STA 106 Project II

Analysis of Variance Report

Natalie Aceves, Adam Hetherwick, Aum Shah

Instructor: Maxime Guiffo Pouokam

03/17/2023

Figure 1: A collection of one hundred dollar bills (TheGuardian.com)

Figure 2: A Red Tailed Hawk from below (SacramentoAudobon.org)

Topic I: Transformation of Variables

I. Introduction

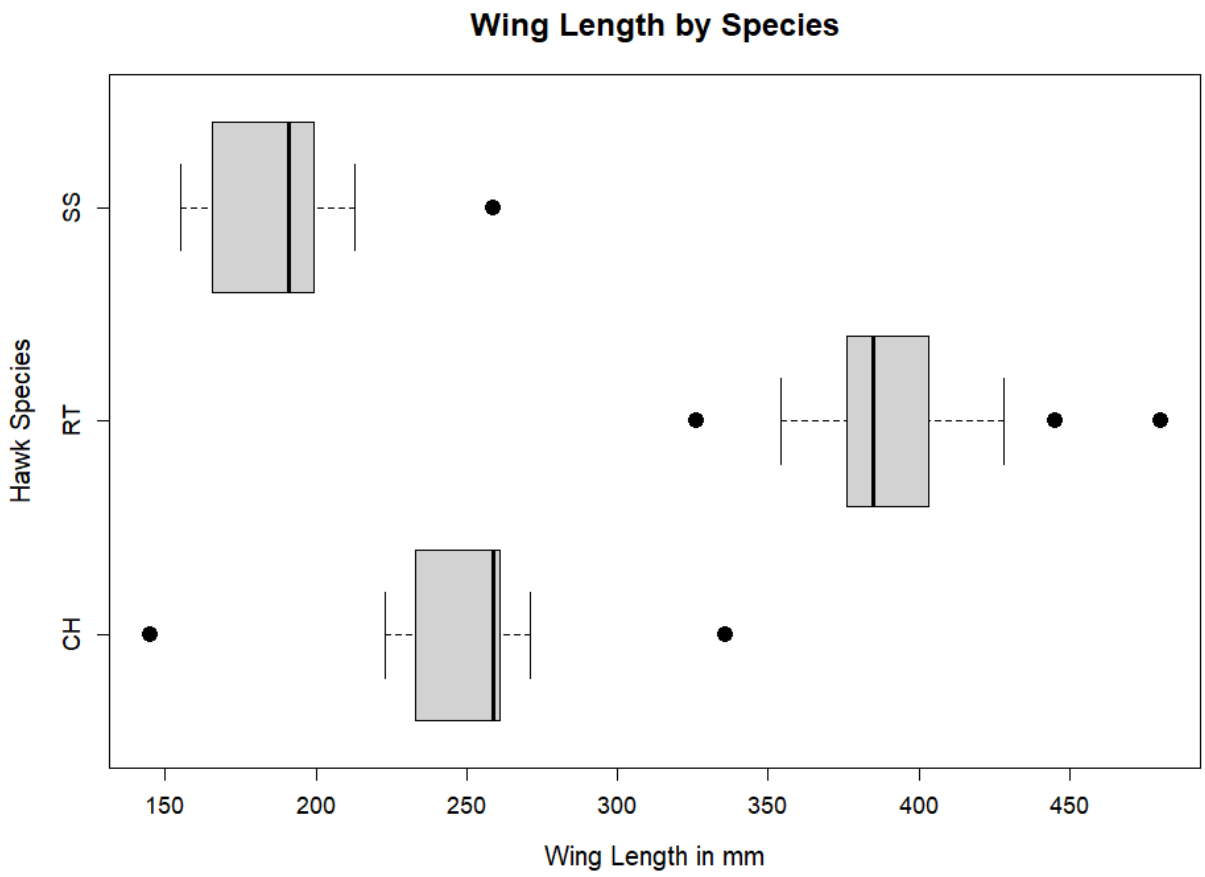
For our first topic, Transformation of Variables, we chose to analyze the NewHawk dataset. In this dataset, there are two variables: wing length, which represents the primary wing feather length measured in millimeters from tip to wrist, and species of hawk, which has three categories including Cooper's Hawk (CH), Red-tailed Hawk (RT), and Sharp-Shinned Hawk (SS).

II. Original Data Diagnostic Plots and Tests

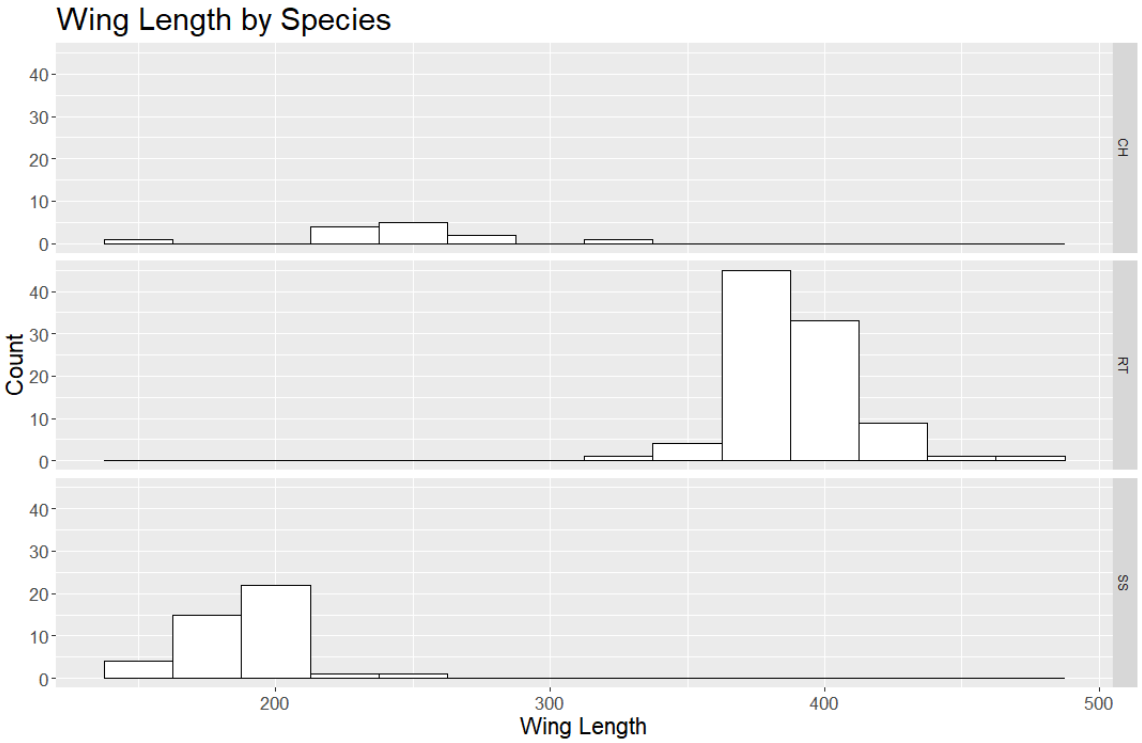
To visualize our data, we can use various plots and tables that represent the differences between each hawk species' primary wing length in mm, including standard deviations, means, distribution, minimums and maximums etc. Visualizing and performing diagnostic tests on the dataset assists us in satisfying the assumptions for single factor ANOVA. The single factor ANOVA model can be found below.

$$Y_{ij} = \mu_i + \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

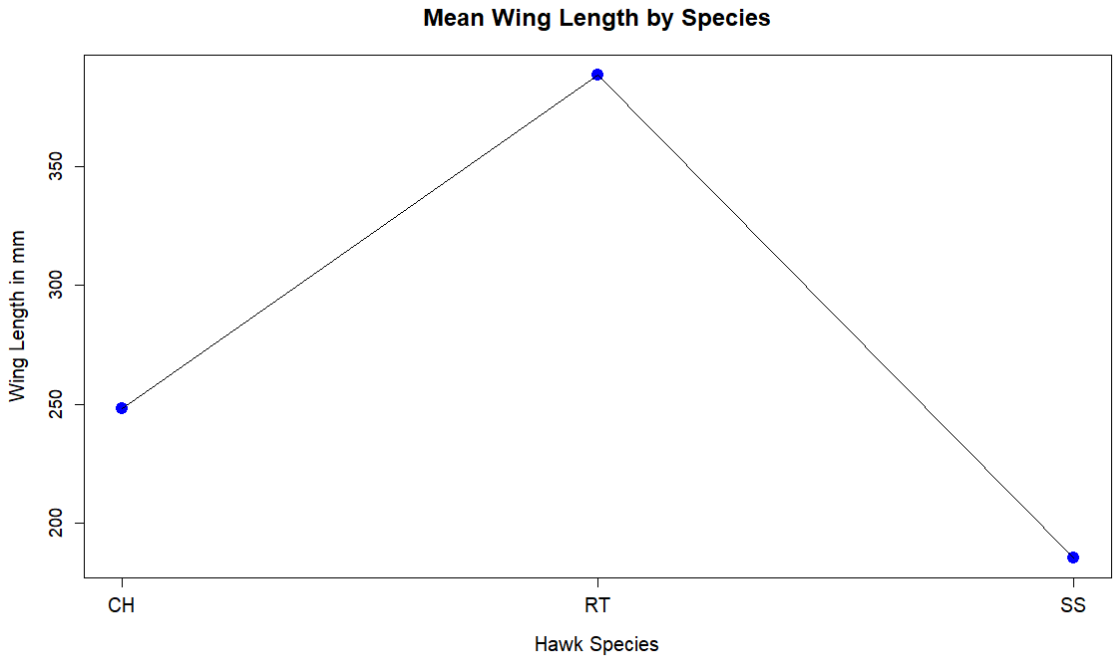
We will choose single factor ANOVA because our dataset includes three predictor variables which don't interact and one response variable, thus only one factor. For single factor ANOVA, we can first examine the variance between groups by constructing boxplots by group to examine the first and third quartiles, the medians, the minimums and maximums, and identify the outliers for each hawk species. Below is a boxplot of the distribution of each hawk species.



