midterm_2

Alexa Aguirre, Selam Berekat

3/7/2021

Topic 1: Transformation of Variables

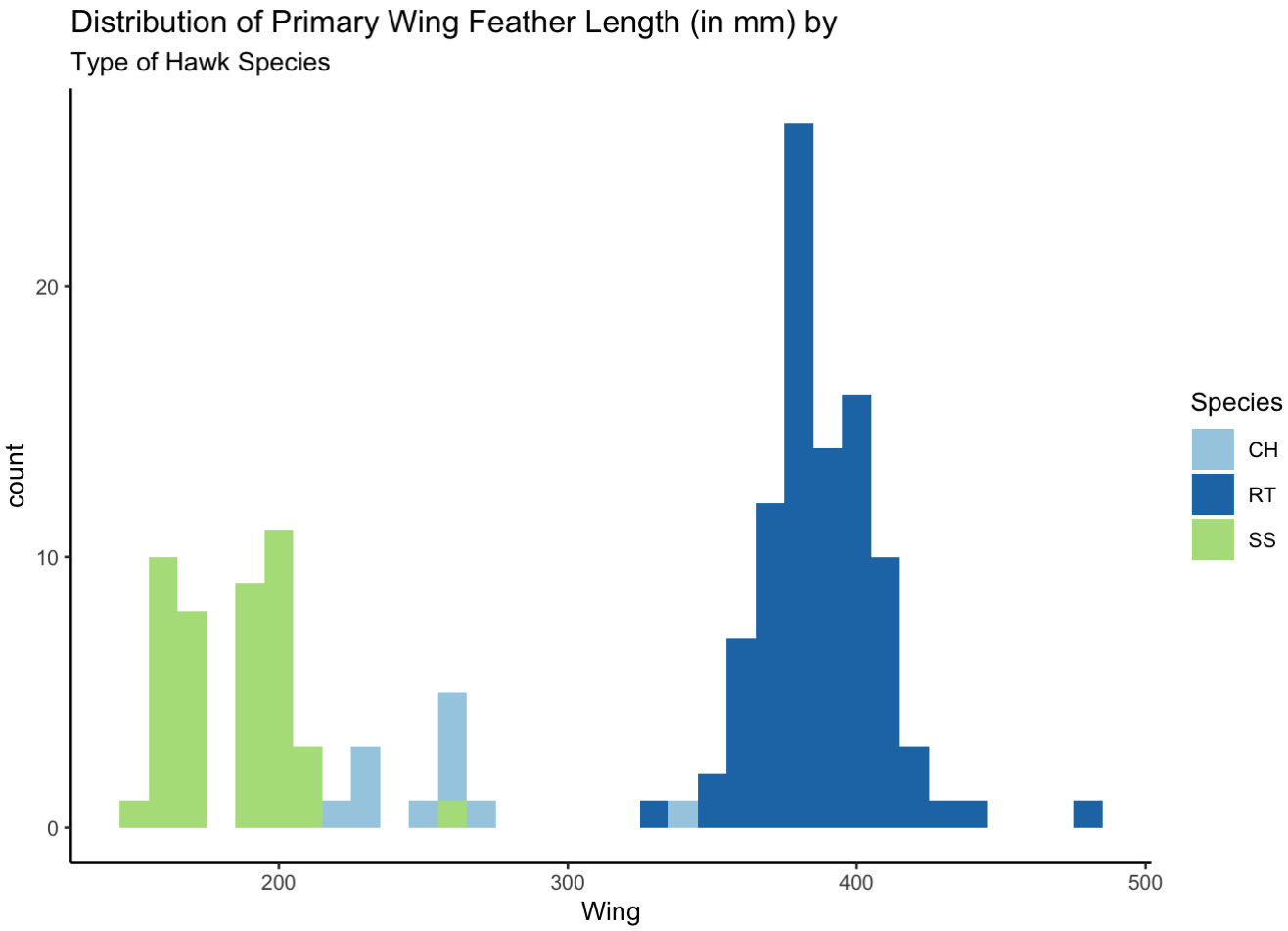
a. Introduction: We will be working on a randomly sampled dataset “Hawks” that consist of 150 observation for 1 numeric column/variable “Wing” and 1 categorical column/variable “Species”. We will be checking if we need to transform the data in order to check if it meets the Single Factor ANOVA assumptions.

b. 1. Plot of original data

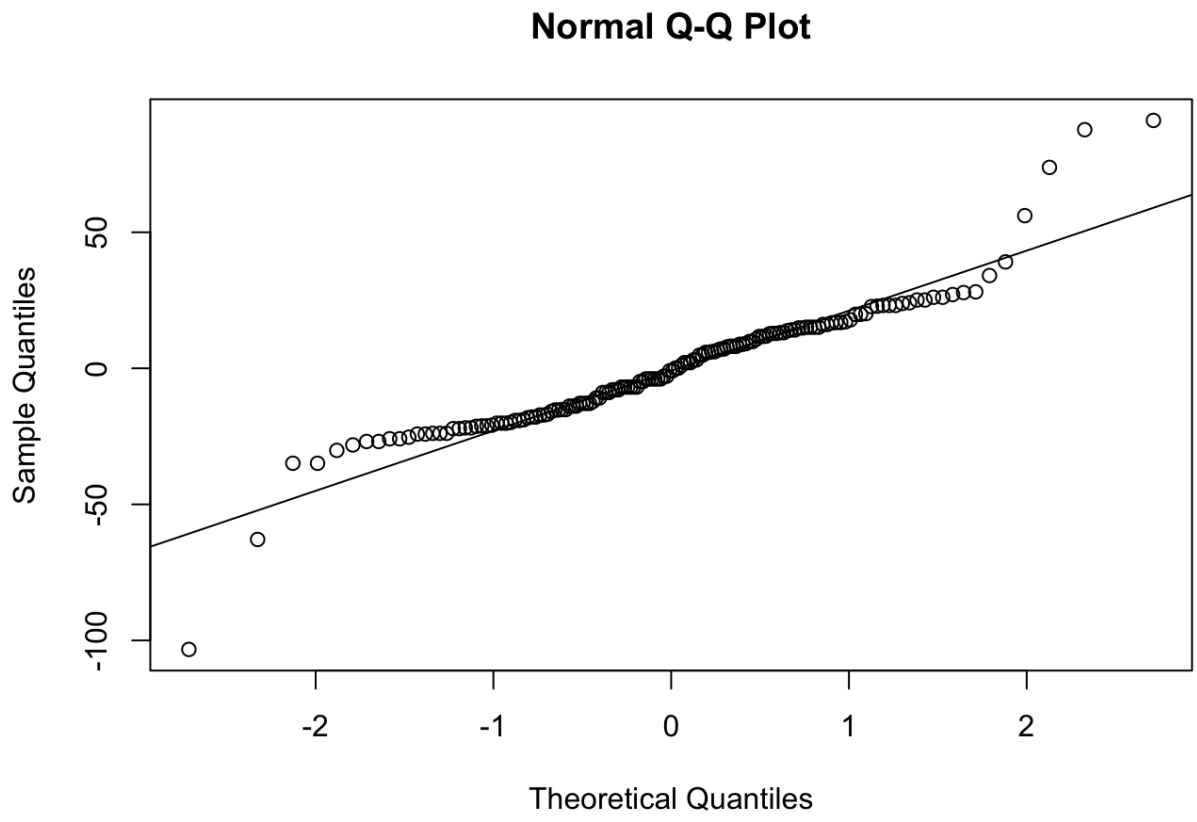
```
##
## Attaching package: 'EnvStats'
```

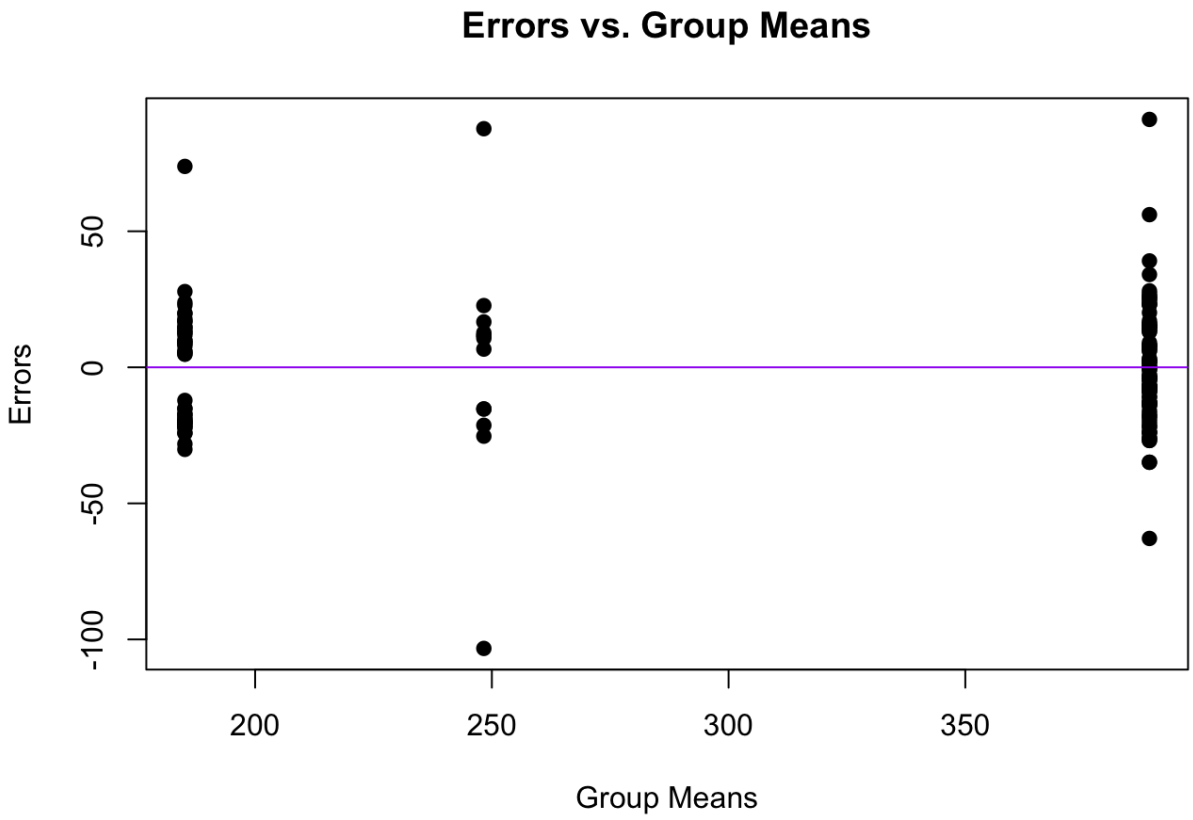
```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

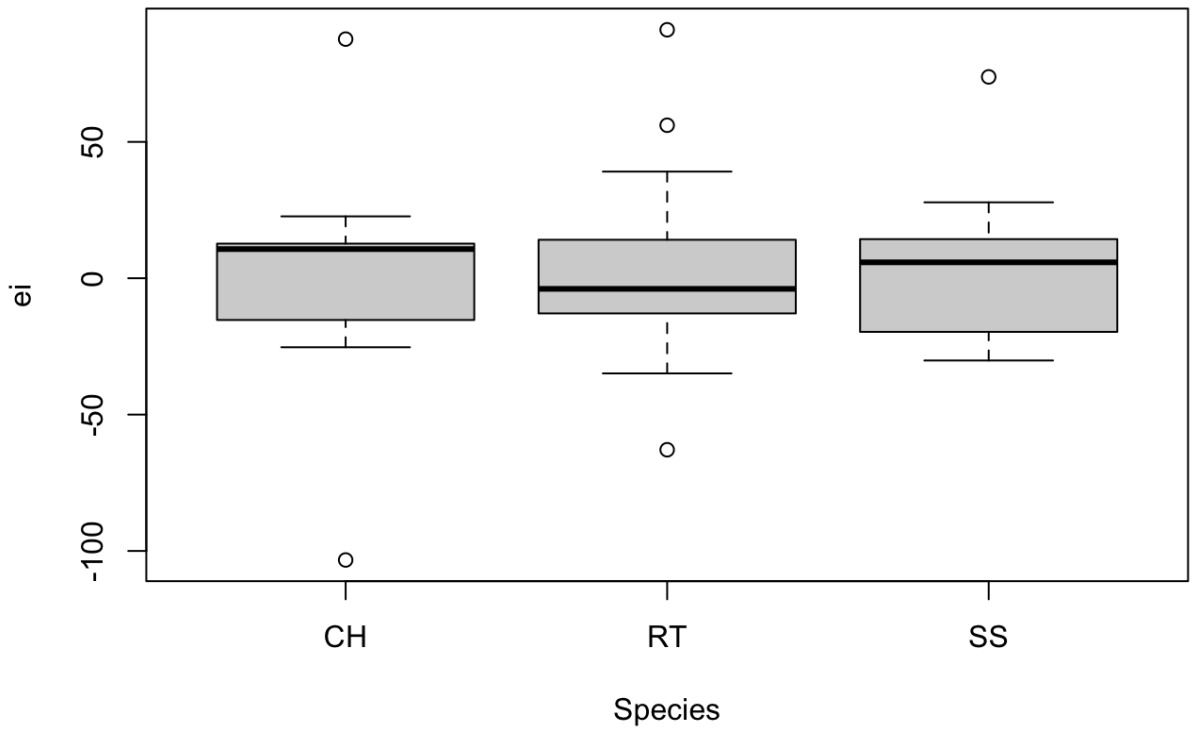
```
## The following object is masked from 'package:base':
##
##     print.default
```



b. 2. Diagnostic plots/tests of original data.




Errors vs. Group Means
Errors
50
0
-50
-100
200
250
300
350
Group Means


ei
50
0
-50
-100
CH
RT
SS
Species

