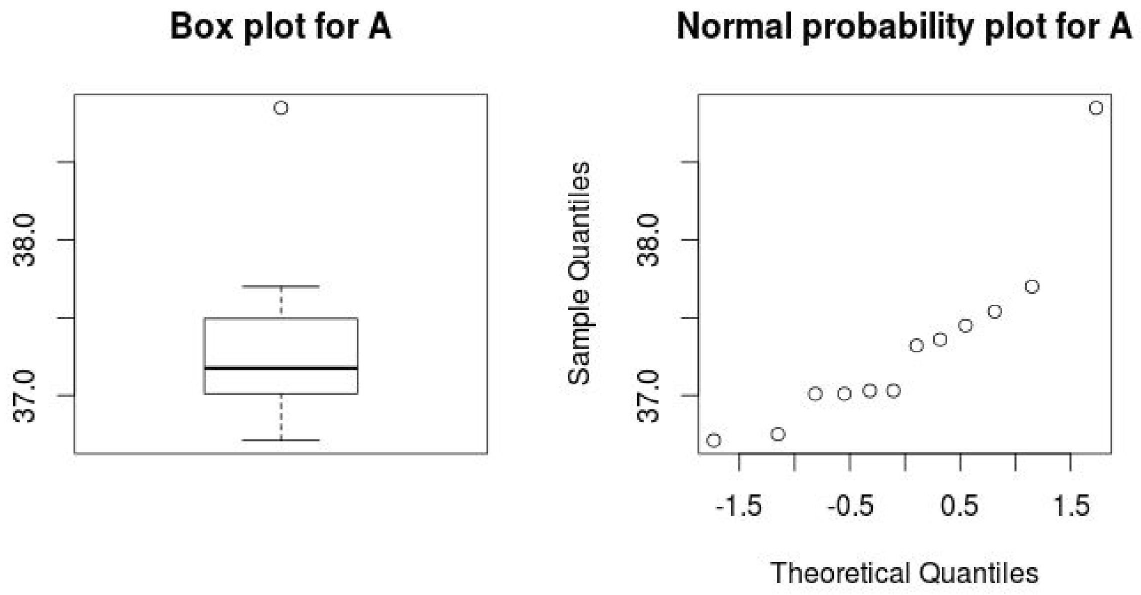
STAT S 520
HOMEWORK 10
RAMPRASAD BOMMAGANTY(rbommaga)

12.6 Problem Set A:

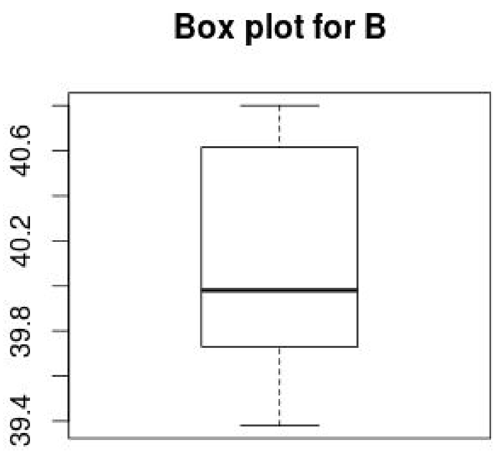
Solution:

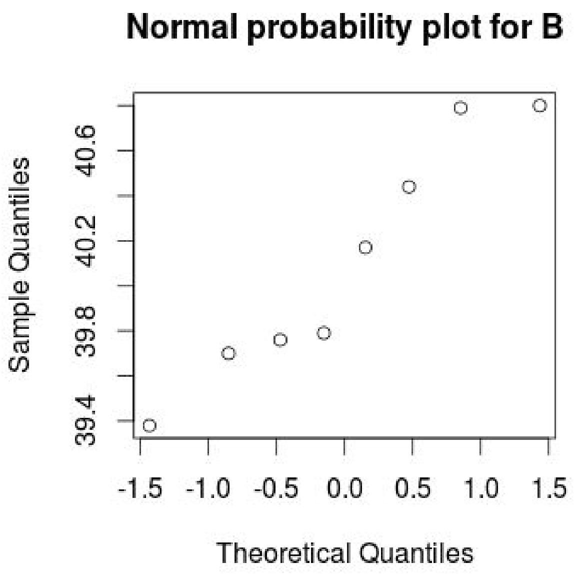
1. The R code for the computation of boxplots and normal probability plots are as follows:

```
> A<-c(37.54,37.01,36.71,37.03,37.32,37.01,37.03,37.70,37.36,36.75,37.45,38.85)
> B<- c(40.17,40.80,39.76,39.70,40.79,40.44,39.79,39.38)
> C<-c(39.04,39.21,39.05,38.24,38.53,38.71,38.89,38.66,38.51,40.08)
> par(mfrow=c(1,2))
> boxplot(A,main="Box plot for A")
> qqnorm(A,main="Normal probability plot for A")
```

