PROJECT 2

Class: STA106 - WQ 2021

Group Member: Shih-Chi Chen

Instructor: Professor Maxime Pouokam

Part I: Report for Topic I

I. Introduction

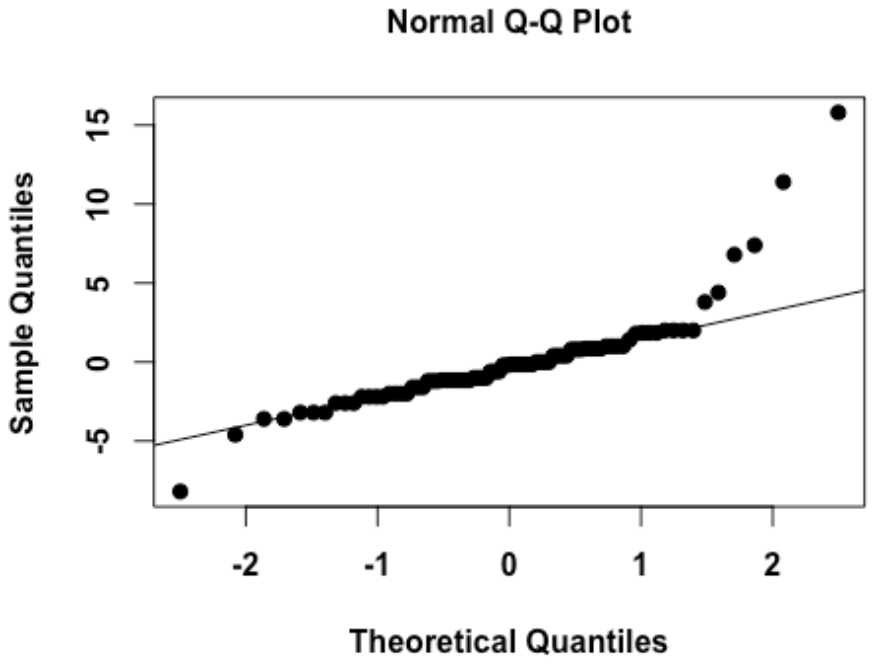
Topic I aims at looking at some plots of Helicopter data to analyze the normality, equal variance assumptions and outliers. Then some tests are conducted to check the assumptions. Finally, a decision of moving outliers or transformation of data is made to meet ANOVA assumptions.

The Helicopter dataset is useful to check different amounts of helicopters requested for a sheriff's office in different times of a day.

II. Diagnostics

- Check Normality: Q-Q Plot

From the Q-Q Plot, the data look approximately normal because most of dots are close to the line or on the line although there are some outliers exist.



- Check Normality: Shaprio-Wilks test

Since plots are subjective, the Shaprio-Wilks test is used to check.

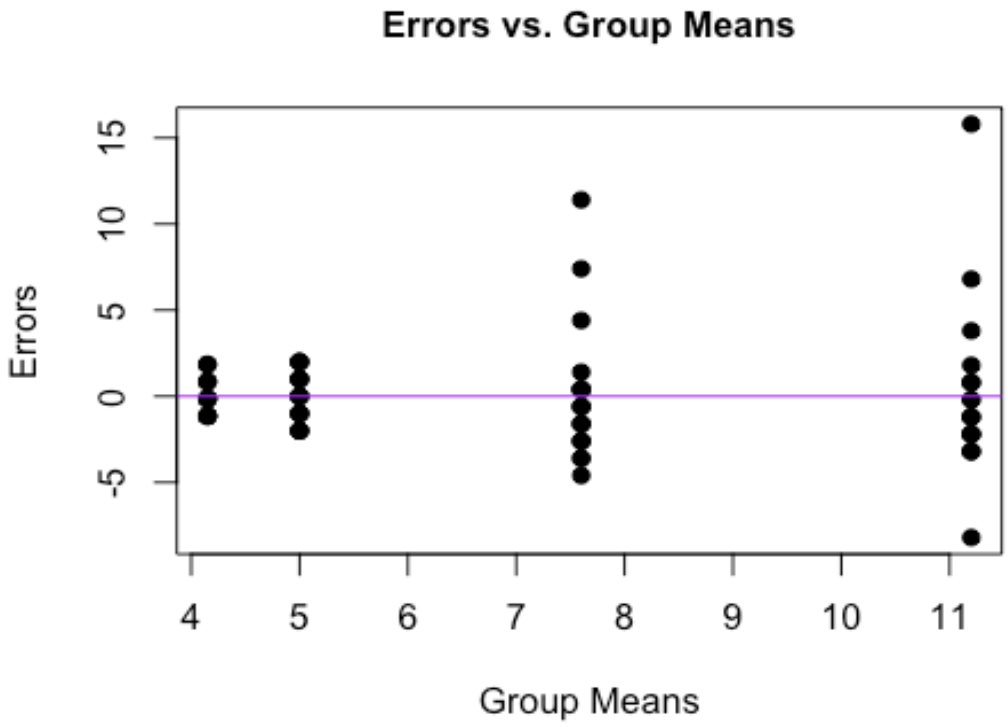
H_0 = the data is normally distributed.

H_a = the data is not normally distributed.

Since p-value is very small (3.945357e-09), the H_0 is reject and the data is not normally distributed. Therefore, the normality assumption is violated.

- Check Equal Variance: Error vs. Group Means Plot

This plot shows that the vertical spreads of variance are not equal. Especially, the variance of Shift I group (rightmost on the plot) is larger than that of others. Therefore, the equal variance assumption is violated.



- Check Equal Variance: Brown-Forsythe test

The Brown-Forsythe test is used to check the equal variance.

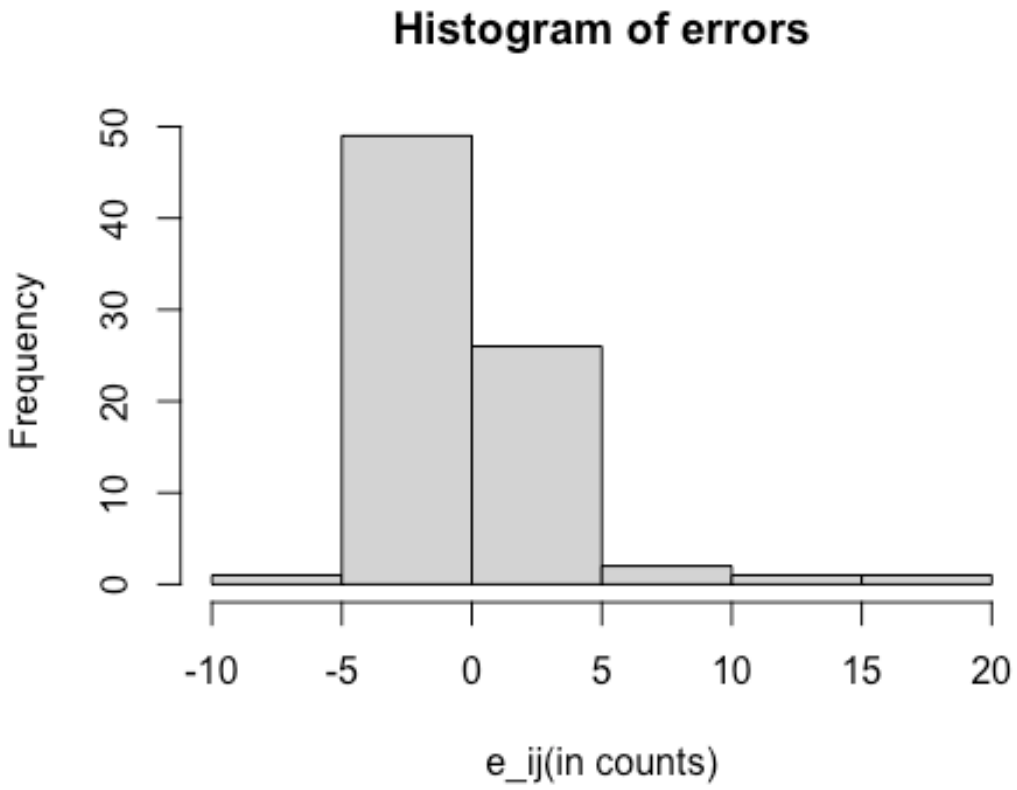
H_0 = All groups variances are equal.