These boxplots represent the wing length by hawk species for the original dataset. For Cooper's Hawk (CH), we see the distribution is evidently left skewed and there are also two outliers within the distribution. We observe the Red-Tailed Hawk (RT) having the highest values of wing length with a distribution that's slightly skewed right with three outliers present. The last boxplot represents the Sharp-Shinned Hawk (SS) distribution, which seems slightly left skewed with one outlier. Another way to visualize the distributions, which incorporates sample size, is a histogram of wing length in mm compared by species.



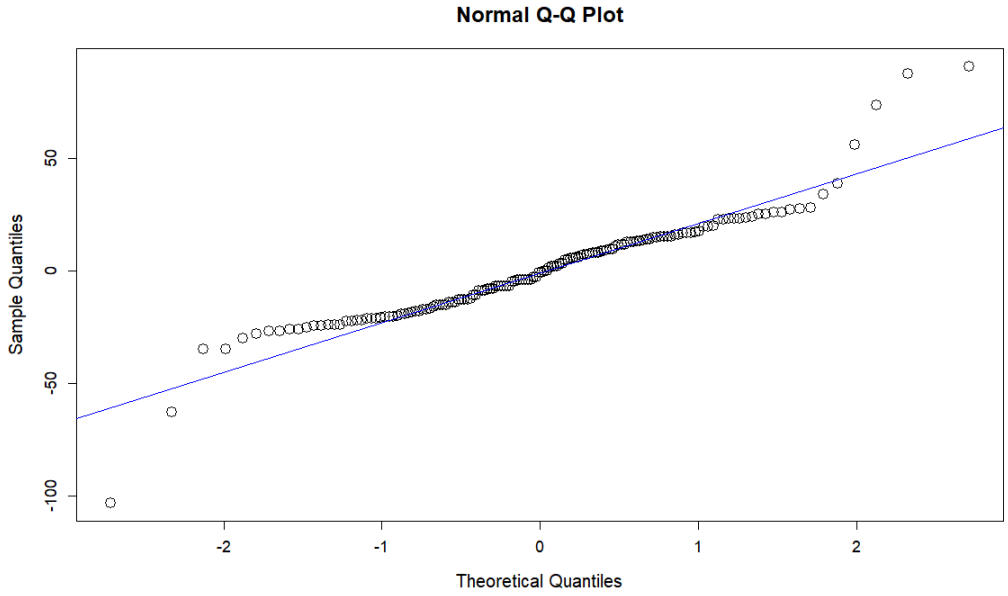
As we can see from the histogram, our observations from the boxplots align. Because all the groups have outliers and none of the species have close to a normal distribution, ANOVA assumptions appear to be violated and therefore we would first consider removing the outliers or making a transformation. We do notice that the mean wing length in mm for each species is different by a significant margin. We can plot the mean wing length in mm by species to visualize the difference in means.



This plot suggests there is a fairly large difference in means between the Sharp-Shinned Hawk and the Red Tailed Hawk, while the Cooper's Hawk is in between the two. This plot hints at the means possibly being statistically different, though before doing ANOVA, we must do further testing on the data to satisfy the assumptions. Below is a table of the mean wing length in mm for each hawk species, along with the minimums, maximums, medians, first and third quartiles, standard deviations, and sample sizes.

	Cooper's	Red-Tailed	Sharp-Shinned	Overall
min	145	326	155	145
Q1	233	376	165.5	203
median	259	385	191	374
Q3	261	403	199.5	391
max	336	480	259	480
mean	248.3077	388.8936	185.1628	318.3067
standard deviation	42.1434	20.8417	21.1828	96.0594
size	13	94	43	150

The mean values for each hawk species supports our claim that the Red-Tailed Hawk has the highest mean, the Sharp-Shinned Hawk has the lowest mean, and the Cooper's Hawk's mean is in between the other two. In order to test this hypothesis, we must first satisfy the assumptions for ANOVA. We can construct a Normal QQ plot to examine whether or not the errors are normally distributed.



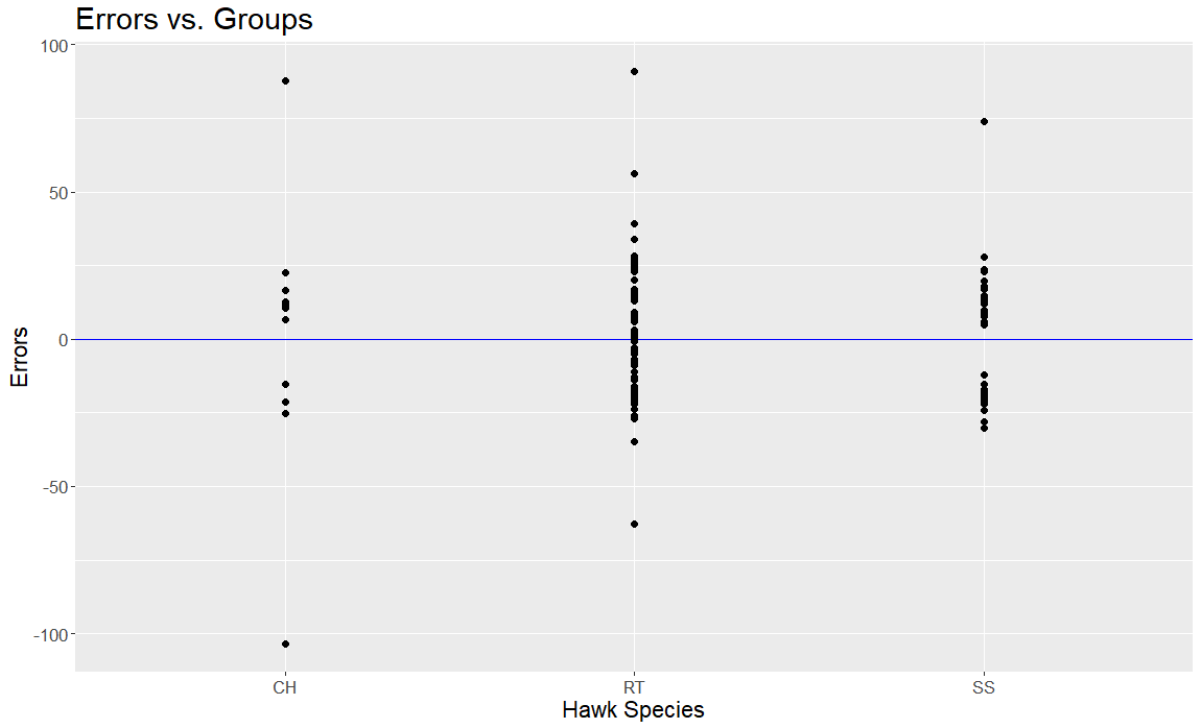
The normal QQ plot represents the Sample Quantiles over the Theoretical Quantiles. This plot calculates the actual centered percentiles of the wing length data, and compares it to what the data should look like through a normal distribution. It is encouraging to see that most of the data resides near the line of best fit, with the exception of a few points near the tail of each end. These plots are subjective, so in order to satisfy the assumptions we must commit to a hypothesis test on normality.

The Shapiro-Wilks test, which calculates the statistical correlation between the theoretical and hypothesized values hypothesis can be found below.

H_0: The data is normally distributed

H_A: The data is not normally distributed

Since the p-value from the Shapiro-Wilk normality test is 1.431e-07 and that is less than alpha at 0.05, we have sufficient evidence to reject the null hypothesis and conclude that the errors are non-normal and therefore violate one of the assumptions for ANOVA testing. To visualize the variance between hawk species, we can construct a plot that examines the errors vs. groups.



