Mini Project #4

Group 2.

Group Members:

Manav Gupta [MXG220027]

Shalin Ronakkumar Kaji [SXK220263]

Contribution:

Manav : Solved Question 1, 3.

Shalin : Solved Question 2 and prepared the final draft of the documentation.

Question 1:

In this problem we are considering two variables : ACT score and GPA score for a particular student. We are plotting the values of x = ACT against y = GPA to develop a scatterplot.

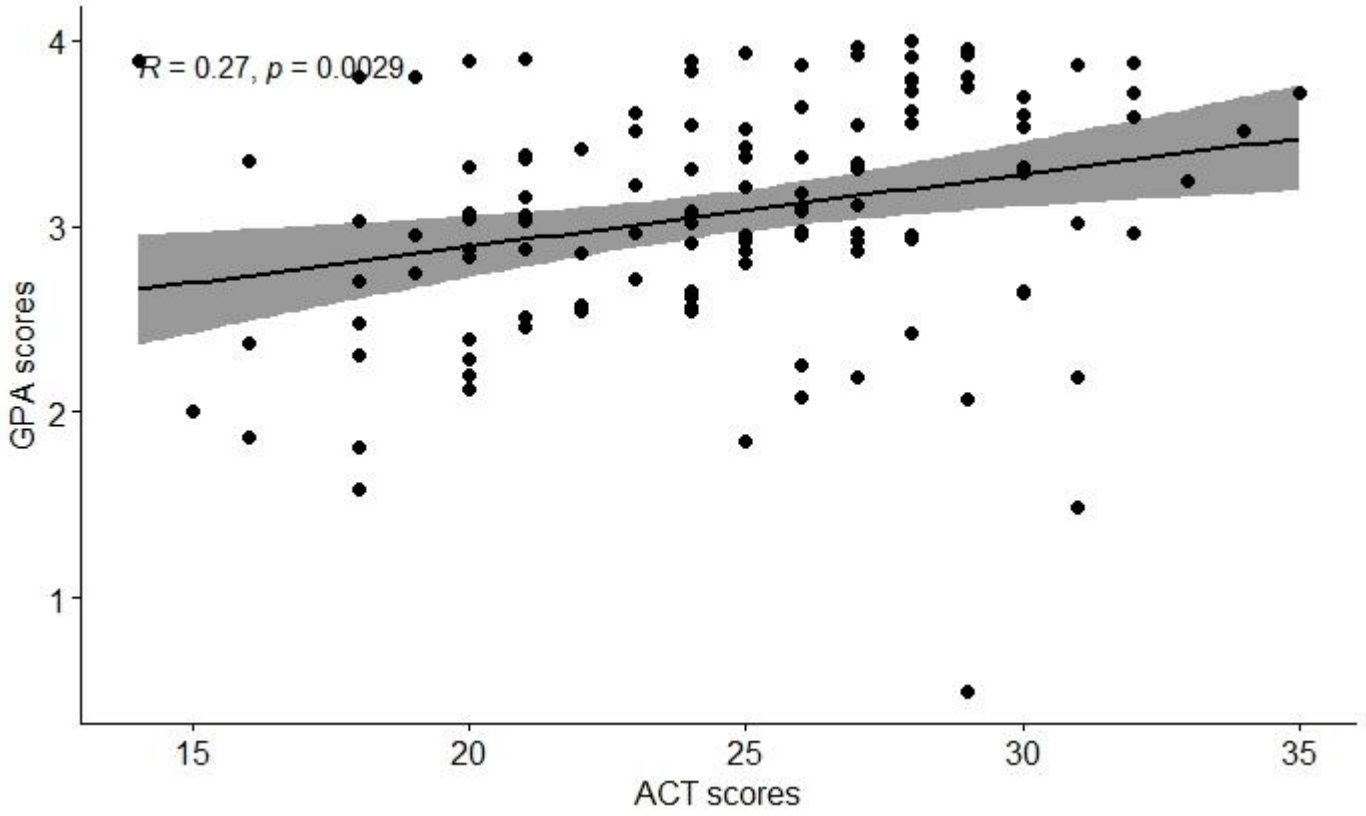
```
# Setting working directory to Proj-4 folder.
> setwd("C:/Users/Shalin Kaji/Desktop/UT-Dallas-Spr23/STATS-DS-Min.Chen/Mini-Proj-4")
> getwd()
# we need to load the 'boot' package first to load it.
> library(boot)
```

