Mini Project 04
Mini Project Group 31
Group Members
Dhairya Pinakin Desai: DPD140130

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.

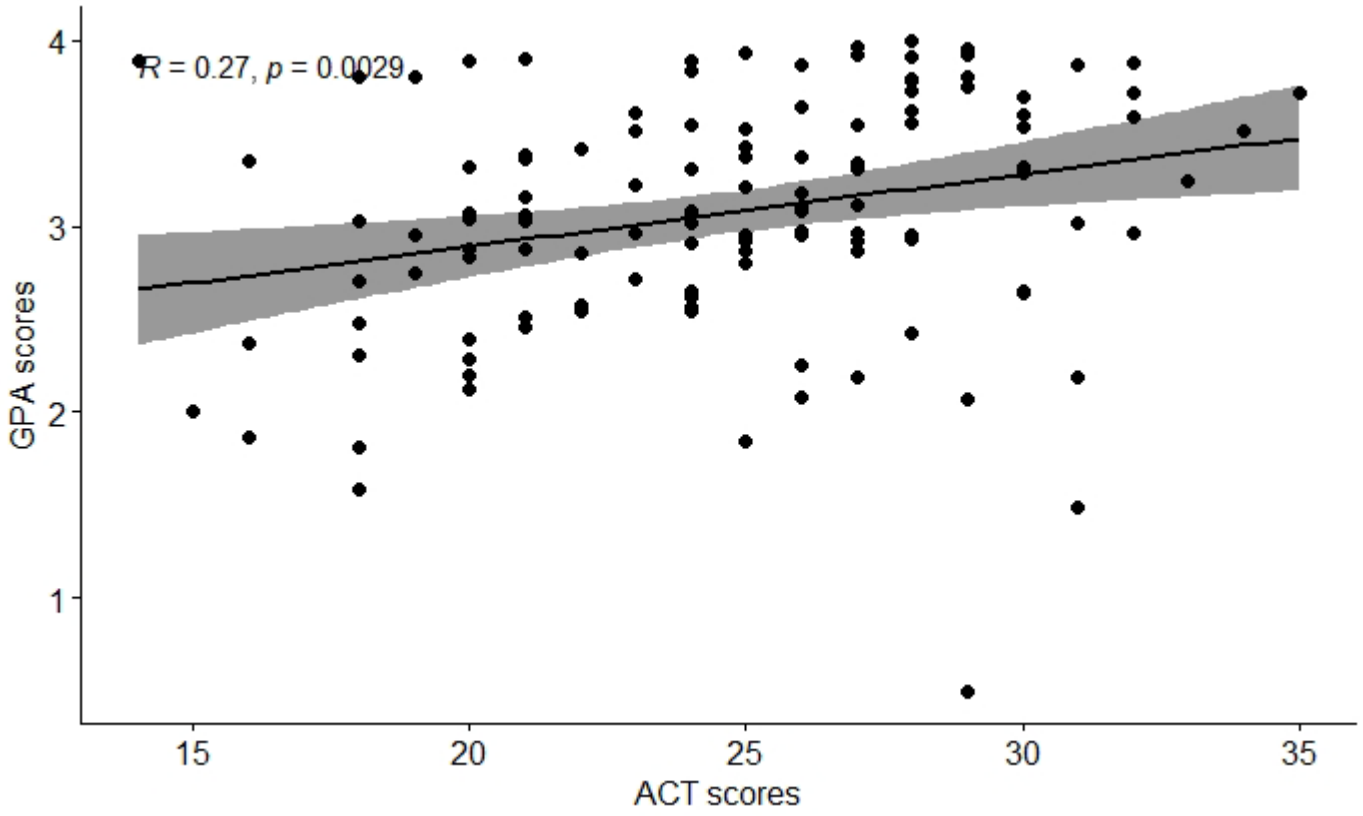
```
# Setting working directory to Proj-4 folder.
> setwd("C:/Users/dpd140130/OneDrive - The University of Texas at Dallas/CS 6313/Projects/04")
> getwd()
# we need to load the 'boot' package first to load it.
> library(boot)
```

Creating 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.
```

Creating the Scatterplot using these two vectors.

```
> 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") # plotting scatterplot
```



Pearson's Correlation Coefficient:

The strength and direction of the linear relationship between two variables are measured using this method. It can take values between -1 and 1, and is represented by the symbol " ρ".

If the value is 1, there is a perfect positive correlation, which means that as one variable rises, the other rises linearly as well. If the value is 1, there is a perfect negative correlation, which means that as one measure rises, the other one falls linearly. No correlation, or a value of 0, denotes the absence of a linear relationship between the variables.

To determine the correlation between the ACT and GPA results, we use R's cor() function.

```
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.

According to our analysis of the value of "rho," which is close to zero, there is only a weakly positive link between the ACT and GPA scores.