Plotting the errors vs each hawk species shows us that the error distribution appears to be different for each species. This plot, like the Normal-QQ plot, is subjective and in order to make a decision we need to commit to a statistical test to determine equal variances between groups.

The Brown-Forsythe test examines the absolute deviations from the median and compares it to the F-distribution to assess whether or not the groups have equal variances. The hypotheses for this test are below:

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$

H_A: At least one σ_i^2 is not equal

By the Brown-Forsythe test with p-value of 0.09912968 and alpha at 0.05, we fail to reject the null hypothesis and conclude that the mean wing length variances in mm for each hawk species are equal.

Since the Brown Forsythe test and Shapiro-Wilks test lead us to believe that the errors are not normally distributed and the groups have constant variance, the assumptions for ANOVA are not met. Two methods we can do to alleviate this is to remove outliers or transform the response variables, which would potentially make the wing length for each species fit a normal distribution and make the errors for each species have equal variance.

III. Transformations and Removing Outliers

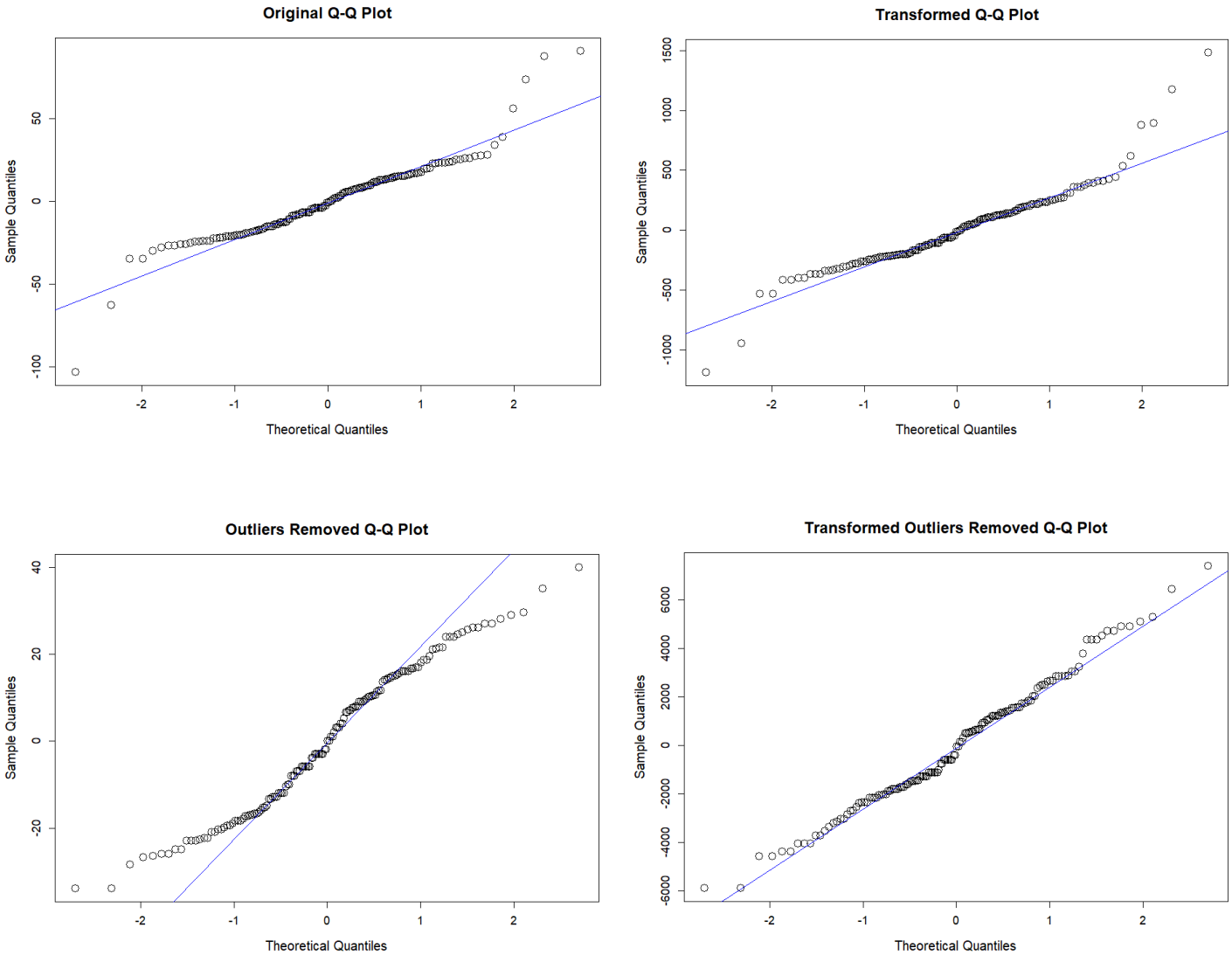
We will do the Box-Cox transformation, which consists of finding the best value of λ from the QQ plot method, the Shapiro-Wilks method, and the Log Likelihood method. We will choose the QQ plot method which measures the correlation between the best line of the QQ plot and the transformed data. We will perform these transformations on both the original dataset and the dataset with removed outliers.

To remove the outliers for each hawk species, we must compare the errors to a *t* distribution. If the following criteria for the errors occur, then we will deem them an outlier and remove them from the dataset:

$$e^{*}_{ij} = \frac{(e_{ij} - 0)}{\sqrt{MSE}} \sim t_{1-\frac{0.05}{300}, 148}$$

$$|e^{*}_{ij}| > 3.674391$$

After removing the outliers and transforming the data, we can plot the errors over the fitted values in a QQ plot for each new dataset. The QQ plots for the original, transformed, outliers removed, and outliers removed and transformed datasets can be found below.



As we can see from the QQ plots, combining the removed outliers with the data transformation results in the most normal looking plot. In order to make a conclusion however, we must test

each dataset with the Brown-Forsythe test and the Shapiro-Wilks test, to examine which dataset results in the most normal errors and most equal group variances.

	Original	Transformed	Outliers Removed	Outliers Removed, Transformed
Brown-Forsythe	0.09912968	0.2000961	0.349729	0.09714334
Shapiro-Wilks	1.43E-07	3.473E-07	0.003435	0.2162

The table suggests that removing outliers and transforming the response variables both have an effect on the error normality and the variances between groups. We will choose the dataset with the highest p-value for the Shapiro-Wilks test and the lowest p-value for the Brown-Forsythe test.

IV. Results

Based on our p-values and diagnostic plots from the Brown-Forsythe and Shapiro-Wilks tests, we conclude that the best data to continue the ANOVA is the transformed dataset with outliers removed. We used a BoxCox transformation with lambda chosen from the "PPCC" method, along with removing the outliers based on if the absolute value of the error is greater than the *t* cutoff.

$$e_{ij}^{*} = \frac{(e_{ij} - 0)}{\sqrt{MSE}} \sim t_{1-\frac{0.05}{300}, 148}$$

$$|e_{ij}^{*}| > 3.674391$$

This transformed dataset with the outliers removed satisfies the assumptions for ANOVA. The downsides of continuing with this dataset is that it may be harder to interpret conclusions and difficult to reverse the effects of the transformation. For anyone going forward with ANOVA models with transformed data, it is important to include the fact that the data was transformed when interpreting the results.

Topic II: Two Factor ANOVA

I. Introduction

The dataset we chose to analyze measures technology worker's annual salary in thousands of dollars grouped by region and profession. Region is split into either San Francisco (SF) or Seattle (S), with professions data scientist, software engineer, and bioinformatics engineer. To analyze the means and variance of this dataset, we will commit to a two factor ANOVA. We are not sure whether or not the interaction model or no interaction model should be used yet, as we have to perform tests to declare whether or not the factors have an effect. The two factor interaction ANOVA model can be found below, along with the two factor no interaction ANOVA model.

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

First we will test for interaction effects, to see whether or not there is a significant difference in salary depending on location. Then we will assess which predictor variables are most important and include those in the model, to answer the question which city results in higher salary for each profession. This information is important for current or aspiring data scientists, software engineers, and bioinformatics engineers. Before constructing ANOVA models, we will learn more about the dataset through plots and tables.