H_a = At least one group variance is unequal.

The p-value (0.03185955) is small than $\alpha = 0.05$ but larger than $\alpha = 0.01$. Therefore, the conclusion depends on what value of α used. However, the equal variance is still a problem because the Error vs. Group Mean plot shows the significant difference.

- Check Outliers: Histogram of Errors

  Based on the Histogram, it is found that the residuals fall more than 5 standard deviations away from 0, which means that there are some outliers existing.



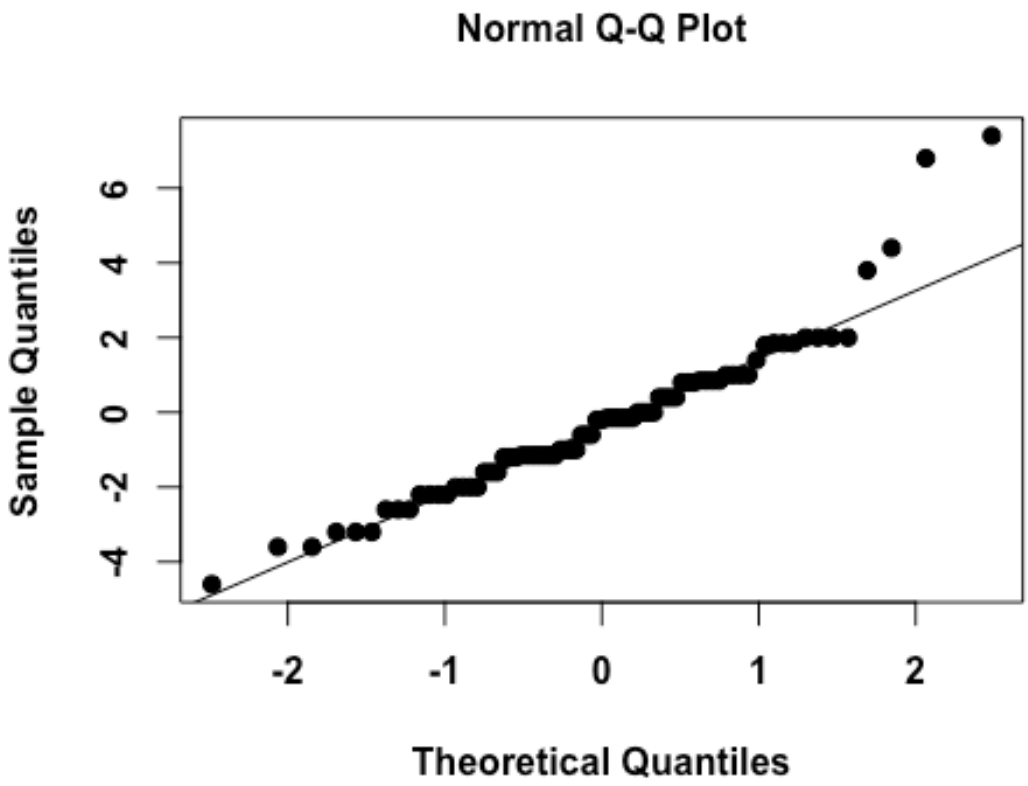
Since assumptions are violated in the analysis of original data, the data are fixed to meet the assumptions. There are three ways to fix data: (1) Remove outliers (2) Data Transformation (3) Both (Remove outliers and Transformation).

1. **Remove outliers**

Any outliers can be removed according to Standardized Residuals, with a cutoff of $t_{1-0.01;nt-a}$

- Check Normality: Q-Q Plot (Remove outliers)

After the outliers are removed, all the dots look closer to the line or on the line. However, there are still some dots away from the line.

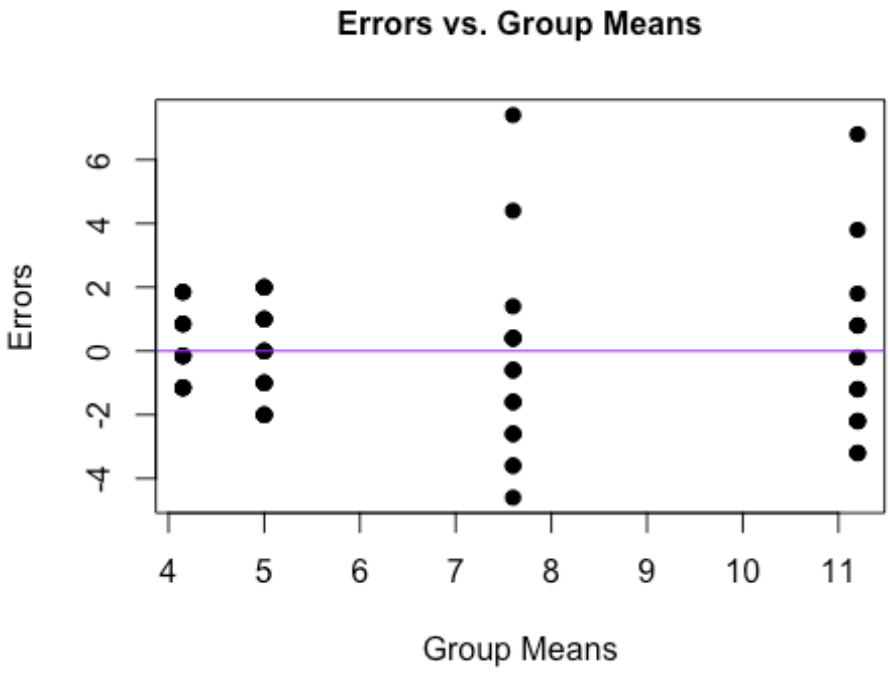


- Check Normality: Shaprio-Wilks test (Remove outliers)

The p-value(0.0004651306) in Shaprio-Wilks test becomes larger than before, but it is still very small. Therefore, H_0 is rejected and it is concluded that the data is still not normally distributed after the outliers are removed.

- Check Equal Variance: Error vs. Group Means Plot (Remove outliers)

After the outliers are removed, the vertical spreads are still significantly different. However, the vertical spreads decrease.



- Check Equal Variance: Brown-Forsythe test (Remove outliers)

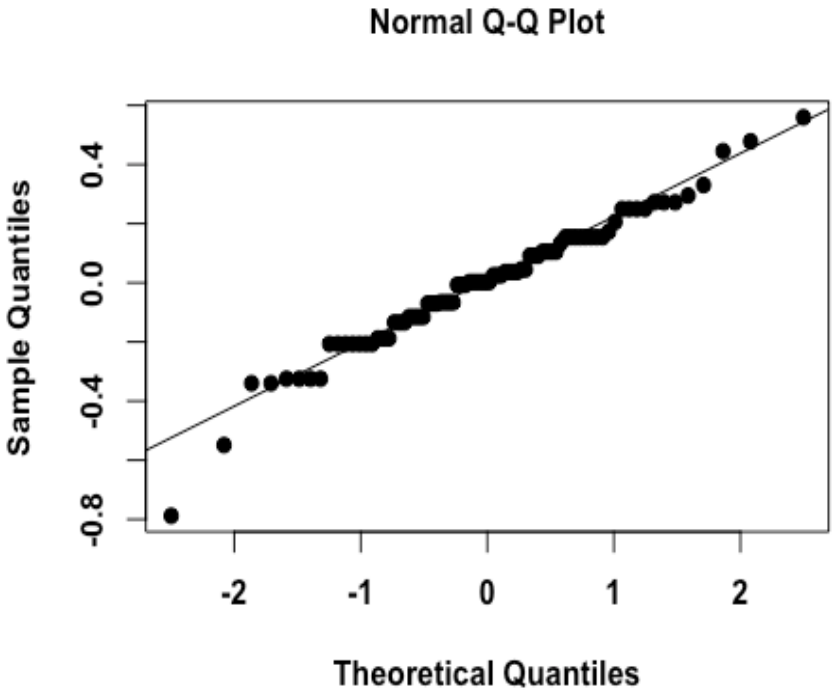
The p-value(0.04614186) in Brown-Forsythe test becomes larger than before. However, the conclusions depend on what value of α used. The equal variance is still a problem because the plot shows the significant different.