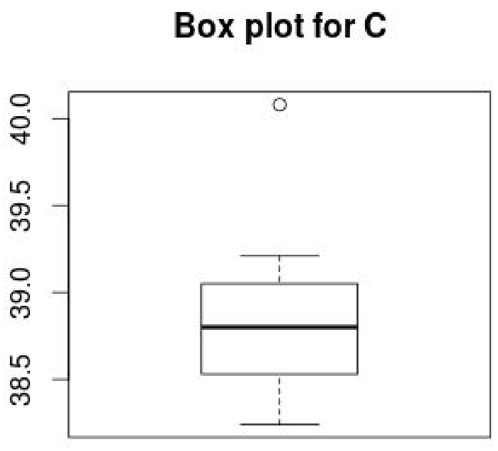


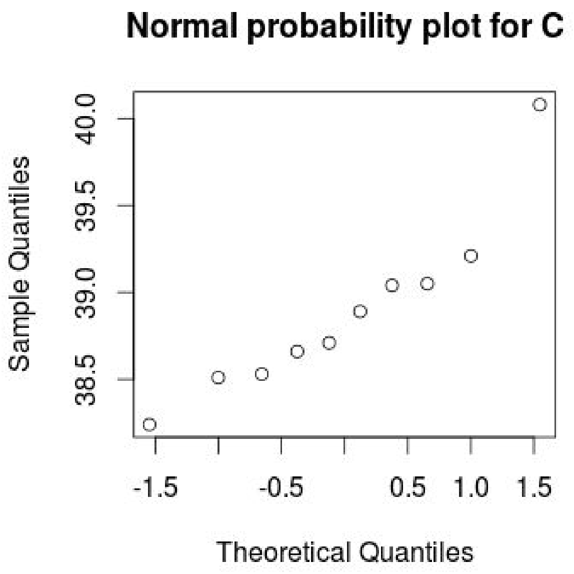
```
> boxplot(B,main="Box plot for B")
> qqnorm(B,main="Normal probability plot for B")
```





```
> boxplot(C,main="Box plot for C")
> qqnorm(C,main="Normal probability plot for C")
```





```
> var(A)
[1] 0.328097
> var(B)
[1] 0.2823696
```

```
> var(C)
[1] 0.2609289
```

From the above plots and calculations, the following observations can be concluded:

a. The assumptions of normality seem plausible, but there are two outliers in A and C. If we can ignore those two outliers, they look part of a normal distribution for the most part.
b. The assumptions of homoscedasticity seem plausible because the variances of all the three samples A,B and C are more or less close to each other, thereby we can classify them to same and assume that the variances are equal.

2. Given that the null hypothesis is that the mean salinity of all the three sites is the same.

Also, given that $\alpha = 0.05$

The R code for the computation of significance probability is as follows:

```
> n1=12
> n2=8
> n3=10
> mean(A)
[1] 37.31333
> mean(B)
[1] 40.10375
> mean(C)
[1] 38.892
> var(A)
[1] 0.328097
> var(B)
[1] 0.2823696
> var(C)
[1] 0.2609289
> grand_mean = (mean(A)*(n1/30))+(mean(B)*(n2/30))+(mean(C)*(n3/30))
> grand_mean
[1] 38.58367
> ss_b=(n1*(mean(A)-grand_mean)^2)+(n2*(mean(B)-grand_mean)^2)+(n3*(mean(C)-grand_mean)^2)
> ss_b
[1] 38.80088
> ss_w=(n1-1)*var(A)+(n2-1)*var(B)+(n3-1)*var(C)
> ss_w
[1] 7.934014
> ss_t=ss_b+ss_w
> ss_t
[1] 46.7349
> f=(ss_b/2)/(ss_w/27)
> f
[1] 66.02105
> p = 1-pf(f,df1=2,df2=27)
> p
[1] 4.008649e-11
```

Since we can note that the significance probability of the F test is very low, we can safely reject our null hypothesis and say that the three cities do not have the same mean salinity.

```
> ms_b=ss_b/2
> ms_b
[1] 19.40044

> ms_w=ss_w/27
> ms_w
[1] 0.2938524
```

The ANOVA table is as follows:

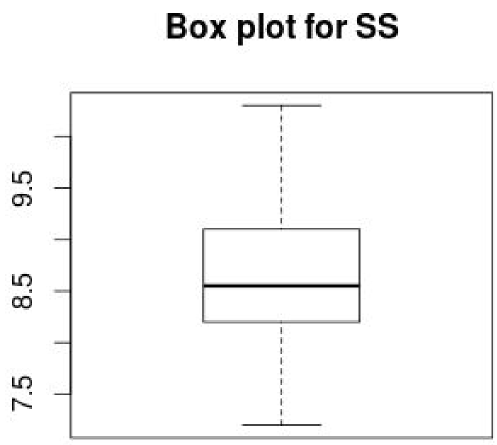
Source	**SS**	**df**	**MS**	**F**	**p**
Between	38.80088	2	19.40044	66.02105	4.008649e-11
Within	7.934014	27	0.2938524		
Total	46.7349	29			

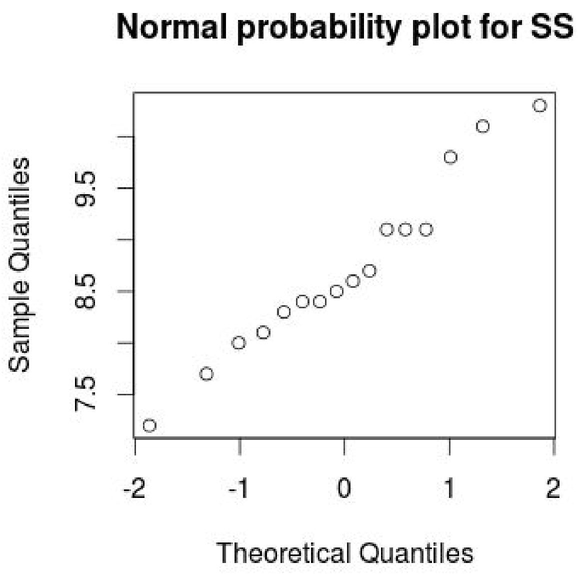
12.6 Problem Set B:

1. The R code for the computation of boxplots and normal probability plots are as follows:

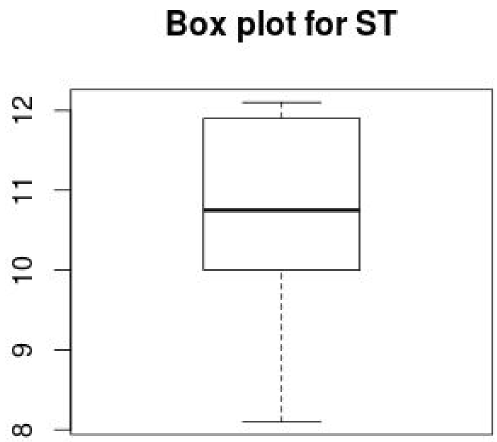
    ```
    > SS<-c(7.2,7.7,8.0,8.1,8.3,8.4,8.4,8.5,8.6,8.7,9.1,9.1,9.1,9.8,10.1,10.3)
    > ST<-c(8.1,9.2,10.0,10.4,10.6,10.9,11.1,11.9,12.0,12.1)
    ```

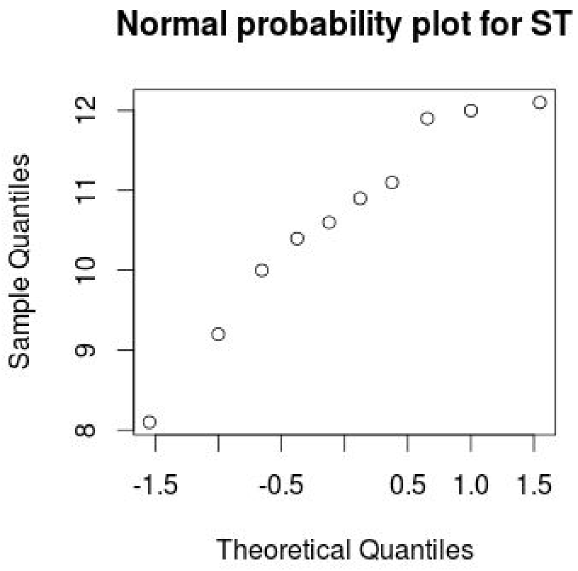
```
> SC<-c(10.7,11.3,11.5,11.6,11.7,11.8,12.0,12.1,12.3,12.6,12.6,13.3,13.3,13.8,13.9)
> par(mfrow=c(1,2))
>boxplot(SS,main="Box plot for SS")
>qqnorm(SS,main="Normal probability plot for SS")
```


