Statistical Methods for Data Science
Mini Project #4

Group Members: Arpita Kumane, Harshali Dube

Contribution: Both the team members collaborated to learn R and worked on the project together. We analyzed, discussed, and efficiently worked to submit the two questions.

Question 1:

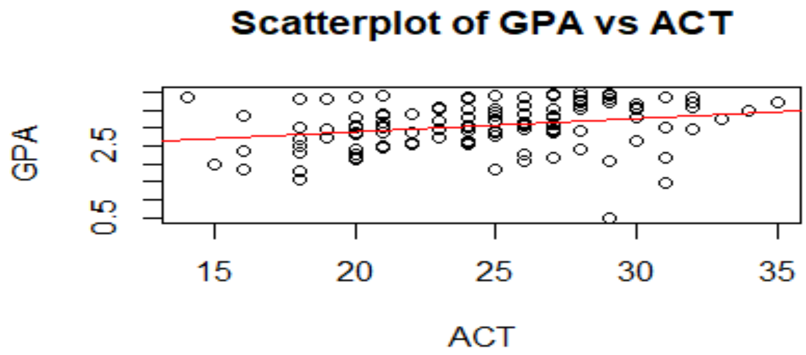
First, we used R's read.csv function to read our gpa dataset into a variable. Following data reading, we create a scatter plot of the dataset's gpa and act. To see the relationship between the two quantities, we utilized the abline function. We utilized the built-in R function cor, which returns the correlation of two variables supplied in its input list, to determine the correlation's true value.

Code:

```
#reading data from csv file
data = read.csv('miniproject_4/gpa.csv')
#extracting gpa column from data
gpa_data = data$gpa
#extracting act column from data
act_data = data$act
plot(act_data,gpa_data,xlab = "ACT",
      ylab = "GPA",main = "Scatterplot of GPA vs ACT")
abline(lm(gpa_data~act_data),col = "red")
#correlation b/w ACT and GPA
corr_gpa_act = cor(act_data,gpa_data)
print(corr_gpa_act)
library(boot)
cor.func <- function(data,indices)
{
  result <- cor(data[indices,1],data[indices,2])
  return(result)
}
cor_boot <- boot(data,statistic= cor.func, R = 10000)
cor_boot
#Point estimate of Boostrap value
mean(cor_boot$t)

#95% CI using Percentile BS
boot.ci(cor_boot,conf=0.95,type="perc")
```

```
> data = read.csv('miniproject_4/gpa.csv')
> #extracting gpa column from data
> gpa_data = data$gpa
> #extracting act column from data
> act_data = data$act
> plot(act_data,gpa_data,xlab = "ACT",
+       ylab = "GPA",main = "Scatterplot of GPA vs ACT")
> abline(lm(gpa_data~act_data),col = "red")
> #correlation b/w ACT and GPA
> corr_gpa_act = cor(act_data,gpa_data)
> print(corr_gpa_act)
[1] 0.2694818
> library(boot)
> cor.func <- function(data,indices)
+ {
+   result <- cor(data[indices,1],data[indices,2])
+   return(result)
+ }
> cor_boot <- boot(data,statistic= cor.func, R = 10000)
> cor_boot

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = data, statistic = cor.func, R = 10000)


Bootstrap Statistics :
     original       bias    std. error
t1* 0.2694818 0.004883417    0.105522
> #Point estimate of Boostrap value
> mean(cor_boot$t)
[1] 0.2743652
> #95% CI using Percentile BS
> boot.ci(cor_boot,conf=0.95,type = "perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

CALL :
boot.ci(boot.out = cor_boot, conf = 0.95, type = "perc")

Intervals :
Level     Percentile
95%   ( 0.0658,  0.4786 )
```

Plot:



The abline used to determine the relationship between gpa and act shows a positive slope in the graph above.
This slope's positive value denotes a positive relationship between the two variables, meaning that act grows as gpa rises and vice versa.
The correlation coefficient was then calculated, and the result was 0.26948. It is confirmed that there is a positive correlation between the quantities because the correlation value likewise turned out to be positive.
However, the correlation value was low, indicating a weak relationship between the two variables.

We utilized a boot function to estimate a sample's correlation value. Point estimates are sometimes used as predicted values in bootstrap samples.
The point estimate of correlation that we arrived at using the mean function was 0.2743652.

Using the boot.ci function, the confidence interval is generated using a 95% confidence level and a range of (0.0658, 0.4786).

Conclusion:

• The estimated value is fairly close to the actual value since the replications are sufficiently large (10000), which reduces bias.
• Because the SE is lower, the estimated values depart from the true value less, indicating that the estimator is accurate.
• After looking at CI, we can say that GPA and ACT scores have a good relationship.

