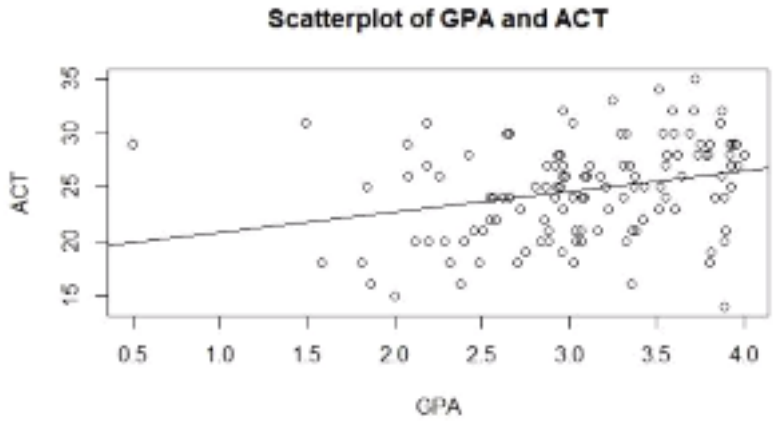
Mini Project #4
Tanushri Singh, Nikhil Pareek
Contribution of each group member:
Nikhil worked on problem number one and Tanu worked on questions 2 and 3 then went onto Document questions and report all the findings. Both partners worked efficiently to complete the project requirements!
NOTE: ALL CODES ARE ATTACHED IN SECTION 2!

Section #1

Question 1:
First the data file is read in and contents get separated. Once the two datasets are separated a scatter plots are drawn. In order to see the correlation between the two datasets abline is used. Scatter plot that gets generated:



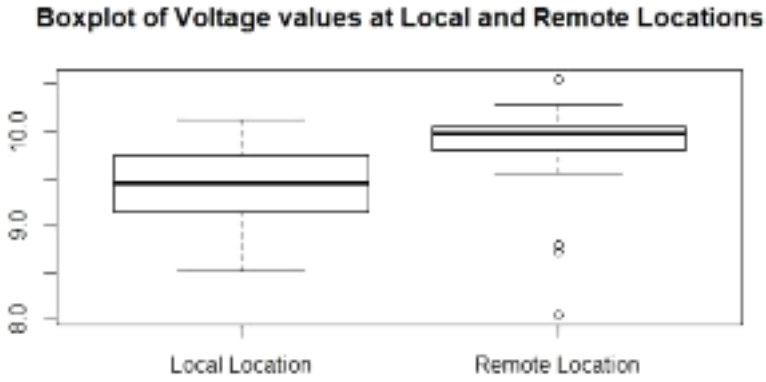
It is clear that the line that is drawn in the scatter plot has a positive slope greater than zero. This means that there is a positive association amongst the gpa and act. This would mean that the strength of the linear relationship is weak.
Now, we went on to find the correlation of the two functions by using the cor functionality. Based on the given datasets, the correlation ended up being 0.2694818.
Next the boot functions is used in order to resample and find estimates for the correlation. Then, statistical functions are made in order to calculate correlation using the cor function once again which is returned. Point estimate is also taken as expected value t* from samples of bootstrap. Values returned from the functions for the data is:
Estimate = 0.2739676, Bias = 0.004485845 and Standard Error = 0.1028841.
Now the boot.ci function is used in order to get the confidence interval. The resulting value after running the code snippets is: (0.0672, 0.4668). Then the bootstrap correlation is sorted and the 1st and 3rd quartiles result in (0.0672043, 0.4667759) which verify that the confidence interval is correct.

It was interpreted that the point estimate of correlation from bootstrap is approximately close to the correlation value from the samples. Also, the confidence interval from boot.ci is approximately close to the quantile values from sorted bootstrap data. Lastly, the correlation value is approximately 0.3 which means there is a positive association in the scatter plot.