II. Summary of data

To familiarize ourselves with the dataset, we can construct boxplots and histograms of salary ranges for certain jobs at certain locations. This assists us in the decision process of determining which factors are important to include in the final model. We will assess what salaries have the

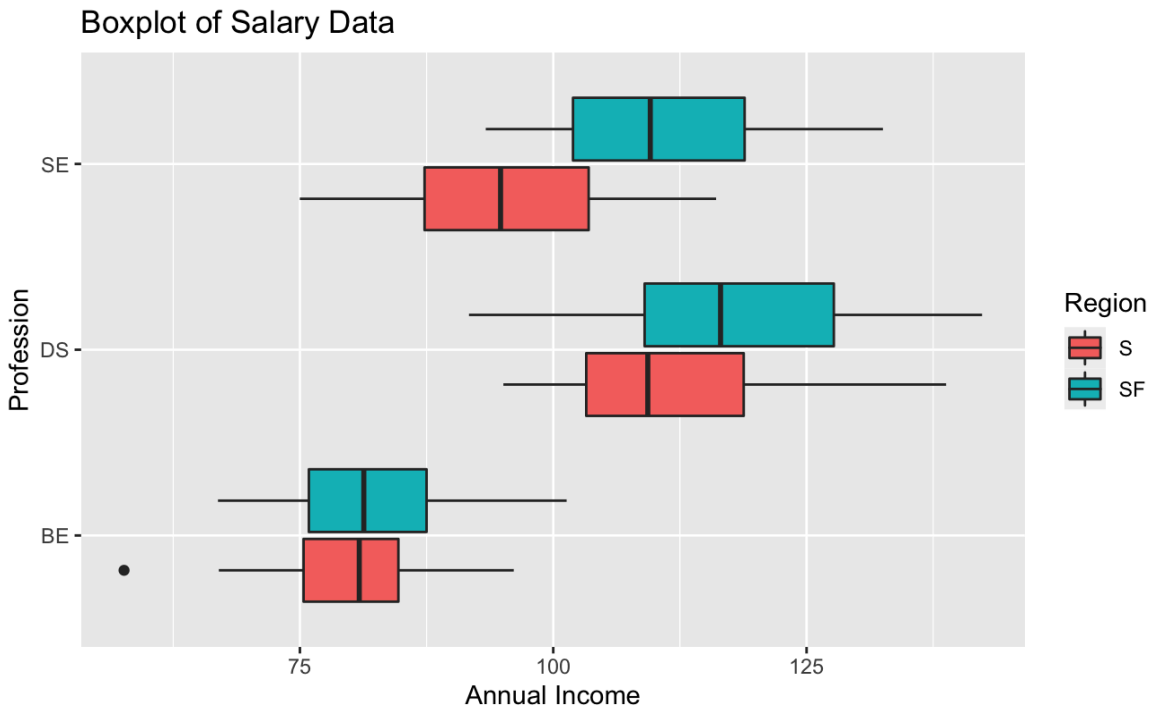
largest difference and make a decision on whether or not the region affects that salary difference. We can start by observing the means and standard deviations for each job in each region.

Means			
Region / Profession	Bioinformatics Engineer	Data Scientist	Software Engineer
Seattle	79.755	112.527	95.549
San Francisco	82.419	117.769	110.264

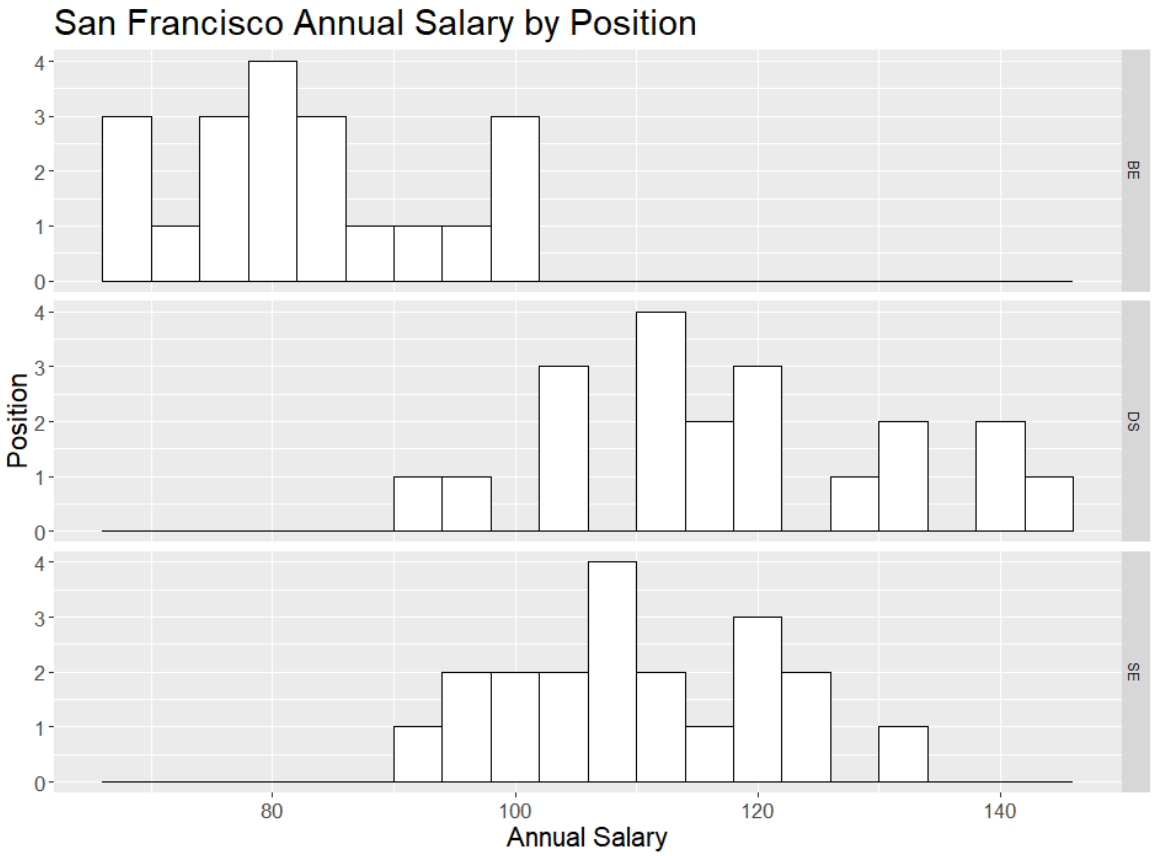
As we can see from the table of mean salaries for each job type in each region, the means seem to stay fairly similar despite a larger difference at software engineer. This leads us to believe that the salary of a person's job may be dependent on the region. To gain an understanding of the distribution and spread of each job and region, we can construct a table of standard deviations.

Standard Deviations			
Region / Profession	Bioinformatics Engineer	Data Scientist	Software Engineer
Seattle	8.786	12.839	11.599
San Francisco	10.521	14.289	10.552

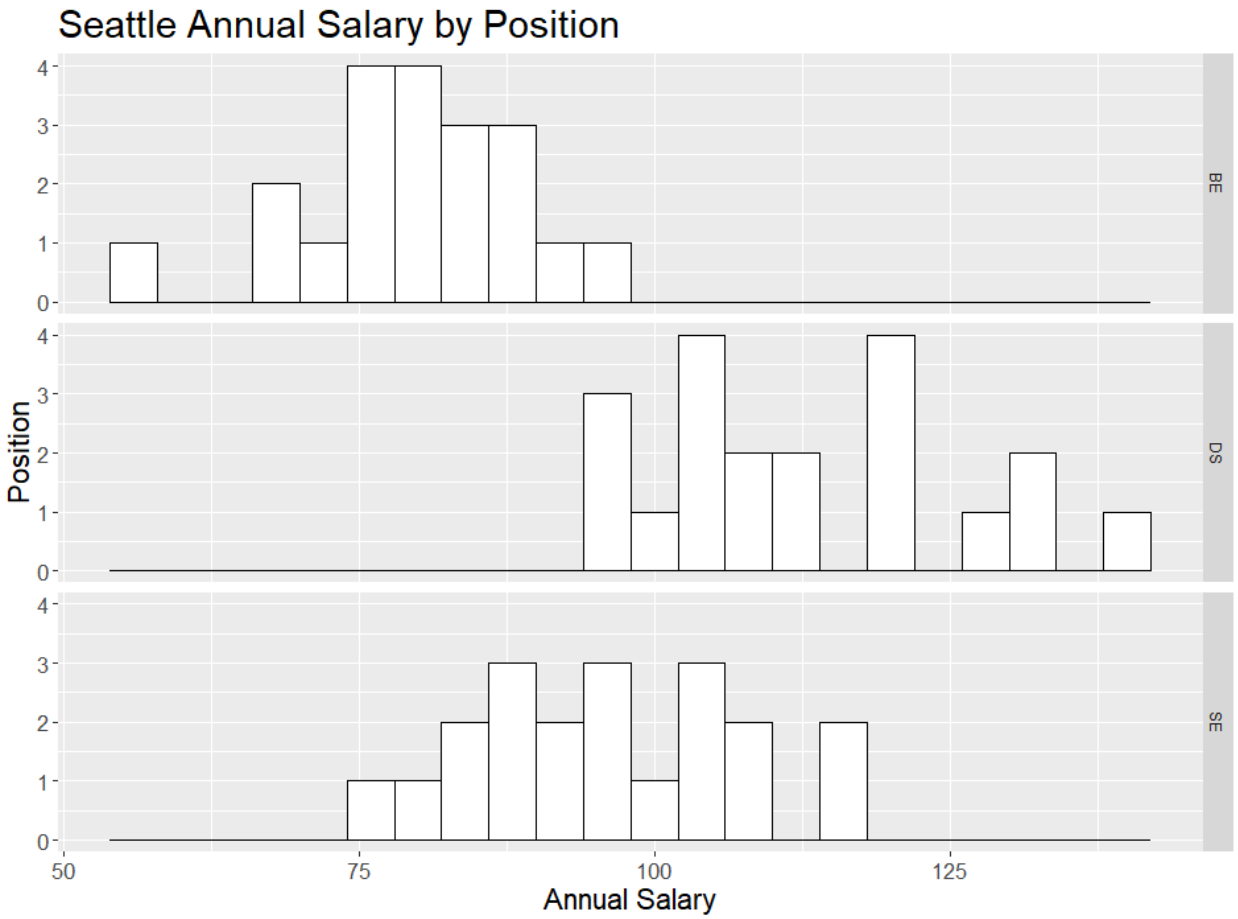
There does seem to be a minor difference in the standard deviation between bioinformatics engineer in San Francisco and bioinformatics engineer in Seattle, though further testing such as the Brown-Forsythe test will let us know if the variances are truly equal or not. We can observe the spreads of each distribution via boxplots for each job and each location shown below.