Additionally, it can be deduced that the correlation value from the samples and the point estimate of correlation from the bootstrap are rather nearby.

Question 2:

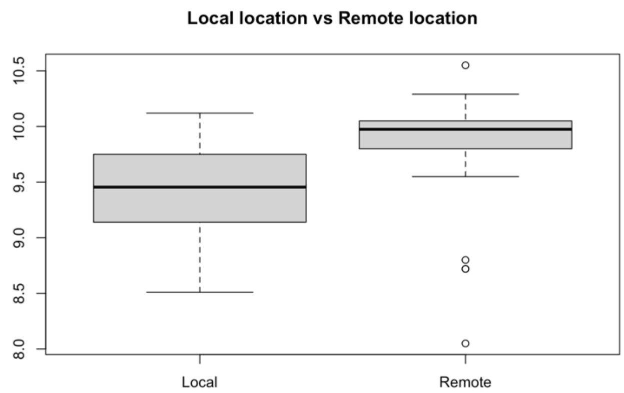
a)

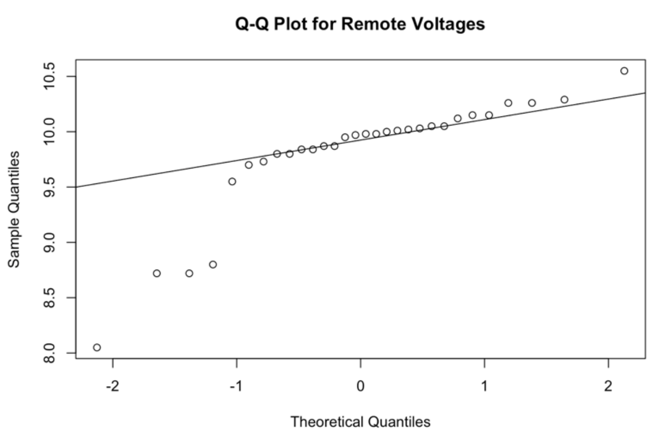
Code:

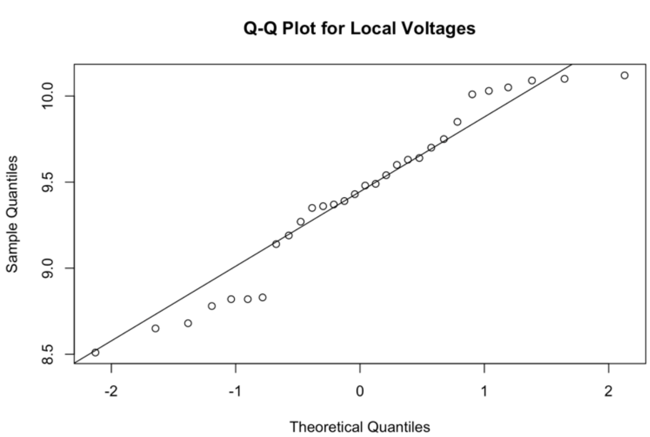
```r
#Extracting data from csv file
data=read.csv('voltage.csv')
#Obtaining remote and local locations voltages
remote_voltage = data$voltage[which(data$location == 0)]
local_voltage = data$voltage[which(data$location == 1)]

#Q-Q Plots of remote and local voltages
qqnorm(local_voltage,main="Q-Q Plot for Local Voltages")
qqline(local_voltage)
qqnorm(remote_voltage,main="Q-Q Plot for Remote Voltages")
qqline(remote_voltage)
par(mfrow=c(1,1))

#Boxplot of remote and local voltages
boxplot(local_voltage, remote_voltage, main="Local location vs Remote location"
         ,names=c("Local","Remote"))

summary(local_voltage)
summary(remote_voltage)
t.test(remote_voltage, local_voltage, alternative = "two.sided",
        conf.level = 0.95, var.equal = FALSE)
```

Result:

```
> #Extracting data from csv file
> data=read.csv('voltage.csv')
> #Obtaining remote and local locations voltages
> remote_voltage = data$voltage[which(data$location == 0)]
> local_voltage = data$voltage[which(data$location == 1)]
> #Side by side boxplot of remote and local voltages
> boxplot(local_voltage, remote_voltage, main="Local location vs Remote location"
+         ,names=c("Local","Remote"))
> summary(local_voltage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  8.510   9.152   9.455   9.422   9.738  10.120 
> summary(remote_voltage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  8.050   9.800   9.975   9.804  10.050  10.550 
> t.test(remote_voltage, local_voltage, alternative = "two.sided",
+        conf.level = 0.95, var.equal = FALSE)

	Welch Two Sample t-test

data:  remote_voltage and local_voltage
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 

> qqline(local_voltage)
> #Q-Q Plots of remote and local voltages
> qqnorm(local_voltage,main="Q-Q Plot for Local Voltages")
> qqline(local_voltage)
> qqnorm(remote_voltage,main="Q-Q Plot for Remote Voltages")
> qqline(remote_voltage)
> 
```


Local location vs Remote location
10.5
10.0
9.5
9.0
8.5
8.0
Local
Remote


Q-Q Plot for Remote Voltages
Sample Quantiles
Theoretical Quantiles


Q-Q Plot for Local Voltages
Sample Quantiles
Theoretical Quantiles

b)

2.