```
##
##  Shapiro-Wilk normality test
##
## data:  ei
## W = 0.91732, p-value = 1.431e-07
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:EnvStats':
##
##     qqPlot
```

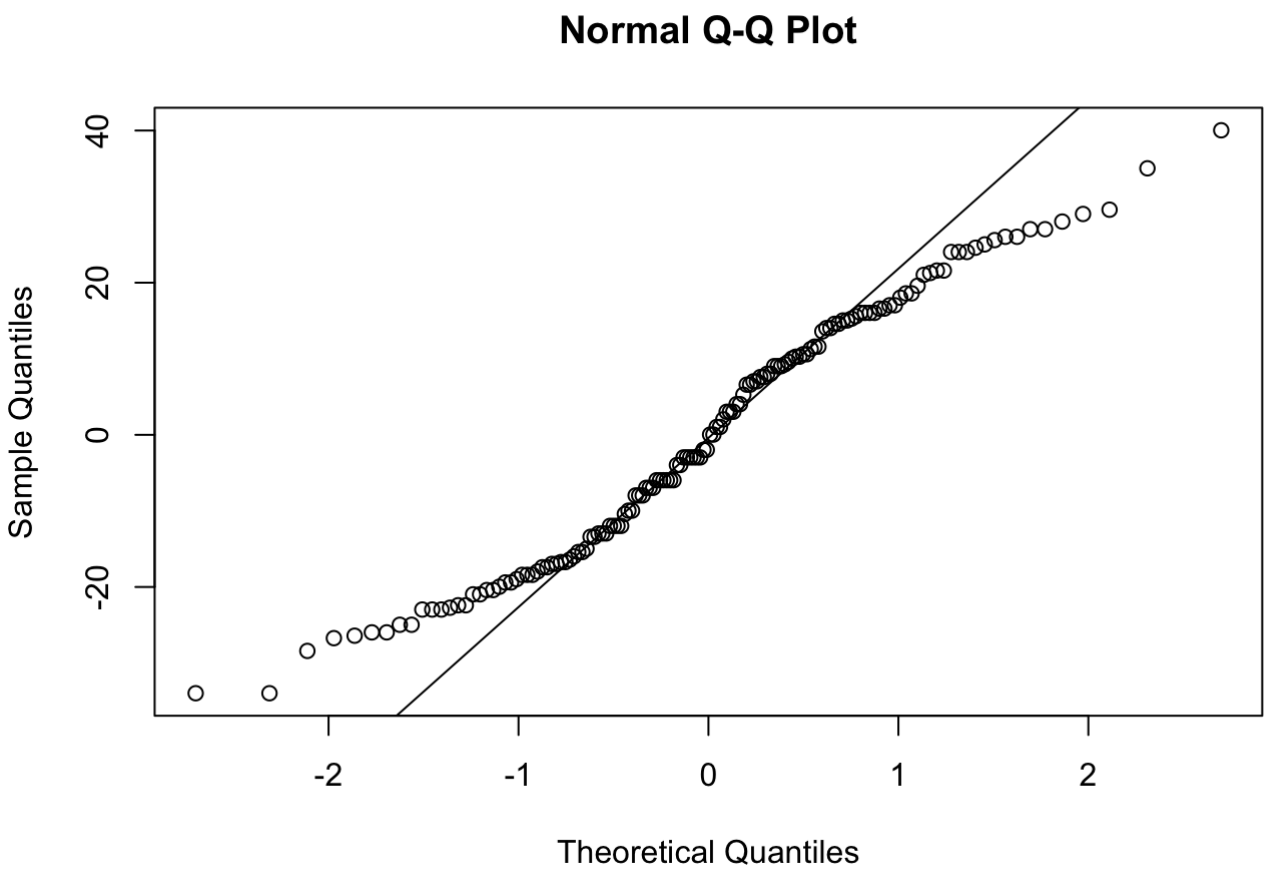
```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   2  2.3481 0.09913 .
##       147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

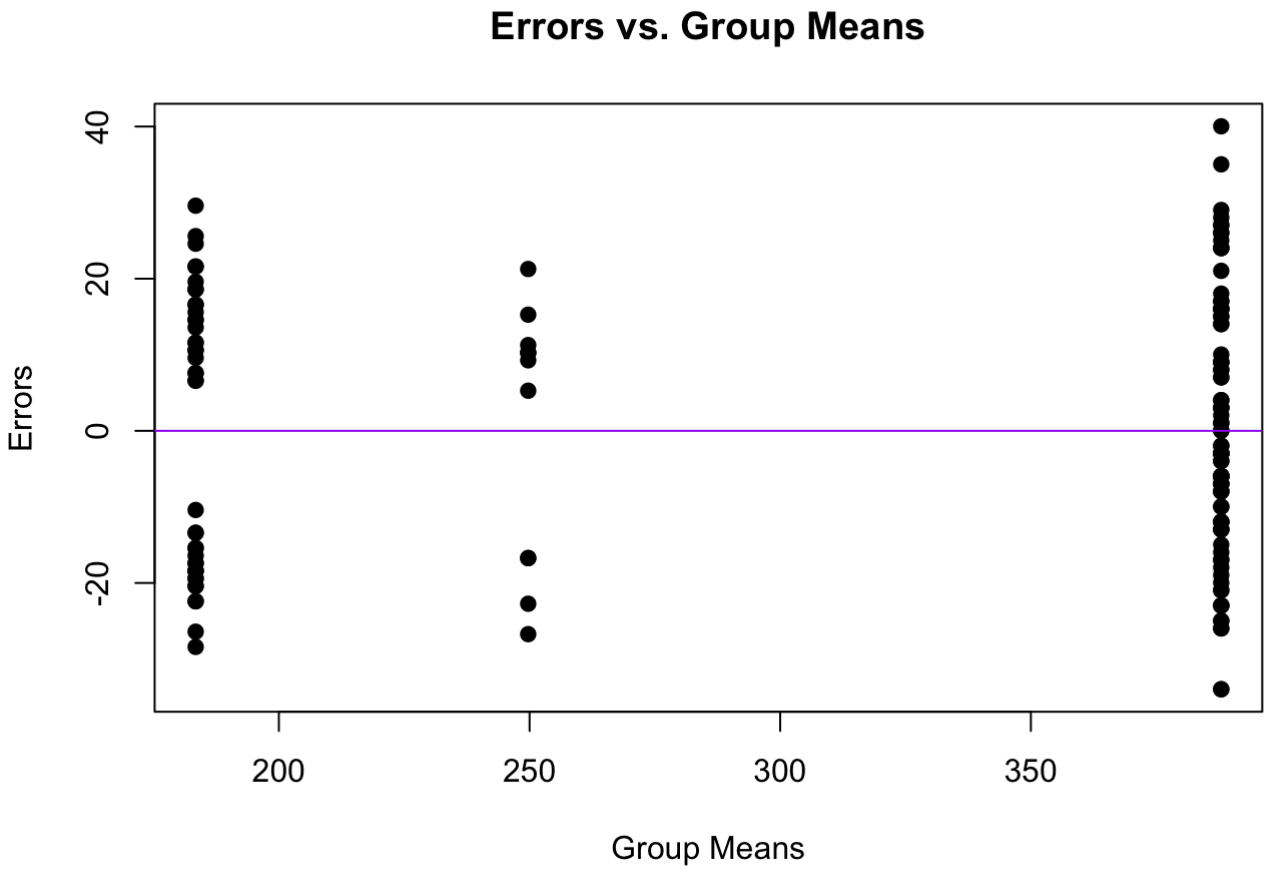
b. 3. Report the model fit of original data.