2. Data Transformation

Box-Cox Transformation (Use QQ plot method)

- Check Normality: QQ Plot (Transformation)

After data are transformed by QQ plot method, all the dots look closer to the line or on the line.

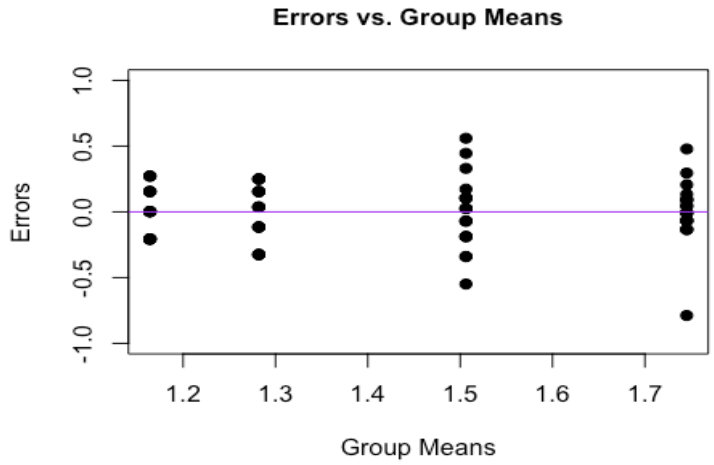


- Check Normality: Shaprio-Wilks test (Transformation)

The p-value (0.1075406) in Shaprio-Wilks test becomes larger than before. Since the p-value is large, fail to reject H_0 and it is concluded that the data is normally distributed after transformation.

- Check Equal Variance: Error vs. Group Means Plot (Transformation)

After data are transformed by QQ plot method, the vertical spreads of variance are less different than before.



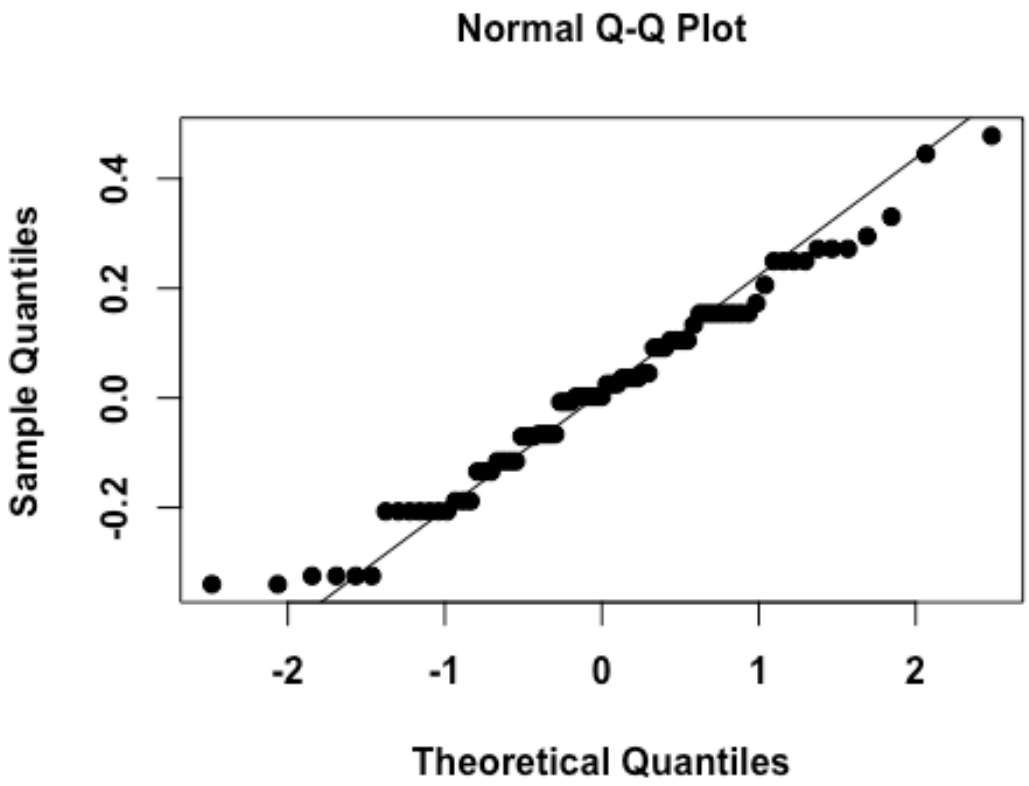
- Check Equal Variance: Brown-Forsythe test (Transformation)

The p-value (0.654481) in Brown-Forsythe test becomes larger than before. Since the p-value is large, fail to reject H_0 and it is concluded that the equal variance assumption is hold after transformation.

3. Both (Remove outliers and use QQ-Plot transfomation)

- Check Normality: QQ Plot (Both)

After outliers are removed and data are trnsformed by QQ plot method, all the dots look closer to the line or on the line.

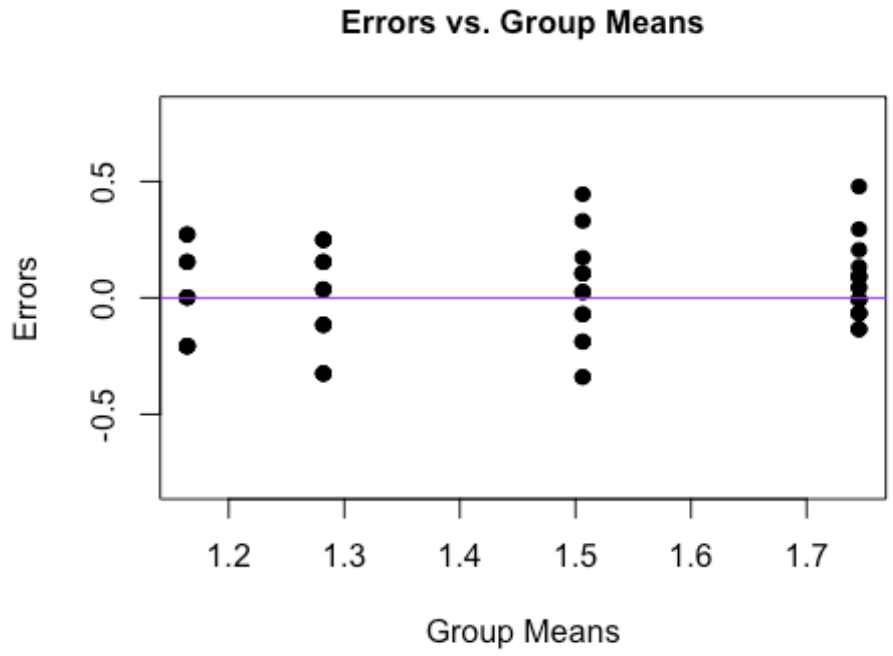


- Check Normality: Shaprio-Wilks test (Both)

The p-value (0.2744395) in Shaprio-Wilks test becomes larger than before. Since the p-value is large, fail to reject H_0 and it is concluded that the data is normally distributed after removement of outliers and transformation .

- Check Equal Variance: Error vs. Group Means Plot (Both)

After outliers are removed and data are transformed by QQ plot method, the vertical spreads of variance are approximately equal.



- Check Equal Variance: Brown-Forsythe test (Both)

The p-value(0.5055236) in Brown-Forsythe test becomes larger than before. Since the p-value is large, fail to reject H_0 and it is concluded that the equal variance assumption is hold after removement of the outliers and transformation.