(b) Assume that the samples are independent
From the plots, it can be assumed that plots are normal.
Since IQR is different, variance cannot be same

∴ Using Satterthwaite's formula with T-distribution

Lower bound = $(\underset{mean}{remote} - \underset{mean}{local}) - Z_{\alpha/2} \times \hat{SE}(\underset{mean}{remote} - \underset{mean}{local})$

Upper bound = $(\underset{mean}{remote} - \underset{mean}{local}) - Z_{\alpha/2} \times \hat{SE}(\underset{mean}{remote} - \underset{mean}{local})$

Now,

remote mean − local mean = 0.3813

$S^2_{remote} = 0.292589$

$S^2_{local} = 0.229322$

$\therefore \hat{SE}(\underset{mean}{remote} - \underset{mean}{local}) = \sqrt{\frac{S^2_{remote}}{n_r} + \frac{S^2_{local}}{n_l}}$

$= \sqrt{\frac{0.292589}{30} + \frac{0.229322}{30}}$

$$= \sqrt{\frac{0.521911}{30}}$$

$$= 0.1318$$

$$z_{\alpha/2} = 1.96$$

Doing calculations,

lower bound = 0.122818

Upper bound = 0.639848

Hence, confidence interval is-

(0.122818, 0.639848)

The confidence interval from t-test is (0.117228, 0.645438)

As the confidence intervals are similar, our assumption of normal distribution holds current.

Since all values in the confidence interval are all positive and not zero, the difference b/w the mean is not equal to zero.

$\therefore$ Manufacturing process cannot be locally establis

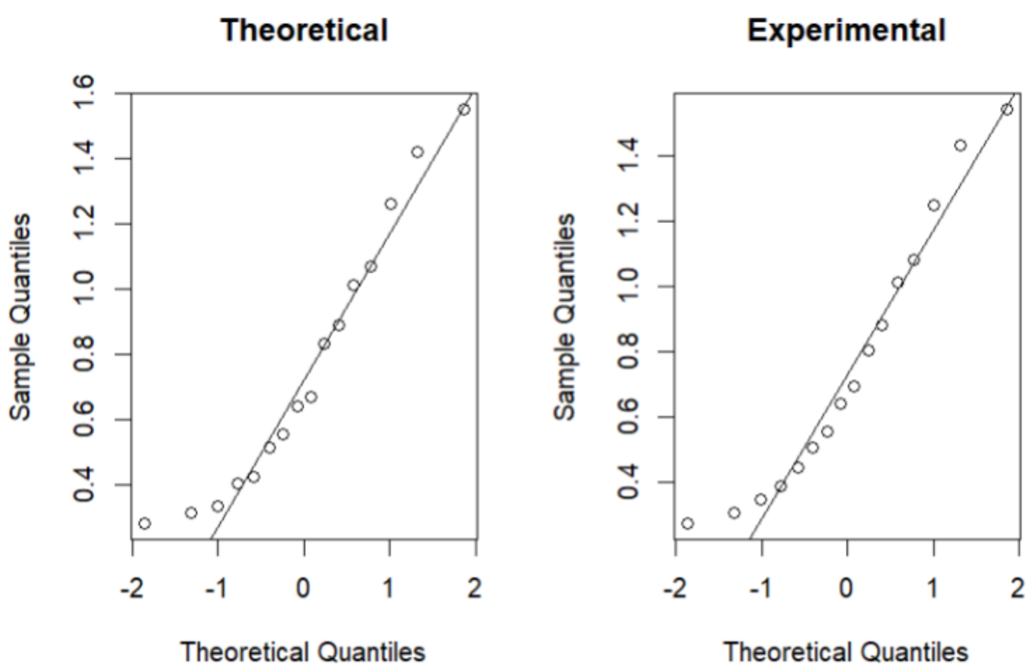
c)

So, from (a) we already noticed that voltage readings at remote are higher than those at local. Hence, we can easily say for any manufacturing process high voltage is required to fuel heavy equipment. Therefore, based on parts (a) and (b) we can easily conclude that manufacturing process have to be located into the remote location.