Box plot for SS

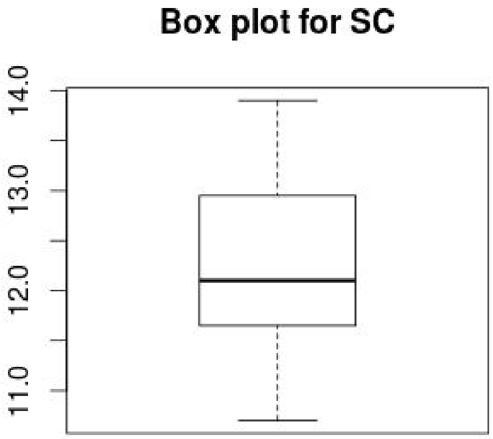

Normal probability plot for SS

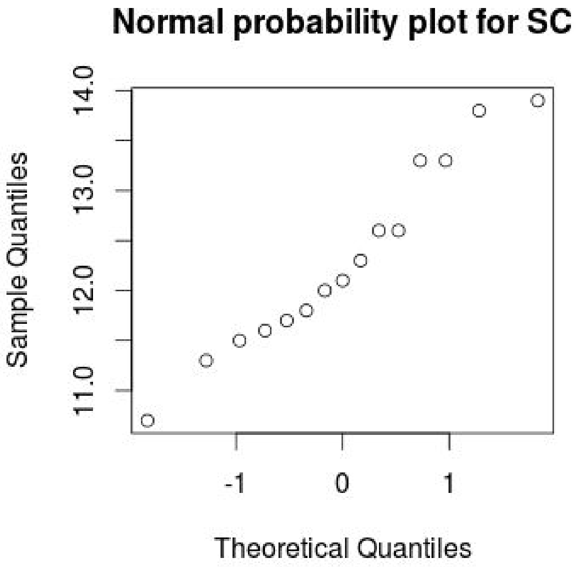
```
> boxplot(ST,main="Box plot for ST")
> qqnorm(ST,main="Normal probability plot for ST")
```


Box plot for ST


Normal probability plot for ST

```
> boxplot(SC,main="Box plot for SC")
> qqnorm(SC,main="Normal probability plot for SC")
```





From the above plots and calculations, the following observations can be concluded:

c. The assumptions of normality seem plausible, but the second plot does not accurately represent a normal distribution. However, considering all the three, they look part of a normal distribution for the most part.

d. The assumptions of homoscedasticity seem plausible because the variances of all the three samples SS, ST and SC are more or less close to each other, except that the variance of ST seems way off. However, we can classify them to same and assume that the variances are equal.

```
> var(SS)
[1] 0.7131667
> var(ST)
[1] 1.649
> var(SC)
[1] 0.8871429
```

2.

Given that the null hypothesis is that the three types of sickle cell disease have the same mean hemoglobin levels.

Also given that the significance level $\alpha = 0.05$

The R code for the computation of significance probability is as follows:

```
> n1=length(SS)
> n2=length(ST)
> n3=length(SC)
> N=n1+n2+n3
> N
[1] 41
> k=3
> mean(SS)
[1] 8.7125
> mean(ST)
[1] 10.63
> mean(SC)
[1] 12.3
> var(SS)
[1] 0.7131667
> var(ST)
[1] 1.649
> var(SC)
[1] 0.8871429
> grand_mean = (mean(SS)*(n1/N))+(mean(ST)*(n2/N))+(mean(SC)*(n3/N))
> grand_mean
[1] 10.49268
>
ss_b=(n1*(mean(SS)-grand_mean)^2)+(n2*(mean(ST)-grand_mean)^2)+(n3*(mean(SC)-grand
_mean)^2)
> ss_b
[1] 99.8893
> ss_w=(n1-1)*var(SS)+(n2-1)*var(ST)+(n3-1)*var(SC)
> ss_w
[1] 37.9585
> ss_t=ss_b+ss_w
> ss_t
[1] 137.8478
> f=(ss_b/(k-1))/(ss_w/(N-k))
> f
[1] 49.99926
> p = 1-pf(f,df1=(k-1),df2=(N-k))
> p
[1] 2.281786e-11
```

Since the computed p-value is very low when compared to our significant level, we reject our null hypothesis that the three types of sickle cell disease have the same mean hemoglobin levels.