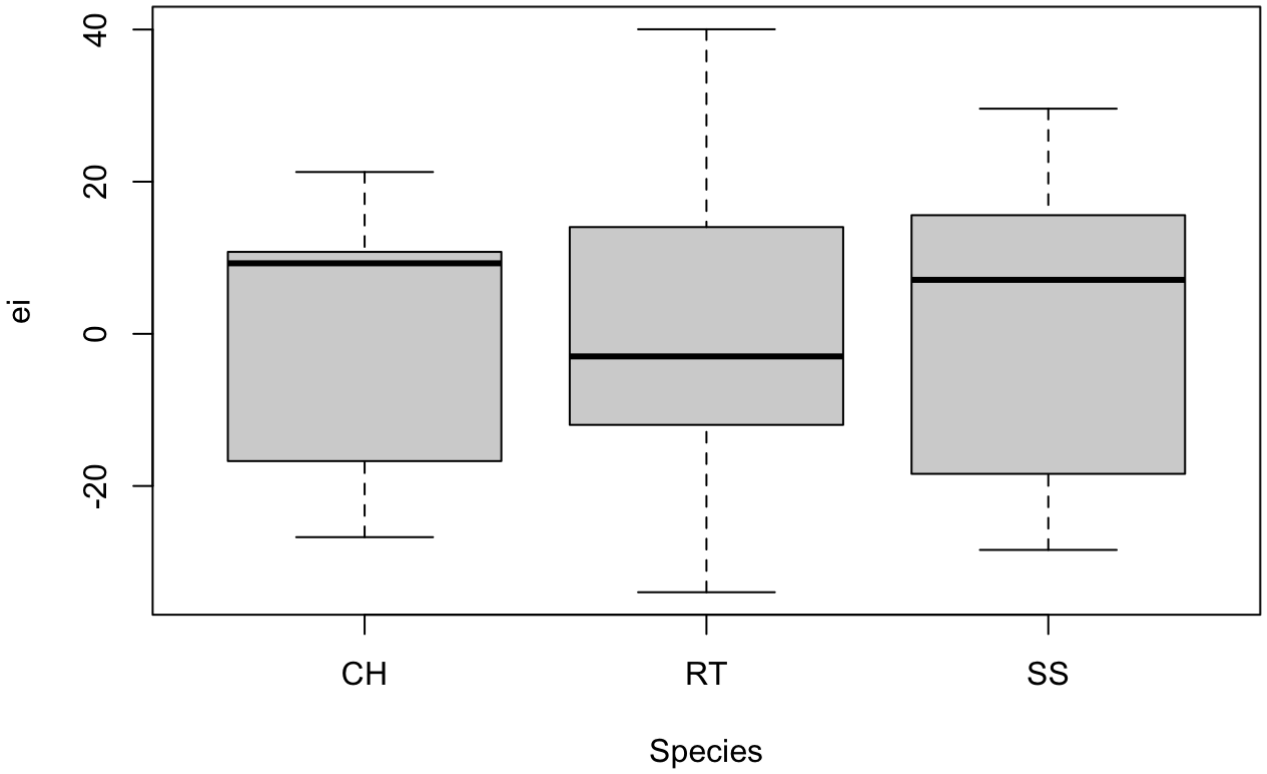
```
## [1] "We first plot the original dataset to study and apply diagnostic to ensure the a
ssumptions of Single Factor ANOVA holds. Our SFA Assumptions are: (1) Yij are randomly s
ampled and independent, (2) The (i) group are independent (i= 1,..,a), and (3) Epsilon(i
j) ~ N(0, sigma^2 epsilon) and errors are independent and normally distributed with mean
0 and constant variance. Our dignostic plot of normality (QQ plot) and plots of equal va
riance (both Residual and Box plot) indicate that SFA assumption appear to be violated b
y the severe departure from ideal line in the QQ plot. Also since plot can be subjective
and can be hard to interpret, we conducted diagnostic test to check the p-value for the
hypothesis test of Shapiro-Wilks is: 1.43079866000663e-07 and the p-value for the hypoth
esis test of Brown-Forsythe is: 0.0991296806779467 . This indicates that with significan
ce level alpha = 0.01, it appears that our data is not normally distributed however, our
group variances are equal."
```

c. 1. Consider removing Outliers.

```
##   7  58  68 113 122 124
##   7  58  68 113 122 124
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
40
20
0
-20
-2
-1
0
1
2


Errors vs. Group Means
Errors
Group Means
40
20
0
-20
200
250
300
350



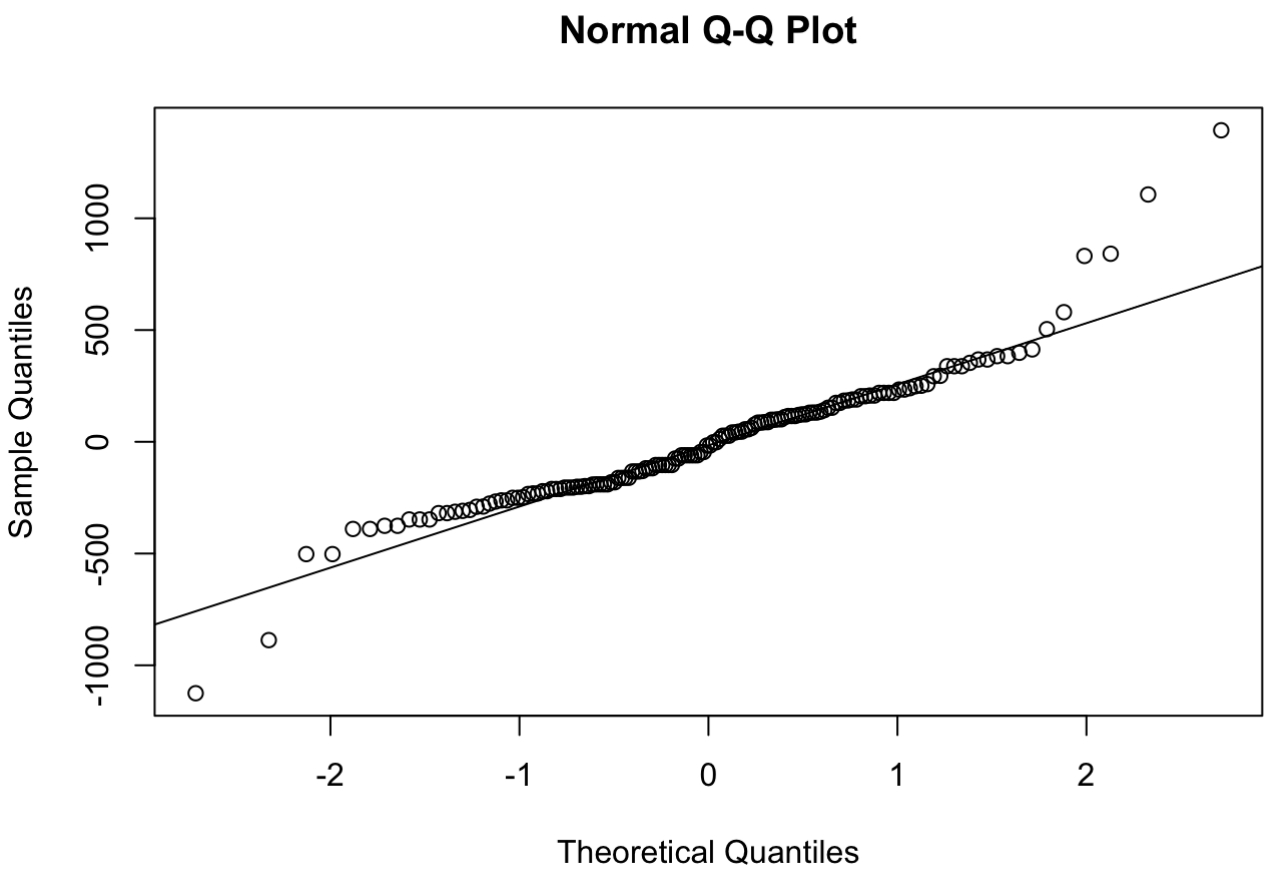

```
## [1] "These are our outliers and we were able to remove the 6 outliers from 150 observ
ations, that equals 4 % of the data. We applied studentized residual method to compare t
he outliers with the cutoff percentiles. According to our plots, removing the outliers h
elped in vertical spread being closer to constant along with the QQ plot have points tha
t are more close to the fitted line. Also, the p-value for the hypothesis test of Shapir
o-Wilks is:  0.00343501102551551 and the p-value for the hypothesis test of Brown-Forsyt
he is: 0.349729032095766 Because the p-value for the Shapiro Wilks test is less than alp
ha (0.01) , the values eij are still non-normally distributed after removing the outlier
s and the group variances are equal since the p-value of the Brown-Forsythe test is grea
ter than alpha. The variance and normality have both become more significant since the t
est produces higher p-values after removing the outliers."
```

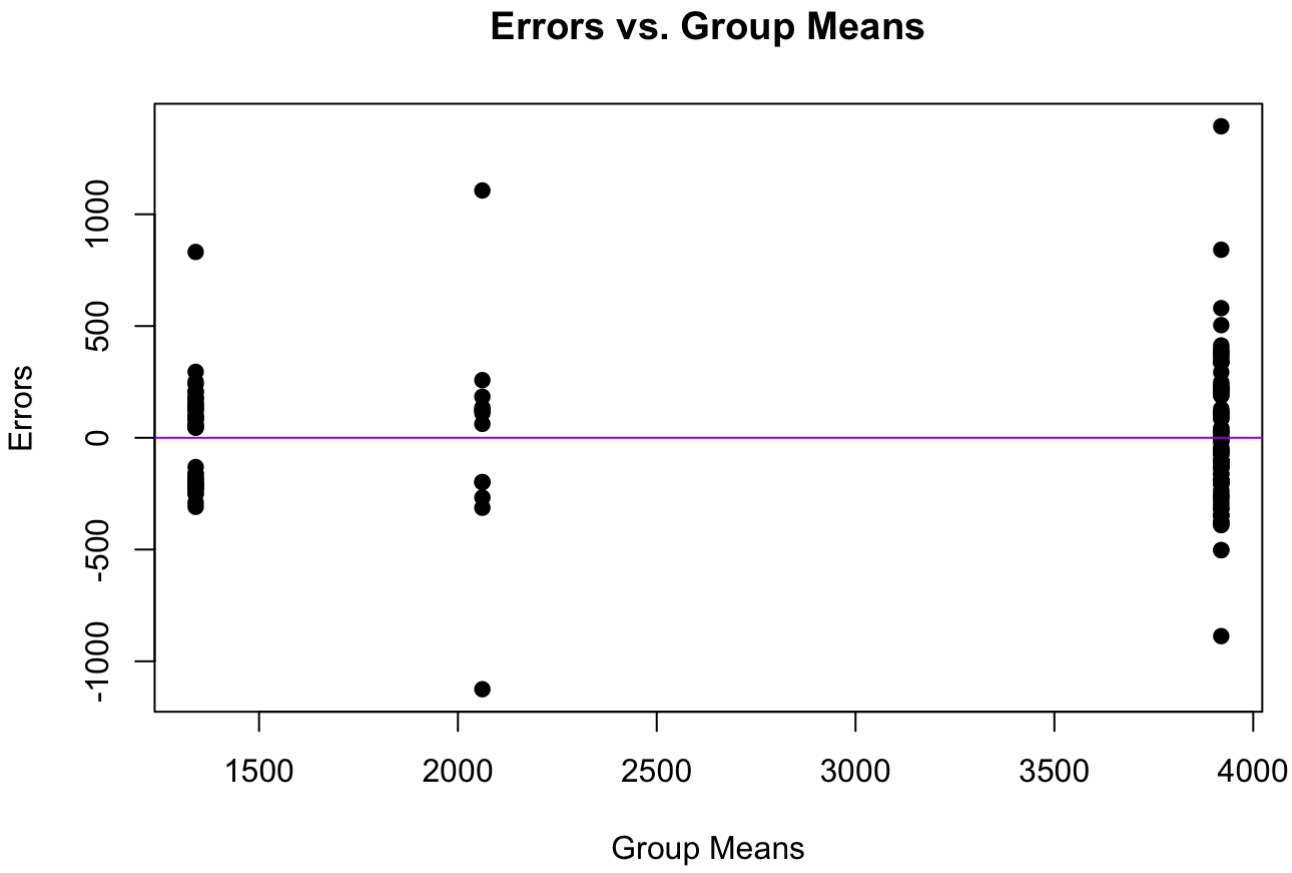
c. 2. Consider transforming Y.

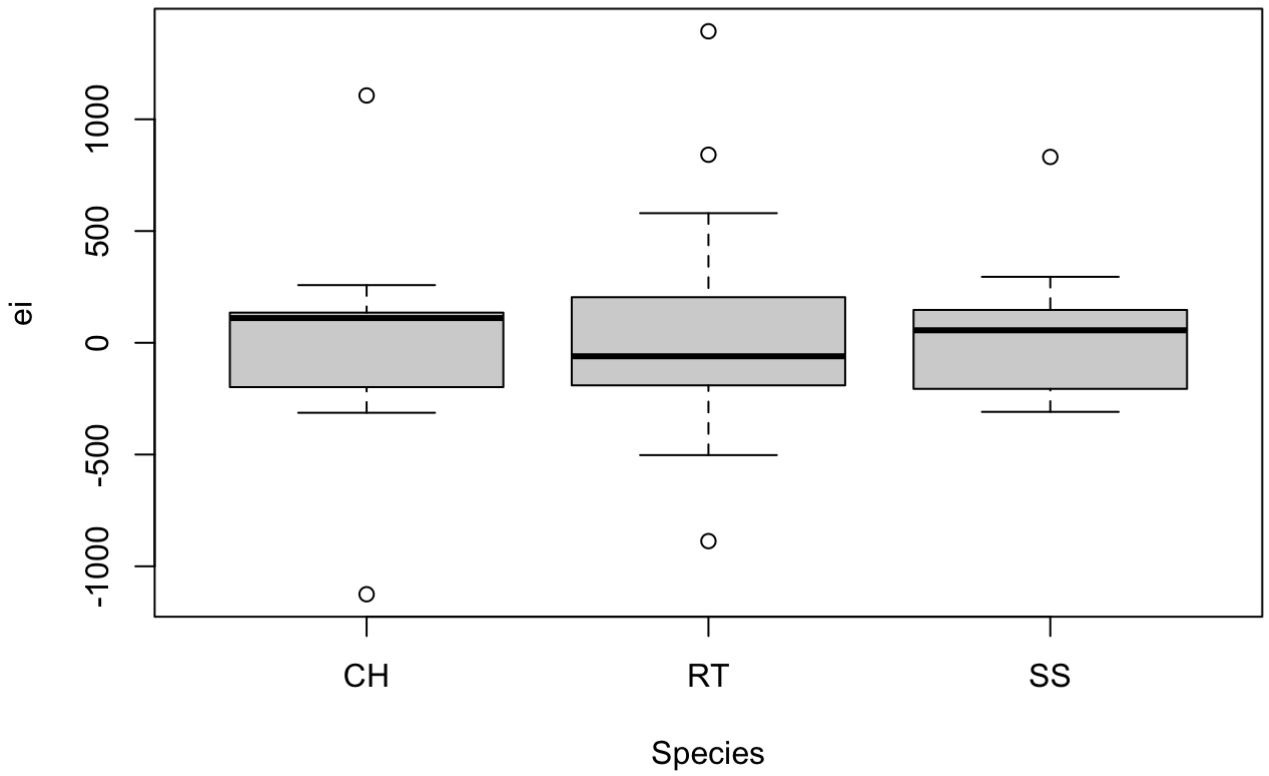
```
## $lambda
## [1] -2.0 -1.5 -1.0 -0.5  0.0  0.5  1.0  1.5  2.0
##
## $objective
## [1] 0.8851343 0.9001209 0.9150790 0.9290534 0.9406579 0.9475409 0.9527222
## [8] 0.9559540 0.9508992
##
## $objective.name
## [1] "PPCC"
##
## $optimize
## [1] FALSE
##
## $optimize.bounds
## lower upper
##    NA    NA
##
## $eps
## [1] 2.220446e-16
##
## $lm.obj
##
## Call:
## lm(formula = Wing ~ Species, data = the.data, y = TRUE, qr = TRUE)
##
## Coefficients:
## (Intercept)    SpeciesRT    SpeciesSS
##      248.31       140.59       -63.14
##
##
## $sample.size
## [1] 150
##
## $data.name
## [1] "the.model"
##
## attr(,"class")
## [1] "boxcoxLm"
```

