MINI PROJECT #4
Statistical Methods for Data Science
Group - 46

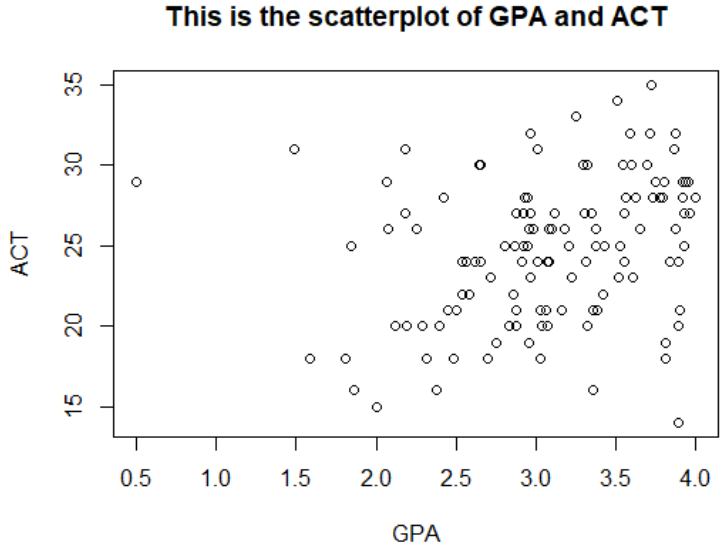
Anish Joshi(2021591978) and Aneena Manoj(2021623362)

Contribution – Firstly, both the teammates discussed the questions together and wrote the required R code for all the three questions. Further, the R code was executed by both, and the findings were reported. Both the teammates worked efficiently to finish the project.

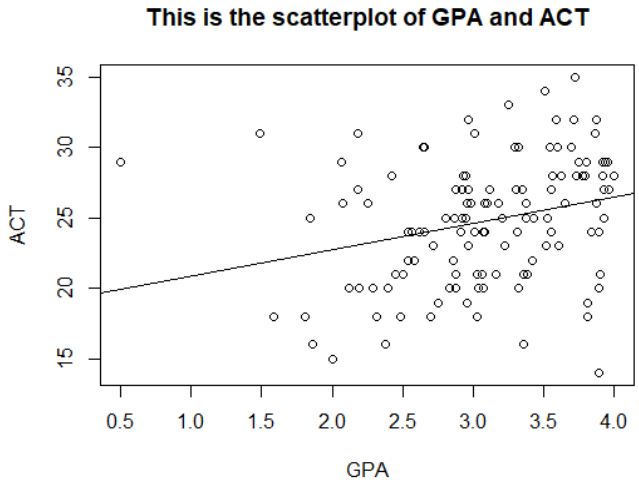
SECTION - 1

QUESTION – 1

The first step is to read in the data from the gpa.csv file which is available on eLearning. We use the function "read.csv" to perform the above task. The next step which we follow is to draw a scatter plot between ACT and GPA to analyze and determine the linear relationship between them.



Now we will draw an abline to determine the correlation.



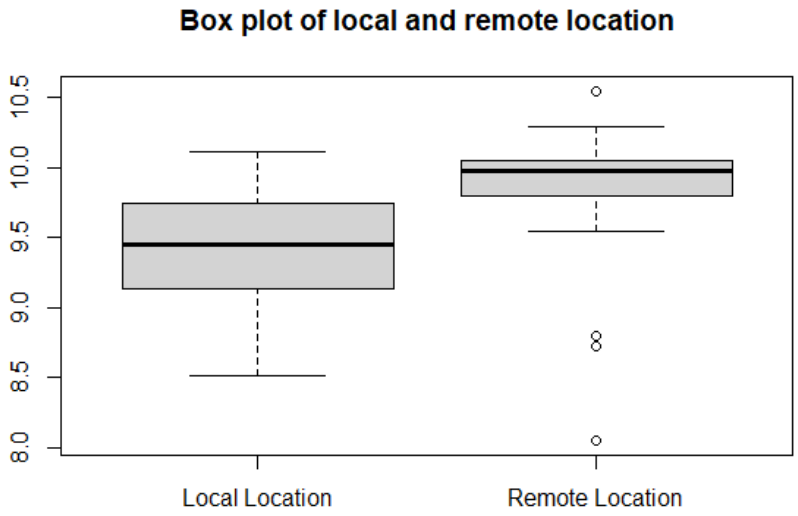
From the above plot, we can clearly see that the line drawn in the scatter plot has a positive slope which is greater than zero. The above statement clearly implies that there is a positive association between the GPA and ACT. The linear relationship is weak. Now the next step is to find the correlation. Upon computing, the correlation value is found to be 0.2694818. R code for this is in SECTION - 2