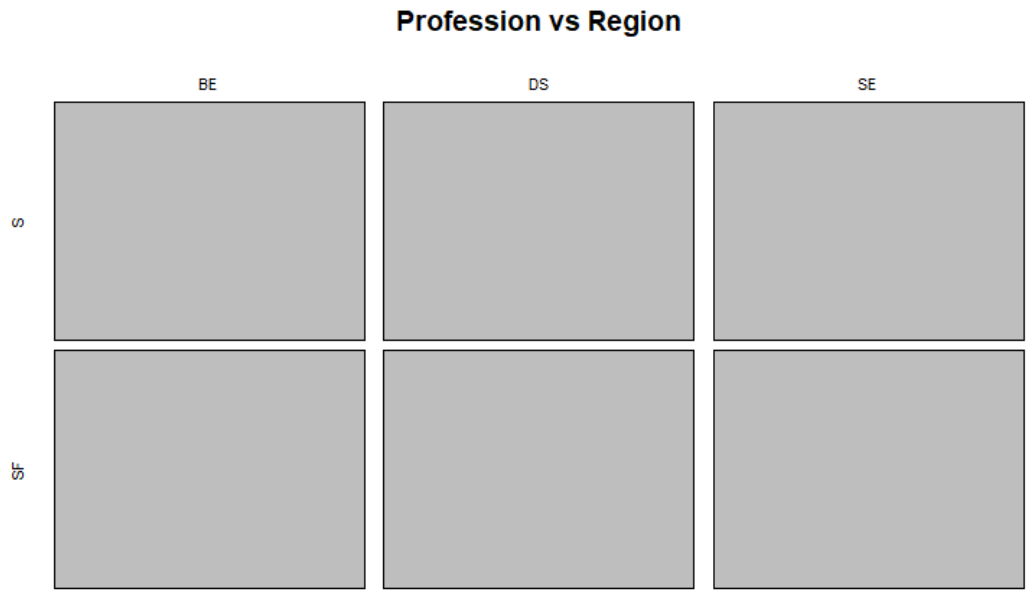
The boxplots for each job in each location suggest that the largest difference in salary is for the software engineer position. We also assume that the variances appear to be equal here, as each boxplots error bars appear to be of similar length. Another method of observing the spread of each dataset is to construct histograms for each group.



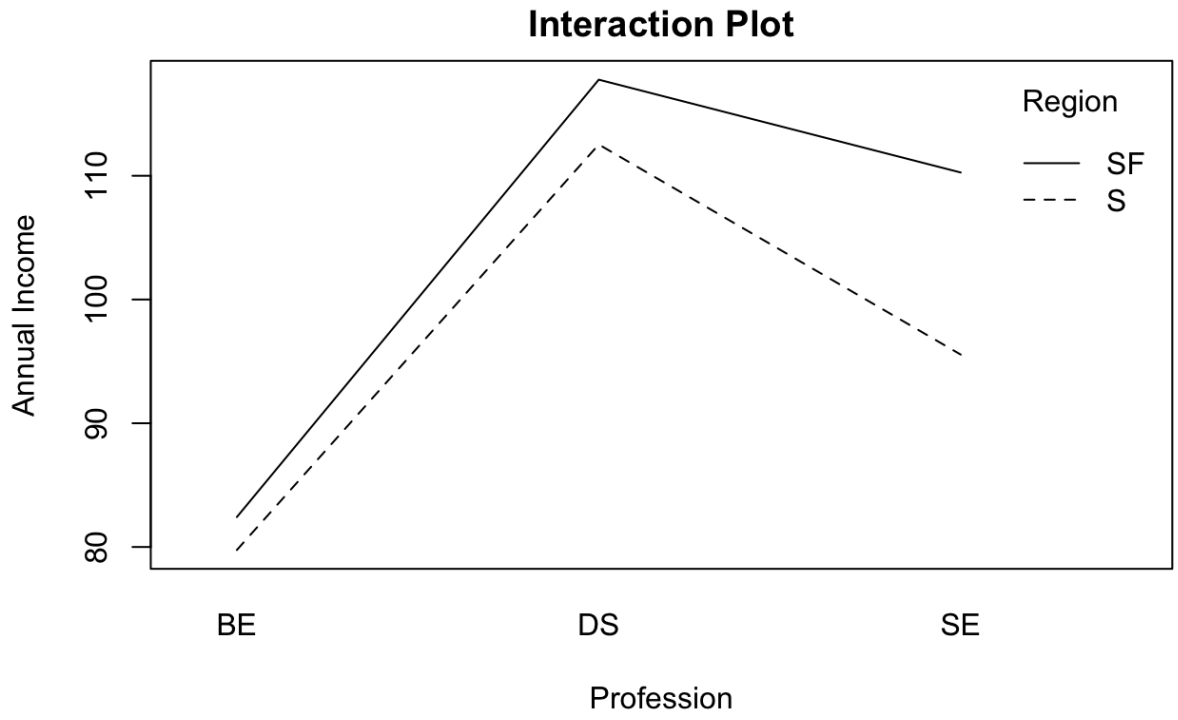
As we can see from the histograms, the distribution appears to be a bit different for each group, as for data scientists it is much wider than for software engineers and bioinformatics engineers. Another takeaway from this plot is that the software engineer appears to have the most normal distribution, which helps for further analysis. The histograms for annual salary by job for Seattle can be found below.



The boxplots show the higher variation for data scientists and relatively even variation for software engineers and bioinformatics engineers. It is reassuring to see that none of the groups have variances that are very spread out because that may violate some of the assumptions for ANOVA. To observe the sample sizes and total contributions between groups, we can create a mosaic plot that resembles the percentage of total data that resides in each group.



This mosaic plot does not offer us much information other than that each sample size is equal and thus contributing to the total data equally. This is why each area within the mosaic plot has equal size. This is however important for ANOVA because if the sample sizes were different, extra calculations would have to be done to find confidence intervals and further analysis.



This interaction plot helps us visualize whether or not there is an interaction effect. Given that the lines are fairly parallel, we can infer that there is no interaction effect and the annual salary for each position does not change by region.

It is important to note that these plots offer us visual information on what to expect for each segment of analysis, though it does not provide us with evidence to make conclusions. In order to satisfy the assumptions for ANOVA, we must perform statistical tests to ensure the assumptions are met.

III. Diagnostics

In order to determine which model is appropriate for two factor ANOVA, we will create a full and reduced model, then compare the SSE's to find a test statistic. This information will help us understand which model is important for further analysis. The test for interaction effects can be

found below, including the models, degrees of freedom, hypotheses, test statistic, p-value and conclusion.