```
## $lambda
## [1] -2.0 -1.5 -1.0 -0.5  0.0  0.5  1.0  1.5  2.0
##
## $objective
## [1] 0.7978999 0.8249011 0.8520476 0.8773074 0.8976770 0.9089629 0.9173193
## [8] 0.9231847 0.9135114
##
## $objective.name
## [1] "Shapiro-Wilk"
##
## $optimize
## [1] FALSE
##
## $optimize.bounds
## lower upper
##    NA    NA
##
## $eps
## [1] 2.220446e-16
##
## $lm.obj
##
## Call:
## lm(formula = Wing ~ Species, data = the.data, y = TRUE, qr = TRUE)
##
## Coefficients:
## (Intercept)    SpeciesRT    SpeciesSS
##      248.31       140.59       -63.14
##
##
## $sample.size
## [1] 150
##
## $data.name
## [1] "the.model"
##
## attr(,"class")
## [1] "boxcoxLm"
```

```
## [1] "The best value of Lambda using the Box-Cox transformation and the Shapiro-Wilks
criteria is 1.45 ."
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-1000
-500
0
500
1000
-2
-1
0
1
2


Errors vs. Group Means
Errors
Group Means
-1000
-500
0
500
1000
1500
2000
2500
3000
3500
4000




```
##
##  Shapiro-Wilk normality test
##
## data:  ei
## W = 0.92328, p-value = 3.476e-07
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  1.6094 0.2035
##       147
```

```
## [1] "It appears from our diagnostic plots and tests, that with transformed dataset we
have all our SFA assumptions hold. Also, the p-value for the hypothesis test of Shapiro-
Wilks is:  3.4758865442054e-07 and the p-value for the hypothesis test of Brown-Forsythe
is:  0.203515870902367 . Compared to the original data the p-value for the normality tes
t has increased after transforming the data but it is still not high enough to reject th
e null hypothesis. The errors are still non-normally distributed and the p-value is less
significant compared to the test for removing outliers. For the test for equal variance,
the p value of the transformed dataset is more significant than the original dataset but
less significant than the removal of outliers data set."
```

c. 3. Report back appropriate values/plots for every combination of transformations and removing Outliers considered. Pick the “best” combination of transformed variables.

```
## [1] "Comparing the p-value for both the transformed dataset 1.63904616845809e-89 , th
e original dataset (with outliers) 2.72565159316942e-91 ; we observe the diagnostic plot
approximately improving however the p-value is not reliable. And, since removing the out
liers may not be a good approach because it is taking away 4% of our data."
```

d. Discuss your results. Did the transformation help? What are the downsides? Do you believe the transformed data is a better fit? What would you suggest for a client who wants to use this data set for ANOVA (which transformations / removal of outliers would you use, if any)?

```
## [1] "Transforming the data may also not be a good approach because the interpretation
of the new data may be difficult. Since both removing the outliers and transforming the
data set did not help meet both of the assumptions of normally distributed errors and eq
ual variance simultaneously, neither model should be used and the data cannot be used wi
th ANOVA."
```

Topic 2: Two-factor ANOVA

a. Introduction:

a(1): Research question:

What type of Profession have higher salary based on Region?

a(2): why is it a question of interest (why might we be interested in the answer)?

Because, we believe profession with higher salary live in SF region. Thus, we are interested in comparing the annual salary (in Dollars) for a random sample of 120 subjects. It can be argued that tech workers are among the highest paid occupations given that technology is become an increasing priority in society. We wish to see if subject title and working region influence the subjects annual salary in thousands of dollars. The three titles that we will consider are Data Scientist, Software Engineer, and Bioinformatics Engineer. The two regions that we are going to consider are San Francisco and Seattle. We will use the approach of Two Factor ANOVA to run a hypothesis test for interactions in order to determine the best model for our data. Additionally, we will consider and evaluate six confidence interval tests to determine which combination of region and title is the highest paid. This information can be useful for people looking to enter the technology workforce if they want to work in the region with the title that is the highest paid.

a(3): what approach you are going to take (just name?)

We will use the approach of Two-way factor ANOVA.

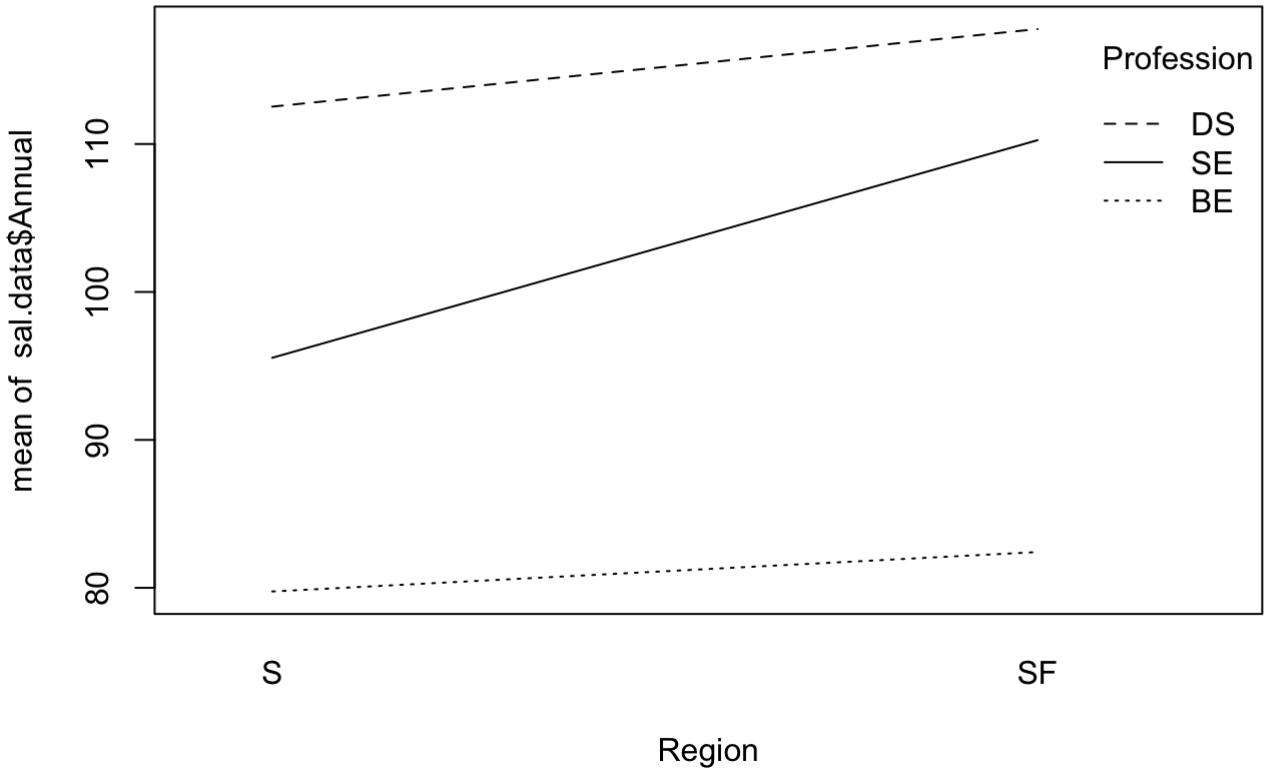
b. Summary

```
## [1] 120
```

```
## [1] "The first 6 rows of our dataset."
```

```
##      Annual Prof Region
## 1 131.32303   DS     SF
## 2 115.06994   DS     SF
## 3 103.88400   DS     SF
## 4  95.83112   DS     SF
## 5 112.46040   DS     SF
## 6 112.21134   DS     SF
```

```
## [1] "Interaction Plot of our dataset."
```




```
## [1] "Group mean for Factor A - Profession"
```

```
##       BE       DS       SE
##  81.0870 115.1480 102.9064
```

```
## [1] "Group mean for Factor B - Region"
```

```
##         S        SF
##  95.94358 103.48403
```

```
## [1] "Group mean for all Treatment Levels"
```

```
##          BE       DS        SE
## S  79.75485 112.5272  95.54875
## SF 82.41914 117.7688 110.26412
```

```
## [1] "Values of Sample Size (treatment and factor levels)"
```

```
## $A
## BE DS SE
## 40 40 40
##
## $B
##  S SF
## 60 60
##
## $AB
##    BE DS SE
## S  20 20 20
## SF 20 20 20
```

```
## [1] "Standard Deviation for Factor A - Profession"
```

```
##        BE        DS        SE
##  9.662515 13.668190 13.240313
```

```
## [1] "Standard Deviation for Factor B - Region"
```

```
##        S       SF
## 17.41791 19.29842
```

