STA 106 Project 2

Instructor: Dr. Melcon

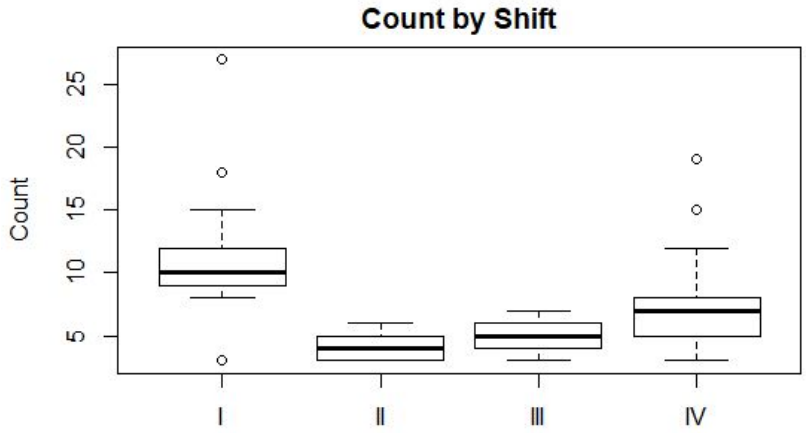
Xin Ye & Yutian Yang

Report for Topic I Question Two

Introduction:

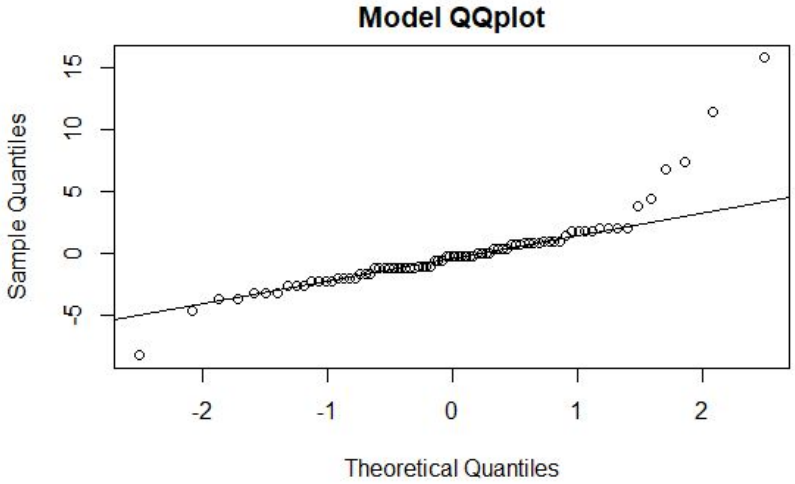
This report focuses on the diagnostics and transformation of the data about the number of times a helicopter was called to an emergency in one year by different shift.

Boxplot for Original Data:



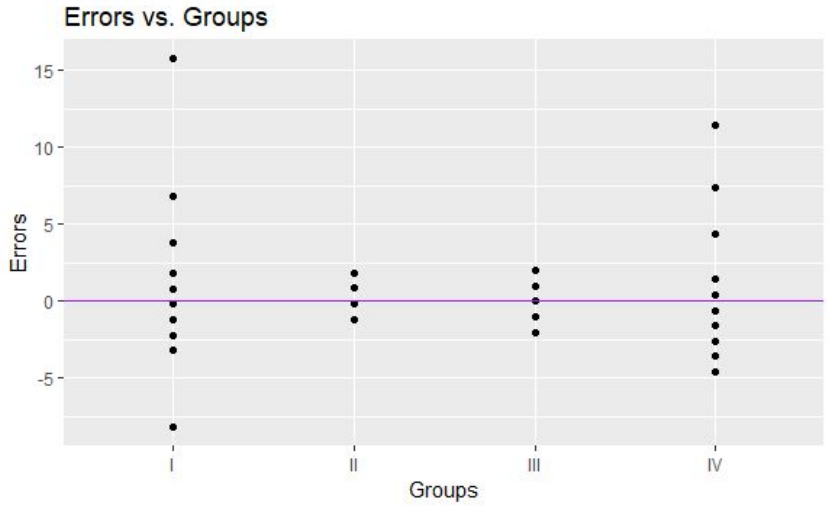
Diagnostics for Original Data

QQplot



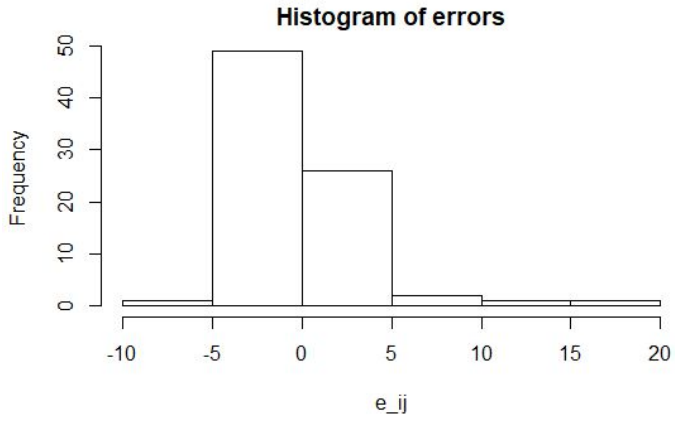
The end tail of the data derives far away from the line, suggesting the violation of the assumption of normality

Errors vs. Groups



The plot suggests unequal vertical spread. Therefore, we assume the assumption of equal variance is violated.

Histogram of Errors



The plot does not suggest the distribution of errors being "bell-shaped", therefore we assume the assumption of normality is violated.

For the Original Data:

Assumptions for Single-Factor ANOVA:

1. Random samples from all treatments
2. Groups are independent.
3. $\varepsilon_{ijk} \sim N(0,\sigma^2_{\varepsilon})$ (independent & identically distributed) (violated by the data)

The Shapiro-Wilk Test for the original data gives a p-value = 3.945e-09 (equals to 0). (Violation of normality at any significant level)
The Brown-Forsythe Test for the original data gives the p-value = 0.0319. (Violation of equal variance at alpha = 0.1, alpha = 0.05, fits the constant variance assumption at alpha = 0.01)
Therefore, the third assumption of Single-Factor ANOVA has been violated.