III. Conclusion

From the plots and tests, it is found that normality and equal variance assumptions are violated in the analysis of original data. Thus, some outliers should be removed or data transformed. After the removement of some outliers, these assumptions are still violated. Therefore, data are transformed. The correlation to normal distribution (best QQ-Plot) method is used to transform the data. After the transformation, the normality and equal variance assumptions are hold. However, a little different in vertical spread of variance on residual plot is found. So outliers are removed and data are transformed at same time. Afterwards, it is found

that the vertical spreads of variance are approximately equal on residual plot, and normality assumptions are also hold from the QQ-plot and Shaprio-Wilks test.

In conclusion, transforming the data is helpful to make data meet the assumptions. For more better fitting, transformation and removement of the outliers at same time is needed because it makes the residual plot look more approximately equal than transformation alone. In addition, this data only has 3 outliers in total sample size 80, which means removing outliers is removing 3.75% of data and it is not large. Therefore, removing ouliers and then transforming data (Both) are worth recommending to clients. Besides, it is worth noting that transformation has some downsides, for example, transformation is irreversible.

Part II: Report for Topic II

I. Introduction

This project aims at analyzing whether there are interaction effect between technology workers title and region. It also analyzes the differences of average annual salary of technology workers between Seattle and San Francisco. By looking at some plots of this data, the normality, equal variance assumption and outliers are analyzed at first. Then, the ANOVA table is used to analyze the interaction effect. Based on the p-value, if there are no interaction effects, the hypothesis test is conducted to find whether there are main effects. Then, the conditional R^2 is calculated to find the percentage of error reduction in each model. Finally, the confidence interval is used to compare the salary of technology workers title and region.

II. Summary of data

- Sample mean and Standard deviation

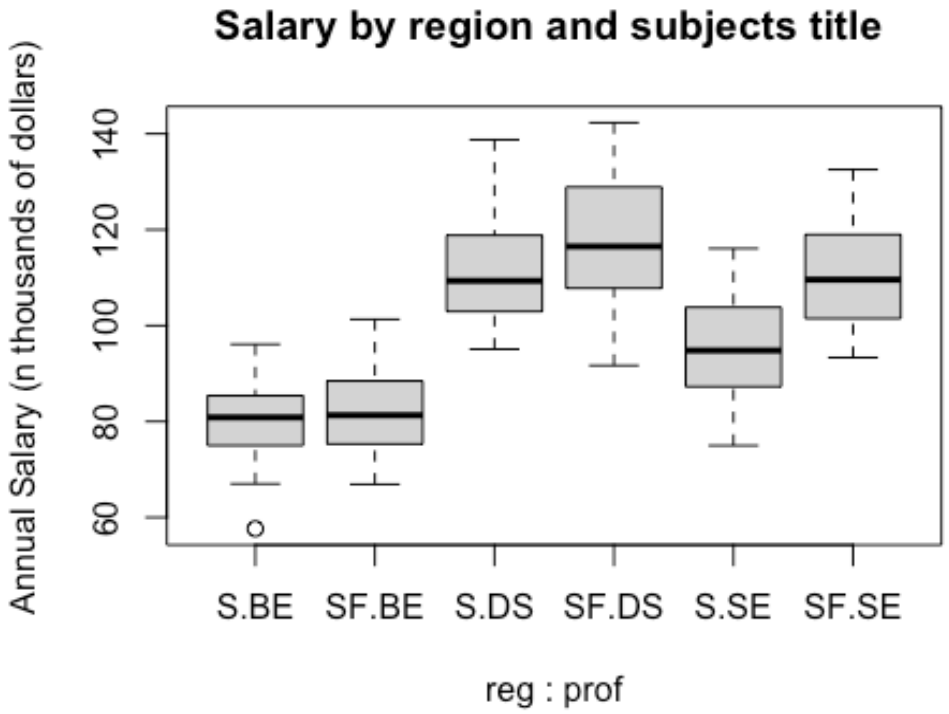
According to the mean table, the Data Scientist and San Francisco groups tend to have the highest salary. However, these two groups also have the largest standard deviations, which means that these two group data are more scattered. For the region, technology workers from San Francisco seem to have higher salaries then those from Seattle. However, salaries of San Francisco have the largest standard deviation, which means that the salaries in this region have large difference. Overall, Data Scientists from San Francisco tend to have highest salaries but large standard deviation. In contrast, Bioinformatics Engineers from Seattle tend to have lowest salaries but small standard deviation.

Sample Mean Table				
	BE (i =1) (Bioinformatics Engineer)	DS (i =2) (Data Scientist)	SE (i=3) (Software Engineer)	
Seattle (j=1)	79.75485	112.5272	95.54875	95.94358
San Francisco (j=2)	82.41914	117.7688	110.26412	103.48403
	81.087	115.148	102.9064	

Standard Deviation Table				
	BE (i =1) (BioinformaticsEngineer)	DS (i =2) (Data Scientist)	SE (i=3) (Software Engineer)	
Seattle (j=1)	8.786628	12.83857	11.59872	17.41791
San Francisco (j=2)	10.521476	14.28923	10.55171	19.29842
	9.662515	13.668190	13.240313	

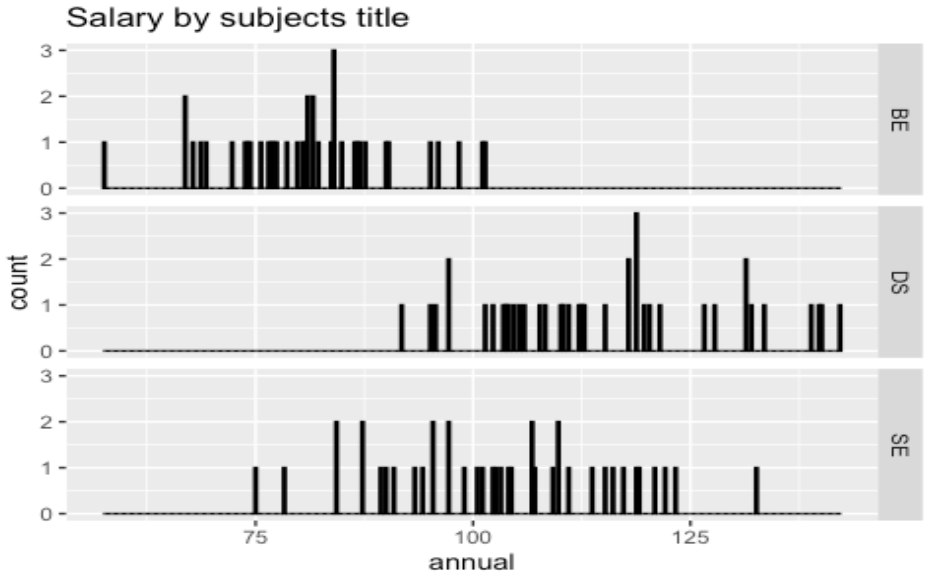
- Boxplots

It is found that people who are Data Scientists and from San Francisco have the highest median of salary, which is higher than other groups. Also, they have the largest IQR (interquartile range), which means that this group data are more scattered.



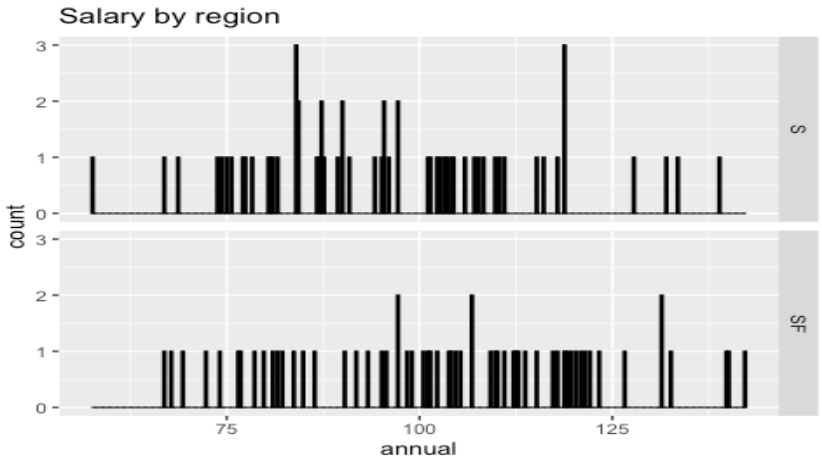
- Histogram (Salary by subject title)

It is found that the smallest salary occurs in the BE group while the largest salary occurs in the DS group.