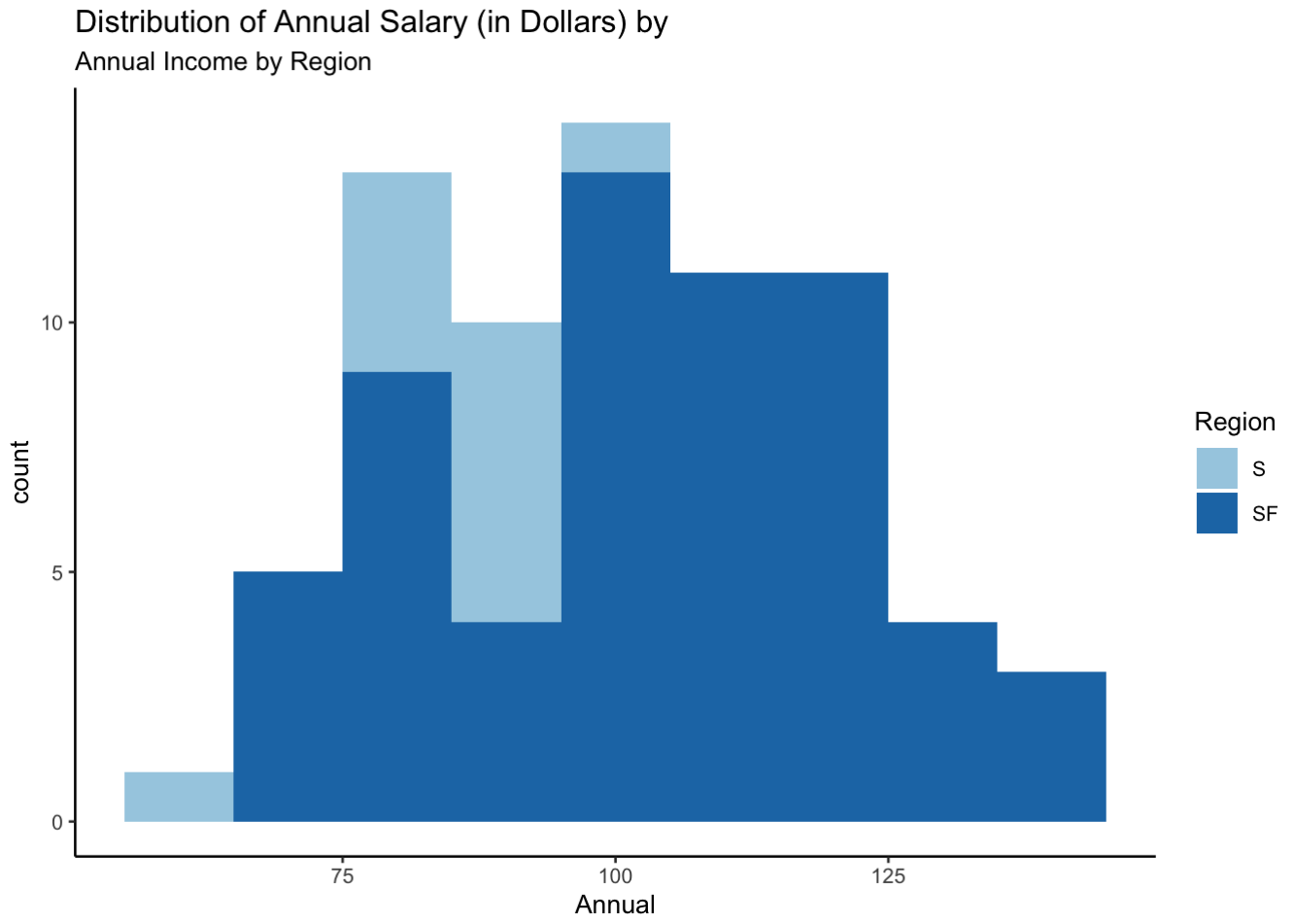
```
## [1] "Standard Deviation for all Treatment Levels"
```

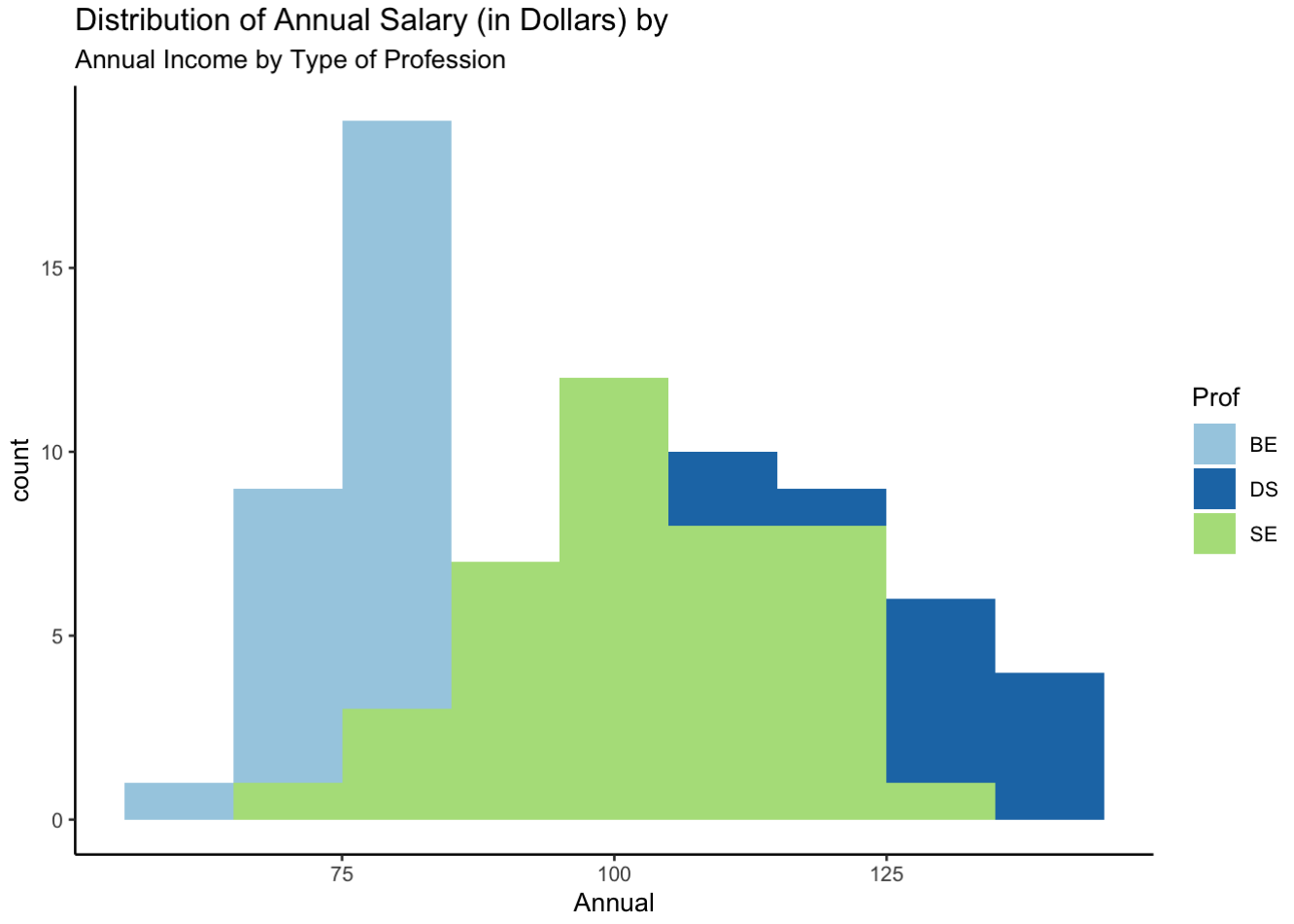
```
##           BE       DS       SE
## S   8.786628 12.83857 11.59872
## SF 10.521476 14.28923 10.55171
```

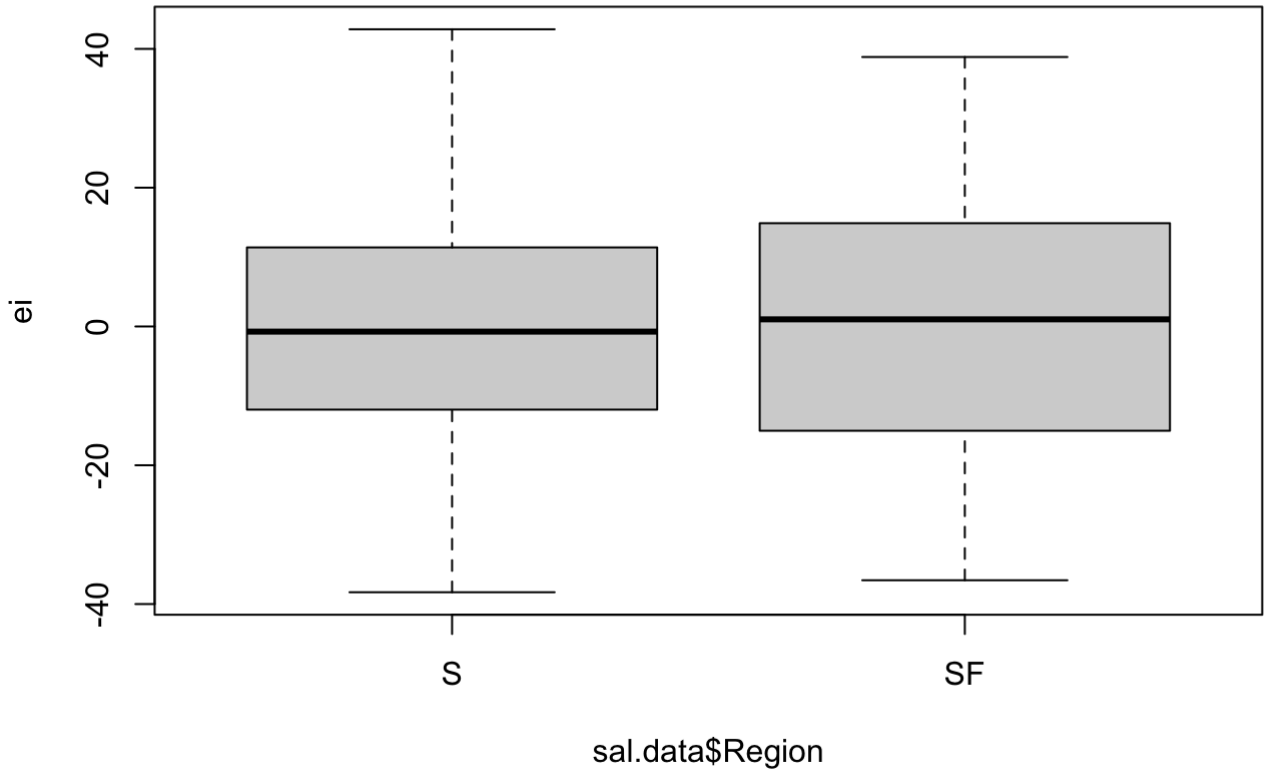
The interaction plot appears to have parallel lines of interaction and this indicates there might not be an interaction between both factor A and factor B effect. Thus, we will follow with our diagnostic, hypothesis testing and corrected confidence interval for our equally weighted data.