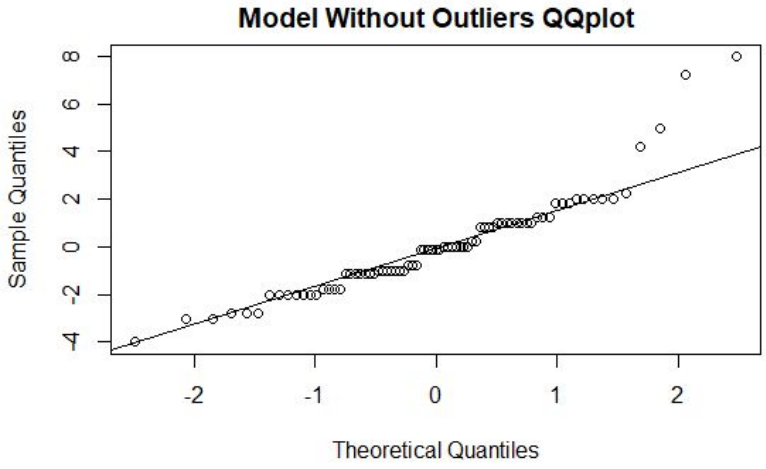
Summary:
The diagnostics suggest that the original data violated both the assumptions of normality and equal variance in the ANOVA model, thus it cannot fit into the ANOVA model. Therefore, we need transformation to alleviate the violation of assumptions and helps the data to fit into the ANOVA model.

Model without Outliers:

We use semi-studentized to find potential outliers in our data. By choosing alpha = 0.01 and the t statistics for cutoff point = 2.37642, we conclude three potential outliers in our data, which takes roughly 3.75% of our data. After removing outliers, the Shapiro-Wilk normality test gives a p-value = 2.023e-05 (equals to 0) for our new model, which has violated the ANOVA assumption of normality at any significant level. The Brown-Forsythe test for the new model gives the p-value = 0.04614186. We reject H_0: the data has constant variance at alpha = 0.1 and alpha = 0.05, and we fail to reject H_0 at alpha = 0.01. This indicates that our data can only be assumed having constant variance at a small alpha level.

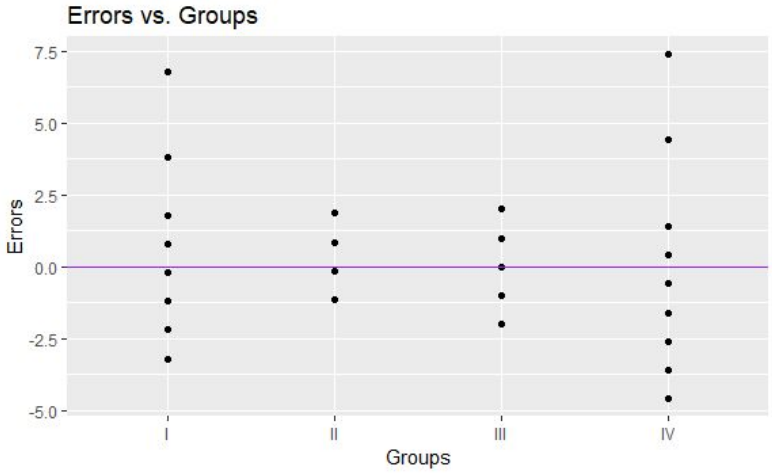
Diagnostics for New Data

QQplot



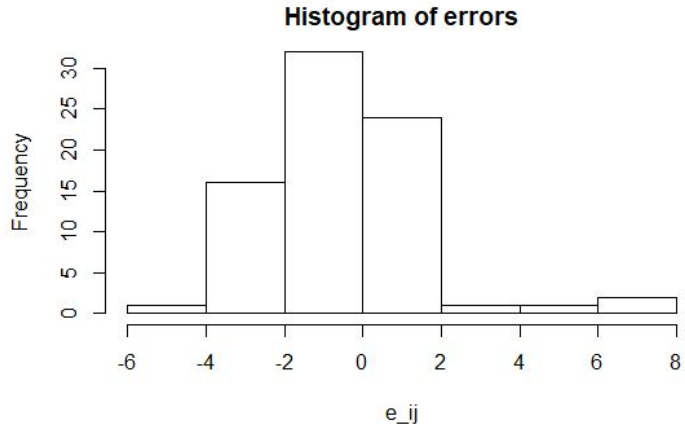
The end of the tail of the new data derives far away from the line, suggesting the violation of the assumption of normality.

Errors vs. Groups



The plot does not suggest equal vertical spread of the new data. Therefore, we conclude the assumption of equal variance is violated.

Histogram of Errors



The plot does not suggest the distribution of errors of the new data being “bell-shaped”, therefore we assume the assumption of normality is violated.

Transformed Model without Removing Outliers:

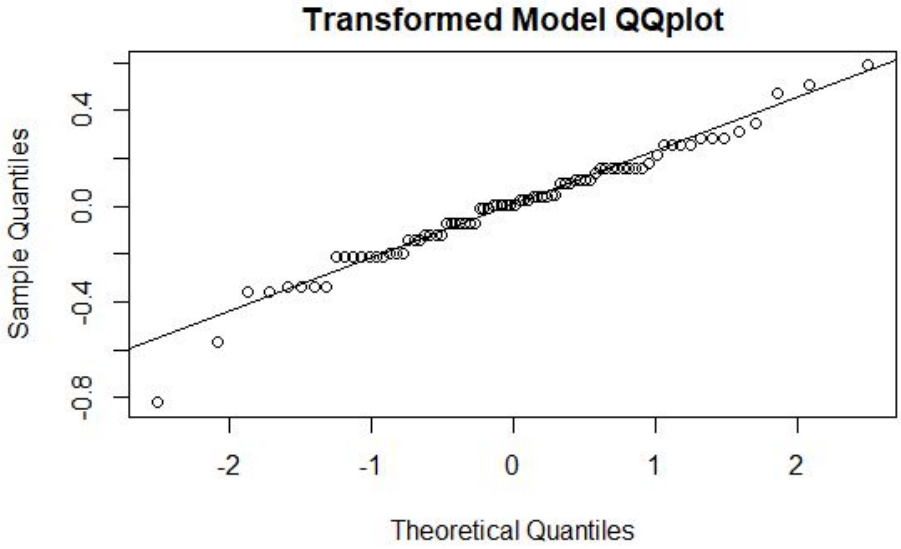
We use all three methods of Box-Cox Transformations, Correlation of normal distribution(QQ plot), the largest p-value for Shapiro-Wilks and Log-likelihood. After doing each transformation, we did the Shapiro-Wilks test and Brown-Forsythe test for transformed datasets and chose the best λ for this equation:

Y_{ij}*= (Y_{ij}^λ-1)/λ for λ≠ 0
Y_{ij}*= ln(Y_{ij}) for λ=0

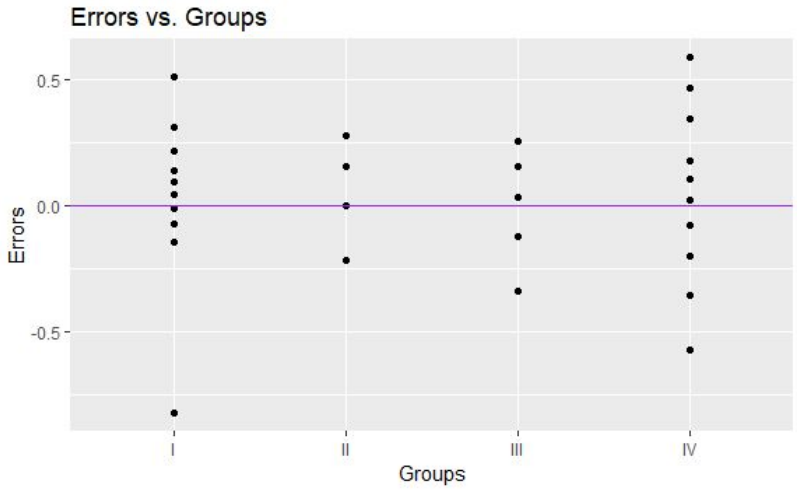
QQ-plot Method
λ = -0.2568
The Shapiro-Wilk Test gives a p-value = 0.1075
The Brown-Forsythe Test p-value = 0.6545
The largest p-value for the Shapiro-Wilks Test Method
λ = -0.2322
The Shapiro-Wilk Test gives a p-value = 0.1090
The Brown-Forsythe Test p-value = 0.6467
The Log-likelihood Method
λ = -0.3964
The Shapiro-Wilk Test gives a p-value = 0.0674
The Brown-Forsythe Test p-value = 0.6597

Choice of Lamda and interpretation of p-values:
We decided to use the Shapiro-Wilks Test Method and choose the Lamda= -0.2322 and our new model is Y_{ij}*=[Y_{ij}^(-0.2322)-1]/(-0.2322).
Since the Shapiro-Wilks p-value is larger than the specific alpha at any significant level(0.1, 0.05, 0.01), we failed to reject our null hypothesis. We can conclude that our transformed data is normally distributed.
Since the Brown-Forsythe p-value is larger than the specific alpha at any significant level(0.1, 0.05, 0.01), we failed to reject our null hypothesis. We can conclude that our transformed data has equal variance.

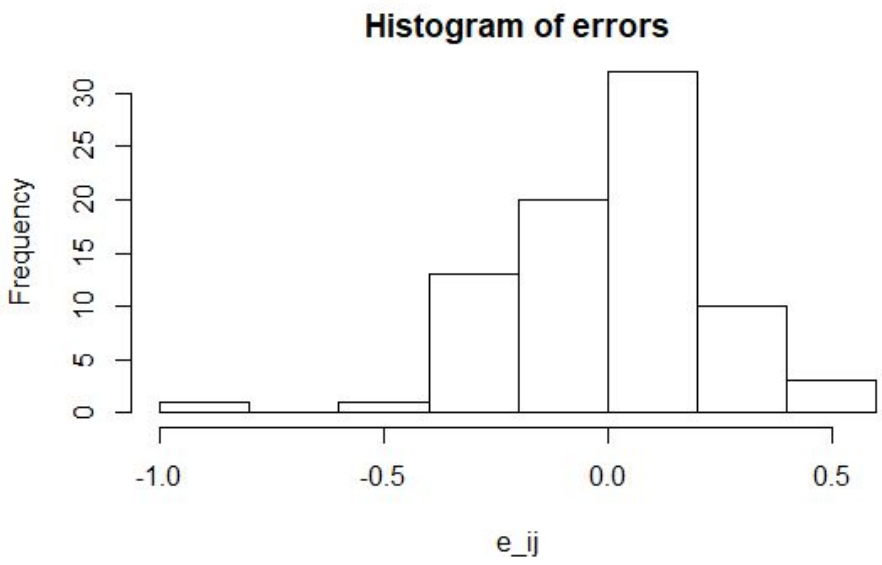
Diagnostics for transformed data without removing outliers:



From the QQplot, we see the transformed data almost distributed close to the line, suggesting the normality of the transformed data.



From this graph, we see that the vertical spread of the transformed data is almost equal, suggesting equal variance.



From this graph, we see that our transformed data are almost a “bell-shaped”, suggesting the normal distribution.

The Shapiro-Wilk p-value = 0.1090

The Brown-Forsythe Test p-value = 0.6467

The Shapiro-Wilk p-value represents that if in reality our errors is normally distributed, we would observe our data or more extreme with the probability of 0.1090. Thus,

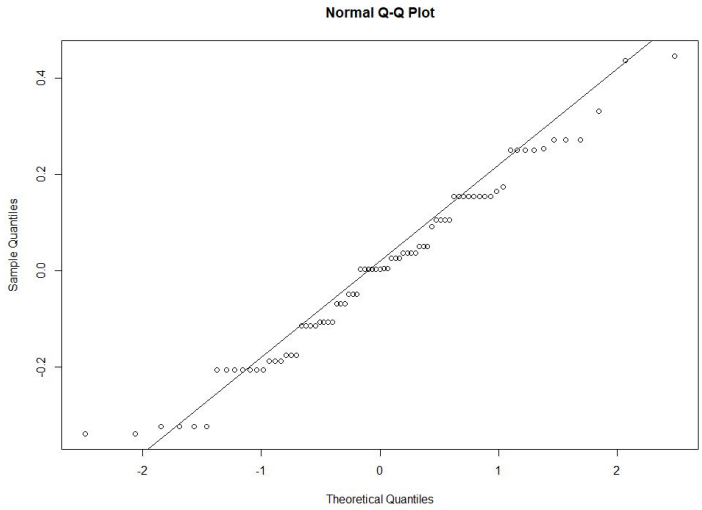
The Brown-Forsythe Test p-value represents that if in reality our errors have equal variance, we would observe our data or more extreme with the probability of 0.6467.