	Model	Df
Full	$Y_{ijk} = \mu_{..} + \gamma_i + \delta_j + (\gamma\delta)_{ij} + \varepsilon_{ijk}$	$nT - ab$
Reduced	$Y_{ijk} = \mu_{..} + \gamma_i + \delta_j + \varepsilon_{ijk}$	$nT - a - b + 1$

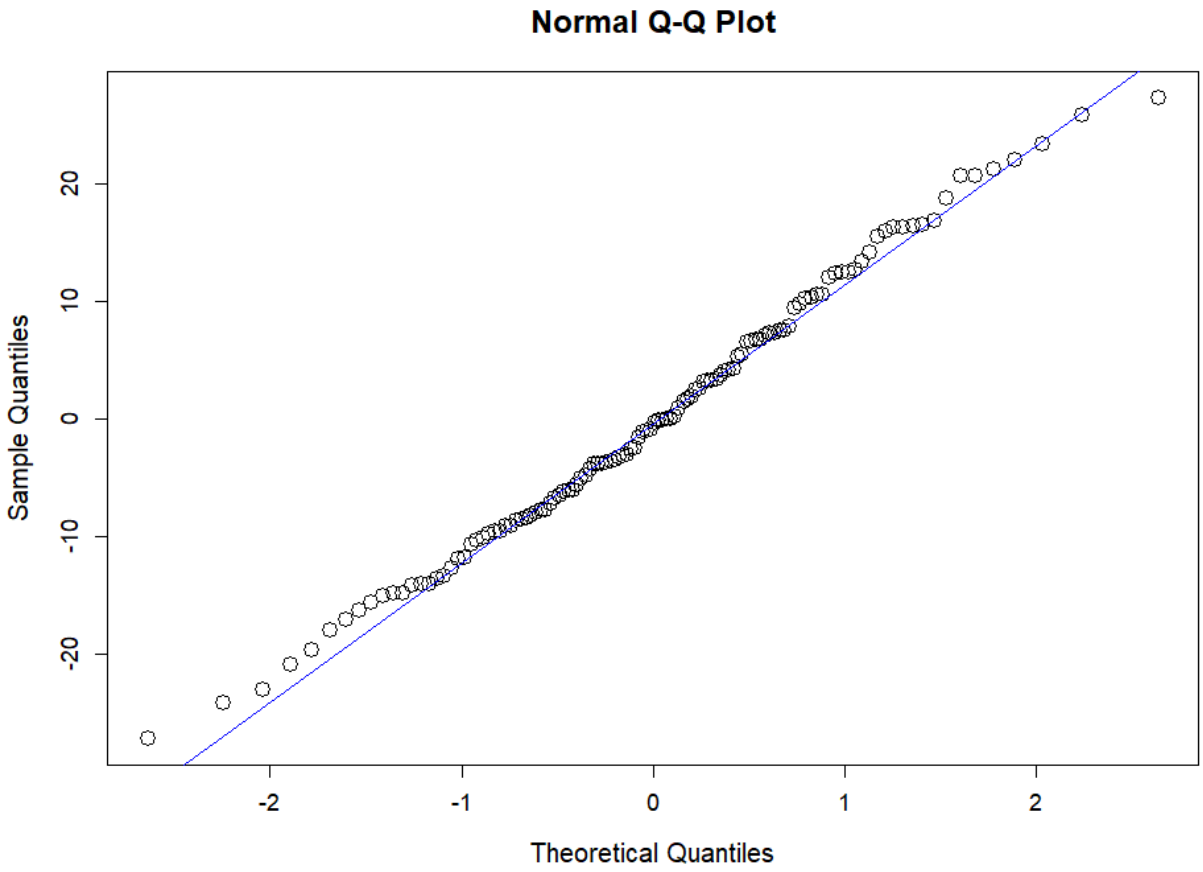
$$H_0: \ All \ \ (\gamma\delta)_{ij} = 0$$

$$H_A: \ At\ least\ one\ (\gamma\delta)_{ij} \neq 0$$

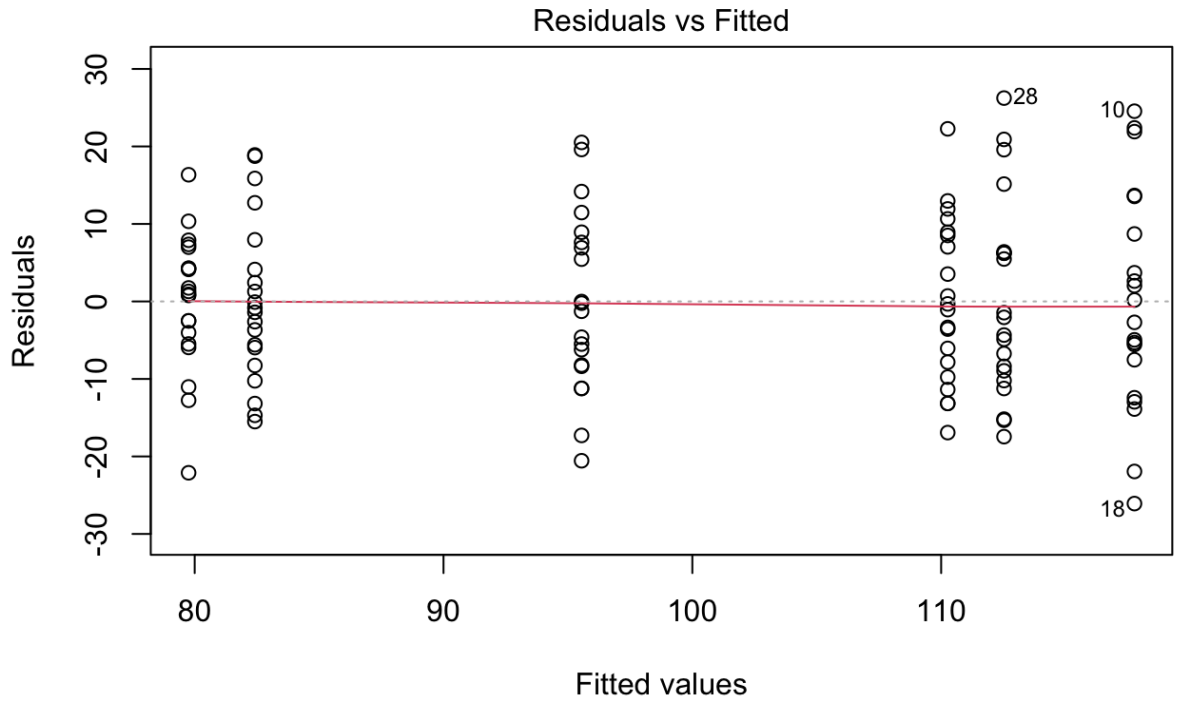
$$F_s = \frac{\left[\frac{SSE_R - SSE_F}{df(SSE_R) - df(SSE_F)}\right]}{MSE_F} = \frac{\left[\frac{16058.3 - 15252.9}{116 - 114}\right]}{133.7974} = 3.009775$$

$$Pr\{F > F_s\}\ at\ df\{SSE_R\} - df\{SSE_F\},\ df\{SSE_F\} = Pr\{F > Fs\}\ at\ 2,\ 114 =$$
$$0.1 < p\ value < 0.05$$

With $F_{critical} = 3.06$, and $0.1 < p\ value < 0.05$, at 5% level of significance ($\alpha = 0.05$) we fail to reject the null hypothesis and conclude that all $(\gamma\delta)_{ij} = 0$. This means the dataset does not have interaction effects and we should continue with the no interaction model. The assumptions for no interaction two factor ANOVA are that all subjects are randomly sampled and independent, all levels of factor A and B are independent, and $\varepsilon_{ijk} \sim N(0,\ \sigma_\varepsilon^2)$. We can construct a QQ plot to visualize the errors for the model.



The plot suggests that the ANOVA model fits a normal distribution because all points are very close to the line of best fit. In order to make a statistical conclusion however we have to perform the Shapiro-Wilks test to satisfy the assumption that $\varepsilon_{ijk} \sim N(0,\ \sigma_{\varepsilon}^{2})$. Since the p-value from the Shapiro-Wilk normality test is 0.6698 and that is greater than alpha at 0.05, we fail to reject the null hypothesis and conclude that the errors are normally distributed. To determine whether the groups have constant variances, we can construct the Errors vs group means plot and Brown-Forsythe test.



By the Brown-Forsythe test with p-value of 0.7149488 and alpha at 0.05, we fail to reject the null hypothesis and conclude that the annual salary variances (in thousands of dollars) for each region are equal. This is supported by the Errors vs Group Means plot above. Since we are continuing with the no interaction model, we can perform a hypothesis test to determine if both factors have an effect on the model. The test to determine if the region factor is significant in the model is shown below.

	Model	Df
Full	$Y_{ijk} = \mu_{..} + \gamma_i + \delta_j + \varepsilon_{ijk}$	$nT - a - b + 1$
Reduced	$Y_{ijk} = \mu_{..} + \delta_j + \varepsilon_{ijk}$	$nT - b$

$$H_0: \gamma_i = 0 \text{ for all } i$$
$$H_A: At\ least\ one\ \gamma_i \neq 0$$

$$F_s = \frac{\left[\frac{SSE_R - SSE_F}{df(SSE_R) - df(SSE_F)}\right]}{MSE_F} = 86.014$$

$$Pr\{F > F_s\}\, at\ df\{SSE_R\} - df\{SSE_F\},\ df\{SSE_F\} = 2.2e\text{-}16$$

Since our p-value is less than alpha (2.2e-16 < 0.05), we reject the null hypothesis and conclude that profession has an effect on salary. This information helps us in deciding which model is most appropriate for analysis. The dataset satisfies the assumptions so we can continue to perform further hypothesis tests.