According to the question, we need to use the bootstrap method in order to resample and find the estimates for the correlation. To do this, the first step is to create a statistical function for covariance/correlation and then use the boot library to perform further operations.
The values returned from the functions are –
Point Estimate – 0.2699557; Bias – 0.0004738617; Standard Error – 0.1061572
To find the Confidence Interval (CI), we will use boot.ci. The confidence intervals are found to be [0.0623,0.4627]. To verify the above CIs, we sort the bootstrap correlation and find out the quantiles which come out to be [0.06231969,0.46268901].
The correlation value is approximately 0.3 which confirms the positive association in the scatter plot.

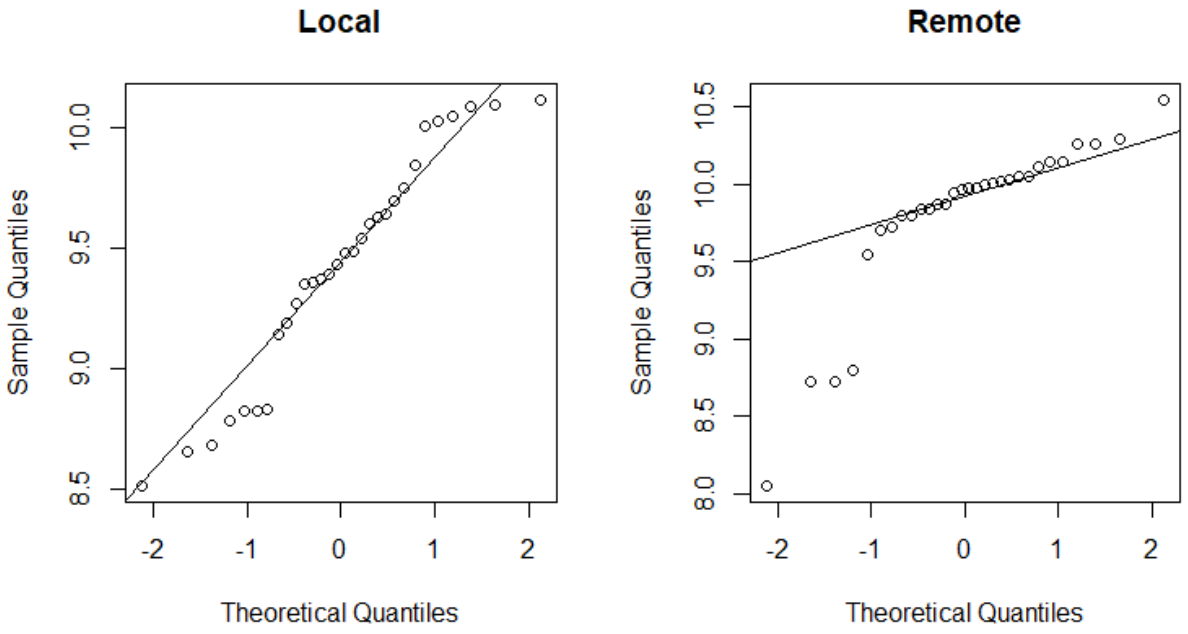
QUESTION - 2

To solve this question, the first step is to read the "voltage.csv" and store the data into two separate variables.

a) The two variables "val_local" and "val_remote" are compared using boxplot.



The above boxplot clearly shows evidence that the voltage readings at remote locations are greater than those at local locations. Both the plots are left-skewed as the median is greater than the mean which is evident from the 5 point summary. The plot of the remote location shows the existence of some outliers. The following is the qq plot for the two datasets.



The dataset can be assumed to be normalized since for some values the data points coincide with the line.

b) The manufacturing process will be established locally if there is no difference between the population means.

Null Hypothesis : Difference between sample mean of remote and sample mean of local is 0.

