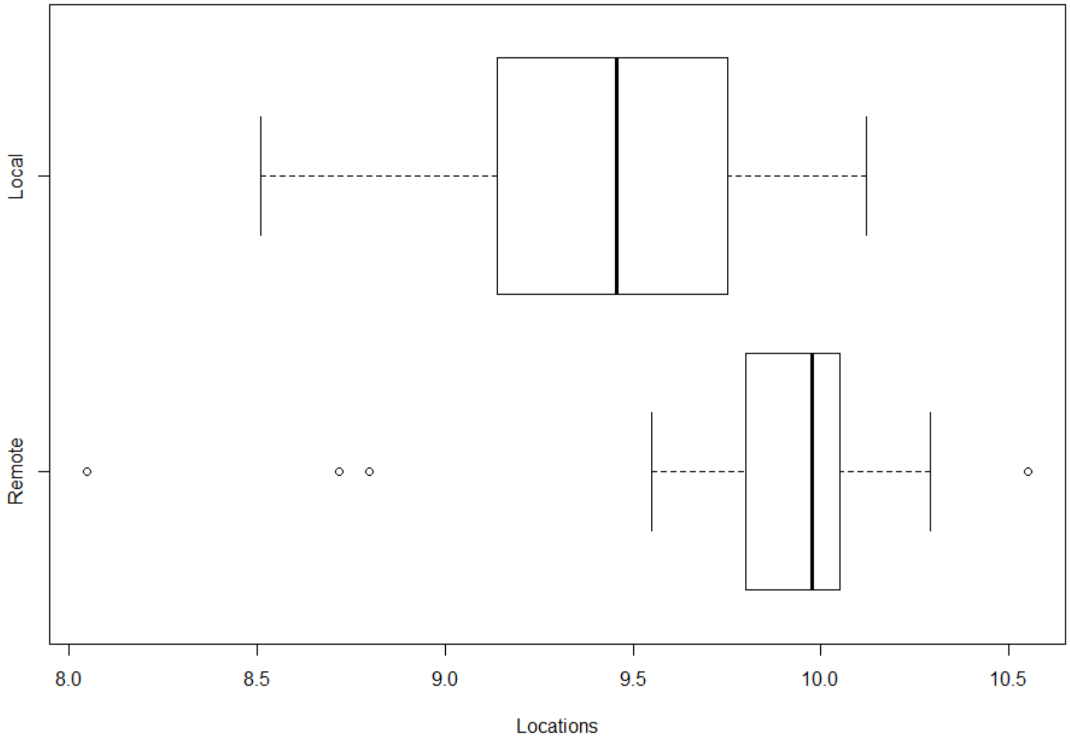
```
> # Reading data
> data = read.csv(file="C:/Users/Parashar Parikh/Desktop/UTD/Sem3/stats/Miniprojects/Miniproject_4/VOLTAGE.csv")
> remote = subset(data, location == "0")
> local = subset(data, location == "1")
> remote = as.numeric(remote$voltage)
> local = as.numeric(local$voltage)
>
> #Histograms for data
> hist(remote, main='Voltage of remote', xlab = 'Remote')
> hist(local, main='Voltage of local', xlab = 'local')
```

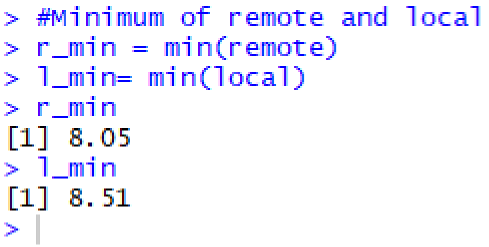

Voltage of remote


Voltage of local

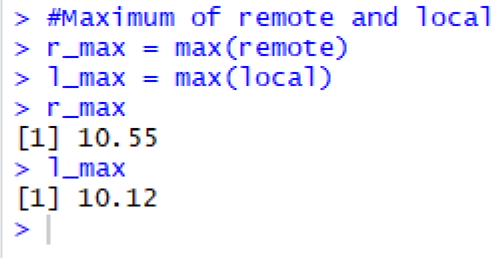


```
> #Side by side boxplot of Remote and local locations
> boxplot(remote,local,horizontal=TRUE,names=c("Remote","Local"),xlab="Locations")
> 
```