Now we create two vectors to store our Array of ACT and GPA scores.

```
# Observing our .csv file
> records <- read.csv("gpa.csv")
> xact <- records$act # vector of ACT scores.
> ygpa <- records$gpa # vector of GPA scores.
```

Now we create the Scatterplot using these two vectors.

```
> plot(xact,ygpa,xlab="ACT scores",ylab="GPA",pch=20) # plotting scatterplot
```



Pearson's Correlation Coefficient:

Used to measure of the strength and direction of the linear relationship between two variables. It is denoted by the symbol "ρ" and can take values between -1 and 1.

A value of +1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable also increases in a linear fashion. A value of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other variable decreases in a linear fashion. A value of 0 indicates no correlation, meaning that there is no linear relationship between the variables.

We make the use of the *cor()* function in R to find the correlation between the ACT and GPA scores.

```
rho <- cor(records$gpa,records$act) # finding Pearson's correlation coeff.
```

```
[1] 0.2694818
```

Thus, the value of the point estimate is given by rho.

After studying the value of 'rho' we can say that there exists a weak positive relationship between the ACT and GPA scores, as the value is near to zero.

The Bootstrap method allows us to produce multiple samples from the given sample using re-sampling techniques. This allows us to calculate the bootstrap estimates of bias and standard error for statistic of interest = correlation.

```
# bootstrap estimates of bias and standard error for corr coeff.
> correl <- function(records,subsample){
>  X <- records[subsample,1]; Y <- records[subsample,2];
  return( cor(X,Y) )
}
```

```
> BootR <- boot(data=records, statistic=correl, R=9999)
```

```
> BootR <- boot(data=records, statistic=correl, R=9999)
> BootR

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = records, statistic = correl, R = 9999)


Bootstrap Statistics :
     original      bias    std. error
t1* 0.2694818 0.001764192   0.1048731
>
```

Calculating Confidence Intervals using the built-in boot.ci() function in R:

```
> BootStat <- boot(data=records, statistic=correl, R=9999)
> BootStat
> boot.ci(boot.out = BootStat)
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 9999 bootstrap replicates

CALL :
boot.ci(boot.out = BootStat)

Intervals :
Level      Normal              Basic
95%   ( 0.0599,  0.4726 )   ( 0.0617,  0.4727 )

Level     Percentile            BCa
95%   ( 0.0663,  0.4773 )   ( 0.0483,  0.4602 )
Calculations and Intervals on Original Scale
Warning message:
In boot.ci(boot.out = BootStat) :
  bootstrap variances needed for studentized intervals
>
```

Now verifying the CI calculated using the Percentile Bootstrap formula.

Percentile Bootstrap:

$$\text{CI: } \left[\hat{\theta}^*_{\left((b+1)(\alpha/2)\right)},\ \hat{\theta}^*_{\left((b+1)(1-\alpha/2)\right)}\right].$$

```
> BootR <- boot.ci(data=records, statistic=correl, R=9999)

# 95% CI using percentile bootstrap method.

> sort(BootR$t)[c(250, 9750)]
```

```
> # 95% CI using percentile bootstrap method.
> sort(BootR$t)[c(250, 9750)]
[1] 0.06674067 0.47327543
>
```

Interpreting the results, we can say that, on drawing n samples – 95% of the Confidence Intervals will contain the true value of the parameter ie. Correlation, and since it is [0.067, 0.473] there exists only a weak positive relationship from which we can establish anything concretely.