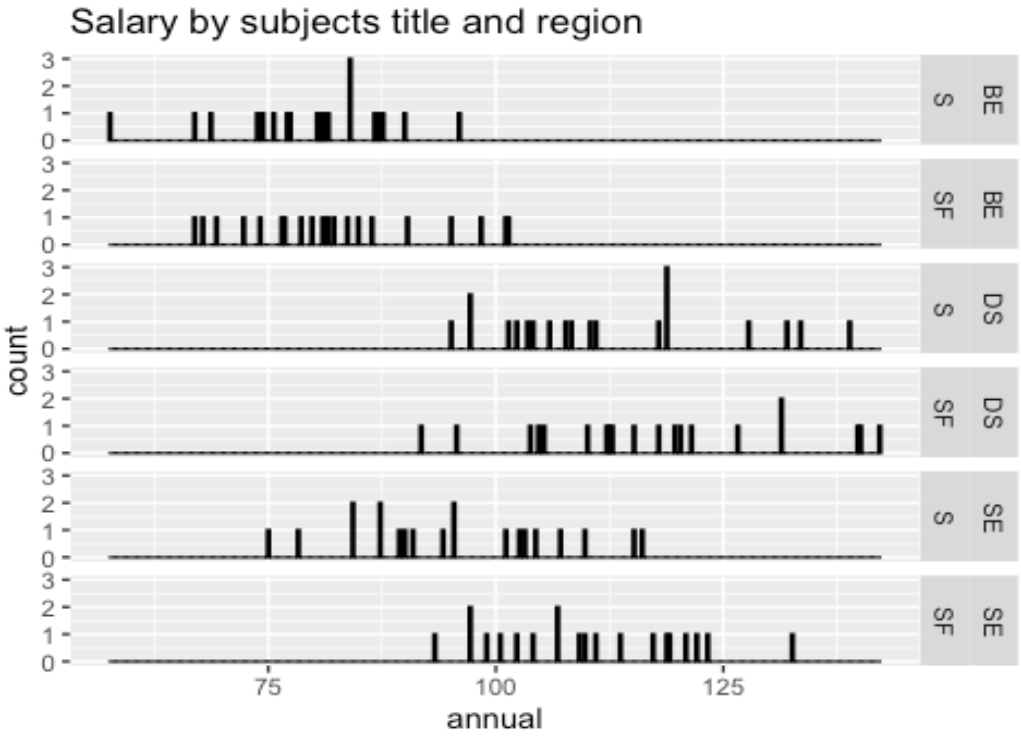
- Histogram (Salary by region)

It is found that the smallest salary occurs in the S group while the largest salary occurs in the SF group.



- Histogram (Salary by subjects title and region)

It is found that people who are data scientists and from San Francisco have the highest salary. On the other hand, people who are bioinformatics engineers and from Seattle have the lowest salary.

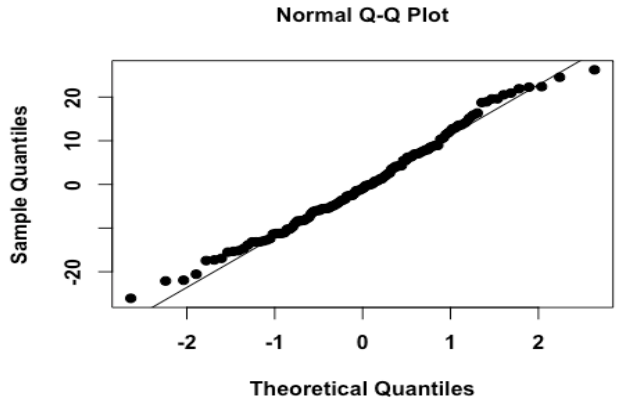


III. Diagnostics

Assume the interaction model is appropriate.

- Check Normality: QQ Plot

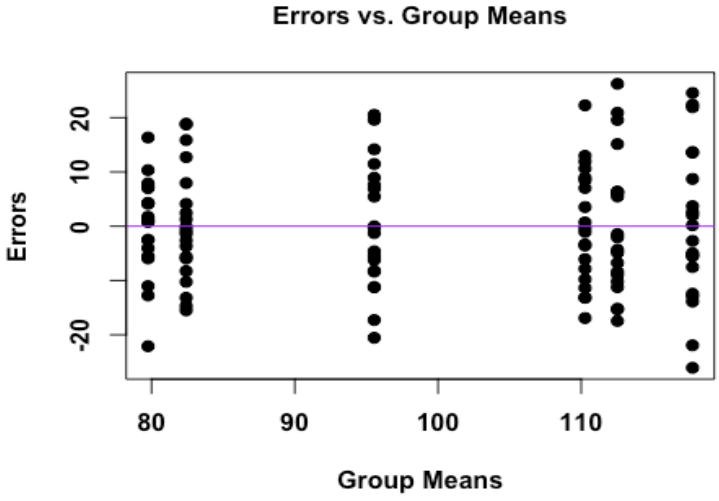
From the Q-Q Plot, the data look approximately normal because most of dots are close to the line or on the line.



- Check Normality: Shaprio-Wilks test

Since p-value is large (0.3237441), fail to reject H_0 and it is concluded that the data are normally distributed. Therefore, the normality assumption is hold.

- Check Equal Variance: Error vs. Group Means Plot



From this plot, it is found that the vertical spreads of variance are approximately equal.

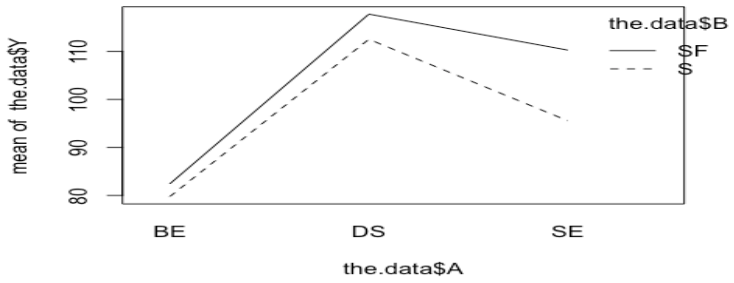
- Check Equal Variance: Brown-Forsythe test

Since the p-value is large (0.7106751), fail to reject H_0 and it is concluded that the equal variance assumption is hold.

IV. Analysis and V.Interpretation

- Interaction plot

The interaction plot shows that these two lines are not parallel, which means there are interactions.



Now, in order to find the best model of data, the hypothesis test is used and the conditional R^2 is calculated:

- Hypothesis testing

H_0: The model with no interactions is a statistically better fit than one with interactions.

H_a: The model with no interactions is not a statistically better fit than one with interactions.

	Res. Df	**RSS**	**Df**	**Sum of Sq**	**F**	**Pr (>F)**
1	116	16058				
2	114	15253	2	805.41	3.0098	0.05324

In the ANOVA table, the p-value = 0.05324. Since p-value is larger than $\alpha = 0.05$, fail to reject H_0 and it is concluded that the model with no interactions is a statistically better fit than one with interactions.

- Calculate $R^2\{AB|(A+B)\}$

$R^2\{AB|(A+B)\}$=0.0501551

When interaction effects are added to a model with subject title and region effects, the reduction in error is 5.01551%. Since add interaction model only reduce SSE 5.01551%, the model with no interactions is recommended, which has the same result as our previous hypothesis test.

Now, in order to test main effects of subject title and region, the hypothesis test is conducted, and the conditional R^2 is calculated.

- Hypothesis testing (Factor A - subject title effect)

H_0: $\gamma_i = 0$ for all i

H_a: At least one $\gamma_i \neq 0$

	Res. Df	**RSS**	**Df**	**Sum of Sq**	**F**	**Pr (>F)**
1	118	39873				
2	116	16058	2	23815	86.014	< 2.2e-16

In the ANOVA table, the p-value < 2.2e-16. Since p-value is very small, H_0 is rejected and it is concluded that the subject title effect (Factor A) exists.