Question 3

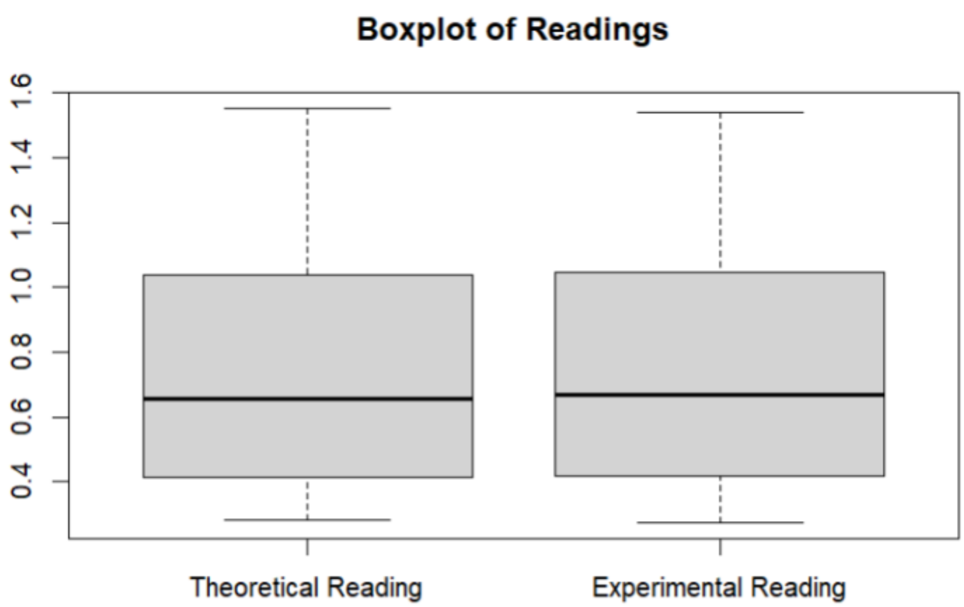
Here is code snippet of QQPlot's theoretical and experimental values.

```
> #Read data from csv
> vapor <- read.csv("/users/14699/documents/VAPOR.csv")
> #Draw qqplots
> par(mfrow= c(1,2))
> qqnorm(vapor$theoretical, main = "Theoretical")
> qqline(vapor$theoretical)
> qqnorm(vapor$experimental, main = "Experimental")
> qqline(vapor$experimental)
```



From QQplots we can say that samples treated as approximately normal.
Here is boxplot code snippet and its output

```
> #Draw boxplots and summaries
> boxplot(vapor$theoretical, vapor$experimental, names = c("Theoretical Reading", "Experimental Reading"),
+         main = "Boxplot of Readings")
> summary(vapor$theoretical)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2820  0.4175  0.6555  0.7606  1.0250  1.5500 
> summary(vapor$experimental)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2760  0.4305  0.6675  0.7599  1.0275  1.5400 
```



The boxplot shown above demonstrates how similar the two datasets are. Additionally, this conclusion is supported by the IQR & 5-Plot summary. As their means are higher than their medians, both distributions are right-skewed.

Let's compare the mean deviation between theoretical and experimental results in this instance.
True mean difference between t(bar) and e(bar) ==0 is the null hypothesis.
A different hypothesis is that there is no true mean difference between t(bar) and e(bar)!

Mean and standard deviation are calculated as follows: 0.0006875, 0.01421604, and 2.13145, respectively.
Lower Bound= 0.008262694
Upper Bound= -0.006887694

So, the Confidence interval calculated is (-0.006887694, 0.008262694)
These values we got in experimental values too.

```
> #Mean, Standard deviation, t(n-1) val, and confidence interval
> vapor.difference = vapor$theoretical - vapor$experimental
> mean(vapor.difference)
[1] 0.0006875
> sd(vapor.difference)
[1] 0.01421604
> qt(0.975, 15)
[1] 2.13145
> mean(vapor.difference) + c(-1,1) * qt(0.975, 15) * sd(vapor.difference)/ sqrt(16)
[1] -0.006887694  0.008262694
> #Confidence interval using t test
> t.test(vapor$theoretical, vapor$experimental, alternative= "two.sided", paired = TRUE, var.equal
+          = FALSE, conf.level = 0.95)

        Paired t-test

data:  vapor$theoretical and vapor$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
```

Since the value 0 lies within found interval, that means T(bar)-E(bar)=0. Hence the Null Hypothesis is accepted, and the true mean difference of theoretical and experimental values is zero. That is also supported by the boxplot.

By this we can conclude that there is no or minimal difference between the population means of theoretical and experimental pressures. So, we can state that theoretical model can be a good model of reality.