Since the Shapiro-Wilks p-value is larger than the specific alpha at any significant level(0.1, 0.05, 0.01), we failed to reject our null hypothesis. We can conclude that our transformed data is normally distributed.

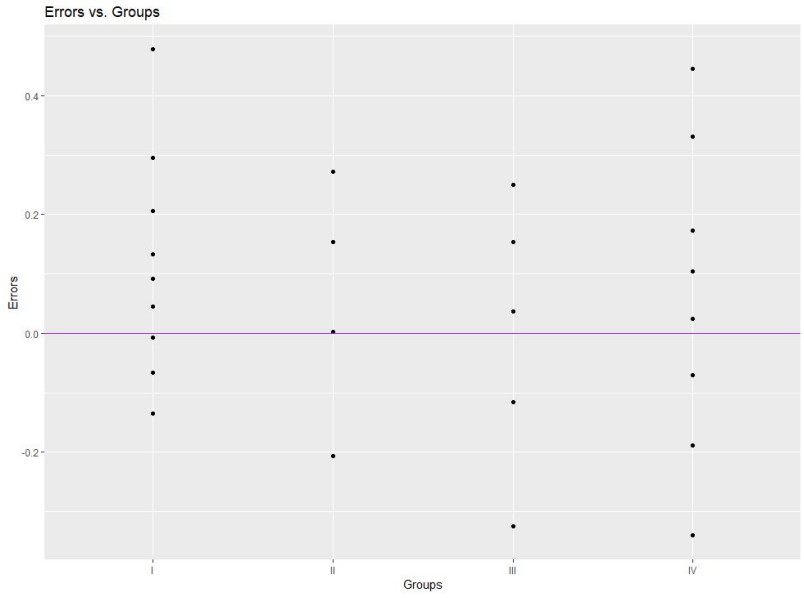
Since the Brown-Forsythe p-value is larger than the specific alpha at any significant level(0.1, 0.05, 0.01), we failed to reject our null hypothesis. We can conclude that our transformed data has equal variance.

Diagnostics for transformed data with removing outliers:

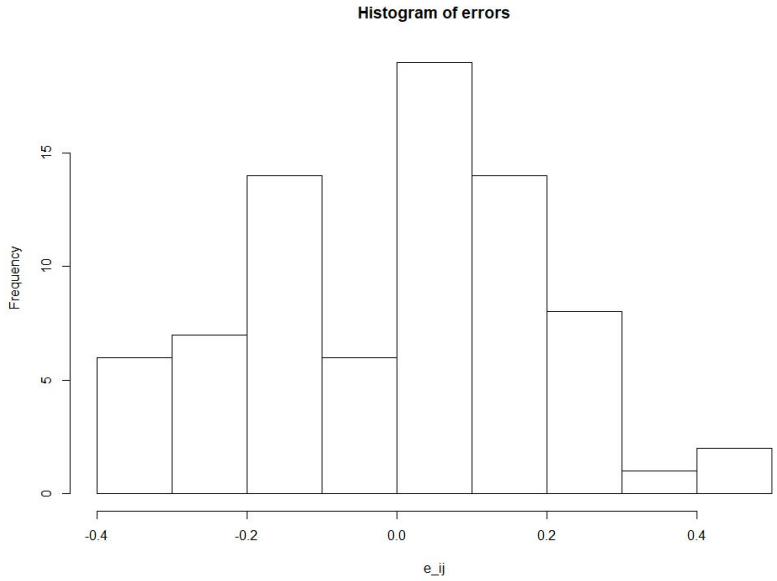
We found three outliers after doing the transformation and removed them from the transformed data.



From the QQplot, we see the transformed data almost distributed close to the line, suggesting the normality of the transformed data.



From this graph, we see that the vertical spread of the transformed data is almost equal, suggesting equal variance.



From this graph, we see that our transformed data are almost a "bell-shaped", suggesting the normal distribution.

Shapiro-Wilk normality test p-value = 0.162
Brown Forsythe p-value =0.5055
The Shapiro-Wilk p-value represents that if in reality our data is normally distributed, we would observe our data or more extreme with the probability of 0.162. Since the Shapiro-Wilks p-value is larger than the specific alpha at any significant level(0.1, 0.05, 0.01), we failed to reject our null hypothesis. We can conclude that our transformed data is normally distributed.

The Brown-Forsythe p-value represents that if in reality our data is normally distributed, we would observe our data or more extreme with the probability of 0.5055. Since the Brown-Forsythe p-value is larger than the specific alpha at any significant level(0.1, 0.05, 0.01), we failed to reject our null hypothesis. We can conclude that our transformed data has equal variance.

Conclusion for Topic One:

The transformation of our data helped our data to become normally distributed and having constant variance, which fits the assumptions of the ANOVA model. ($Y_{ij}= \mu_i+\varepsilon_{ij}$)

However, the downside of our transformed data is that we have to interpret our values and test results in terms of the transformed data(Y_{ij}*=[Y_{ij}^(-0.2322)-1]/(-0.2322)), which may create difficulties of understanding the meaning for some interpretation.

I would suggest for a client who wants to use this data set for ANOVA to first do a Box-cox Transformation using the Shapiro-Wilks Test Method, which gives result $\lambda = -0.2322$. Then our client can either remove the three outliers from the transformed data or not, because both data meet the assumptions of normality and equal variance of errors. Thus, all assumptions of ANOVA can be met for the data.

Report for Topic II Question One

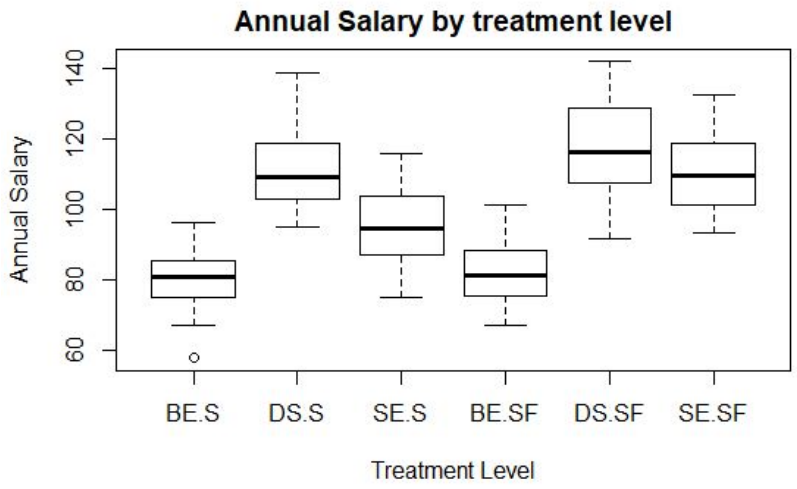
Introduction:

With the development of technology, technology workers have taken a very role in the industry. Many technology companies are desiring to hire many technology workers such as data scientist, software engineering, and bioinformatics engineering to promote their product and increase profit. We are interested in the average salary for technology workers and its relationship with these three professions (data scientist, software engineer, and bioinformatics engineer) and the region they are working in.