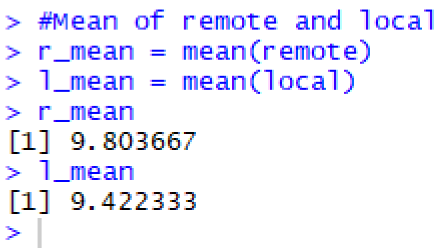




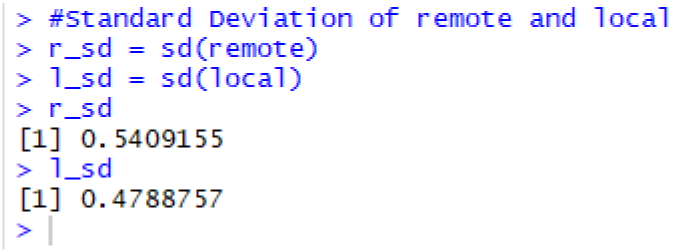
```
> #Minimum of remote and local
> r_min = min(remote)
> l_min= min(local)
> r_min
[1] 8.05
> l_min
[1] 8.51
> 
```



```
> #Maximum of remote and local
> r_max = max(remote)
> l_max = max(local)
> r_max
[1] 10.55
> l_max
[1] 10.12
> 
```



```
> #Mean of remote and local
> r_mean = mean(remote)
> l_mean = mean(local)
> r_mean
[1] 9.803667
> l_mean
[1] 9.422333
> 
```



```
> #Standard Deviation of remote and local
> r_sd = sd(remote)
> l_sd = sd(local)
> r_sd
[1] 0.5409155
> l_sd
[1] 0.4788757
> 
```

Analysis:

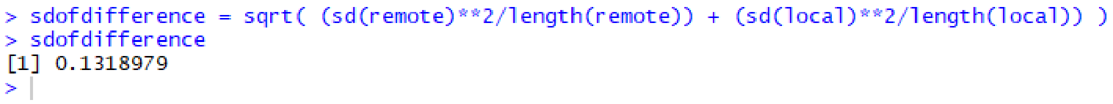
From the above, we can write the following table

	Remote	Local
No of Observation	30	30
Minimum	8.05	8.51
Maximum	10.55	10.12
Mean	9.8	9.42
Standard Deviation	0.54	0.48

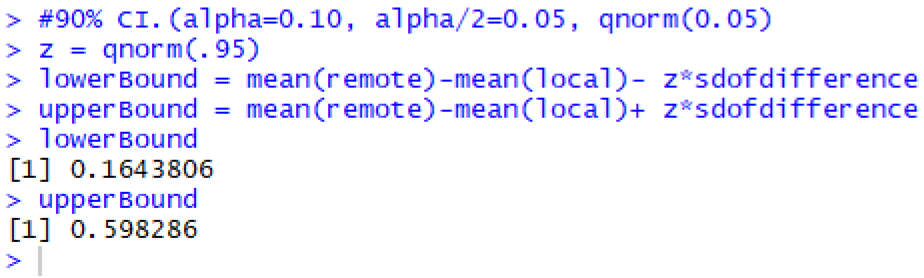
Conclusion:

From the histogram and box plot, we can conclude that, distributions are not similar. From the above side by side box plot, we can say that distribution range of the data isn't similar and there are many outliers. Also, from the above table for analysis, the mean is almost similar. Thus, we can conclude that the given distributions are different from each other.

b.



```
> sdofdifference = sqrt( (sd(remote)**2/length(remote)) + (sd(local)**2/length(local)) )
> sdofdifference
[1] 0.1318979
> 
```



```
> #90% CI.(alpha=0.10, alpha/2=0.05, qnorm(0.05)
> z = qnorm(.95)
> lowerBound = mean(remote)-mean(local)- z*sdofdifference
> upperBound = mean(remote)-mean(local)+ z*sdofdifference
> lowerBound
[1] 0.1643806
> upperBound
[1] 0.598286
> 
```