Using resampling techniques, the Bootstrap approach enables us to generate additional samples from the original sample. This enables us to compute the bootstrap estimates of bias and standard error for the correlation statistic that is of interest.

```
# 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)[c(250, 9750)]
[1] 0.07056891 0.47832157
```

We may interpret the findings by saying that, after drawing n samples, 95% of the Confidence Intervals will include the real value of the parameter, i.e., correlation, and because it is [0.067, 0.473], there exists only a weak positive association from which we can make any specific conclusions.

R Code:

```
# STATS-Mini-Project-4
# 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)

# 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)


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

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)
```

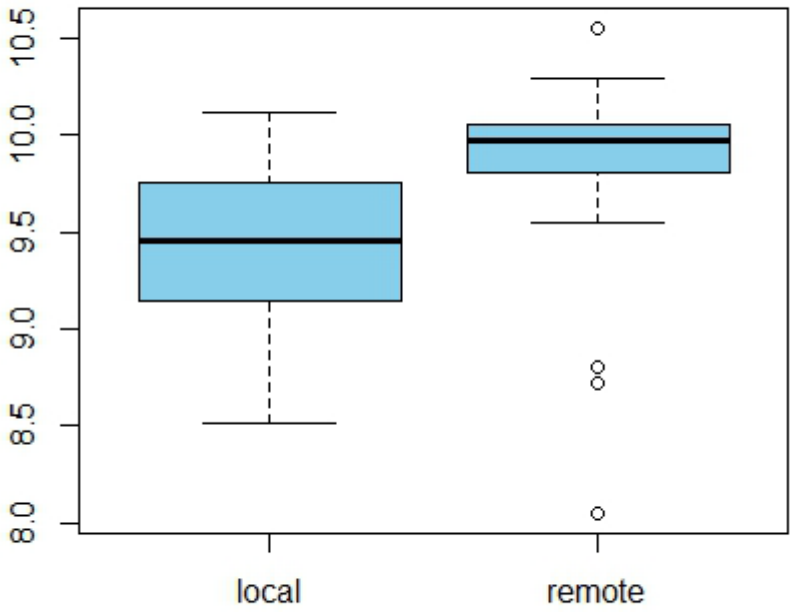
Now we create two lists to store remote and local locations which are given in the dataset.

```
# 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')
```

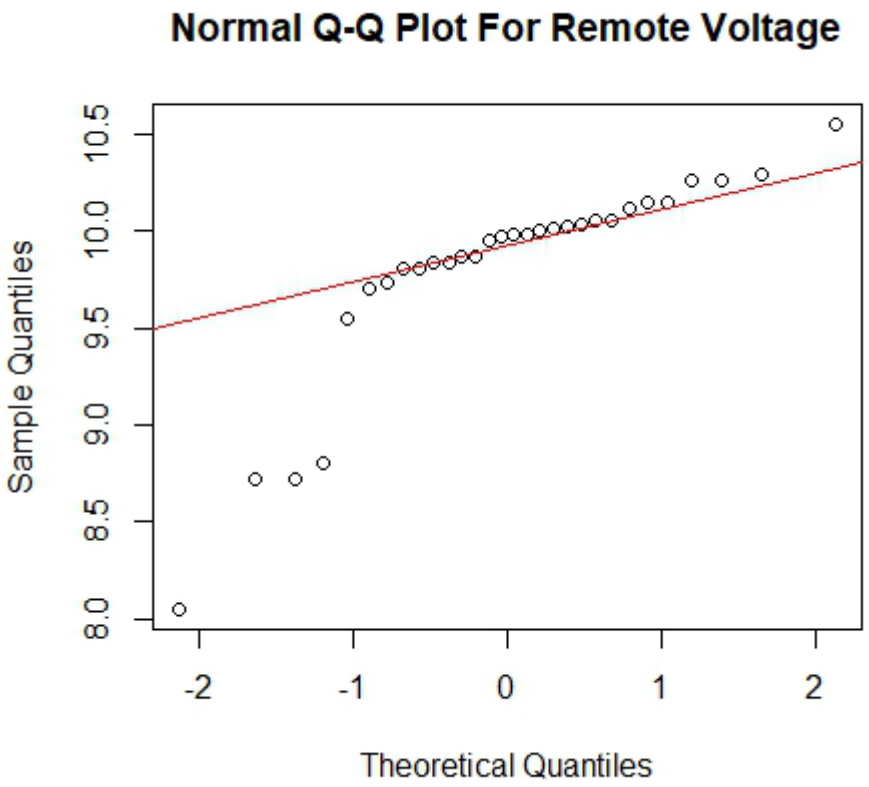
```
> summary(remotebr$voltage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.050   9.800   9.975   9.804  10.050  10.550
> summary(localbr$voltage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.510   9.152   9.455   9.422   9.738  10.120
```