R Code:

```
# STATS-Mini-Project-4

# Setting working directory to Proj-4 folder.

setwd("C:/Users/Shalin Kaji/Desktop/UT-Dallas-Spr23/STATS-DS-Min.Chen/Mini-Proj-4")

getwd()

# we need to load the 'boot' package first to load it.

library(boot)


# Observing our .csv file

records <- read.csv("gpa.csv")

xact <- records$act # vector of ACT scores.

ygpa <- records$gpa # vector of GPA scores.


rho <- cor(records$gpa,records$act) # finding Pearson's correlation coeff.

rho

# found the point estimate of rho.


ggpubr::ggscatter(records, x = "act", y = "gpa",

          add = "reg.line", conf.int = TRUE,

          cor.coef = TRUE, cor.method = "pearson",

          xlab = "ACT scores", ylab = "GPA scores")


# bootstrap estimates of bias and standard error for corr coeff.

correl <- function(records,subsample){

  X <- records[subsample,1]; Y <- records[subsample,2];

  return( cor(X,Y) )

}


BootStat <- boot(data=records, statistic=correl, R=9999)

BootStat

boot.ci(boot.out = BootStat)

BootR <- boot.ci(data=records, statistic=correl, R=9999)
```

Question 2:

(a)

To perform an exploratory analysis of the data to examine the distributions of the voltage readings at the two locations we are using the boxplot.
We are using following R code.

```
# Setting working directory to Proj-4 folder.

> setwd("C:/Users/dpd140130/OneDrive - The University of Texas at Dallas/CS 6313/Projects/04")

> getwd()

# we need to load the 'boot' package first to load it.

> library(boot)
```

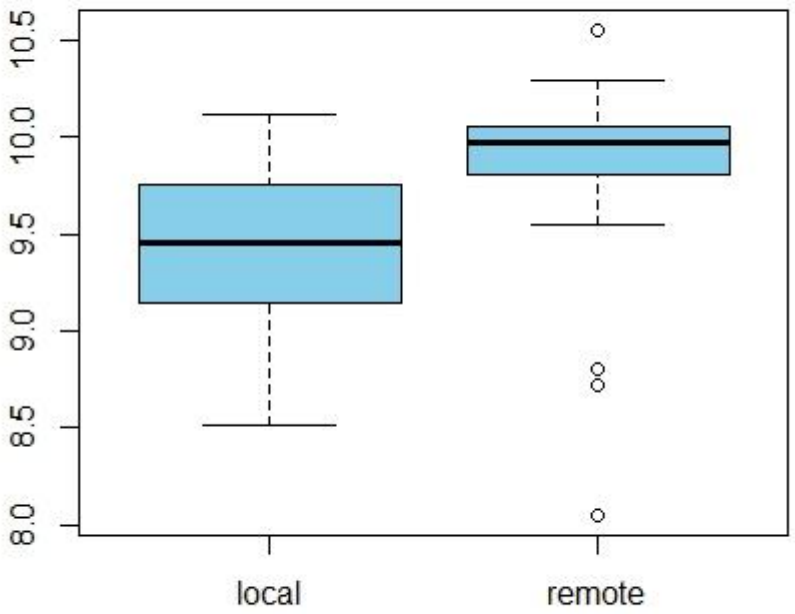
Now we create two lists to store remote and local locations which are given in the dataset.

```
# Observing our .csv file

> companyrecs <- read.csv("voltage.csv")

> remotebr <- companyrecs[companyrecs$location==0,]

> localbr <- companyrecs[companyrecs$location==1,]
```

Now we compare remote and local data using side by side box plots and determine the summary statistics.

```
> boxplot (localbr$voltage, remotebr$voltage, names = c("local","remote"), col='skyblue')
```