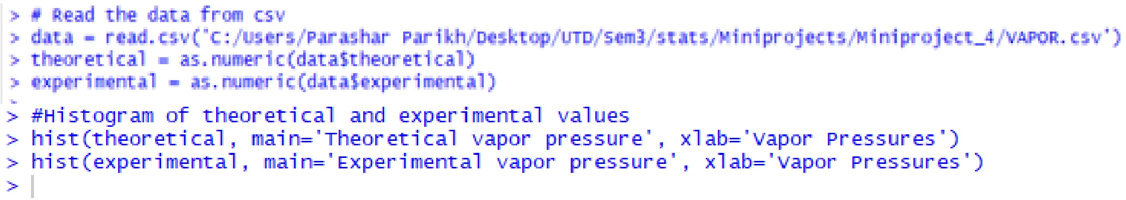
Observations:

We cannot manufacture locally. There is a difference in population means of voltage reading at two locations. We are getting confidence interval (0.1644, 0.5983) and 0 does not fall in this interval. Therefore, the voltage reading at local is smaller than the remote location.

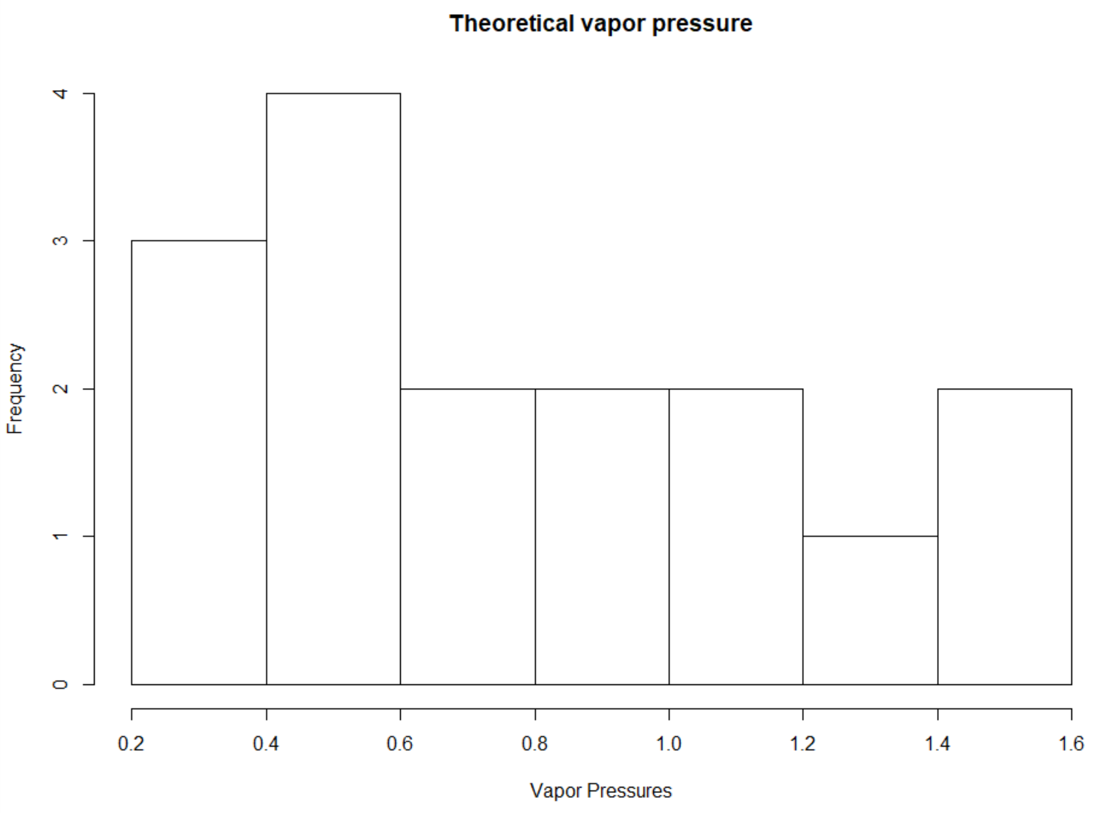
c.