Alternative Hypothesis: Difference between the sample mean of remote and sample mean of local is not 0.

The two samples are treated as independent samples and the variances are not known to be equal.

Since the n value is greater than or equal to 30, we will treat it as a large sample.

Calculation –

$\overline{R} - \overline{L} = 0.3813333$

$S_r^2 = 0.2925895$

$S_l^2 = 0.229322$

$$\overline{SE}\left(\overline{R} - \overline{L}\right) = \sqrt{\left(\frac{S_r^2}{n_r}\right) + \left(\frac{S_l^2}{n_l}\right)}$$

$$= \sqrt{\left(\frac{0.2925895}{30}\right) + \left(\frac{0.229322}{30}\right)}$$

$$= \sqrt{\frac{0.5219115}{30}}$$

$$= 0.1318979$$

95% Confidence Interval for the Z value is 1.96. *So now computing the CI.*

$$\textit{Lower Bound} = \left(\overline{R} - \overline{L}\right) - Z_{\frac{\alpha}{2}} \times \overline{SE}\left(\overline{R} - \overline{L}\right)$$

$$= 0.3813333 - 1.96 \times 0.1318979$$

$$= 0.1228182$$

$$\textit{Upper Bound} = \left(\overline{R} - \overline{L}\right) + Z_{\frac{\alpha}{2}} \times \overline{SE}\left(\overline{R} - \overline{L}\right)$$

$$= 0.3813333 - 1.96 \times 0.1318979$$

$$= 0.6398485$$

The Confidence Interval is $[0.1228182, 0.6398485]$

It can be concluded that the CI is appropriate. The null hypothesis is rejected since 0 does not lie in the above calculated CI. This means that the manufacturing process cannot be established at local locations.

c) On the basis of subquestion a) and b), it is clear that the manufacturing process must be located at a remote location.

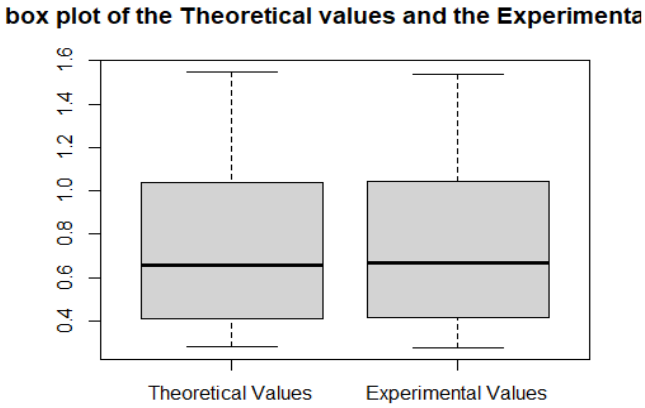
QUESTION - 3

The first step is to read the values in the VAPOR.csv file using the read.csv function. We will be storing the theoretical values in vap_theoritical and the experimental values in vap_experimental. We will use the following notations for the calculations –

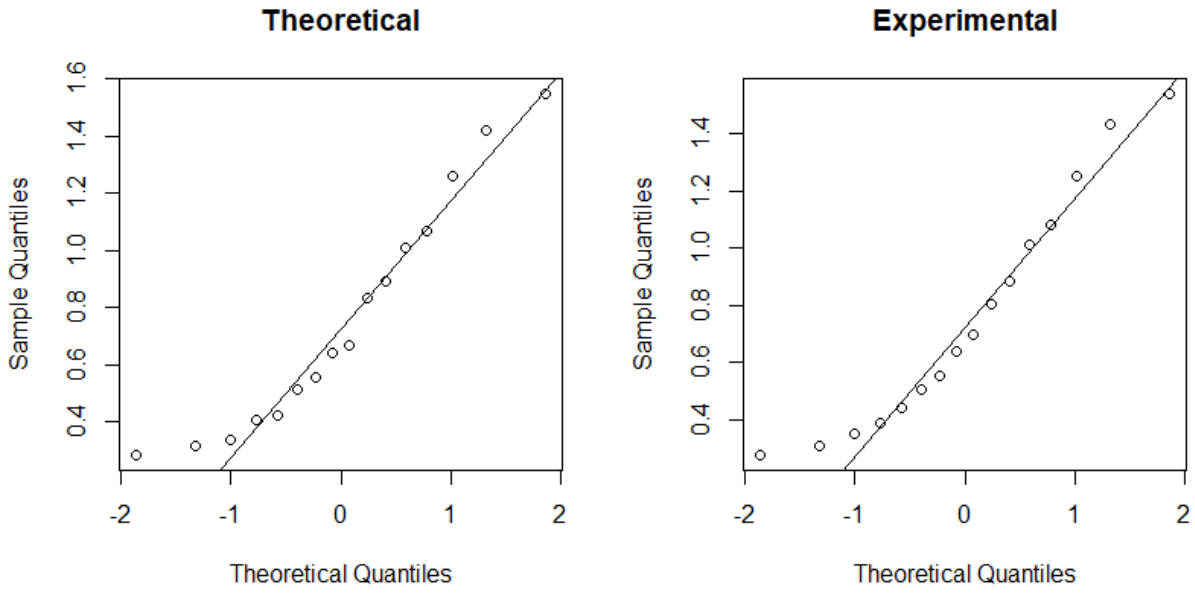
T(bar) – Sample mean of Theoretical values; This is used to estimate the population mean of Theoretical values.