IV. Analysis

We concluded that the best model to continue with our analysis is the two factor no interaction ANOVA model with both profession and region included. To analyze the dataset and determine difference in means, we can construct pairwise and contrast based confidence intervals. Below is a table representing mean values for each group and factor.

Profession \ Region	San Francisco	Seattle	Overall
Data Scientist	117.76883	112.52715	115.148
Software Engineer	110.26412	95.54875	102.9064
Bioinformatics Engineer	82.41914	79.75485	81.087
	103.48403	95.94358	

To construct our confidence intervals, we will use the Bonferroni multiplier for contrasts, and the Tukey multiplier for pairwise intervals. The equation for contrast confidence intervals can be found below, with pairwise confidence intervals below that.

$$\Sigma_i \Sigma_j c_{ij} \overline{Y}_{ij.} \pm t_{1-\frac{\alpha}{2}} \sqrt{MSE \bullet \Sigma_i \Sigma_j \frac{c_{ij}^2}{n_{ij}}}$$

$$(\overline{Y}_{ij.} - \overline{Y}_{i'j'.}) \pm t_{1-\frac{\alpha}{2}} \sqrt{MSE \bullet (\frac{1}{n_{ij}} + \frac{1}{n_{i'j'}})}$$

The 95% pairwise confidence intervals can be found below:

Variables of Interest	Confidence Interval	Notation
Data Scientists - Software Engineers	(-49.23054, -18.89145)	$\mu_{1.} - \mu_{2.}$
Data Scientists - Bioinformatics Engineers	(-36.988983, -6.649898)	$\mu_{1.} - \mu_{3.}$
Software Engineers - Bioinformatics Engineers	(-2.927988, 27.411097)	$\mu_{2.} - \mu_{3.}$
San Francisco - Seattle	(-19.926329, 4.845431)	$\mu_{.1} - \mu_{.2}$

The 95% contrast confidence intervals can be found below.

Variables of Interest	Confidence Interval	Notation
Software Engineer (S) - Software Engineer (SF)	(-22.053174, 11.569809)	$\mu_{21} - \mu_{22}$
Bioinformatics Engineer (S) - Bioinformatics Engineer (SF)	(-31.526865, 2.096118)	$\mu_{31} - \mu_{32}$

We will also calculate the partial R^2 for each factor to examine the general importance of each variable.

$R^2\{B|A + B\} = 0.5972622$
$R^2\{A|A + B\} = 0.09602243$

V. Interpretation

The interaction test found that factors A and B were independent, and that is why we fit the no interaction two factor ANOVA model. The Shapiro-Wilks test for normality found that the errors in the data set were distributed normally, and the Brown-Forsythe Test for constant variance concluded that all the groups in the data set had equal variance. Since the assumptions for two

factor ANOVA were satisfied, we created pairwise and contrast confidence intervals, along with partial R^2 values.

Our 95% confidence intervals show that in both San Francisco and Seattle, the annual salary in thousands of dollars for data scientists is larger than that of software engineers and data scientists. The annual salary for bioinformatics engineers and software engineers are not statistically different however.

We are 95% confident that the true average salary in thousands of dollars for technology employees in San Francisco is less than that of technology employees in Seattle by between -19.926329 and 4.845431 thousand dollars. Since this confidence interval includes 0, it supports our claim that region and profession do not have an interaction effect, because if they did, then the salaries for technology employees would be statistically different and not include 0. Our contrast confidence intervals show that the annual salary for software engineers and bioinformatics engineers is not statistically different from Seattle to San Francisco.

Our partial R^2 values show that when examining the model without interaction effects, when we add 59.72% reduction in error when adding region to the model, and a 9.6% reduction in error when adding profession to the model.

VI. Conclusion

To conclude, we examined how salary is affected by profession and location, for bioinformatics engineers, software engineers, and data scientists in San Francisco and Seattle. Upon performing

an interaction test and testing the effects of the profession factor on the annual salary, we decided to fit the Two-Factor ANOVA model with no interaction, including factors region and profession. We then analyzed our data set constructing pairwise and contrast confidence intervals for difference of means, partial R-squared testing. This information helped us understand the dataset better and come to statistical conclusions on whether salary is dependent on region or profession. We concluded that there are no interaction effects between region and profession, though each profession's salary was statistically different from the others with the exception of bioinformatics and software engineers.