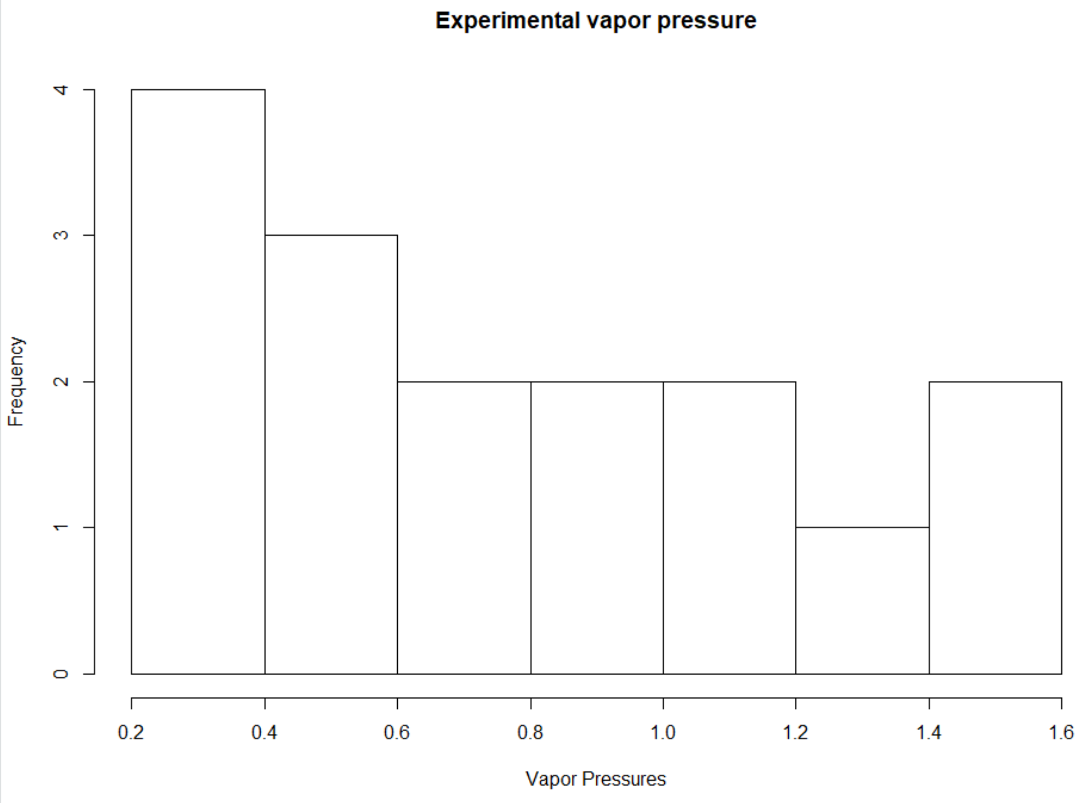
In (a) analysis we expected that the confidence interval doesn't contain 0 and in (b) we obtained that 0 doesn't fall in the 90% confidence interval. So, this conclude our conclusion and we can say that, the voltage reading at local is smaller than the remote location.

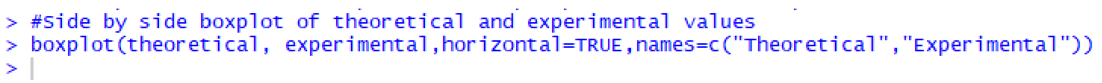
2.