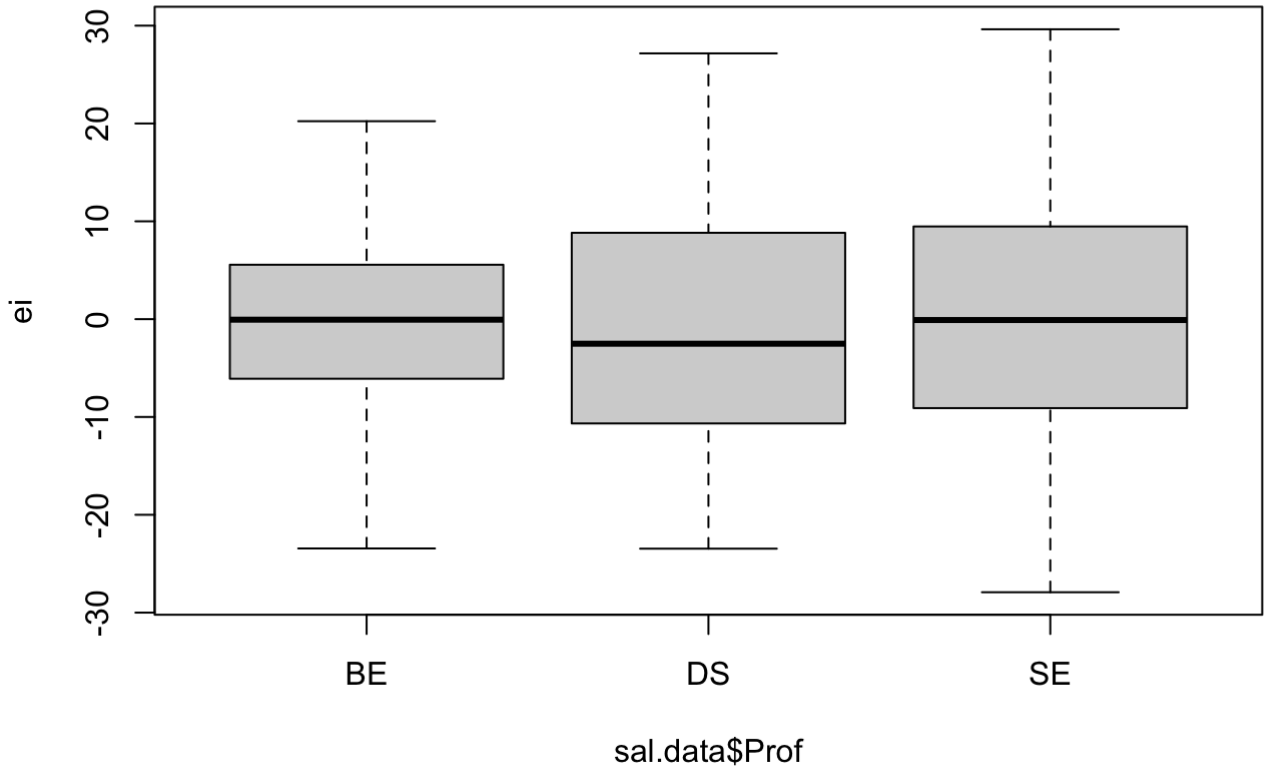
```
## [1] "Two-way table used before plotting categorical variables."
```

```
##
##       S SF
##   BE 20 20
##   DS 20 20
##   SE 20 20
```


Distribution of Annual Salary (in Dollars) by
Annual Income by Region
count
0
5
10
75
100
125
Annual
Region
S
SF


Distribution of Annual Salary (in Dollars) by
Annual Income by Type of Profession
count
0
5
10
15
75
100
125
Annual
Prof
BE
DS
SE

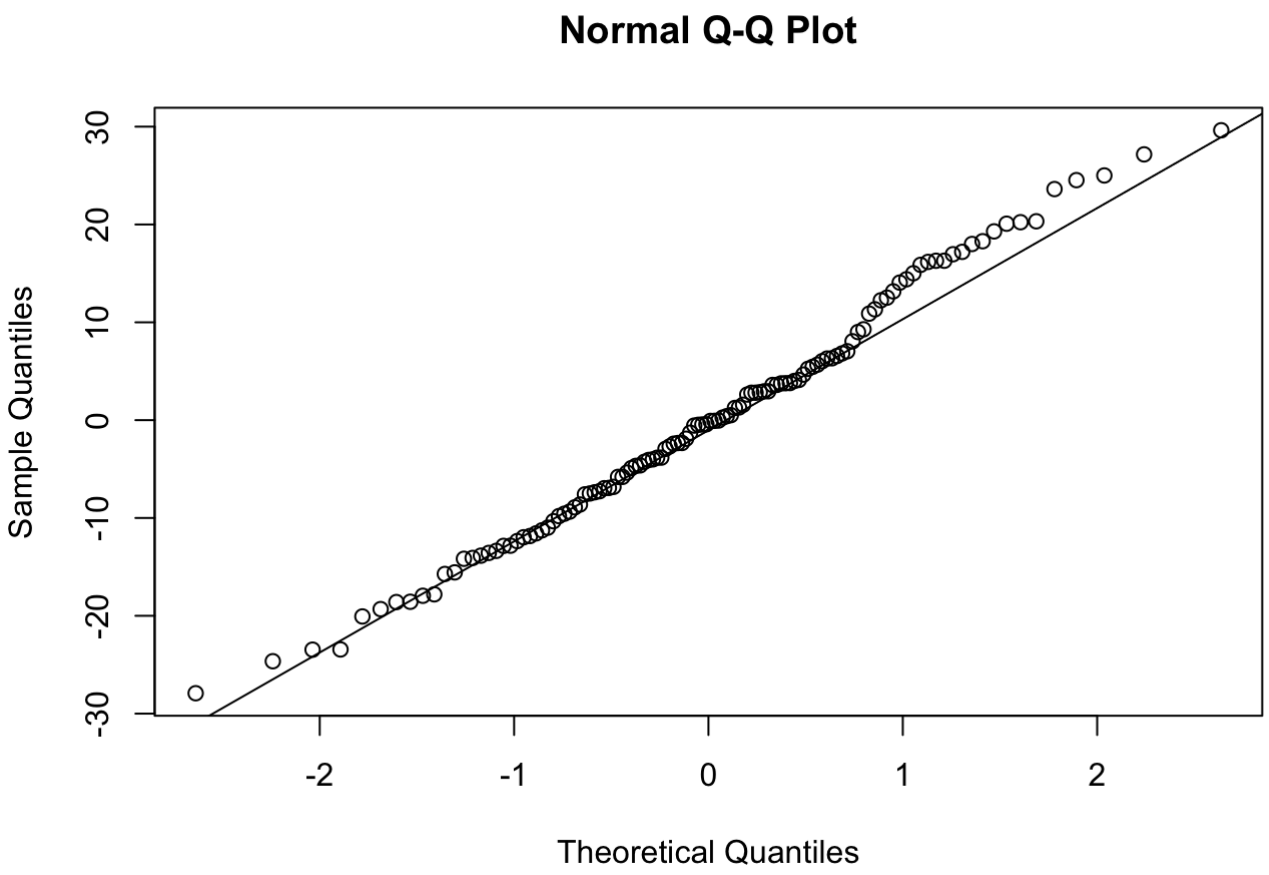

ei
40
20
0
-20
-40
S
SF
sal.data$Region

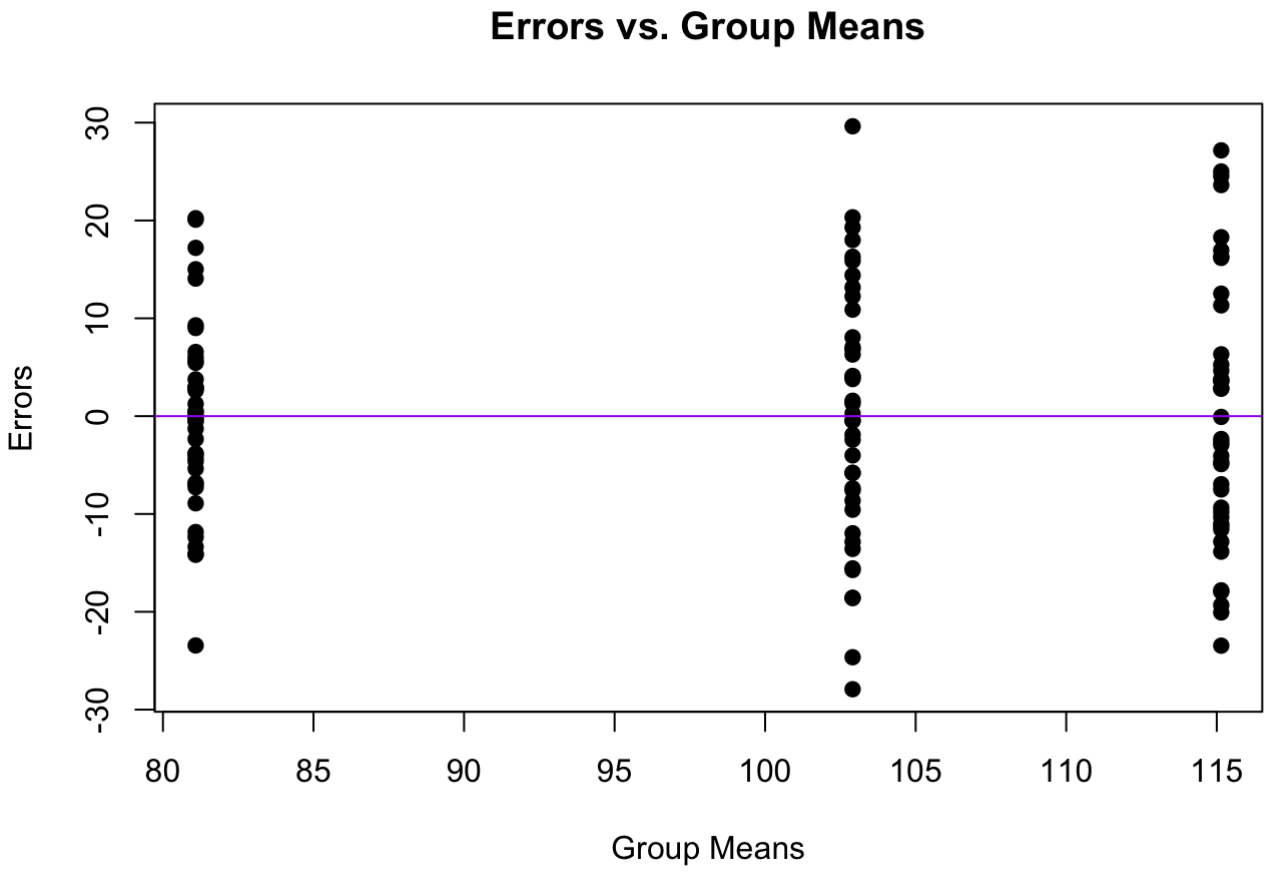

ei
30
20
10
0
-10
-20
-30
BE
DS
SE
sal.data$Prof

By looking at the histogram of income by region we can analyze the shape of the data. In both San Francisco and Seattle regions the income levels are approximately normally distributed and it doesn't seem like one region typically has a higher average salary than the other. By looking at the first box plot, we can determine that there are no outliers and the range of data for Seattle is larger than that of San Francisco. The medians are approximately the same and the IQR for the San Francisco region is slightly larger than that of the Seattle region.