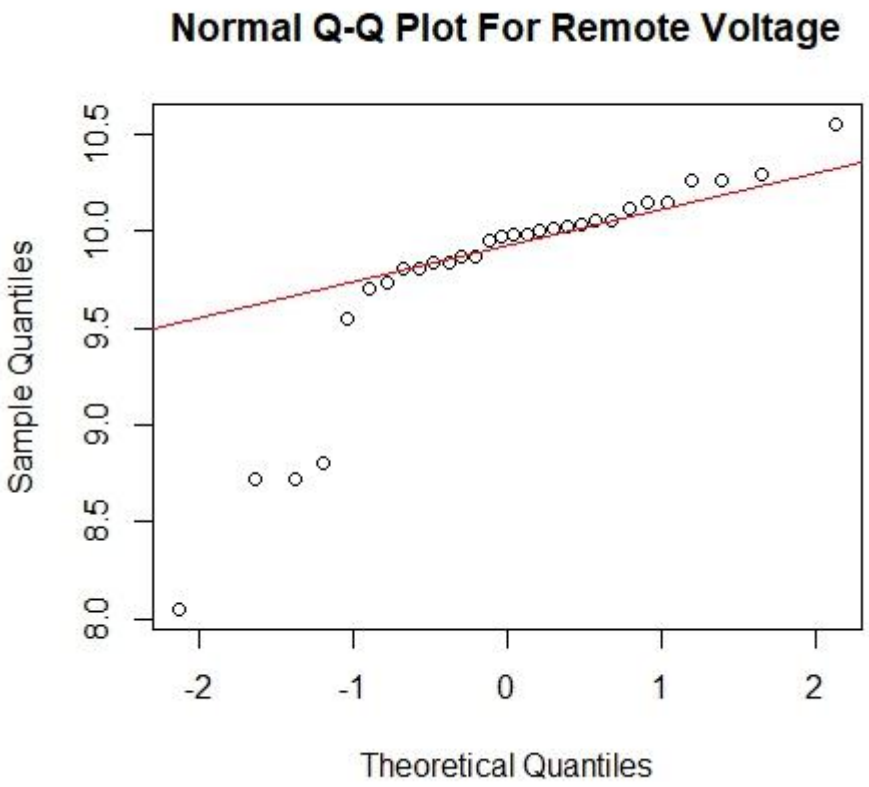
```
> summary(remotebr$voltage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.050   9.800   9.975   9.804  10.050  10.550
> summary(localbr$voltage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.510   9.152   9.455   9.422   9.738  10.120
```



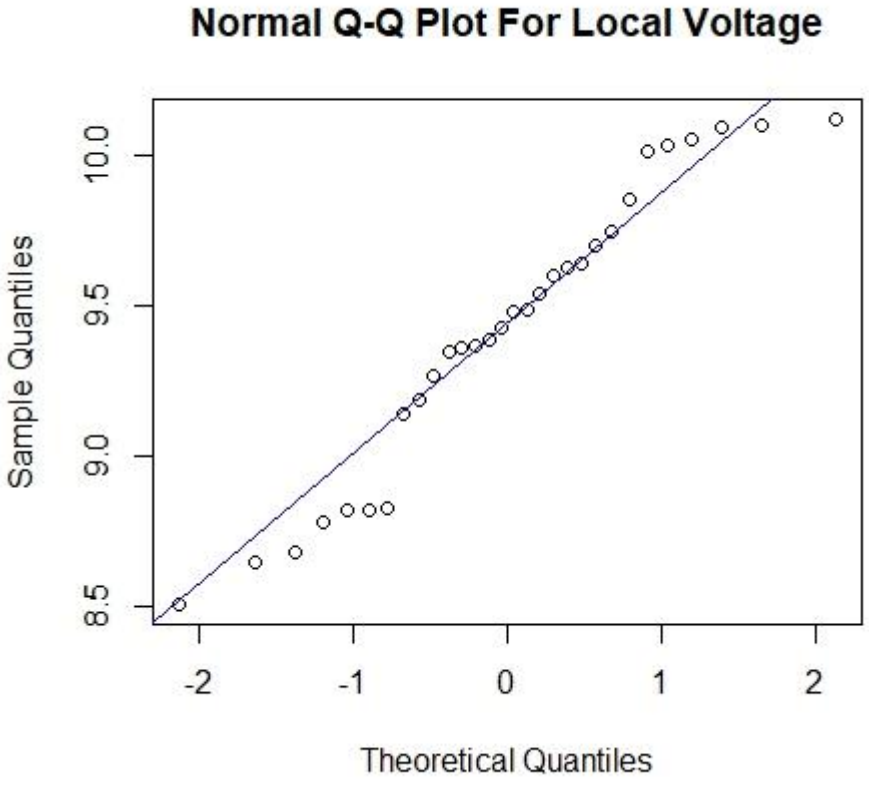
The summary statistics show that both station's mean and median are not equal. From the boxplot, one can see that the Inter quartile range (IQR) for voltage values manufactured in remote areas is smaller than the Inter quartile range(IQR) for voltage values in local areas. Thus, we can conclude from the boxplot that remote location voltage distribution is right skewed.