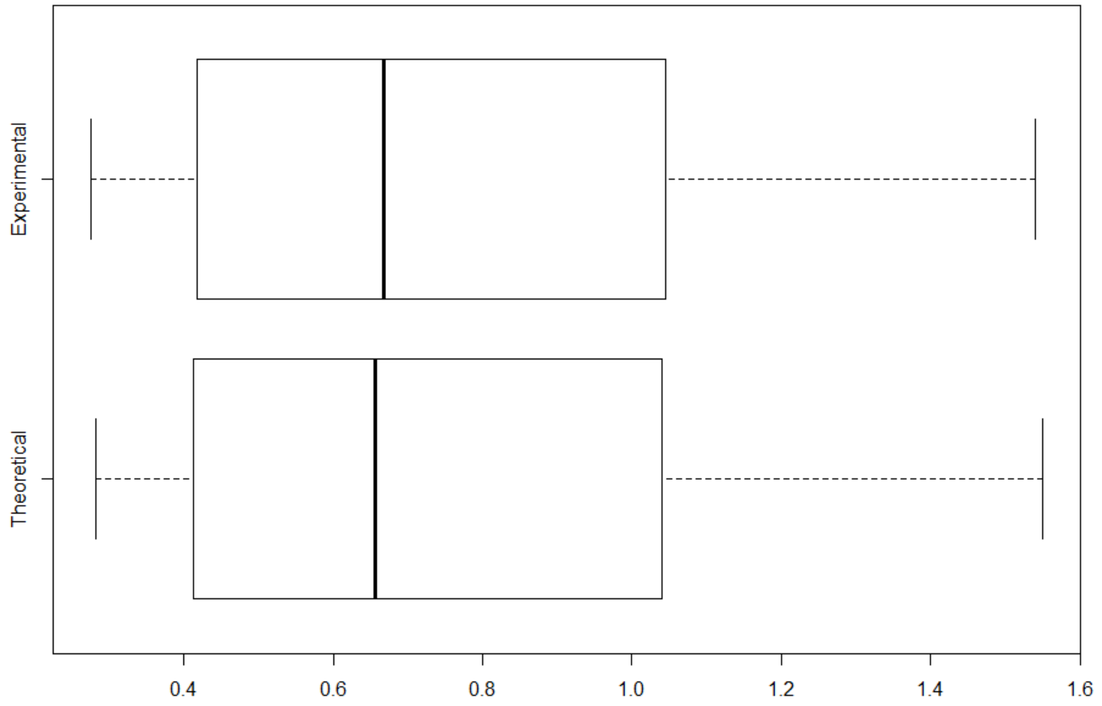
```
> # Read the data from csv
> data = read.csv('C:/Users/Parashar Parikh/Desktop/UTD/Sem3/stats/Miniprojects/Miniproject_4/VAPOR.csv')
> theoretical = as.numeric(data$theoretical)
> experimental = as.numeric(data$experimental)
> #Histogram of theoretical and experimental values
> hist(theoretical, main='Theoretical vapor pressure', xlab='Vapor Pressures')
> hist(experimental, main='Experimental vapor pressure', xlab='Vapor Pressures')
> 
```

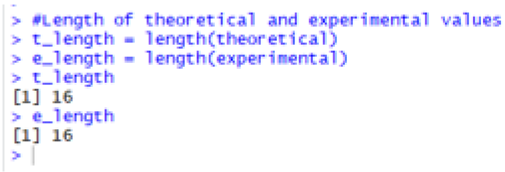

Theoretical vapor pressure
Frequency
0
1
2
3
4
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Vapor Pressures


Experimental vapor pressure
Frequency
0
1
2
3
4
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Vapor Pressures

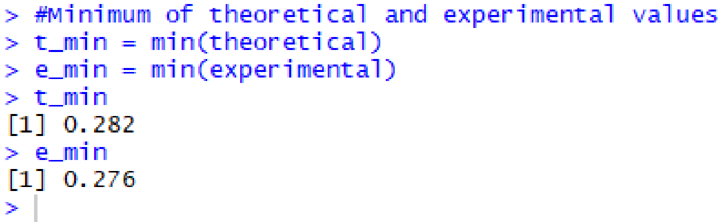


```
> #Side by side boxplot of theoretical and experimental values
> boxplot(theoretical, experimental,horizontal=TRUE,names=c("Theoretical","Experimental"))
> 
```

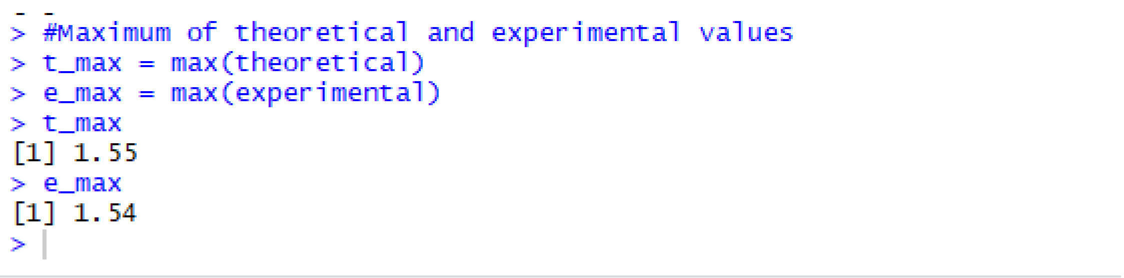




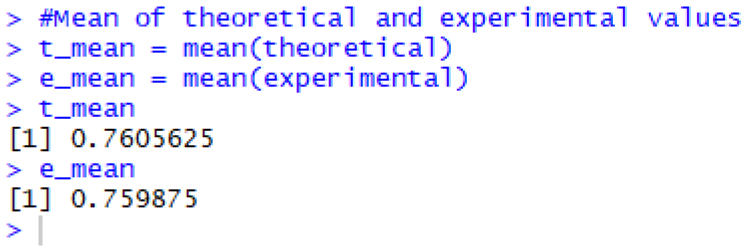
```
> #Length of theoretical and experimental values
> t_length = length(theoretical)
> e_length = length(experimental)
> t_length
[1] 16
> e_length
[1] 16
> 
```