The ANOVA table is as follows:

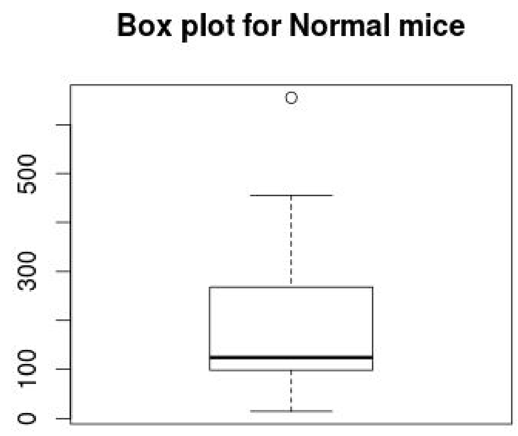
Source	**SS**	**df**	**MS**	**F**	**p**
Between	99.8893	2	49.94465	49.99926	2.281786e-11
Within	37.9585	38	0.9989079		
Total	137.8478	40			

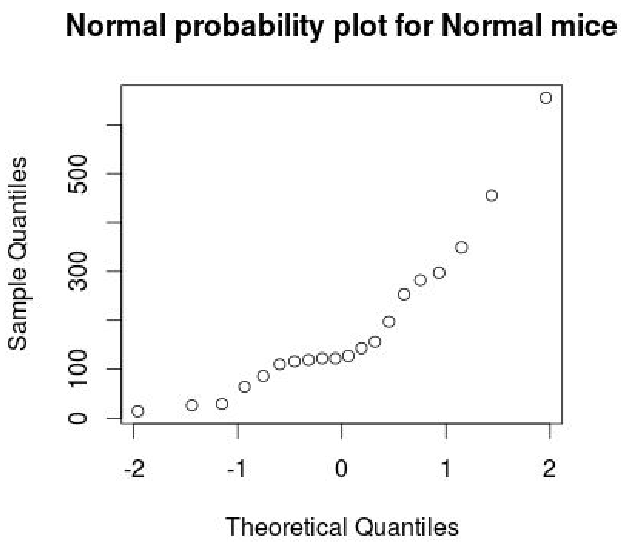
12.6 Problem Set G:

1.

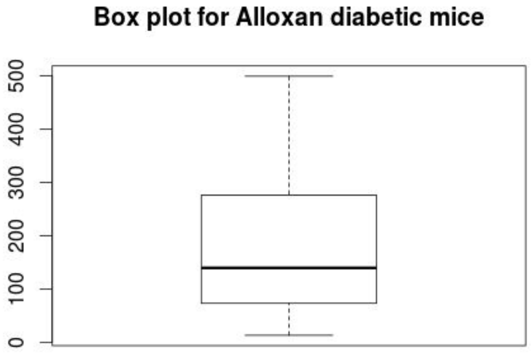
The R code for the computation of boxplots and normal probability plots is as follows:

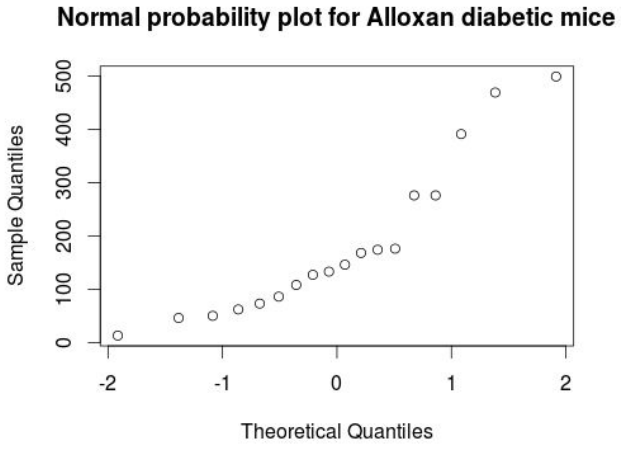
```
> N<-c(156,282,197,297,116,127,119,29,253,122,349,110,143,64,26,86,122,455,655,14)
> Al<-c(391,46,469,86,174,133,13,499,168,62,127,276,176,146,108,276,50,73)
> Al_i<-c(82,100,98,150,243,68,228,131,73,18,20,100,72,133,465,40,46,34,44)
> par(mfrow=c(1,2))
> boxplot(N,main="Box plot for Normal mice")
> qqnorm(N,main="Normal probability plot for Normal mice")
```