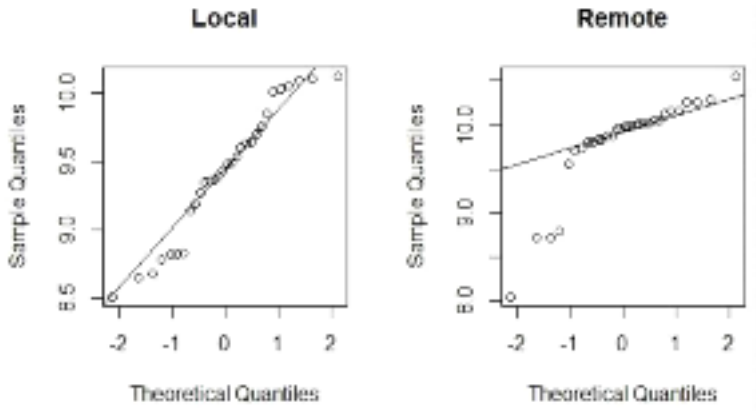
Question 2:

Firstly the data gets read in and it is separated based on locations into 2 separate variables.

a) In order to compare the two distributions the boxplots are examined:



It is evident that the voltage readings at remote locations are greater than those at local locations. With the 5 point summary it is evident that both graphs are left skewed since the medians are greater than the mean. It can also be clearly seen that outliers exist in the remote location graph.
QQplots for the two datasets:



It can be noticed that for some values the data points and line coincide hence it can be assumed that the data sets are normalized.

b) Its given that the manufacturing process will be established locally if no difference persists between the population means. So, the null hypothesis would be:
Difference = 0 => sample mean of remote - sample mean of local = 0
And the Alternative Hypothesis would be:
Difference != 0 => sample mean of remote - sample mean of local != 0

To start, the two samples must be treated as independent samples. It is assumed that data is normal based on the QQ plots that are graphed. Now, since the IQR are vastly

distinct population variances are equal cannot be assumed. So, satterthwaite's approximation and t-distributions must be done.
Calculating the variance and mean:

$$\bar{r} - \bar{l} = 0.3813333$$
$$S_r^2 = 0.2925895$$
$$S_l^2 = 0.229322$$
$$\widehat{SE}(\bar{r} - \bar{l}) = \sqrt{\left(\frac{S_r^2}{n_r}\right) + \left(\frac{S_l^2}{n_l}\right)}$$
$$= \sqrt{\frac{0.2925895}{30} + \frac{0.229322}{30}}$$
$$= \sqrt{\frac{0.5219115}{30}}$$
$$= 0.1318979$$

It is known that the 95% confidence interval for the Z value is 1.96, so calculating the confidence interval:

Lower Bound: $(\bar{r} - \bar{l}) - Z_{\frac{\alpha}{2}} \times \widehat{SE}(\bar{r} - \bar{l})$
: $0.3813333 - 1.96 * 0.1318979$
: 0.1228182

Upper Bound: $(\bar{r} - \bar{l}) + Z_{\frac{\alpha}{2}} \times \widehat{SE}(\bar{r} - \bar{l})$
: $0.3813333 + 1.96 * 0.1318979$
: 0.6398485

The confidence interval we calculated is (0.1228182, 0.6398485)