Now we need to see whether the values of remote and local data has a normal distribution or not. For this we use qqnorm() and qqline() functions.

```
> qqnorm(remotebr$voltage, main = 'Normal Q-Q Plot For Remote Voltage')
> qqline(remotebr$voltage, col='red')
```




```
> qqnorm(localbr$voltage, main = 'Normal Q-Q Plot For Local Voltage')
> qqline(localbr$voltage, col='blue')
```



One can see from above plots that most of the points are lying near or on the qqline. This indicates that both the dataset follow the normal distribution.

R Code:

```
# Setting working directory to Proj-4 folder.
setwd("C:/Users/dpd140130/OneDrive - The University of Texas at Dallas/CS
6313/Projects/04")
getwd()

#Load the 'Boot' Package
library(boot)

# Observing our .csv file
companyrecs <- read.csv("voltage.csv")
remotebr <- companyrecs[companyrecs$location==0,]
localbr <- companyrecs[companyrecs$location==1,]

#Creating Boxplot
boxplot(localbr$voltage,remotebr$voltage,names = c("local","remote"),
col='skyblue')

#Creating Summary Statistics
summary(remotebr$voltage)
summary(localbr$voltage)

#Creating QQ Plots
qqnorm(remotebr$voltage, main = 'Normal Q-Q Plot For Remote Voltage')
qqline(remotebr$voltage, col='red')
qqnorm(localbr$voltage, main = 'Normal Q-Q Plot For Local Voltage')
qqline(localbr$voltage, col='blue')
```

(b)

Assumptions that we are making:

- We are assuming that the data of both remote and local manufactured have normal distribution because the sample size(n) is large enough.
- Since information about variances is not given we assume they are unknown and not equal.

Verification of Assumption:

- From the qqplots drawn in Q2(a) we can verify that our assumption of normal distribution is correct.
- We will now find the variance of the local and remote voltage.

```
> #Finding Variance
> remotebr_var <- var(remotebr$voltage)
> remotebr_var
[1] 0.2925895
> localbr_var <- var(localbr$voltage)
> localbr_var
[1] 0.229322
```

So, variances are not equal, and our second assumption is also correct.

Now, we find the 95% confidence interval for the given problem statement because the distribution is normal, and variance is unequal.

For that we consider applying null hypothesis on the given problem statement. Then,
H0: Population mean of voltage readings at the two stations are same.
i.e.; mean(remotedata) -mean(localdata) = 0. This means that the process can be established locally because there is no difference of mean.