- Hypothesis testing (Factor B - region effect)

H_0 : $\delta_j = 0$ for all j

H_a: At least one $\delta_j \neq 0$

	Res. Df	RSS	Df	Sum of Sq	F	Pr (>F)
1	117	17764				
2	116	16058	1	1705.8	12.322	0.0006385

In the ANOVA table, the p-value=0.0006385. Since p-value is small, H_0 is rejected and it is concluded that the region effect exists.

- Calculate $R^2\{A+B|B\}$ and $R^2\{A+B|A\}$:

$R^2\{A+B|B\}$=0.5972622

When information on subject title is added to a model with information on region, the reduction in error is 59.72%.

$R^2\{A+B|A\}$=0.09602243

When information on region is added to a model with information on subject title, the reduction in error is 9.6%.

- **Report best model**

  The model with subject title effect is recommended because the ANONA test finds that subject title and region have effects. However, from the calculation of R^2, the error reduction is found to be larger when the subject title effect is added. Besides, the reduction of error is only 9.6% when the region effect is added. Therefore, the model with subject title effect is the best one.

After find the best model, some average annual salary difference between subject title and region are worth analyzing. Therefore, some confidence intervals are analyzed.

- **Confidence intervals**

1) Find the 95% confidence interval for $\mu_{1.} - \mu_{2.}$ and $\mu_{.1} - \mu_{.2}$ (g=2):

- $\mu_{1.} - \mu_{2.}$

Bonferroni	Tukey	Scheffe
2.271	2.375	2.480

Use smallest multipliers: Bonferroni (2.271)

The 95% confidence interval for $\mu_{1.} - \mu_{2.} = (-39.93489, -28.18710)$:

We are 95% confident that the true average of annual salary for Bioinformatics Engineers are less than that for Data Scientists by between 28.18710 and 39.93489 thousands of dollars.

- $\mu_{.1} - \mu_{.2}$

Bonferroni	Tukey	Scheffe
2.271	1.981	1.982

Use smallest multipliers: Tukey(1.981)

The 95% confidence interval for $\mu_{.1} - \mu_{.2} = (-11.724029, -3.356869)$:

We are 95% confident that the true average of annual salary for technology workers from Seattle are less than those from San Francisco by between 3.356869 and 11.724029 thousands of dollars.

2) Find the 95% confidence interval for $\mu_{11} - \mu_{12}$ and $\mu_{21} - \mu_{22}$ (g=2):

- $\mu_{11} - \mu_{12}$

Bonferroni	Tukey	Scheffe
2.271	2.899	3.387

Use smallest multipliers: Bonferroni(2.271)

The 95% confidence interval for $\mu_{11} - \mu_{12} = (-10.971238, 5.642654)$:

We are 95% confident that the true average of annual salary for Bioinformatics Engineers from Seattle and Bioinformatics Engineers from San Francisco are not significantly different.

- $\mu_{21} - \mu_{22}$

Bonferroni	**Tukey**	**Scheffe**
2.271	2.899	3.387

Use smallest multipliers: Bonferroni(2.271)

The 95% confidence interval for $\mu_{21} - \mu_{22} = (-13.548628, 3.065263)$:

We are 95% confident that the true average of annual salary for Data Scientists from Seattle and Data Scientists from San Francisco are not significantly different.

3) Find the 95% confidence interval for $\frac{\mu_{21}+\mu_{22}}{2} - \frac{\mu_{11}+\mu_{12}}{2}$ and $\frac{\mu_{31}+\mu_{32}}{2} - \frac{\mu_{11}+\mu_{12}}{2}$ (g=2):

- $\frac{\mu_{21}+\mu_{22}}{2} - \frac{\mu_{11}+\mu_{12}}{2}$

Bonferroni	**Tukey**	**Scheffe**
2.271	2.899	3.387

Use smallest multipliers: Bonferroni(2.271)

The 95% confidence interval for $\frac{\mu_{21}+\mu_{22}}{2} - \frac{\mu_{11}+\mu_{12}}{2} = (28.18710, 39.93489)$:

We are 95% confident that the mean of Data Scientists from Seattle and San Francisco average annual salary are higher than the mean of Bioinformatics Engineer from Seattle and San Francisco by between 28.18710 and 39.93489 thousands of dollars.

- $\frac{\mu_{31}+\mu_{32}}{2} \_ \frac{\mu_{11}+\mu_{12}}{2}$

Bonferroni	Tukey	Scheffe
2.271	2.899	3.387

Use smallest multipliers: Bonferroni(2.271)

The 95% confidence interval for $\frac{\mu_{31}+\mu_{32}}{2} - \frac{\mu_{11}+\mu_{12}}{2} = (15.94554, 27.69334)$:

We are 95% confident that the mean of Software Engineers from Seattle and San Francisco average annual salary are higher than the mean of Bioinformatics Engineers from Seattle and San Francisco by between 15.94554 and 27.69334 thousands of dollars.

VI.Conclusion

From the plots (Q-Q plot and Error vs. Group Means Plot) and tests, the normality and equal variance assumptions of this data are hold. Therefore, the hypothesis test is conducted and the conditional R^2 is calculated to find the best model. Firstly, the interaction test is conducted. The p-value and conditional R^2 show that there are no interactions. Next, the main effects of Factor A (subject title) and Factor B (region) are tested. It is found that both factors have effects individually. Accordingly, conditional R^2 is calculated to see which factor is important. Afterwards, it is found that Factor A reduces error by 59% while Factor B only reduces error by 9.6%. Therefore, it is concluded that the model with subject title is the best model. Furthermore, the confidence interval shows that the largest difference of average annual salary in subject title occurs between Bioinformatics Engineers and Data Scientists, which is 39.93489 thousands of dollars. Another meaningful finding is that the average annual salary of Bioinformatics Engineers and Data Scientists from Seattle and San Francisco are all not significantly different. Moreover, the contrast (not pairwise) confidence interval shows that the smallest difference of annual salary occurs between the mean of Software Engineers Engineers from Seattle & San Francisco and the mean of Bioinformatics Engineers from Seattle & San Francisco, which is 15.94554 thousands of dollars.