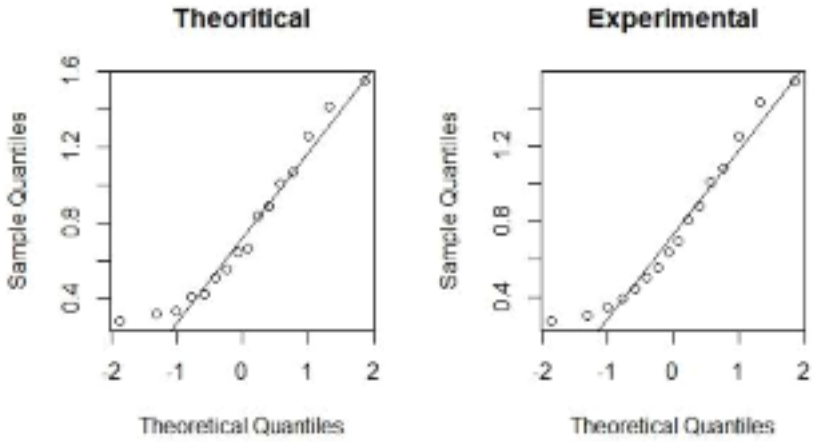
In order to verify the confidence interval a t test is performed. The resultant value from the t test gives the values (0.1172284, 0.6454382). So, it can be concluded that the confidence interval is appropriate and normal assumptions hold. Since 0 does not lie in the confidence interval (from t test) the null hypothesis is rejected. This implies that the difference between the means at the two locations is not zero. This means that the manufacturing process cannot be established at local locations.

c) From part a it is observed that voltage readings at remote are higher than those at local. It is obvious that for any manufacturing process high voltage is required to fuel heavy equipment. So, based on parts A and B it is clear that the manufacturing process must be located in a remote location.

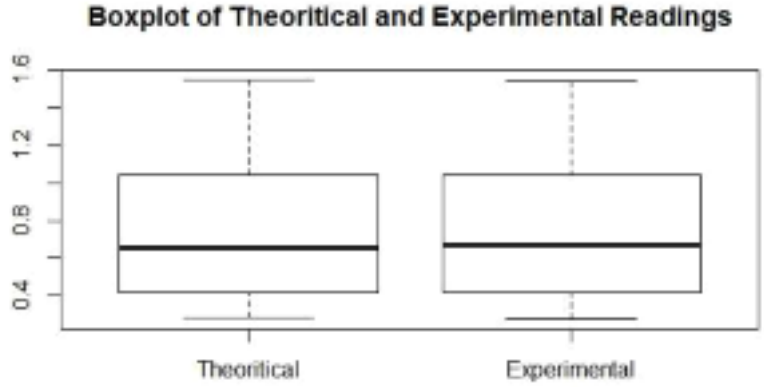
Question 3:

Values in csv are read in. Then the file gets separated. ‘T’ is used to denote theoretical values and ‘E’ will be used to denote experimental values. T(bar) denotes the sample value of T that estimates the population of the mean and e(bar) denote the sample mean of E that estimates the population mean.

Result of the qqplot’s for theoretical and experimental values:



So, from the qqplots it is evident that the samples can be treated as approximately normal.
Boxplot for the theoretical and experimental values:



It is evident that the two datasets are very similar, and the differences are nearly negligible. The IQR, and 5-plot summary also supports this conclusion. Both the distributions are right skewed since their mean is greater than their median.

Now, let's test the mean difference between theoretical and experimental values.
Null Hypothesis: True mean difference between t(bar) and e(bar) == 0
Alternative Hypothesis: True mean difference between t(bar) and e(bar) != 0.
Now, the confidence interval is calculated using the t distribution.
Calculating the mean, standard dev results in:
mean = 0.0006875, standard dev = 0.01421604, t = 2.13145

Lower Bound: $\bar{d} - t_{\frac{\alpha}{2},n-1} \times \frac{S_d}{\sqrt{n}} = 0.0006875 - 2.13145 \times \frac{0.01421604}{4} = 0.008262694$

Upper Bound: $\bar{d} + t_{\frac{\alpha}{2},n-1} \times \frac{S_d}{\sqrt{n}} = 0.0006875 + 2.13145 \times \frac{0.01421604}{4} = -0.006887694$

Therefore the confidence interval calculated is (-0.006887694, 0.008262694)

In order to verify the confidence interval a t test is conducted. The observed interval is (-0.006887694, 0.008262694). This means that the interval is appropriate.

Since the value 0 lies within the found interval, it means that the t(bar) - e(bar) = 0. So, the null hypothesis is accepted , so the true mean difference of theoretical and experimental values is zero. This is also supported by the boxplot.