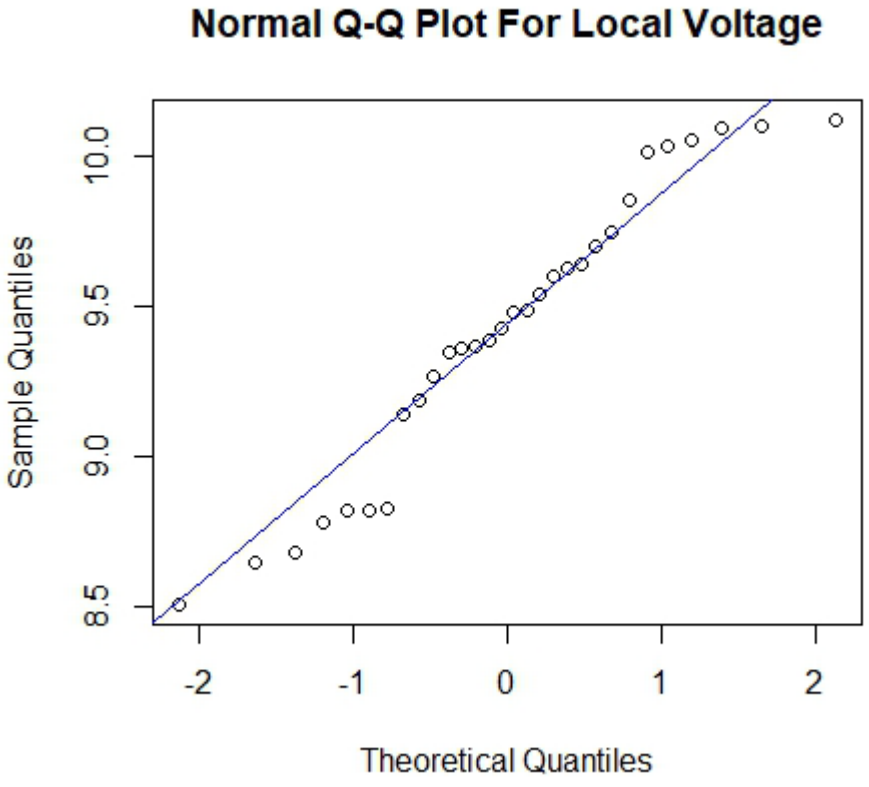
The summary data demonstrate that the median and mean of the station are not equal. The boxplot demonstrates that the interquartile range (IQR) for voltage values produced at remote locations is lower than the IQR for voltage values produced locally. Consequently, we may infer from the boxplot that the voltage distribution at the remote point is right-skewed.

We must now determine whether the values of the distant and local data follow a normal distribution. We employ the qqnorm() and qqline() tools for this.

```
> 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:

- Because the sample size(n) is sufficiently high, we are supposing that the data from both remote and local manufactured have a normal distribution.

• As there is no information provided regarding variances, we presume that they are both unknown and unequal.

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.

Given that the distribution is normal and the variance is unequal, we can get the 95% confidence interval for the provided issue statement.
To that end, we consider using the null hypothesis on the specified issue statement. Hence,
H0: The population mean of the voltage values at the two stations is equal.
The result is that mean(remotedata) - mean(localdata) = 0. As a result, the procedure may be implemented locally because there is no mean difference.
H1: is a possible alternative, which states that there is a data difference between distant and local produced voltages that is not equal to 0.
Namely, mean(remotedata) - mean(localdata) 0

The null hypothesis is disregarded as being non-significant if the p-value from the t-test is less than 0.05. The t-test is used to calculate the p value and confidence interval values.

```
>#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
```

We will accept the alternative hypothesis and reject the null hypothesis since the p-value is smaller than the significant p-value of 0.05. The population averages of the voltage readings at the two sites in the 95% confidence interval differ, according to this statement.
As a result, the null hypothesis is rejected, and the production process cannot be developed locally.

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 voltage sample means at the nearby and distant locations were very different. The means in the given issue differ by a significant 0.381334 (9.803667 - 9.422333) amount for the values. We may infer from this that the genuine population means will vary considerably. So, we may draw the conclusion that the production process cannot be built locally as was anticipated from examination of (a). Also, as shown in (b) portion, the manufacturing method cannot be created locally since the p-value is below the significance level of 0.05.

Question 3:

In order to address the given issue, we will first load the "vapor.csv" file in R studio and then conduct the following analysis on the information on the theoretical and experimental values of the vapor pressure for dibenzothiophene, a heterocycloaromatic compound related to those found in coal tar, at the specified temperatures:

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

The true mean difference between the theoretical and empirical estimates of vapor pressure will be zero if the theoretical model for vapor pressure is an accurate representation of reality.

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 are unable to remark on the distribution of the sample since we are unsure of whether the population from which the sample was obtained adhered to the Normal distribution or not.

2) We cannot assume Normal Distribution using the Central Limit Theorem since the sample size of the vapor.dat is just n = 16, which is not large enough (n>30).

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:

```
#Setting up working directory
setwd("C:/Users/dpd140130/OneDrive - The University of Texas
at Dallas/CS 6313/Projects/04")
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"))
```