VII. Appedix

```
####Part I:  Report for Topic I####

#Data setting
Helicopter=read.csv("~/Downloads/Helicopter.csv")
the.data = Helicopter
count=Helicopter$Count
shift=Helicopter$Shift
group.means = by(count,shift,mean)

#QQ plot (original)
the.model = lm(count ~ shift, data = Helicopter)
Helicopter$ei = the.model$residuals
qqnorm(the.model$residuals,pch = 19,font = 2,font.lab =2,cex =1,cex.lab=1, ce
x.axis=1, cex.main=1, cex.sub=1)
qqline(the.model$residuals)

#Shaprio-Wilks test (original)
sp.pval = shapiro.test(the.model$residuals)$p.val
sp.pval

# Errors vs. Group Means plot (original)
plot(the.model$fitted.values, the.model$residuals, main = "Errors vs. Group M
eans",xlab = "Group Means",ylab = "Errors",pch = 19,font = 1,font.lab =1,cex
=1,cex.lab=1, cex.axis=1, cex.main=1, cex.sub=1)
abline(h = 0,col = "purple")

#Brown-Forsythe test (original)
library(car)
the.BFtest = leveneTest(the.model$residuals~shift , data=the.data, center=med
ian)
bf.p.val = the.BFtest[[3]][1]
bf.p.val

#check outliers (original)
hist(Helicopter$ei,main = "Histogram of errors",xlab = "e_ij(in counts)")

#find outliers
alpha = 0.01
a = length(unique(Helicopter$Shift))
nt = nrow(Helicopter)
rij = rstandard(the.model)
t.cutoff= qt(1-alpha, nt-a)
CO.rij = abs(rij) > t.cutoff

#QQ plot (Remove any outliers)
qqnorm(the.model$residuals[!CO.rij],pch = 19,font = 2,font.lab =2,cex =1,cex.
lab=1, cex.axis=1, cex.main=1, cex.sub=1)
qqline(the.model$residuals)
```

```r
sp.pval.2 = shapiro.test(the.model$residuals[!CO.rij])$p.val
sp.pval.2

# Errors vs. Group Means plot (Remove any outliers)
plot(the.model$fitted.values[!CO.rij], the.model$residuals[!CO.rij], main = "
Errors vs. Group Means",xlab = "Group Means",ylab = "Errors",pch = 19,font =
1,font.lab =1,cex =1,cex.lab=1, cex.axis=1, cex.main=1, cex.sub=1)
abline(h = 0,col = "purple")

#Brown-Forsythe test (Remove any outliers)
the.BFtest2 = leveneTest(the.model$residuals[!CO.rij]~shift[!CO.rij] , data=t
he.data, center=median)
bf.p.val.2 = the.BFtest2[[3]][1]
bf.p.val.2

#Create a new data set with the transformed Y (QQ plot method)
library(EnvStats)
the.model = lm(count ~ shift,data = the.data)
boxcox(the.model ,objective.name = "PPCC")
L1 =boxcox(the.model ,objective.name = "PPCC",optimize = TRUE)$lambda
YT = (the.data$Count^(L1)-1)/L1
t.data.pc = data.frame(count = YT, shift = the.data$Shift)
t.model.pc = lm(count ~ shift,data = t.data.pc)

#QQ plot (Transformation qq)
t.data.pc$ei = t.model.pc$residuals
qqnorm(t.model.pc$residuals,pch = 19,font = 2,font.lab =2,cex =1,cex.lab=1, c
ex.axis=1, cex.main=1, cex.sub=1)
qqline(t.model.pc$residuals)

#Shaprio-Wilks test  (Transformation qq)
sp.pval = shapiro.test(t.model.pc$residuals)$p.val
sp.pval

# Errors vs. Group Means plot (Transformation qq)
plot(t.model.pc$fitted.values, t.model.pc$residuals, main = "Errors vs. Group
 Means",xlab = "Group Means",ylab = "Errors",ylim=c(-1,1),pch = 19,font = 1,f
ont.lab =1,cex =1,cex.lab=1, cex.axis=1, cex.main=1, cex.sub=1)
abline(h = 0,col = "purple")

#Brown-Forsythe test (Transformation qq)
the.BFtest2 = leveneTest(t.model.pc$residuals~shift , data=t.data.pc, center=
median)
bf.p.val = the.BFtest2[[3]][1]
bf.p.val

#Create a new data set with the transformed Y (QQ plot method)
the.model = lm(count ~ shift,data = the.data)
boxcox(the.model ,objective.name = "PPCC")
```

```r
L1 =boxcox(the.model ,objective.name = "PPCC",optimize = TRUE)$lambda
YT = (the.data$Count^(L1)-1)/L1
t.data.pc = data.frame(count = YT, shift = the.data$Shift)
t.model.pc = lm(count ~ shift,data = t.data.pc)

#QQ plot (Remove outliers and transformation)
alpha = 0.01
a = length(unique(t.data.pc$shift))
nt = 80
rij = rstandard(t.model.pc)
t.cutoff= qt(1-alpha, nt-a)
CO.rij = abs(rij) > t.cutoff
qqnorm(t.model.pc$residuals[!CO.rij],pch = 19,font = 2,font.lab =2,cex =1,cex.
lab=1, cex.axis=1, cex.main=1, cex.sub=1)
qqline(t.model.pc$residuals)

#Shaprio-Wilks test (Remove outliers and transformation)
sp.pval.2 = shapiro.test(t.model.pc$residuals[!CO.rij])$p.val
sp.pval.2

#Errors vs. Group Means plot (Remove outliers and transformation)
plot(t.model.pc$fitted.values[!CO.rij], t.model.pc$residuals[!CO.rij], main =
 "Errors vs. Group Means",xlab = "Group Means",ylab = "Errors",ylim=c(-0.8,0.
8),pch = 19,font = 1,font.lab =1,cex =1,cex.lab=1, cex.axis=1, cex.main=1, ce
x.sub=1)
abline(h = 0,col = "purple")

#Brown-Forsythe test (Remove outliers and transformation)
the.BFtest2 = leveneTest(t.model.pc$residuals[!CO.rij]~shift[!CO.rij] , data=
t.data.pc, center=median)
bf.p.val.2 = the.BFtest2[[3]][1]
bf.p.val.2

####Part II:  Report for Topic II####
#data setting
Salary=read.csv("~/Downloads/Salary.csv")
the.data = Salary
annual=Salary$Annual
prof=Salary$Prof
reg=Salary$Region

#boxplot
boxplot(annual ~ reg+prof, main = "Salary by region and subjects title ",ylab
 = "Annual Salary (n thousands of dollars)")

#Histogram (Salary by subjects title)
library(ggplot2)
ggplot(the.data, aes(x = annual)) + geom_histogram(binwidth = 0.3,color = "bl
ack",fill = "white") +facet_grid(the.data$Prof ~.) +ggtitle("Salary by subjec
```

```
ts title")

#Histogram (Salary by region)
ggplot(the.data, aes(x = annual)) + geom_histogram(binwidth = 0.3,color = "bl
ack",fill = "white") +facet_grid(the.data$Region ~.) +ggtitle("Salary by regi
on")

#Histogram (Salary by subjects title and region)
library(ggplot2)
ggplot(the.data, aes(x = annual)) + geom_histogram(binwidth = 0.3,color = "bl
ack",fill = "white") +facet_grid(the.data$Prof+Region ~.) +ggtitle("Salary by
 subjects title and region")

#find sample mean and sd
find.means = function(the.data,fun.name = mean){
    a = length(unique(the.data[,2]))
    b = length(unique(the.data[,3]))
    means.A = by(the.data[,1], the.data[,2], fun.name)
    means.B = by(the.data[,1],the.data[,3],fun.name)
    means.AB = by(the.data[,1],list(the.data[,2],the.data[,3]),fun.name)
    MAB = matrix(means.AB,nrow = b, ncol = a, byrow = TRUE)
    colnames(MAB) = names(means.A)
    rownames(MAB) = names(means.B)
    MA = as.numeric(means.A)
    names(MA) = names(means.A)
    MB = as.numeric(means.B)
    names(MB) = names(means.B)
    results = list(A = MA, B = MB, AB = MAB)
    return(results)
}

the.means = find.means(the.data)
the.sd=find.means(the.data,sd)

nt = nrow(the.data)
a = length(unique(the.data[,2]))
b = length(unique(the.data[,3]))
names(the.data) = c("Y","A","B")
nt = nrow(the.data)

#All model
AB = lm(Y ~ A*B,the.data)
A.B = lm(Y ~ A + B,the.data)