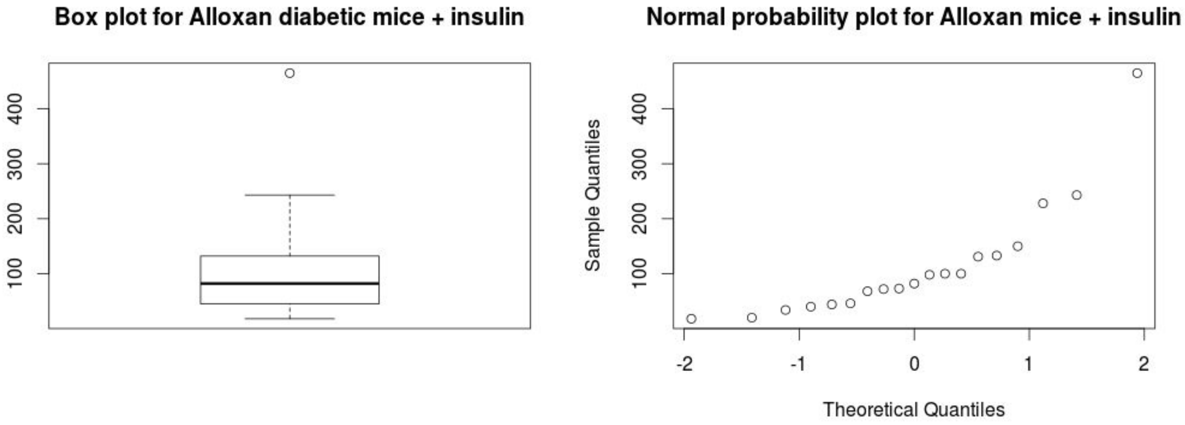


```
> boxplot(Al,main="Box plot for Alloxan diabetic mice")
> qqnorm(Al,main="Normal probability plot for Alloxan diabetic mice")
```





```
> boxplot(Al_i,main="Box plot for Alloxan diabetic mice + insulin")
> qqnorm(Al_i,main="Normal probability plot for Alloxan mice + insulin")
```



From the above plots for all the three samples, it can be concluded that the ANOVA assumptions of normality and homoscedasticity are not plausible. None of the boxplots and normal probability plots represent normal distributions.

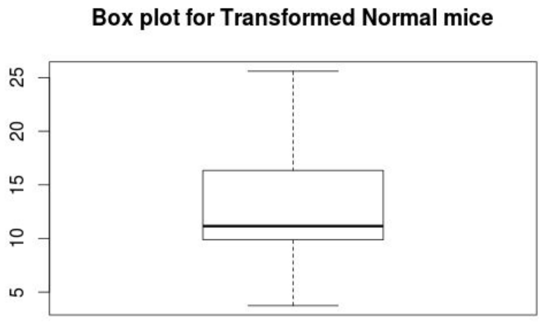
```
> var(N)
[1] 25228.52
> var(Al)
[1] 20981.32
> var(Al_i)
[1] 11191.43
> plot(var(N))
```

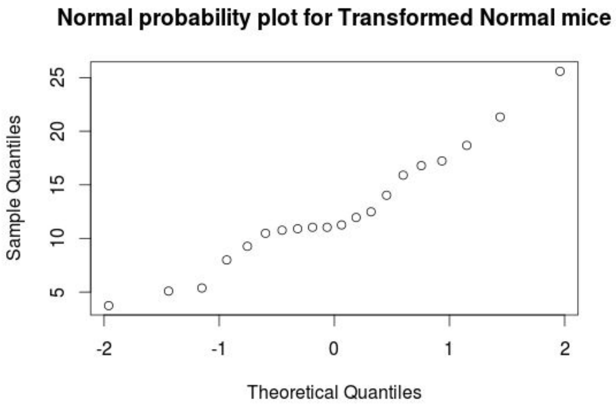
From the above variance calculations, it is safe to say that homoscedasticity is also not plausible, considering the wide disparity in the variance values of the samples.

2.

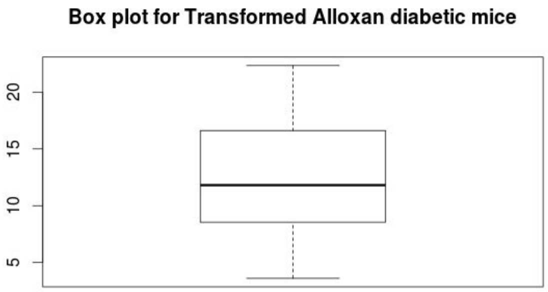
The R code for transformation of the data is as follows:

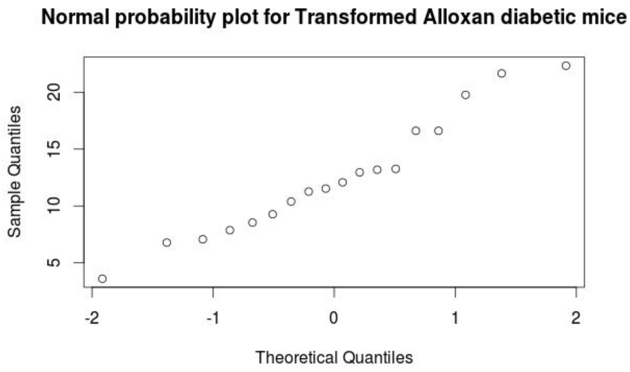
```
> N<-c(156,282,197,297,116,127,119,29,253,122,349,110,143,64,26,86,122,455,655,14)
> Al<-c(391,46,469,86,174,133,13,499,168,62,127,276,176,146,108,276,50,73)
> Al_i<-c(82,100,98,150,243,68,228,131,73,18,20,100,72,133,465,40,46,34,44)
> trN=sqrt(N)
> trAl=sqrt(Al)
> trAl_i=sqrt(Al_i)
> par(mfrow=c(1,2))

> boxplot(trN,main="Box plot for Transformed Normal mice")
> qqnorm(trN,main="Normal probability plot for Transformed Normal mice")
```



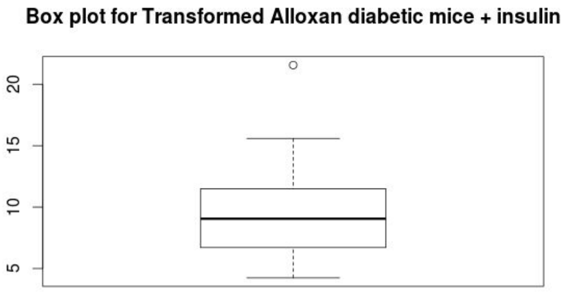


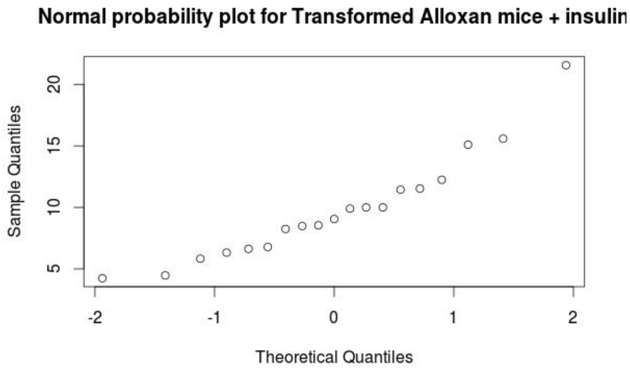
```
> boxplot(trAl,main="Box plot for Transformed Alloxan diabetic mice")
> qqnorm(trAl,main="Normal probability plot for Transformed Alloxan diabetic mice")
```





```
> boxplot(trAl_i,main="Box plot for Transformed Alloxan diabetic mice + insulin")
> qqnorm(trAl_i,main="Normal probability plot for Transformed Alloxan mice + insulin")
```





From the above plots it can be clearly seen that the transformed data belongs to a normal distribution. Though the plots of trN are a little off with respect to the other two, overall they look like they are from a normal distribution. Hence the assumption of normality is plausible.

```
> var(trN)
[1] 30.03866
> var(trAl)
[1] 27.32089
> var(trAl_i)
[1] 18.01556
```

From the above variance values, we can conclude that the assumption of homoscedasticity is plausible since the difference in variances is not as huge as the ones obtained before the transformation of data.

3.

The null hypothesis is that the means of the nitrogen-bound bovine serum albumen produced by the three groups of mice: normal, alloxan diabetic and alloxan diabetic treated with insulin have the same measurement levels.

The alternative hypothesis is that the means of the nitrogen-bound bovine serum albumen produced by the three groups of mice: normal, alloxan diabetic and alloxan diabetic treated with insulin do not have the same measurement levels.