E(bar) – Sample mean of Experimental values; This is used to estimate the population mean of Experimental values.

The following is the box plot of Theoretical and Experimental readings –



The following is the qq plot comparing the theoretical and the experimental values –



From the above qq plot, we can conclude that the samples are **approximately normal** since for some values the line and the data points coincide.

From the R code for Question 3 in Section - 2, we calculate the summary statistics. According to the summary statistics, both the distributions are almost similar and are right skewed since the mean is greater than the median.

Now the next part of the question tells us to test the difference between theoretical values and the experimental values.
Firstly, we will write the Null hypothesis and the Alternative Hypothesis.
Null Hypothesis: True mean difference between T(bar) and E(bar) = 0
Alternative Hypothesis: True mean difference between T(bar) and E(bar) is not equal to 0.
From the R code for Question - 3 in Section -2, the following are the computed values –
Mean of difference = 0.0006875
SD of the difference = 0.01421604
Value of t for 95% CI and degree of freedom (n-1) = 2.13145
Now calculating the CI
Lower Bound and Upper Bound:

$$\overline{D} + t_{\frac{\alpha}{2};n-1} \times \frac{S_d}{\sqrt{n}}$$

$$= 0.0006875 + 2.13145 \times \left(\frac{0.01421604}{4}\right) = 0.008262694$$

$$\overline{D} - t_{\frac{\alpha}{2};n-1} \times \frac{S_d}{\sqrt{n}}$$

$$= 0.0006875 - 2.13145 \times \left(\frac{0.01421604}{4}\right) = -0.006887694$$

So, the CI is [-0.006887694,0.008262694]

The Null hypothesis is accepted since the given CI contains the value 0 so the true mean difference value is 0.

SECTION - 2

Question - 1

#Reading the data of gpa.csv into val_gpa
val_gpa<-
read.csv("/Users/axj200101/Desktop/UTD/1stSemester/CS6313_StatisticalMethods/MiniProject4/gpa.csv
")
gp_val<-as.numeric(val_gpa$gpa)
act_val<-as.numeric(val_gpa$act)
#Plotting the scatterplot
plot(gp_val,act_val,main="This is the scatterplot of GPA and ACT",xlab='GPA',ylab='ACT')
abline(lm(act_val~gp_val))

```
> cor(gp_val,act_val)
[1] 0.2694818
> library(boot)
> #Creating a statistical function for correlation/covariance
> cov_fn<-function(val_gpa,indices)
+ {
+   fn_gpa<-val_gpa$gpa[indices]
+   fn_act<-val_gpa$act[indices]
+   ans<-cor(fn_gpa,fn_act)
+   return(ans)
+ }
> #using bootstrap for estimation of the statistical function of correlation
> cov_fn_boot<-boot(val_gpa,cov_fn,R=999,sim="ordinary",stype='i')
> cov_fn_boot

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = val_gpa, statistic = cov_fn, R = 999, sim = "ordinary",
    stype = "i")


Bootstrap Statistics :
     original       bias    std. error
t1* 0.2694818 0.0004738617   0.1061572
```

```
> #calculating the point estimate
> names(cov_fn_boot)
 [1] "t0"        "t"         "R"         "data"      "seed"      "statistic"
 [7] "sim"       "call"      "stype"     "strata"    "weights"
> mean(cov_fn_boot$t)
[1] 0.2699557
> #calculating the confidence interval
> boot.ci(cov_fn_boot)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

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

Intervals :
Level      Normal              Basic
95%   ( 0.0609,  0.4771 )   ( 0.0763,  0.4766 )

Level     Percentile            BCa
95%   ( 0.0623,  0.4627 )   ( 0.0430,  0.4532 )
Calculations and Intervals on Original Scale
Warning message:
In boot.ci(cov_fn_boot) :
  bootstrap variances needed for studentized intervals
> #verifying the computed confidence intervals
> sort(cov_fn_boot$t)[c(25,975)]
[1] 0.06231969 0.46268901
>
```