The histogram for annual income by profession shows that the Bioinformatics Engineer makes on average the least amount of money followed by the Software Engineer and then by the Data Scientists who tend to make the most. The second boxplot shows that the medians for each group are approximately the same and the BE profession has the smallest IQR while the SE and DS IQR's are approximately the same.

Diagnostic


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
30
20
10
0
-10
-20
-30
-2
-1
0
1
2


Errors vs. Group Means
Errors
Group Means
30
20
10
0
-10
-20
-30
80
85
90
95
100
105
110
115

```
##
##  Shapiro-Wilk normality test
##
## data:  ei
## W = 0.99027, p-value = 0.5585
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1   0.134 0.7149
##       118
```

```
## [1] "Our assumptions for two-way ANOVA is (1) All subjects are randomly sampled and i
ndependent, (2) All levels of factor A and factor B are independent, (3) Epsilon(ijk) ~
N(mean=0, variance = sigma^2 epsilon)"
```

```
## [1] "According to the diagnostic plot and tests, our p-value for hypothesis test of S
hapiro-Wilk that is 0.558547676457733 and of Brown-Forsythe test that is 0.7149487772330
01 indicates that our data(errors) are normal and have an equal variance. Therefore, we
do not suggest transforming the data."
```

```
## [1] "normal QQ plot analysis: The plot shows us that most points are fairly close to
the line meaning that the assumption of normality for Two Factor ANOVA has been met. err
or vs group means analysis: The Errors vs. Group Means plot shows us that since the vert
ical height distributions are approximately the same, the assumption of equal variance h
as been met. "
```

Analysis/Interpretations

```
##           AB    (A+B)        A        B     Null
## SSE 15252.93 16058.34 17764.09 39872.94 41578.69
```

We will conduct a hypothesis test of ANOVA if there is an interaction effect or not.

Hypothesis Test:

Ho, null hypothesis: The model with no interactions is statistically a better fit model than the model with interactions. All Gamma*Delta(ij) = 0

Ha, alternative hypothesis: The model with interactions is a statistically better fit than the model with no interactions. At least one Gamma*Delta(ij) != 0

Full model: Yijk = mu.. + gamma(i) + delta(j) + Gamma*Delta(ij) + epsilon(ijk) with df{SSE} = 120-(32) = 114 constraints: summation of gamma(i) = 0, summation of delta(j) = 0, summation for all i of Gamma*Delta(ij) = 0 = summation for all j of Gamma*Delta(ij)

Reduced model: Yijk = mu.. + gamma(i) + delta(j) + epsilon(ijk) with df{SSE} = 120-3-2+1 = 116 constraints: summation of gamma(i) = 0, summation of delta(j) = 0

```
## The test statistic is 3.0098
## The p-value is 0.0532
```

Since the p-value is greater than the alpha value of 0.01, we fail to reject the null hypothesis and conclude that the model with no interaction effect between Factor A and Factor B does not improve the statistical fit of the model. In other words, we do not need an interaction model.

1. Before we conclude, we will check the partial R^2 that results:

```
## [1] "The proportion of reduction in error when adding an interaction effect to a mode
l with factor A,B effects is 5.02 % meaning that we can use a no interaction model becau
se the model with interactions does not significantlly reduce errors."
```

```
## [1] "The proportion of reduction in error when adding factor A effects to a model wit
h factor B effects is 59.73 %. While the proportion of reduction in error when adding fa
ctor B effects to a model with factor A effects is 9.6 %."
```

2. We will conduct Hypothesis Test for each interaction effect.

```
## Analysis of Variance Table
##
## Model 1: Y ~ A
## Model 2: Y ~ A + B
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    117 17764
## 2    116 16058  1    1705.8 12.322 0.0006385 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "When conducting hypothesis test for Factor A effects, we get test statistic 12.3
217683676166 and p-value 0.000638465488539829 . Since our p-value is smaller than alpha
(0.01), then we reject the null hypothesis and conclude that Factor A exist therefore th
e model with Factor A effect is statistically better fit. If in reality there is no inte
raction effect, then we would observe our data/more exteme with p-value between 0.000638
465488539829"
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ B
## Model 2: Y ~ A + B
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    118 39873
## 2    116 16058  2     23815 86.014 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "When conducting hypothesis test for Factor B effects, we get test statistic 86.0
142968469385 and p-value 1.23395222447536e-23 . Since our p-value is less than alpha (0.
01), then we reject the null hypothesis and conclude that Factor B exists, therefore the
model with Factor B effect is statistically a better fit. If in reality there is no inte
raction effect, then we would observe our data/more exteme with p-value between 1.233952
22447536e-23"
```

Since there is evidence of both factor A and B effects we will include both in our final model. Our final model will be the reduced model with both factor A and B effects.

From the above table of sample size for our dataset, we know that we have equally weighted data. Since we are asked to consider 6 confidence intervals total that consist 4 pairwise and 2 contrasts we will be using the corrected multipliers to prevent an increase of Type 1 error due to multiple CI's.

```
## Bonferroni      Tukey    Scheffe
##      2.430      2.899      3.387
```

Two-way table of means by region and profession

```
##      [,1] [,2]
## [1,]    0    0
## [2,]    0    0
## [3,]    0    0
```

```
##            S        SF
## BE  79.75485  82.41914
## DS 112.52715 117.76883
## SE  95.54875 110.26412
```

mu1.-mu2.

```
## (1)BE+(-1)DS  lower bound  upper bound
##    -34.06100    -40.08798    -28.03401
```

We are overall 95% confident that the average annual income for the Bioinformatics Enigneer is less than the income for the Data Scientist by between 28.03 and 40.09 thousand dollars.

mu2.-mu3.

```
## (1)DS+(-1)SE  lower bound  upper bound
##    12.241555     6.214571    18.268539
```

We are overall 95% confident that the average annual income for the Data Scientist Engineer is more than the income for the Software Engineer by between 6.21 and 18.27 thousand dollars.

mu1.-mu3.

```
## (1)BE+(-1)SE  lower bound  upper bound
##    -21.81944    -27.84642    -15.79246
```

We are overall 95% confident that the average annual income for the Bioinformatics Enigneer is less than the income for the Software Engineer by between 15.79 and 27.85 thousand dollars.

mu.1-mu.2

```
## (1)S+(-1)SF lower bound upper bound
##   -7.540449  -12.461461   -2.619437
```

We are overall 95% confident that the average annual income for tech workers in Seattle is less than the income of tech workers in San Francisco by between 2.6 and 12.5 thousand dollars.

Contrast mu2.-((mu1.+mu3.)/2)

```
## (-0.5)BE+(1)DS+(-0.5)SE            lower bound            upper bound
##                23.15127               17.93175               28.37080
```

We are overall 95% confident that Data Scientists make more annually than the average income of Bioinformatics Engineers with the Software Engineer by between 17.93 and 28.37 thousand dollars.

Contrast mu1.-((mu2.+mu3)/2)

```
## (1)BE+(-0.5)DS+(-0.5)SE            lower bound            upper bound
##               -27.94022              -33.15974              -22.72070
```

We are overall 95% confident that Bioinformatic Engineers make less annually than the average income of Data Scientists and Software Engineers by between 22.72 and 33.16 thousand dollars.

Conclusion

Our original data set did not have to be transformed since the diagnostic plots and tests our data(errors) are normally distributed and the group variance of the means were equal.After conducting a hypothesis test for interactions we concluded that the test with no interactions is a statistically better fit than the model with interactions. We went on to test for individual factor effects and determined that there were factor A and B effects. Therefore our final model will be the reduced model with factor A and B effects. In order to better interpret our data we ran confidence interval tests to compare between different means on the factor levels. It appears that Data Scientists in San Francisco have a higher income than the other professions that we analyzed. On average the region of San Francisco had a higher overall income than the Region of Seattle. On average Data Scientists had the highest paid profession over Bioinformatics Engineers and Software Engineers.