Approach:

In this research, we first visually and numerically summarize the data by summary table and boxplot. Then, we use semi-studentized residuals to remove the outliers. To assess the normality, we use QQplot to see if the new model is a good plot and then use the Shapiro-Wilks test, which is non-subjective and gives statistical evidence. Then, we use the Brown-Forsythe test to assess constant variance. We are going to construct interaction and factor effect hypothesis tests to decide which ANOVA model we are going to use. After that, we choose the smallest multiplier to construct four pairwise confidence intervals to determine the relationship between the average salary of the two factors using Tukey Correction, Scheffe Correction, and Bonferroni Correction. We also choose the smallest multiplier to calculate two contrast confidence intervals by using Scheffe Correction and Bonferroni Correction.

Summary:



From this boxplot, we see that there may be an outlier in the BE.S group.

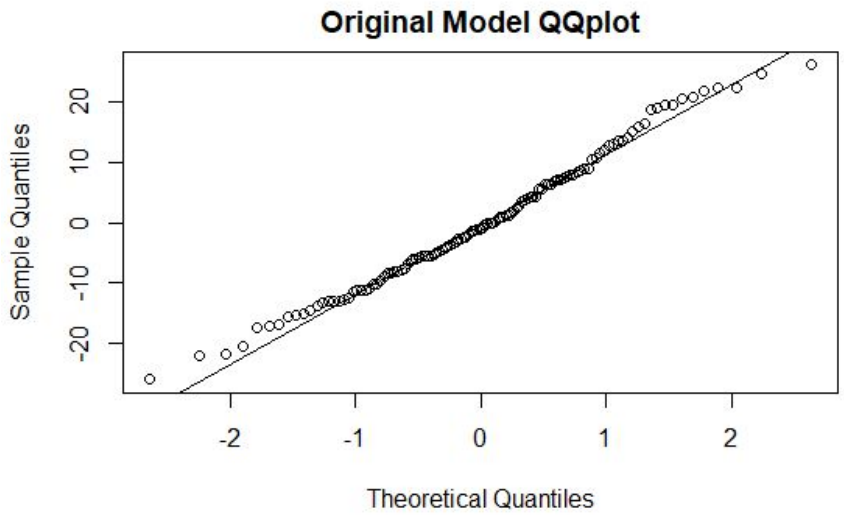
Mean(Sample size)

	BE	DS	SE	Total
SF	82.41914(20)	117.7688(20)	110.26412(20)	103.48403 (60)
S	79.75485(20)	112.5272 (20)	95.54875(20)	95.94358(60)
Total	81.0870(40)	115.1480 (40)	102.9064 (40)	99.71(120)

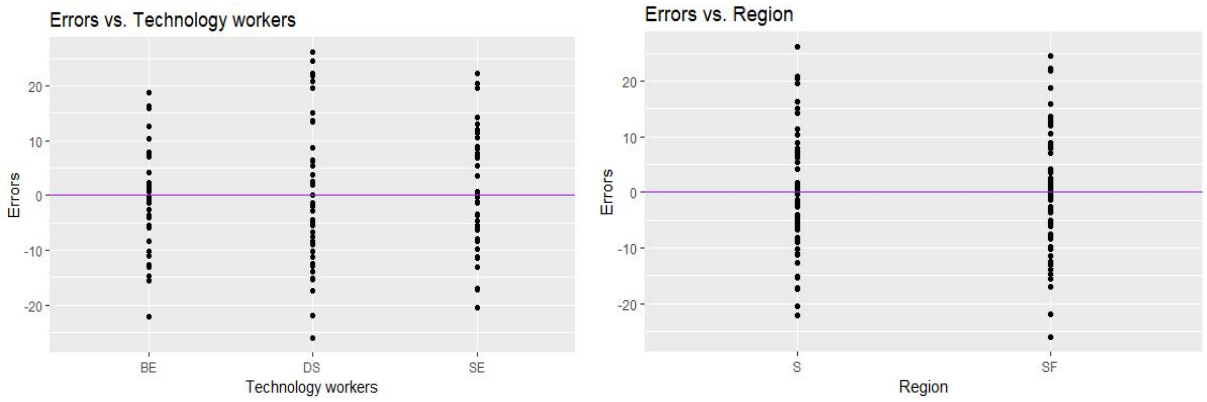
Standard Deviation(Sample Size)

	BE	DS	SE	Total
SF	10.52148(20)	14.28923(20)	10.55171(20)	19.29842 (60)
S	8.786628(20)	12.838566(20)	11.598722(20)	17.41791(60)
Total	9.662515(40)	13.668190(40)	13.240313 (40)	18.6923(120)

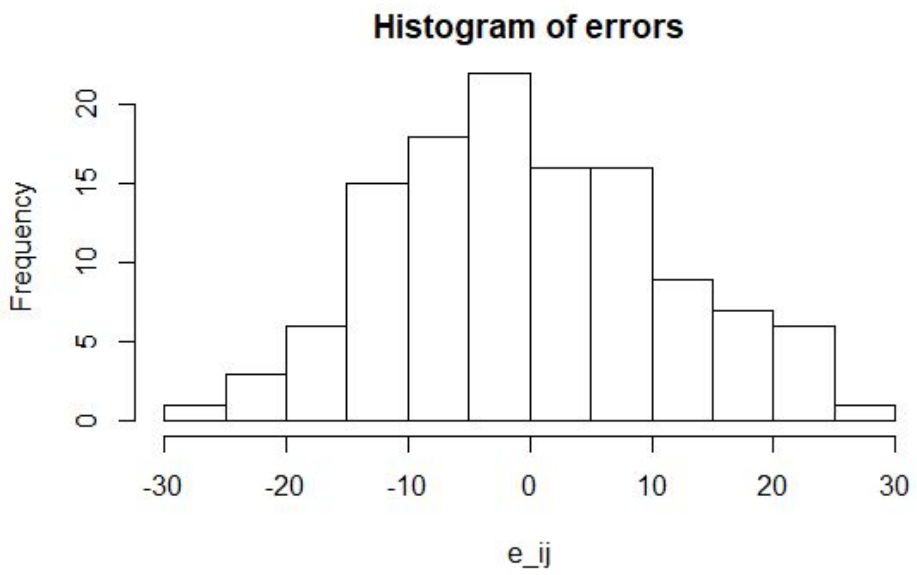
Diagnostics:



From this QQ-plot, we find that the data is almost distributed along the line, suggesting the normality of the errors.