Question - 2

#Question - 2
#Reading the data from VOLTAGE.csv and store in separate variables
val_volt<-read.csv("/Users/axj200101/Desktop/UTD/1st Semester/CS6313_StatisticalMethods/MiniProject4/VOLTAGE.csv")
val_remote<-val_volt$voltage[which(val_volt$location == 0)]
val_local<-val_volt$voltage[which(val_volt$location == 1)]

#Drawing boxplot
boxplot(val_local,val_remote,main="Box plot of local and remote location",names=c("Local Location","Remote Location"),range=1.5)

```
> #Summary statistics of voltage values at local and remote location
> summary(val_local)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.510   9.152   9.455   9.422   9.738  10.120
> summary(val_remote)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.050   9.800   9.975   9.804  10.050  10.550
>
> #Now drawing qqplots for the voltage values at local location and remote location"
> par(mfrow=c(1,2))
> qqnorm(val_local,main='Local')
> qqline(val_local)
> qqnorm(val_remote,main='Remote')
> qqline(val_remote)
>
>
> #Calculation of mean,variance, standard error and CI of the dataset
> var(val_remote)
[1] 0.2925895
> var(val_local)
[1] 0.229322
> se<-sqrt(var(val_local)/30 + var(val_remote)/30)
> se
[1] 0.1318979
> dif_mean<-mean(val_remote)-mean(val_local)
> dif_mean+c(-1,1)*qnorm(0.975)*0.1318979
[1] 0.1228182 0.6398485
> dif_mean
[1] 0.3813333
>
```

Question - 3

```
#Question3
#Reading the data in VAPOR.csv into val_vap
val_vap<-read.csv("/Users/axj200101/Desktop/UTD/1st Semester/CS6313_StatisticalMethods/MiniProject4/VAPOR.csv")
vap_theoretical<-val_vap$theoretical
vap_experimental<-val_vap$experimental

#Drawing the qqplots
par(mfrow=c(1,2))
qqnorm(vap_theoretical,main="Theoretical")
qqline(vap_theoretical)
qqnorm(vap_experimental,main="Experimental")
qqline(vap_theoretical)

#Drawing the boxplot and printing the summary statistics
boxplot(vap_theoretical,vap_experimental,names=c("Theoretical Values","Experimental Values"),main="The box plot of the Theoretical values and the Experimental values")
```

```
>
> #summary statistics
> summary(vap_theoretical)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2820  0.4175  0.6555  0.7606  1.0250  1.5500
> summary(vap_experimental)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2760  0.4305  0.6675  0.7599  1.0275  1.5400
>
> #Calculation of mean,sd, CI and t(n-1) value of the difference between theoretical and experim
ental values
> vap_diff<-vap_theoretical-vap_experiment
Error: object 'vap_experiment' not found
> vap_diff<-vap_theoretical-vap_experimental
> vap_diff
 [1]  0.006  0.007 -0.015  0.014 -0.022  0.008  0.000  0.002 -0.026  0.029  0.008  0.000
[13] -0.010  0.010 -0.010  0.010
> mean(vap_diff)
[1] 0.0006875
> sd(vap_diff)
[1] 0.01421604
> n=16
> qt(0.975,n-1)
[1] 2.13145
> #Calculating the CI now
> mean(vap_diff)+c(-1,1)*qt(0.975,(n-1))*(sd(vap_diff)/sqrt(n))
[1] -0.006887694  0.008262694
>
```