A = lm(Y ~ A,the.data)
B = lm(Y ~ B,the.data)
N = lm(Y ~ 1, the.data)

#QQ-plot (interaction model)
Salary$ei = AB$residuals
```

```r
qqnorm(AB$residuals,pch = 19,font = 2,font.lab =2,cex =1,cex.lab=1, cex.axis=
1, cex.main=1, cex.sub=1)
qqline(AB$residuals)

#Shaprio-Wilks test
sp.pval = shapiro.test(AB$residuals)$p.val
sp.pval

#Error vs group means plot
plot(AB$fitted.values, AB$residuals, main = "Errors vs. Group Means",xlab = "
Group Means",ylab = "Errors",pch = 19,font = 2,font.lab =2,cex =1,cex.lab=1,
cex.axis=1, cex.main=1, cex.sub=1)
abline(h = 0,col = "purple")

#Brown-Forsythe test
the.BFtest = leveneTest(AB$residuals~reg , data=the.data, center=median)
bf.p.val = the.BFtest[[3]][1]
bf.p.val

#Interaction plot
interaction.plot(the.data$A, the.data$B,the.data$Y)

#test interactions
anova(A.B,AB)

#R2 for interactions
anova(A.B,AB)
Partial.R2 = function(small.model,big.model){
    SSE1 = sum(small.model$residuals^2)
    SSE2 = sum(big.model$residuals^2)
    PR2 = (SSE1 - SSE2)/SSE1
    return(PR2)
}
rab=Partial.R2(A.B,AB)
rab

#test factor A effect
anova(B,A.B)

#test factor B effect
anova(A,A.B)

#R2 for A and B
ra=Partial.R2(B,A.B)
ra
rb=Partial.R2(A,A.B)
rb

#function of ci
```

```
scary.CI = function(the.data,MSE,equal.weights = TRUE,multiplier,group,cs){
    if(sum(cs) != 0 & sum(cs !=0 ) != 1){
        return("Error - you did not input a valid contrast")
    }else{
        the.means = find.means(the.data)
        the.ns =find.means(the.data,length)
        nt = nrow(the.data)
        a = length(unique(the.data[,2]))
        b = length(unique(the.data[,3]))
        if(group =="A"){
            if(equal.weights == TRUE){
                a.means = rowMeans(the.means$AB)
                est = sum(a.means*cs)
                mul = rowSums(1/the.ns$AB)
                SE = sqrt(MSE/b^2 * (sum(cs^2*mul)))
                N = names(a.means)[cs!=0]
                CS = paste("(",cs[cs!=0],")",sep = "")
                fancy = paste(paste(CS,N,sep =""),collapse = "+")
                names(est) = fancy
            } else{
                a.means = the.means$A
                est = sum(a.means*cs)
                SE = sqrt(MSE*sum(cs^2*(1/the.ns$A)))
                N = names(a.means)[cs!=0]
                CS = paste("(",cs[cs!=0],")",sep = "")
                fancy = paste(paste(CS,N,sep =""),collapse = "+")
                names(est) = fancy
            }
        }else if(group == "B"){
            if(equal.weights == TRUE){
                b.means = colMeans(the.means$AB)
                est = sum(b.means*cs)
                mul = colSums(1/the.ns$AB)
                SE = sqrt(MSE/a^2 * (sum(cs^2*mul)))
                N = names(b.means)[cs!=0]
                CS = paste("(",cs[cs!=0],")",sep = "")
                fancy = paste(paste(CS,N,sep =""),collapse = "+")
                names(est) = fancy
            } else{
                b.means = the.means$B
                est = sum(b.means*cs)
                SE = sqrt(MSE*sum(cs^2*(1/the.ns$B)))
                N = names(b.means)[cs!=0]
                CS = paste("(",cs[cs!=0],")",sep = "")
                fancy = paste(paste(CS,N,sep =""),collapse = "+")
                names(est) = fancy
            }
        } else if(group == "AB"){
            est = sum(cs*the.means$AB)
            SE = sqrt(MSE*sum(cs^2/the.ns$AB))
```

```
            names(est) = "someAB"
        }
        the.CI = est + c(-1,1)*multiplier*SE
        results = c(est,the.CI)
        names(results) = c(names(est),"lower bound","upper bound")
        return(results)
    }
}

#function of all multipliers
find.mult = function(alpha,a,b,dfSSE,g,group){
    if(group == "A"){
        Tuk = round(qtukey(1-alpha,a,dfSSE)/sqrt(2),3)
        Bon = round(qt(1-alpha/(2*g), dfSSE ) ,3)
        Sch = round(sqrt((a-1)*qf(1-alpha, a-1, dfSSE)),3)
    }else if(group == "B"){
        Tuk = round(qtukey(1-alpha,b,dfSSE)/sqrt(2),3)
        Bon = round(qt(1-alpha/(2*g), dfSSE ) ,3)
        Sch = round(sqrt((b-1)*qf(1-alpha, b-1, dfSSE)),3)
    }else if(group == "AB"){
        Tuk = round(qtukey(1-alpha,a*b,dfSSE)/sqrt(2),3)
        Bon = round(qt(1-alpha/(2*g), dfSSE ) ,3)
        Sch = round(sqrt((a*b-1)*qf(1-alpha, a*b-1, dfSSE)),3)
    }
    results = c(Bon, Tuk,Sch)
    names(results) = c("Bonferroni","Tukey","Scheffe")
    return(results)
}
the.means = find.means(the.data)
the.model = lm(Y ~ A*B, data = the.data)
SSE = sum(the.model$residuals^2)
MSE = SSE/(nt-a*b)

#CI for u1.-u2.
all.mult = find.mult(alpha = 0.05, a = 3, b = 2, dfSSE = 120 - 3*2, g = 2, gr
oup = "A")
Bon = all.mult[1]
Tuk = all.mult[2]
Sch = all.mult[3]
allmu=c(Bon,Tuk,Sch)
allmu
A.cs.12 = c(1,-1,0)
scary.CI(the.data,MSE,equal.weights = FALSE,Bon,"A",A.cs.12)

#CI for u.1-u.2
all.mult = find.mult(alpha = 0.05, a = 3, b = 2, dfSSE = 120 - 3*2, g = 2, gr
oup = "B")
Bon = all.mult[1]
Tuk = all.mult[2]
Sch = all.mult[3]
```

```r
allmu=c(Bon,Tuk,Sch)
allmu
B.cs.12 = c(1,-1)
scary.CI(the.data,MSE,equal.weights = FALSE,Tuk,"B",B.cs.12)

#CI for u11-u12
the.means$AB
all.mult = find.mult(alpha = 0.05, a = 3, b = 2, dfSSE = 120 - 3*2, g = 2, gr
oup = "AB")
Bon = all.mult[1]
Tuk = all.mult[2]
Sch = all.mult[3]
allmu=c(Bon,Tuk,Sch)
allmu
AB.cs.1 = matrix(0,nrow = b, ncol = a)
AB.cs.1[1,1] = 1
AB.cs.1[2,1] = -1
scary.CI(the.data,MSE,equal.weights = FALSE,Bon,"AB",AB.cs.1)

#CI for u21-u22
the.means$AB
all.mult = find.mult(alpha = 0.05, a = 3, b = 2, dfSSE = 120 - 3*2, g = 2, gr
oup = "AB")
Bon = all.mult[1]
Tuk = all.mult[2]
Sch = all.mult[3]
allmu=c(Bon,Tuk,Sch)
allmu
AB.cs.2 = matrix(0,nrow = b, ncol = a)
AB.cs.2[1,2] = 1
AB.cs.2[2,2] = -1
scary.CI(the.data,MSE,equal.weights = FALSE,Bon,"AB",AB.cs.2)

#CI for 2122/2-1112/2
the.means$AB
all.mult = find.mult(alpha = 0.05, a = 3, b = 2, dfSSE = 120 - 3*2, g = 2, gr
oup = "AB")
Bon = all.mult[1]
Tuk = all.mult[2]
Sch = all.mult[3]
allmu=c(Bon,Tuk,Sch)
allmu
AB.cs.3 = matrix(0,nrow = b, ncol = a)
AB.cs.3[1,2] = 1/2
AB.cs.3[2,2] = 1/2
AB.cs.3[1,1] = -1/2
AB.cs.3[2,1] = -1/2
scary.CI(the.data,MSE,equal.weights = FALSE,Bon,"AB",AB.cs.3)

#CI for 3132/2-1112/2
```

```
the.means$AB
all.mult = find.mult(alpha = 0.05, a = 3, b = 2, dfSSE = 120 - 3*2, g = 2, gr
oup = "AB")
Bon = all.mult[1]
Tuk = all.mult[2]
Sch = all.mult[3]
allmu=c(Bon,Tuk,Sch)
allmu
AB.cs.4 = matrix(0,nrow = b, ncol = a)
AB.cs.4[1,3] = 1/2
AB.cs.4[2,3] = 1/2
AB.cs.4[1,1] = -1/2
AB.cs.4[2,1] = -1/2
scary.CI(the.data,MSE,equal.weights = FALSE,Bon,"AB",AB.cs.4)
```