Alternative hypothesis will be,
H1: Difference in mean of data between remote and local manufactured voltages is not equal to 0.
i.e.; mean(remotedata) -mean(localdata) ≠ 0

If the p-value obtained from the t-test is less than 0.05, the null hypothesis is rejected as not being significant. The p value and confidence interval values are determined using the t-test.

```
>#Evaluating Null Hypothesis
> t.test(remotebr$voltage,localbr$voltage,paired = F, var.equal = F, conf.level = 0.95)

        Welch Two Sample t-test

data:  remotebr$voltage and localbr$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
```

Since the p-value is less than the significant p-value 0.05, we will reject the NULL Hypothesis and accept alternate hypothesis. That means there is a difference in the population means of voltage readings at the two locations in the 95% confidence interval.
Therefore, the manufacturing process cannot be established locally as null hypothesis is rejected.

R Code:

```
#Finding Variance
remotebr_var <- var(remotebr$voltage)
remotebr_var
localbr_var <- var(localbr$voltage)
localbr_var

#Evaluating Null Hypothesis
t.test(remotebr$voltage,localbr$voltage,paired = F, var.equal = F,
conf.level = 0.95)
```

(c)

The sample means of the voltages at the local and remote locations varied significantly. For the values in the given problem, the difference between the means 0.381334 (9.803667 - 9.422333) is considerable. From this, we can predict that the true population means will differ significantly. So, we can conclude that the manufacturing process cannot be established locally which is expected from analysis of (a). Moreover, the manufacturing process cannot be developed locally, as demonstrated in (b) part, as the p-value is below the threshold of significance of 0.05.

Question 3:

To solve the given problem, we will first load the "vapor.csv" file in R studio and then analyze the data provided on the theoretical and experimental values of the vapor pressure for dibenzothiophene, a heterocycloaromatic compound similar to those found in coal tar, at given values of temperature as follows:

```
# Setting working directory to Proj-4 folder.

> setwd("C:/Users/Shalin Kaji/Desktop/UT-Dallas-Spr23/STATS-DS-Min.Chen/Mini-Proj-4")

> getwd()
```

Now we create lists to store theoretical and experimental values which are given in the dataset.

```
# Observing our .csv file

> gasstats <- read.csv("vapor.csv")

> theoval <- gasstats$theoretical

> expval <- gasstats$experimental
```

It is given that, if the theoretical model for vapor pressure is a good model of reality, the true mean difference between the empirical and theoretical of vapor pressure will be zero.

Therefore, the **Null Hypothesis** will be:

$H_0 : \mu_E - \mu_{TH} = 0$

Alternate Hypothesis:

$H_A : \mu_E - \mu_{TH} \neq 0$

We can't assume that the distribution is Normal because:

1) We don't know whether the population from which sample was derived followed Normal distribution or not and hence cannot comment on the distribution of the sample.

2) The sample size of the VAPOR.DAT is n = 16, which is not sufficiently large (n>30) for us to assume Normal Distribution through the Central Limit Theorem.

Therefore, we will use Student's t-distribution for our Hypothesis testing.

> t.test(theoval, expval, paired = FALSE, var.equal = TRUE, alternative = c("two.sided"))

Output:

```
        Two Sample t-test

data:  theoval and expval
t = 0.0048042, df = 30, p-value = 0.9962
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2915711  0.2929461
sample estimates:
mean of x mean of y
0.7605625 0.7598750
```

Based on the results of our t-test, we reject the Null Hypothesis (H_0).

Conclusion:

Hence, the theoretical model for vapor pressure isn't a good model of reality.

R Code:

```
setwd("C:/Users/Shalin Kaji/Desktop/UT-Dallas-Spr23/STATS-DS-Min.Chen/Mini-Proj-4")

getwd()

# Observing our .csv file

gasstats <- read.csv("vapor.csv")

gasstats

theoval <- gasstats$theoretical

expval <- gasstats$experimental


t.test(theoval, expval, paired = FALSE, var.equal = TRUE,alternative = c("two.sided"))
```