From both of the graphs, we find that the vertical spread is almost equal in each factor groups, which suggesting equal variance by groups.



The histogram is bell-shaped, suggesting a normal distribution of our data.
These are the subjective conclusions from the graphs. We will further use SW test and BF test to examine if the data violates the assumptions of the Two Factors ANOVA model.

Assumptions for Two Factors ANOVA:

1. Random samples from all treatments
2. Factor A groups are independent
3. Factor B groups are independent
4. $\varepsilon_{ijk} \sim N(0,\sigma^2_{\varepsilon})$

We believe all the assumptions except for the fourth that required to perform ANOVA are satisfied, and we use the Shapiro-Wilks Test(SW) and the Brown-Forsythe Test(BF) if the fourth assumption(normality of errors) is violated.

We used semi-studentized residuals to detect outliers(set α=0.01) and did not find any outliers. We used the Shapiro-Wilks test to accessing normal and get p-value = 0.3237, which is greater than α at any significant level, therefore we concluded that the data is normal.
We also used the Brown-Forsythe test to accessing constant variance and get p-value = 0.3048, which is greater than α at any significant level, therefore we concluded that the data has constant variance. With these two results, our data satisfied all four assumptions and can fit into the Two-Factor ANOVA Model. Thus, we continue to test the interaction and factor effects.

Analysis:

Interaction plot:

From this plot, we can see these two lines are not parallel, which suggests an interaction effect. Thus, we further use the hypothesis test to prove this point.

Nulls and Alternatives:

H_0: $(\gamma\delta)_{ij} = 0$ for all i,j

H_a: $(\gamma\delta)_{ij} \neq 0$ for at least one pair of i,j

We construct a hypothesis test to test if there is an interaction effect in our data.

We find the F statistics = 3.0098, and the p-value = 0.05324.

This p-value means that if in reality there is no interaction effect(H_0 is true), we will observe our data or more extreme case with probability 0.05324.

Since our p-value is greater than $\alpha = 0.1$, we reject our H_0 and conclude that there is an interaction effect in this model.

We also calculated $R^2\{AB|A+B\} = 0.0501551$, which suggests that when we add interactions between professions and regions to a model with no interactions, we decrease error by 5.01551%.

Based on all the results from our interaction plot, hypothesis test, and R^2, we decide to use the two-way ANOVA model with an interaction effect:

$Y_{ijk} = \mu_{..} + \gamma_{i+} \delta_j + (\gamma\delta)_{ij+} \varepsilon_{ijk}$

Pairwise Confident intervals:

We find all three pairwise Confident Intervals by Subjects Title (Factor A) using the smallest multiplier(Bonferroni Mutiplier= 2.998, g=3).

	lower bound	upper bound
99% Bonferroni CI for $\mu_{1.}$-$\mu_{2.}$	-41.81526	-26.30673
99% Bonferroni CI for $\mu_{1.}$-$\mu_{3.}$	-29.57371	-14.06517
99% Bonferroni CI for $\mu_{2.}$-$\mu_{3.}$	4.487285	19.995824

Interpretation for this table:

99% Bonferroni CI for $\mu_{1.}$-$\mu_{2.}$:

We are overall 99% Confident that the true average annual salary of Bioinformatics Engineers is less than that of Data Scientists by between 26.30673 and 41.81526.

99% Bonferroni CI for $\mu_{1.}$-$\mu_{3.}$:

We are overall 99% Confident that the true average annual salary of Bioinformatics Engineers is less than that of Software Engineers by between 14.06517 and 29.57371.

99% Bonferroni CI for $\mu_{2.}$-$\mu_{3.}$:

We are overall 99% Confident that the true average annual salary of Data Scientist is more than that of Software Engineers by between 4.487285 and 19.995824.

And we find the remaining one pairwise Confident Interval by Region(Factor B) using the smallest multiplier(Bonferroni Multiplier = 2.620, g=1)

	lower bound	upper bound
99% Bonferroni CI for $\mu_{.1}$-$\mu_{.2}$	-13.073503	-2.007395

Interpretation for this table:

99% Bonferroni CI for $\mu_{.1}$-$\mu_{.2}$:

We are overall 99% Confident that the true average annual salary of technology workers in Seattle is less than that of technology workers in San Francisco by between 2.007395 and 13.073503.

Contrast Confident intervals:

	lower bound	upper bound
99% Bonferroni CI for μ_{21}-(½)μ_{11}-(½)μ_{31}	15.80600	33.94470
99% Bonferroni CI for μ_{22}-(½)μ_{12}-(½)μ_{32}	12.35785	30.49655

Interpretation for this table:

99% Bonferroni CI for μ_{21}-(½)μ_{11}-(½)$\mu_{31:}$

We are overall 99% Confident that the true average annual salary of Data scientists in Seattle is more than the mean of the true average annual salary of Bioinformatics Engineers and Software Engineers in Seattle by between 15.80600 and 33.94470.

99% Bonferroni CI for μ_{22}-(½)μ_{12}-(½)$\mu_{32:}$

We are overall 99% Confident that the true average annual salary of Data scientists in San Francisco is more than the mean of the true average annual salary of Bioinformatics Engineers and Software Engineers in San Francisco by between 12.35785 and 30.49655.

Conclusion:

We conclude that the data is normally distributed at any significant level(0.01, 0.05, 0.1), and the data has constant variance at any significance level(0.01, 0.05, 0.1). Thus, our data fit into the two-way ANOVA Model.