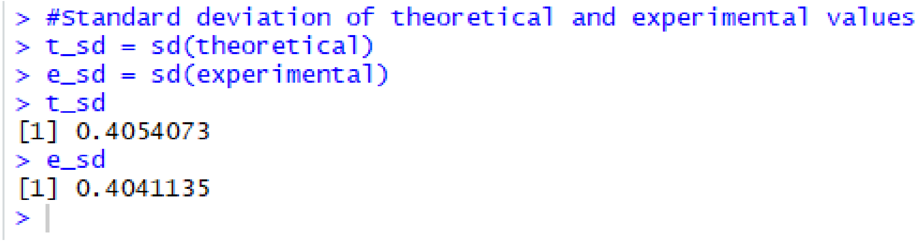
```
> #Minimum of theoretical and experimental values
> t_min = min(theoretical)
> e_min = min(experimental)
> t_min
[1] 0.282
> e_min
[1] 0.276
> 
```



```
> #Maximum of theoretical and experimental values
> t_max = max(theoretical)
> e_max = max(experimental)
> t_max
[1] 1.55
> e_max
[1] 1.54
> 
```



```
> #Mean of theoretical and experimental values
> t_mean = mean(theoretical)
> e_mean = mean(experimental)
> t_mean
[1] 0.7605625
> e_mean
[1] 0.759875
> 
```

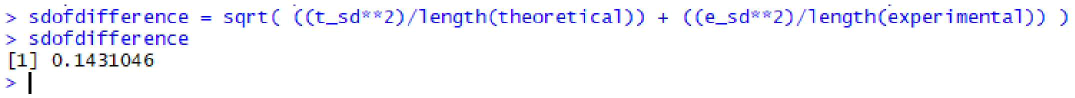


```
> #Standard deviation of theoretical and experimental values
> t_sd = sd(theoretical)
> e_sd = sd(experimental)
> t_sd
[1] 0.4054073
> e_sd
[1] 0.4041135
> 
```

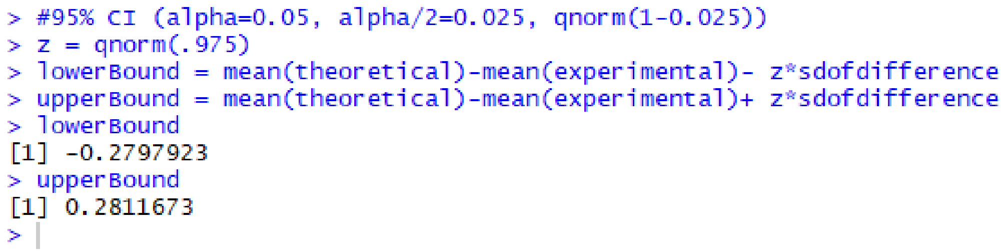
Analysis:

From the above, we can write the following table

	Remote	Local
No of Observation	16	16
Minimum	0.282	0.276
Maximum	1.55	1.54
Mean	0.760	0.759
Standard Deviation	0.405	0.405



```
> sdofdifference = sqrt( ((t_sd**2)/length(theoretical)) + ((e_sd**2)/length(experimental)) )
> sdofdifference
[1] 0.1431046
> 
```



```
> #95% CI (alpha=0.05, alpha/2=0.025, qnorm(1-0.025))
> z = qnorm(.975)
> lowerBound = mean(theoretical)-mean(experimental)- z*sdofdifference
> upperBound = mean(theoretical)-mean(experimental)+ z*sdofdifference
> lowerBound
[1] -0.2797923
> upperBound
[1] 0.2811673
> 
```

Observations:

From both the graphs, histogram and side by side boxplot we can say that theoretical and experimental values are same because distribution are identical.

We calculated the 95% confidence interval and 0 falls in the interval. So, we can conclude that for the large number of samples mean difference between both values (theoretical and experimental) will be almost 0.