The R code for the computation of the significance probability is as follows:

```
> n1=length(trN)
> n2=length(trAl)
> n3=length(trAl_i)
> N=n1+n2+n3
> N
[1] 57
> k=3
> mean(trN)
[1] 12.55242
> mean(trAl)
[1] 12.49121
> mean(trAl_i)
[1] 9.789145
> var(trN)
[1] 30.03866
> var(trAl)
[1] 27.32089
```

```
> var(trAl_i)
[1] 18.01556
> grand_mean = (mean(trN)*(n1/N))+(mean(trAl)*(n2/N))+(mean(trAl_i)*(n3/N))
> grand_mean
[1] 11.612
>
ss_b=(n1*(mean(trN)-grand_mean)^2)+(n2*(mean(trAl)-grand_mean)^2)+(n3*(mean(trAl_i)-gra
nd_mean)^2)
> ss_b
[1] 94.73514
> ss_w=(n1-1)*var(trN)+(n2-1)*var(trAl)+(n3-1)*var(trAl_i)
> ss_w
[1] 1359.47
> ss_t=ss_b+ss_w
> ss_t
[1] 1454.205
> f=(ss_b/(k-1))/(ss_w/(N-k))
> f
[1] 1.881505
> p = 1-pf(f,df1=(k-1),df2=(N-k))
> p
[1] 0.1622134
```

No, the null hypothesis should not be rejected for α = 0.05. Since the significance probability is more than the significance level α, we accept our null hypothesis and conclude that means of the nitrogen-bound bovine serum albumen produced by the three groups of mice: normal, alloxan diabetic and alloxan diabetic treated with insulin have the same measurement levels.

The ANOVA table for the transformed data is as follows:

Source	**SS**	**df**	**MS**	**F**	**p**
Between	94.73514	2	47.36757	1.881505	0.1622134
Within	1359.47	54	25.17537		
Total	1454.205	56			

4.

The research questions framed are as follows:

a. Does the antibody response of alloxan diabetic mice differ from the antibody response of normal mice?
b. Does the antibody response of alloxan diabetic mice treated with insulin differ from the antibody response of normal mice?
c. Does treating alloxan diabetic mice with insulin affect their antibody response?

a. The null hypothesis for the first research problem (a.) is H_0: $\theta_1 = 0$ where $\theta_1 = \mu_1 - \mu_2$ and μ_2 is the population mean of alloxan diabetic mice and μ_1 is the population mean of normal mice.

The alternative hypothesis for the first research problem (a.) is H_1: $\theta_1 \neq 0$ where $\theta_1 = \mu_1 - \mu_2$ and μ_2 is the population mean of alloxan diabetic mice and μ_1 is the population mean of normal mice.

b. The null hypothesis for the second research problem (b.) is H_0: $\theta_2 = 0$ where $\theta_2 = \mu_1 - \mu_3$ and μ_3 is the population mean of alloxan diabetic mice treated with insulin and μ_1 is the population mean of normal mice.

The alternative hypothesis for the second research problem (b.) is H_1: $\theta_2 \neq 0$ where $\theta_2 = \mu_1 - \mu_3$ and μ_3 is the population mean of alloxan diabetic mice treated with insulin and μ_1 is the population mean of normal mice.

c. The null hypothesis for the third research problem (c.) is H_0: $\theta_3 = 0$ where $\theta_3 = \mu_2 - \mu_3$ and μ_3 is the population mean of alloxan diabetic mice treated with insulin and μ_2 is the population mean of alloxan diabetic mice.

The alternative hypothesis for the third research problem (c.) is H_0: $\theta_3 \neq 0$ where $\theta_3 = \mu_2 - \mu_3$ and μ_3 is the population mean of alloxan diabetic mice treated with insulin and μ_2 is the population mean of alloxan diabetic mice.

Given that the FWER is 5% i.e, 0.05.

Therefore, the significance level α = FWER/m.

Here m = 3 pairwise comparisons,
$\alpha = 0.05/3 = 0.016$

The R code for computing the significant probabilities is as follows:

```
> ms_w = ss_w/(N-k)
> ms_w
[1] 25.17537
> t_theta1 <- (mean(trN)-mean(trAl))/sqrt(((1/n1)+(1/n2))*ms_w)
> t_theta1
[1] 0.03755071
> p1 = 2*pt(-t_theta1,df=N-k)
> p1
[1] 0.9701844
> t_theta2 <- (mean(trN)-mean(trAl_i))/sqrt(((1/n1)+(1/n3))*ms_w)
> t_theta2
[1] 1.719079
> p2 = 2*pt(-t_theta2,df=N-k)
> p2
[1] 0.09132825
>
> t_theta3 <- (mean(trAl)-mean(trAl_i))/sqrt(((1/n2)+(1/n3))*ms_w)
> t_theta3
[1] 1.637268
> p3 = 2*pt(-t_theta3,df=N-k)
> p3
[1] 0.1073898
```

From the above computations,

p1 = 0.9701844
p2 = 0.09132825
p3 = 0.1073898

Since all the three significant probabilities are greater than α = 0.016, we decline to reject all the of the three multiple null hypothesis that we have formulated with respect to contrasts.