Appendix

```
#Topic 1:
library(ggplot2)
library(EnvStats)
the.data = read.csv("NewHawk.csv")
the.model = lm(Wing ~ Species, data = the.data)
#Original plot for 1 numeric and 1 categorical variables
ggplot(the.data, aes(x = Wing, fill = Species)) + geom_histogram(binwidth = 10, position
="identity") + ggtitle("Distribution of Primary Wing Feather Length (in mm) by", subtitl
e = "Type of Hawk Species") + theme_classic(base_size = 10) + scale_fill_brewer(palette
 = "Paired")
#Diagnostic plot 1 (QQ Plot) to assess normality
qqnorm(the.model$residuals); qqline(the.model$residuals)
#Diagnostic plot 2 (Residual Plot) to assess equal variance
plot(the.model$fitted.values, the.model$residuals,
     main = "Errors vs. Group Means",
     xlab = "Group Means",
     ylab = "Errors", pch = 19)
abline(h = 0, col = "purple")
#Diagnostic plot 3 (Box Plot) to assess equal variance
ei = the.model$residuals
boxplot(ei ~ Species, data = the.data)
#Diagnostic test for normality
the.SWtest = shapiro.test(ei)
the.SWtest
p.val1 = the.SWtest[[2]][1]
#Diagnostic test for equal variance
library(car)
the.BFtest = leveneTest(ei ~ as.factor(Species), data = the.data, center = median)
the.BFtest
p.val2 = the.BFtest[[3]][1]
alpha = 0.01; nt = nrow(the.data); a = length(unique(the.data$Species));
#Outliers
t.cutoff= qt(1-alpha, nt-a)
rij = rstandard(the.model)
CO.rij = which(abs(rij) > t.cutoff)
outliers = CO.rij
outliers
new.data = the.data[-outliers,]
new.model = lm(Wing ~ Species, data = new.data)
#Proportion of outlier removed
proportions = (6/nt)*100
#Diagnostic plot 1 (QQ Plot) to assess normality
qqnorm(new.model$residuals); qqline(new.model$residuals)
#Diagnostic plot 2 (Residual Plot) to assess equal variance
plot(new.model$fitted.values, new.model$residuals,
     main = "Errors vs. Group Means",
     xlab = "Group Means",
     ylab = "Errors", pch = 19)
abline(h = 0,col = "purple")
#Diagnostic plot 3 (Box Plot) to assess equal variance
ei = new.model$residuals; boxplot(ei ~ Species, data = new.data)
#Diagnostic test for normality
```

```
the.SWtest = shapiro.test(ei); p.val3 = the.SWtest[[2]][1]
#Diagnostic test for equal variance
the.BFtest = leveneTest(ei ~ as.factor(Species), data = new.data, center = median)
p.val4 = the.BFtest[[3]][1]
the.data = read.csv("NewHawk.csv")
the.model = lm(Wing ~ Species, data = the.data)
boxcox(the.model, objective.name = "PPCC"); boxcox(the.model, objective.name = "Shapiro-
Wilk")
BestL = boxcox(the.model, objective.name = "Shapiro-Wilk", optimize = TRUE)$lambda
paste("The best value of Lambda using the Box-Cox transformation and the Shapiro-Wilks c
riteria is", round(BestL, digits = 3), ".")
YT = (the.data$Wing^(BestL)-1)/BestL
t.data = data.frame(Wing = YT, Species = the.data$Species)
t.model = lm(Wing ~ Species, data = t.data)
#Diagnostic plot 1 (QQ Plot) to assess normality
qqnorm(t.model$residuals); qqline(t.model$residuals)
#Diagnostic plot 2 (Residual Plot) to assess equal variance
plot(t.model$fitted.values, t.model$residuals,
     main = "Errors vs. Group Means",
     xlab = "Group Means",
     ylab = "Errors", pch = 19)
abline(h = 0, col = "purple")
#Diagnostic plot 3 (Box Plot) to assess equal variance
ei = t.model$residuals; boxplot(ei ~ Species, data = t.data)
#Diagnostic test for normality
the.SWtest = shapiro.test(ei)
the.SWtest
p.val5 = the.SWtest[[2]][1]
#Diagnostic test for equal variance
the.BFtest = leveneTest(ei ~ as.factor(Species), data = t.data, center = median)
the.BFtest
p.val6 = the.BFtest[[3]][1]
anova.1 = anova(the.model)
pval.1 = anova.1[1,5]
anova.2 = anova(t.model)
pval.2 = anova.2[1,5]


#Topic 2:
sal.data = read.csv("Salary.csv")
nt = nrow(sal.data)
nt
paste("The first 6 rows of our dataset.")
head(sal.data)
paste("Interaction Plot of our dataset.")
Profession = as.factor(sal.data$Prof); Region = as.factor(sal.data$Region)
interaction.plot(Region, Profession, sal.data$Annual)
#Find Group Mean and Standard Deviation
find.means = function(sal.data, fun.name = mean){
  nt = nrow(sal.data)
  a = length(unique(sal.data[,2])); b = length(unique(sal.data[,3]))
  means.A = by(sal.data[,1], sal.data[,2], fun.name)
```

```
  means.B = by(sal.data[,1], sal.data[,3], fun.name)
  means.AB = by(sal.data[,1], list(sal.data[,2], sal.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(sal.data, mean)
the.sizes = find.means(sal.data, length)
the.sds = find.means(sal.data, sd)
paste("Group mean for Factor A - Profession")
the.means$A
paste("Group mean for Factor B - Region")
the.means$B
paste("Group mean for all Treatment Levels")
the.means$AB
paste("Values of Sample Size (treatment and factor levels)")
find.means(sal.data, length)
paste("Standard Deviation for Factor A - Profession")
the.sds$A
paste("Standard Deviation for Factor B - Region")
the.sds$B
paste("Standard Deviation for all Treatment Levels")
the.sds$AB
paste("Two-way table used before plotting categorical variables.")
two.way = table(sal.data$Prof, sal.data$Region)
two.way
#Histogram
ggplot(sal.data, aes(x = Annual, fill = Region)) + geom_histogram(binwidth = 10, positio
n = "identity") + ggtitle("Distribution of Annual Salary (in Dollars) by", subtitle = "A
nnual Income by Region") + theme_classic(base_size = 10) + scale_fill_brewer(palette =
"Paired")
# + facet_wrap(vars(Region))
ggplot(sal.data, aes(x = Annual, fill = Prof)) + geom_histogram(binwidth = 10, position
 = "identity") + ggtitle("Distribution of Annual Salary (in Dollars) by", subtitle = "An
nual Income by Type of Profession ") + theme_classic(base_size = 10) + scale_fill_brewer
(palette = "Paired")
# + facet_wrap(vars(Region))
#Boxplot to assess equal variance
y.model = lm(Annual ~ Region, data = sal.data)
ei = y.model$residuals
boxplot(ei ~ sal.data$Region, data = sal.data)
y.model = lm(Annual ~ Prof, data = sal.data)
ei = y.model$residuals
boxplot(ei ~ sal.data$Prof, data = sal.data)
#Diagnostic plot 1 (QQ Plot) to assess normality
qqnorm(y.model$residuals); qqline(y.model$residuals)
#Diagnostic plot 2 (Residual Plot) to assess equal variance
plot(y.model$fitted.values, y.model$residuals,
     main = "Errors vs. Group Means",
```

```
     xlab = "Group Means",
     ylab = "Errors", pch = 19)
abline(h = 0, col = "purple")
#Diagnostic test for normality
the.SWtest = shapiro.test(ei)
the.SWtest
p.val1 = the.SWtest[[2]][1]
#Diagnostic test for equal variance
library(car)
the.BFtest = leveneTest(ei ~ as.factor(Region), data = sal.data, center = median)
the.BFtest
p.val2 = the.BFtest[[3]][1]
paste("Our assumptions for two-way ANOVA is (1) All subjects are randomly sampled and in
dependent, (2) All levels of factor A and factor B are independent, (3) Epsilon(ijk) ~ N
(mean=0, variance = sigma^2 epsilon)")
paste("According to the diagnostic plot and tests, our p-value for hypothesis test of Sh
apiro-Wilk that is", p.val1, "and of Brown-Forsythe test that is", p.val2, "indicates th
at our data(errors) are normal and have an equal variance. Therefore, we do not suggest
 transforming the data.")
paste("normal QQ plot analysis: The plot shows us that most points are fairly close to t
he line meaning that the assumption of normality for Two Factor ANOVA has been met. erro
r vs group means analysis: The Errors vs. Group Means plot shows us that since the verti
cal height distributions are approximately the same, the assumption of equal variance ha
s been met. ")
names(sal.data) = c("Y","A","B")
AB = lm(Y ~ A*B,sal.data)
A.B = lm(Y ~ A + B,sal.data)
A = lm(Y ~ A,sal.data)
B = lm(Y ~ B,sal.data)
N = lm(Y ~ 1, sal.data)
all.models = list(AB, A.B, A, B, N)
SSE = t(as.matrix(sapply(all.models,function(M) sum(M$residuals^2))))
colnames(SSE) = c("AB","(A+B)", "A", "B", "Null")
rownames(SSE) = "SSE"
SSE
#Test statistic and p-value for testing interaction hypothesis test
results = anova(A.B, AB)
testStat = results[2,5]; p.val = results[2,6]
soln = paste("The test statistic is", round(testStat, digits = 4), "\nThe p-value is", r
ound(p.val, digits = 4))
cat(soln)
Partial.R2 = function(small.model, big.model){
  SSE1 = sum(small.model$residuals^2)
  SSE2 = sum(big.model$residuals^2)
  PR2 = (SSE1 - SSE2)/SSE1
  return(PR2)
}
#Partial R^2 {AB | (A+B)}
RAB = Partial.R2(A.B,AB)
paste("The proportion of reduction in error when adding an interaction effect to a model
with factor A,B effects is", round(RAB, digits = 4)*100, "% meaning that we can use a no
interaction model because the model with interactions does not significantlly reduce err
```

```
ors.")
#R^2{A+B|B}
GivenB = Partial.R2(B, A.B)
#R^2{A+B|A}
GivenA = Partial.R2(A, A.B)
paste("The proportion of reduction in error when adding factor A effects to a model with
factor B effects is", round(GivenB, digits = 4)*100, "%. While the proportion of reducti
on in error when adding factor B effects to a model with factor A effects is", round(Giv
enA, digits = 4)*100, "%.")
#Test for factor A effects
TB = anova(A, A.B)
TB
testStatA = TB[2,5]; p.valueA = TB[2,6]
paste("When conducting hypothesis test for Factor A effects, we get test statistic", tes
tStatA, "and p-value", p.valueA, ". Since our p-value is smaller than alpha (0.01), then
we reject the null hypothesis and conclude that Factor A exist therefore the model with
 Factor A effect is statistically better fit. If in reality there is no interaction effe
ct, then we would observe our data/more exteme with p-value between", p.valueA)
#Test for factor B effects
TA = anova(B, A.B)
TA
testStatB = TA[2,5]; p.valueB = TA[2,6]
paste("When conducting hypothesis test for Factor B effects, we get test statistic", tes
tStatB, "and p-value", p.valueB, ". Since our p-value is less than alpha (0.01), then we
reject the null hypothesis and conclude that Factor B exists, therefore the model with F
actor B effect is statistically a better fit. If in reality there is no interaction effe
ct, then we would observe our data/more exteme with p-value between", p.valueB)
#finds the values of all three multipliers
nt = nrow(sal.data)
a = length(unique(sal.data[,2])); b = length(unique(sal.data[,3]))

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)
}
all.mult = find.mult(alpha = 0.05, a = 3, b = 2, dfSSE = 120 - 2*3, g = 3, group = "AB")
all.mult
Bon = all.mult[1]; Tuk = all.mult[2]; Sch = all.mult[3]
```

```
#factor A given as rows; factor B given as columns
find.means = function(sal.data,fun.name = mean){
  a = length(unique(sal.data[,2]))
  b = length(unique(sal.data[,3]))
  means.A = by(sal.data[,1], sal.data[,2], fun.name)
  means.B = by(sal.data[,1], sal.data[,3], fun.name)
  means.AB = by(sal.data[,1], list(sal.data[,2], sal.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)
  MAB = t(MAB)
  results = list(A = MA, B = MB, AB = MAB)
  return(results)
}
#confidence intervals
scary.CI = function(sal.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(sal.data)
      the.ns = find.means(sal.data, length)
      nt = nrow(sal.data)
      a = length(unique(sal.data[,2]))
      b = length(unique(sal.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 = find.means(sal.data)
the.model = lm(Y ~ A+B, data = sal.data)
SSE = sum(the.model$residuals^2)
MSE = SSE/(nt-a*b)
AB.cs = matrix(0,nrow = a, ncol = b)
AB.cs
the.means$AB
Bon = find.mult(alpha = 0.05, a = 2, b = 3, dfSSE = 120 - 2*3, g = 2, group = "A")[1]
A.cs.1 = c(1,-1,0) #mu1.-mu2.
A.cs.2 = c(0,1,-1) #mu2.-mu3.
A.cs.3 = c(1,0,-1) #mu1.-mu3.
scary.CI(sal.data, MSE, equal.weights = TRUE, Bon, "A", A.cs.1)
scary.CI(sal.data, MSE, equal.weights = TRUE, Bon, "A", A.cs.2)
scary.CI(sal.data, MSE, equal.weights = TRUE, Bon, "A", A.cs.3)
Bon = find.mult(alpha = 0.05, a = 2, b = 3, dfSSE = 120 - 2*3, g = 2, group = "B")[1]
B.cs.1 = c(1,-1) #mu.1-mu.2
scary.CI(sal.data, MSE, equal.weights = TRUE, Bon, "B", B.cs.1)
Bon = find.mult(alpha = 0.05, a = 2, b = 3, dfSSE = 120 - 2*3, g = 2, group = "A")[1]
C.cs.1 = c(-0.5,1,-0.5) #mu2.-((mu1.+mu3.)/2)
scary.CI(sal.data, MSE, equal.weights = TRUE, Bon, "A", C.cs.1)
Bon = find.mult(alpha = 0.05, a = 2, b = 3, dfSSE = 120 - 2*3, g = 2, group = "A")[1]
C.cs.2 = c(1,-0.5,-0.5) #mu1.-((mu2.+mu3)/2)
scary.CI(sal.data, MSE, equal.weights = TRUE, Bon, "A", C.cs.2)
```