Section #2

R code for Question 1

#Read data in

```
> gpavalues <- read.csv("gpa.csv")
```

#Separate the datasets

```
> gpa <- as.numeric(gpavalues$gpa)
> act <- as.numeric(gpavalues$act)
```

#plot of the datasets

```
> plot(gpa, act, main="Scatterplot of GPA and of ACT", xlab = "GPA", ylab = "ACT")
> abline(lm(act~gpa))
```

#Correlation

```
> cor(gpa, act)
[1] 0.2694818
```

#Import/attach the boot library

```
> library(boot)
```

#statistic function for correlation

```
> covarience.npar <- function(gpaset, indices){
+         xgpa <- gpaset$gpa[indices]
+         xact <- gpaset$act[indices]
+         result <- cor(xgpa, xact)
+         return(result)
+ }
```

#execute boot function w/ statistical function

```
> covarience.npar.boot <- boot(gpavalues,covarience.npar, R = 999, sim = "ordinary", stype = "i")
```

```
> covarience.npar.boot

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = gpavalues, statistic = covarience.npar, R = 999,
    sim = "ordinary", stype = "i")


Bootstrap Statistics :
     original      bias    std. error
t1* 0.2694818 0.004485845   0.1028841
```

#Point estimate of bootstrap value

```
> mean(covarience.npar.boot$t)
[1] 0.2739676
```

#getting confidence interval using boot.ci

```
> boot.ci(covarience.npar.boot)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

CALL :
boot.ci(boot.out = covarience.npar.boot)

Intervals :
Level      Normal              Basic
95%   ( 0.0633,  0.4666 )   ( 0.0722,  0.4718 )

Level     Percentile            BCa
95%   ( 0.0672,  0.4668 )   ( 0.0407,  0.4577 )
Calculations and Intervals on Original Scale
```

#verifying confidence interval by extracting quantiles

```
> sort(covarience.npar.boot$t)[c(25,975)]
[1] 0.0672043 0.4667759
```

R code for Question 2

#Read data from csv

> voltage <- read.csv("VOLTAGE.DAT")

#Separate datasets based upon location

> voltage.remote = voltage$voltage[which(voltage$location == 0)]

> voltage.local = voltage$voltage[which(voltage$location == 1)]

#Draw boxplots and summary of datasets

> boxplot(voltage.local, voltage.remove, names = c("Local Location", "Remote Location"), main = "Boxplot of voltage values at Local and Remote Locations", range = 1.5)

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

#Draw qqplots for datasets

> par(mfrow = c(1,2))

> qqnorm(voltage.local, main = "Local")

> qqline(voltage.local)

> qqnorm(voltage.remote, main = "Remote")

> qqline(voltage.remote)

#Calculate mean, variance, standard error and confidence interval

> var(voltage. local)

[1] 0.229322

> var(voltage.remote)

[1] 0.2925895

> se <- sqrt(var(voltage.local)/30 + var(voltage.remote)/ 30)

> se

[1] 0.1318979

> diff = mean(voltage.remote) - mean(voltage.local)

> diff + c(-1,1) * qnorm(0.975) * 0.1318979

[1] 0.1228182 0.6398485

#Calculate confidence interval using t test

> t.test(voltage.remote, voltage.local, alternative = "two.sided", paired = False, var.equal = False, conf.level = 0.95)

```
        Welch Two Sample t-test

data:  voltage.remote and voltage.local
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
```

R code for Question 3

#Read data from csv
> vapor <- read.csv("VAPOR.DAT")
#Draw qqplots
> par(mfrow=(c,2))
> qqnorm(vapor$theoretical, main = "Theoretical")
> qqline(vapor$theoretical)
> qqnorm(vapor$experimental, main = "Experimental")
> qqline(vapor$experimental)
#Draw boxplots and summaries
> boxplor(vapor$theoretical, vapor$experimental, names = c("Theoretical", "Experimental"), main = "Boxplot of Theoretical and Experimental 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
```

#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
```