Conclusion:

From the above two observations we can conclude that the statement is true that, the theoretical model of vapor pressure is a good model of reality.

Source code:

```
# Reading data
data<-read.csv(file="C:/Users/Parashar
Parikh/Desktop/UTD/Sem3/stats/Miniprojects/Miniproject_4/gpa.csv")
# gpa and act scatterplot
plot(data$gpa,data$act,xlab="GPA",ylab="ACT",main="Scatterplot of gpa and act")

# Correlation Coefficient for gpa and act
cor(data$gpa,data$act)

# Bootstrap estimate of bias and standard error
library(boot)
func <- function(data, index){
dt <- data[index,]
return (cor(dt$gpa,dt$act))
}
BootStrap <- boot(data, func, R=999)
BootStrap


#Plotting bootstrap estimation
plot(BootStrap)

# 95% confidence interval for percentile bootstrap
ci <- boot.ci(BootStrap,index=1,type="perc")
ci


# Reading data
data = read.csv(file="C:/Users/Parashar
Parikh/Desktop/UTD/Sem3/stats/Miniprojects/Miniproject_4/VOLTAGE.csv")
remote = subset(data, location == "0")
local = subset(data, location == "1")
remote = as.numeric(remote$voltage)
local = as.numeric(local$voltage)

#Histograms for data
hist(remote, main='Voltage of remote', xlab = 'Remote')
hist(local, main='Voltage of local', xlab = 'Local')

#Side by side boxplot of Remote and local locations
boxplot(remote,local,horizontal=TRUE,names=c("Remote","Local"),xlab="Locations")

#Length of remote and local
```

```
r_length = length(remote)
l_length = length(local)

#Minimum of remote and local
r_min = min(remote)
l_min= min(local)
r_min
l_min

#Maximum of remote and local
r_max = max(remote)
l_max = max(local)
r_max
l_max

#Mean of remote and local
r_mean = mean(remote)
l_mean = mean(local)
r_mean
l_mean

#Standard Deviation of remote and local
r_sd = sd(remote)
l_sd = sd(local)
r_sd
l_sd

sdofdifference = sqrt( (sd(remote)**2/length(remote)) + (sd(local)**2/length(local)) )
sdofdifference

#90% CI.(alpha=0.10, alpha/2=0.05, qnorm(1-0.05)
z = qnorm(.95)
lowerBound = mean(remote)-mean(local)- z*sdofdifference
upperBound = mean(remote)-mean(local)+ z*sdofdifference
lowerBound
upperBound



# Read the data from csv
data = read.csv('C:/Users/Parashar
Parikh/Desktop/UTD/Sem3/stats/Miniprojects/Miniproject_4/VAPOR.csv')
theoretical = as.numeric(data$theoretical)
experimental = as.numeric(data$experimental)
```

```
#Length of theoretical and experimental values
t_length = length(theoretical)
e_length = length(experimental)
t_length
e_length

#Minimum of theoretical and experimental values
t_min = min(theoretical)
e_min = min(experimental)
t_min
e_min

#Maximum of theoretical and experimental values
t_max = max(theoretical)
e_max = max(experimental)
t_max
e_max

#Mean of theoretical and experimental values
t_mean = mean(theoretical)
e_mean = mean(experimental)
t_mean
e_mean

#Standard deviation of theoretical and experimental values
t_sd = sd(theoretical)
e_sd = sd(experimental)
t_sd
e_sd


#Histogram of theoretical and experimental values
hist(theoretical, main='Theoretical vapor pressure', xlab='Vapor Pressures')
hist(experimental, main='Experimental vapor pressure', xlab='Vapor Pressures')

#Side by side boxplot of theoretical and experimental values
boxplot(theoretical, experimental,horizontal=TRUE,names=c("Theoretical","Experimental"))


sdofdifference = sqrt( ((t_sd**2)/length(theoretical)) + ((e_sd**2)/length(experimental)) )

#95% CI (alpha=0.05, alpha/2=0.025, qnorm(1-0.025))
z = qnorm(.975)
lowerBound = mean(theoretical)-mean(experimental)- z*sdofdifference
upperBound = mean(theoretical)-mean(experimental)+ z*sdofdifference
```

lowerBound
upperBound