We then find this data has an interaction effect at significant level of $\alpha = 0.1$; thus, we use the two-way ANOVA Model with interaction term: $Yijk = \mu..+ \gamma i + \delta j + (\gamma\delta)ij + \varepsilon ijk$.

From the pairwise confidence intervals, we are 99% confident that the Data Scientist has the highest annual salary, followed by Software Engineers. And Bioinformatics Engineers have the lowest annual salary. Technology workers receive a higher annual salary working in San Francisco than in Seattle.

From our contrast confident intervals, we infer that data scientists have a higher annual salary than the average of the other two types of technology workers in both regions. From the differences in the bounds of the confidence interval, we can also infer an interaction between the type of technology workers and regions.

R Appendix

```
Helicopter <- read.csv("C:/Users/838748635/Downloads/Helicopter.csv")
boxplot(Helicopter$Count~Helicopter$Shift,main="Count by Shift", ylab="Count")
group.means =  by(Helicopter$Count,Helicopter$Shift,mean)
group.sds = by(Helicopter$Count,Helicopter$Shift,sd)
group.nis = by(Helicopter$Count,Helicopter$Shift,length)
the.summary = rbind(group.means,group.sds,group.nis)
the.summary = round(the.summary,digits = 4)
colnames(the.summary) = names(group.means)
rownames(the.summary) = c("Means","Std. Dev","Sample Size")
the.summary
the.model = lm(Count ~ Shift,data = Helicopter)
Helicopter$ei = the.model$residuals
qqnorm(the.model$residuals,main="Original Model QQplot")
qqline(the.model$residuals)
nt = nrow(Helicopter) #Calculates the total sample size
a = length(unique(Helicopter$Shift)) #Calculates the value of a
SSE = sum(Helicopter$ei^2) #Sums and squares the errors (finds SSE)
MSE = SSE/(nt-a) #Finds MSE
eij.star = the.model$residuals/sqrt(MSE)
ei = the.model$residuals
the.SWtest = shapiro.test(ei)
the.SWtest
library(car)
the.BFtest = leveneTest(ei~ Shift, data= Helicopter, center=median)
p.val = the.BFtest[[3]][1]
p.val
alpha = 0.01
t.cutoff= qt(1-alpha, nt-a)
CO.eij = which(abs(eij.star) > t.cutoff)
CO.eij
outliers = CO.eij
new.data = Helicopter[-outliers,]
new.model = lm(Count ~ Shift, data = new.data)
qqnorm(new.model$residuals,main="Model Without Outliers QQplot")
qqline(new.model$residuals)
ei = new.model$residuals
the.SWtest = shapiro.test(ei)
the.SWtest
library(car)
the.BFtest = leveneTest(ei~ Shift, data= new.data, center=median)
p.val = the.BFtest[[3]][1]
p.val
the.model = lm(Count ~ Shift,data = Helicopter)
Helicopter$ei = the.model$residuals
qqnorm(the.model$residuals,main="Model QQplot")
qqline(the.model$residuals)
library(ggplot2)
qplot(Shift, ei, data = Helicopter) +  ggtitle("Errors vs. Groups") + xlab("Groups") + ylab("Errors") +
geom_hline(yintercept = 0,col = "purple")
hist(Helicopter$ei, main = "Histogram of errors",xlab = "e_ij",ylab = "Frequency")
```

```
qplot(Shift, ei, data = new.data) +  ggtitle("Errors vs. Groups") + xlab("Groups") + ylab("Errors") +
geom_hline(yintercept = 0,col = "purple")
hist(new.data$ei, main = "Histogram of errors",xlab = "e_ij",ylab = "Frequency")
L1 =boxcox(the.model ,objective.name = "PPCC",optimize = TRUE)$lambda
L2 = boxcox(the.model ,objective.name = "Shapiro-Wilk",optimize = TRUE)$lambda
L3 = boxcox(Helicopter$Count,objective.name = "Log-Likelihood",optimize = TRUE)$lambda
YT = (Helicopter$Count^(L1)-1)/L1
t.data = data.frame(Count = YT, Shift = the.data$Shift)
t.model = lm(Count ~ Shift,data = t.data)
qqnorm(t.model$residuals)
qqline(t.model$residuals)
Helicopter$ei = t.model$residuals
library(ggplot2)
t.data$ei = t.model$residuals
qplot(Shift, ei, data = t.data) +  ggtitle("Errors vs. Groups") + xlab("Groups") + ylab("Errors") +
geom_hline(yintercept = 0,col = "purple")
hist(t.data$ei, main = "Histogram of errors",xlab = "e_ij",ylab = "Frequency")
ei = t.model$residuals
the.SWtest = shapiro.test(ei)
the.SWtest
library(car)
the.BFtest = leveneTest(ei~ Shift, data=t.data, center=median)
p.val = the.BFtest[[3]][1]
p.val
library(EnvStats)
L1=boxcox(new.model ,objective.name = "PPCC",optimize = TRUE)$lambda
L2=boxcox(new.model ,objective.name = "Shapiro-Wilk",optimize = TRUE)$lambda
L3=boxcox(new.data$Count,objective.name = "Log-Likelihood",optimize = TRUE)$lambda
L1
L2
L3
YT = (new.data$Count^(L2)-1)/L2
t.data = data.frame(Count = YT, Shift = new.data$Shift)
t.model = lm(Count ~ Shift,data = t.data)
t.data$ei = t.model$residuals
qqnorm(t.model$residuals,main="Transformed Model QQplot")
qqline(t.model$residuals)
library(ggplot2)
qplot(Shift, ei, data = t.data) +  ggtitle("Errors vs. Groups") + xlab("Groups") + ylab("Errors") +
geom_hline(yintercept = 0,col = "purple")
hist(t.data$ei, main = "Histogram of errors",xlab = "e_ij",ylab = "Frequency")
ei = t.model$residuals
the.SWtest = shapiro.test(ei)
the.SWtest
the.BFtest = leveneTest(ei~ Shift, data= t.data, center=median)
p.val = the.BFtest[[3]][1]
P.val

nt = nrow(t.data) #Calculates the total sample size
> a = length(unique(t.data$Shift)) #Calculates the value of a
> SSE = sum(t.data$ei^2) #Sums and squares the errors (finds SSE)
> MSE = SSE/(nt-a) #Finds MSE
> eij.star = t.model$residuals/sqrt(MSE)
```

```
> alpha = 0.01
> t.cutoff= qt(1-alpha, nt-a)
> CO.eij = which(abs(eij.star) > t.cutoff)
> CO.eij
 1 67 72
 1 67 72
> outliers = CO.eij
> newt.data = t.data[-outliers,]
> newt.model = lm(Count ~ Shift,data = newt.data)
> qqnorm(newt.model$residuals)
> qqline(newt.model$residuals)
> qplot(Shift, ei, data = newt.data) +  ggtitle("Errors vs. Groups") + xlab("Groups") + ylab("Errors") +
geom_hline(yintercept = 0,col = "purple")
> ei = newt.model$residuals
newt.data$ei = newt.model$residuals
> hist(newt.data$ei, main = "Histogram of errors",xlab = "e_ij",ylab = "Frequency")
> ei = newt.model$residuals
> the.SWtest = shapiro.test(ei)
> the.SWtest
 the.BFtest = leveneTest(ei~ Shift, data=newt.data, center=median)
> p.val = the.BFtest[[3]][1]
> p.val
Salary <- read.csv("C:/Users/838748635/Downloads/Salary.csv")
the.data = Salary
find.means = function(the.data,fun.name = mean){
  a = length(unique(the.data[,2]))
  b = length(unique(the.data[,3]))
  means.A = by(the.data[,1], the.data[,2], fun.name)
  means.B = by(the.data[,1],the.data[,3],fun.name)
  means.AB = by(the.data[,1],list(the.data[,2],the.data[,3]),fun.name)
  MAB = matrix(means.AB,nrow = b, ncol = a, byrow = TRUE)
  colnames(MAB) = names(means.A)
  rownames(MAB) = names(means.B)
  MA = as.numeric(means.A)
  names(MA) = names(means.A)
  MB = as.numeric(means.B)
  names(MB) = names(means.B)
  results = list(A = MA, B = MB, AB = MAB)
  return(results)
}
boxplot(Annual~Prof*Region,data = Salary,main="Annual Salary by treatment level", ylab="Annual Salary",
xlab="Treatment Level")
the.means =find.means(the.data,mean)
the.sizes=find.means(the.data,length)
the.sds=find.means(the.data,sd)
the.means
the.sds
the.sizes
sd(Salary$Annual)
the.data = Salary
nt = nrow(the.data)
a = length(unique(the.data[,2]))
b = length(unique(the.data[,3]))
names(the.data) = c("Y","A","B")
```

```
the.model = lm(Annual ~ Prof*Region,data = Salary)
Salary$ei = the.model$residuals
qqnorm(the.model$residuals,main="Original Model QQplot")
qqline(the.model$residuals)
SSE = sum(Salary$ei^2) #Sums and squares the errors (finds SSE)
MSE = SSE/(nt-a) #Finds MSE
eij.star = the.model$residuals/sqrt(MSE)
ei = the.model$residuals
the.SWtest = shapiro.test(ei)
the.SWtest
library(car)
the.BFtest = leveneTest(ei~ Prof*Region, data= Salary, center=median)
p.val = the.BFtest[[3]][1]
p.val
alpha = 0.01
t.cutoff= qt(1-alpha, nt-a)
CO.eij = which(abs(eij.star) > t.cutoff)
CO.eij
outliers = CO.eij
library(ggplot2)
qplot(Region, ei, data = Salary) +  ggtitle("Errors vs. Region") + xlab("Region") + ylab("Errors") +
geom_hline(yintercept = 0,col = "purple")
hist(Salary$ei, main = "Histogram of errors",xlab = "e_ij",ylab = "Frequency")
interaction.plot(Salary$Prof, Salary$Region, Salary$Annual)
AB = lm(Y ~ A*B,the.data)
A.B = lm(Y ~ A + B,the.data)
A = lm(Y ~ A,the.data)
B = lm(Y ~ B,the.data)
N = lm(Y ~ 1, the.data)
all.models = list(AB,A.B,A,B,N)
SSE = t(as.matrix(sapply(all.models,function(M) sum(M$residuals^2))))
colnames(SSE) = c("AB","(A+B)","A","B","Empty/Null")
rownames(SSE) = "SSE"
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)
}
  Partial.R2(A.B,AB)
the.data = Salary
a = length(unique(the.data[,2]))
b = length(unique(the.data[,3]))
names(the.data) = c("Y","A","B")
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)
}
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)
 }
}
the.means
all.mult = find.mult(alpha  = 0.01, a = 3, b = 2, dfSSE = 120 - 2*3, g = 2, group = "AB")
Bon = all.mult[1]
Tuk = all.mult[2]
Sch = all.mult[3]
Bon
the.means = find.means(the.data)
the.model = lm(Y ~ A*B, data = the.data)
SSE = sum(the.model$residuals^2)
MSE = SSE/(nt-a*b)
Bon = find.mult(alpha  = 0.01, a = 2, b = 3, dfSSE = 120 - 3*2, g = 3, group = "B")[1]
A.cs.1 = c(1,-1,0)
A.cs.2 = c(1,0,-1)
A.cs.3 = c(0,1,-1)
scary.CI(the.data,MSE,equal.weights = TRUE,Bon,"B",A.cs.1)
scary.CI(the.data,MSE,equal.weights = TRUE,Bon,"B",A.cs.2)
scary.CI(the.data,MSE,equal.weights = TRUE,Bon,"B",A.cs.3)
Bon = find.mult(alpha  = 0.01, a = 3, b = 2, dfSSE = 120 - 3*2, g = 1, group = "B")[1]
B.cs.1 = c(1,-1)
scary.CI(the.data,MSE,equal.weights = FALSE,Bon,"B",B.cs.1)
AB.cs = matrix(0,nrow = a, ncol = b)
AB.cs
the.means$AB
Bon = find.mult(alpha  = 0.01, a = 3, b = 2, dfSSE = 120 - 3*2, g = 2, group = "AB")[1]
AB.cs[2,1] = 1
AB.cs[1,1] = -0.5
AB.cs[3,1] = -0.5
scary.CI(the.data,MSE,equal.weights = TRUE,Bon,"AB",AB.cs)
AB.cs2 = matrix(0,nrow = a, ncol = b)
AB.cs2[1,2] = -0.5
AB.cs2[2,2] = 1
AB.cs2[3,2] = -0.5
scary.CI(the.data,MSE,equal.weights = TRUE,Bon,"AB",AB.cs2)
```