Appendix

```
NewHawk <- read.csv("C:/Users/cheif/RProjects/STA106/NewHawk.csv")
View(NewHawk)

model1 <- lm(NewHawk$Wing ~ NewHawk$Species)
summary(model1)
anova(model1)
nT = nrow(NewHawk)
a = length(unique(NewHawk$Species))
SSE = sum(model1$residuals^2)
MSE = SSE/(nT - a)
eij.star = model1$residuals/sqrt(MSE)

alpha = 0.05
t.cutoff = qt(1-alpha/(2*nT), nT-a)
CO.eij = which(abs(eij.star) > t.cutoff)
outliers = CO.eij
CO1b =  which(NewHawk$Species =="CH" & NewHawk$Wing < 155)
CO1a =  which(NewHawk$Species =="CH" & NewHawk$Wing > 325)
CO2b =  which(NewHawk$Species =="RT" & NewHawk$Wing < 340)
CO2a =  which(NewHawk$Species =="RT" & NewHawk$Wing > 428)
CO3a =  which(NewHawk$Species =="SS" & NewHawk$Wing > 255)
outliers.r = NewHawk[-c(CO1b, CO1a, CO2b, CO2a, CO3a),]
outliers.r.model = lm(Wing ~ Species, data = outliers.r)

boxplot(NewHawk$Wing ~ NewHawk$Species,
        main = "Wing Length by Species",
        ylab = "Hawk Species",
        xlab = "Wing Length in mm",
        cex.axis = 1.1, cex.lab = 1.25, cex.main = 1.5, cex = 1.7,
```

```
pch = 19, horizontal = T)
boxplot(outliers.r$Wing ~ outliers.r$Species)

group.means = by(NewHawk$Wing, NewHawk$Species, mean)
plot(group.means,
     xaxt = "n", pch = 19, col = "blue",
     xlab = "Hawk Species",
     ylab = "Wing Length in mm",
     main = "Mean Wing Length by Species",
     cex.axis = 1.1, cex.lab = 1.25, cex.main = 1.5, cex = 1.7)
axis(1,1:length(group.means), names(group.means), cex.axis = 1.24)
lines(group.means, cex = 1.9)

qqnorm(model1$residuals, cex.axis = 1.1, cex.lab = 1.25, cex.main =
1.5, cex = 1.7)
qqline(model1$residuals, col = "blue", cex = 2)

ei = model1$residuals
the.SWtest = shapiro.test(ei)
the.SWtest

library(ggplot2)
ggplot(NewHawk, aes(x = Wing)) +
  geom_histogram(binwidth = 25,color = "black",fill = "white") +
  facet_grid(Species ~.) +
  ggtitle("Wing Length by Species") +
  xlab("Wing Length") +
  ylab("Count") +
  theme(axis.title = element_text(size = 16.5)) +
  theme(axis.text = element_text(size = 12.5)) +
  theme(title = element_text(size = 18))

library(ggplot2)
qplot(NewHawk$Species, ei, data = NewHawk) +
  ggtitle("Errors vs. Groups") +
  xlab("Groups") + ylab("Errors") +
  geom_hline(yintercept = 0,col = "blue") +
  xlab("Hawk Species") +
  theme(axis.title = element_text(size = 16.5)) +
  theme(axis.text = element_text(size = 12.5)) +
  theme(title = element_text(size = 18)) +
  geom_point(size = 2)
abline(h = 0,col = "purple")

aggregate(NewHawk$Wing, FUN= length, by= list(NewHawk$Species))
aggregate(NewHawk$Wing, FUN= sd, by= list(NewHawk$Species))
aggregate(NewHawk$Wing, FUN= mean, by= list(NewHawk$Species))
aggregate(NewHawk$Wing, FUN= max, by= list(NewHawk$Species))
aggregate(NewHawk$Wing, FUN= median, by= list(NewHawk$Species))
aggregate(NewHawk$Wing, FUN= min, by= list(NewHawk$Species))
```

```
library(car)
the.BFtest = leveneTest(ei ~ Species, data=NewHawk, center=median)
p.val = the.BFtest[[3]][1]
p.val

#problem 1c Transformation, Outlier Removal
#Transformed with outliers
library(EnvStats)
model1 <- lm(NewHawk$Wing ~ NewHawk$Species)
L1.t.original = boxcox(model1, objective.name = "PPCC", optimize =
TRUE)$lambda
L2.t.original = boxcox(model1, objective.name = "Shapiro-Wilk",
optimize = TRUE)$lambda
L3.t.original = boxcox(NewHawk$Wing, objective.name =
"Log-Likelihood", optimize = TRUE)$lambda
YT.t.original = (NewHawk$Wing^(L1.t.original) - 1) / L1.t.original
data.t.original = data.frame(Wing = YT.t.original, Species =
NewHawk$Species)
model.t.original = lm(Wing ~ Species, data = data.t.original)
ei.t.original = model.t.original$residuals
the.SWtest.t.original = shapiro.test(ei.t.original)
the.SWtest.t.original
the.BFtest.t.original = leveneTest(ei.t.original ~ Species,
data=data.t.original, center=median)
p.val.t.original = the.BFtest.t.original[[3]][1]
p.val.t.original

#Outliers removed no transformation
outliers.r.model = lm(Wing ~ Species, data = outliers.r)
ei.outliers.r = outliers.r.model$residuals
the.SWtest.outliers.r = shapiro.test(ei.outliers.r)
the.SWtest.outliers.r
the.BFtest.outliers.r = leveneTest(ei.outliers.r ~ Species,
data=outliers.r, center=median)
p.val.outliers.r = the.BFtest.outliers.r[[3]][1]
p.val.outliers.r

#Outliers removed, transformed
L1.outliers.r = boxcox(outliers.r.model, objective.name = "PPCC",
optimize = TRUE)$lambda
L2.outliers.r = boxcox(outliers.r.model, objective.name =
"Shapiro-Wilk", optimize = TRUE)$lambda
L3.outliers.r = boxcox(outliers.r$Wing, objective.name =
"Log-Likelihood", optimize = TRUE)$lambda
YT.outliers.r = (outliers.r$Wing^(L1.outliers.r) - 1) /
L1.outliers.r
t.data.outliers.r = data.frame(Wing = YT.outliers.r, Species =
outliers.r$Species)
t.model.outliers.r = lm(Wing ~ Species, data = t.data.outliers.r)
ei.t.outliers.r = t.model.outliers.r$residuals
the.SWtest.t.outliers.r = shapiro.test(ei.t.outliers.r)
```

```
the.SWtest.t.outliers.r
the.BFtest.t.outliers.r = leveneTest(ei.t.outliers.r ~ Species,
data=t.data.outliers.r, center=median)
p.val.t.outliers.r = the.BFtest.t.outliers.r[[3]][1]
p.val.t.outliers.r

#qqplots for all datasets
qqnorm(model1$residuals, main = "Original Q-Q Plot", cex.axis =
1.1, cex.lab = 1.25, cex.main = 1.5, cex = 1.7)
qqline(model1$residuals, col = "blue", cex = 2)
qqnorm(model.t.original$residuals, main = "Transformed Q-Q Plot",
cex.axis = 1.1, cex.lab = 1.25, cex.main = 1.5, cex = 1.7)
qqline(model.t.original$residuals, col = "blue", cex = 2)

qqnorm(outliers.r.model$residuals, main = "Outliers Removed Q-Q
Plot", cex.axis = 1.1, cex.lab = 1.25, cex.main = 1.5, cex = 1.7)
qqline(outliers.r.model$residuals, col = "blue", cex = 2)
qqnorm(t.model.outliers.r$residuals, main = "Transformed Outliers
Removed Q-Q Plot", cex.axis = 1.1, cex.lab = 1.25, cex.main = 1.5,
cex = 1.7)
qqline(t.model.outliers.r$residuals, col = "blue", cex = 2)

#topic 2 - Salary data
#Summary Section code
ggplot(salarydata, aes(x=Annual, y=Prof, fill=Region)) +
    geom_boxplot() + ylab("Profession") +xlab("Annual Income")
+ggtitle("Boxplot of Salary Data")
Salary <- read.csv("C:/Users/cheif/RProjects/STA106/Salary.csv")
library(ggplot2)
ggplot(Salary.SF, aes(x = Annual)) +
  geom_histogram(binwidth = 4,color = "black",fill = "white") +
  facet_grid(Prof ~.) +
  ggtitle("San Francisco Annual Salary by Position") +
  xlab("Annual Salary") +
  ylab("Position") +
  theme(axis.title = element_text(size = 16.5)) +
  theme(axis.text = element_text(size = 12.5)) +
  theme(title = element_text(size = 18))

Salary.SF = Salary[c(which(Salary$Region == "SF")),]
Salary.Se = Salary[c(which(Salary$Region == "S")),]

ggplot(Salary.Se, aes(x = Annual)) +
  geom_histogram(binwidth = 4,color = "black",fill = "white") +
  facet_grid(Prof ~.) +
  ggtitle("Seattle Annual Salary by Position") +
  xlab("Annual Salary") +
  ylab("Position") +
  theme(axis.title = element_text(size = 16.5)) +
  theme(axis.text = element_text(size = 12.5)) +
  theme(title = element_text(size = 18))
```

```
two.way = table(Salary$Prof,Salary$Region)
mosaicplot(two.way, main = "Profession vs Region")

#Diagnostic Section
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)
  MAB = t(MAB)
  results = list(A = MA, B = MB, AB = MAB)
  return(results)
}
AB = lm(Salary$Annual ~ Salary$Prof*Salary$Region)
A.B = lm(Salary$Annual ~ Salary$Prof+Salary$Region)
get.gamma.delta = function(the.model,the.data){
  nt = nrow(the.data)
  a = length(unique(the.data[,2]))
  b = length(unique(the.data[,3]))
  the.data$hat = the.model$fitted.values
  the.ns = find.means(the.data,length)
  a.vals = sort(unique(the.data[,2]))
  b.vals= sort(unique(the.data[,3]))
  muij = matrix(nrow = a, ncol = b)
  rownames(muij) = a.vals
  colnames(muij) = b.vals
  for(i in 1:a){
    for(j in 1:b){
      muij[i,j] = the.data$hat[which(the.data[,2] == a.vals[i] &
the.data[,3] == b.vals[j])[1]]
    }
  }
  mi. = rowMeans(muij)
  m.j = colMeans(muij)
  mu.. = sum(muij)/(a*b)
  gammai = mi. - mu..
  deltaj = m.j - mu..
  gmat = matrix(rep(gammai,b),nrow = a, ncol = b, byrow= FALSE)
  dmat = matrix(rep(deltaj,a),nrow = a, ncol = b,byrow=TRUE)
  gamma.deltaij =round(muij -(mu.. + gmat + dmat),8)
  results = list(Mu.. = mu.., Gam = gammai, Del = deltaj, GamDel =
```

```
gamma.deltaij)
  return(results)
}
gd.AB = get.gamma.delta(AB,Salary)
gd.ApB = get.gamma.delta(A.B, Salary)
anova(AB)
anova(A.B)
SSER = 16058.3
SSEF = 15252.9
nT = nrow(Salary)
a = length(unique(Salary$Prof))
b = length(unique(Salary$Region))
dfSSER = nT - a - b + 1
dfSSEF = nT - a*b
MSEF = SSEF/(nT-a*b)
Fs = ((SSER - SSEF) / (dfSSER - dfSSEF)) / MSEF
dfSSER - dfSSEF
dfSSEF

qqnorm(A.B$residuals, cex.axis = 1.1, cex.lab = 1.25, cex.main =
1.5, cex = 1.7)
qqline(A.B$residuals, col = "blue", cex = 2)
plot(A.B$fitted.values ~ A.B$residuals)
library(car)
the.SWtest.tfa = shapiro.test(A.B$residuals)
the.SWtest.tfa
the.BFtest.tfa = leveneTest(A.B$residuals ~ Salary$Region,
center=median)
p.val.tfa = the.BFtest.tfa[[3]][1]
p.val.tfa
the.BFtest.tfa2 = leveneTest(A.B$residuals ~ Salary$Prof,
center=median)
p.val.tfa2 = the.BFtest.tfa[[3]][1]
p.val.tfa2

A = lm(Salary$Annual ~ Salary$Prof)
B = lm(Salary$Annual ~ Salary$Region)
#factor A test
anova(B, A.B)

#analysis section
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)
}
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)
  MAB = t(MAB)
  results = list(A = MA, B = MB, AB = MAB)
  return(results)
}

Salary.mean = find.means(Salary)
Salary.mean$AB
Salary.mean$A
Salary.mean$B
mean(Salary$Annual)

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)
  }
}
all.mult = find.mult(alpha = 0.05, a = 3, b = 2, dfSSE = 116, g =
2, group = "AB")
all.mult
Tuk = all.mult[2]
Bon = all.mult[1]

#contrast confidence intervals
A.cs1 = c(1, -1, 0)
A.cs2 = c(1, 0, -1)
A.cs3 = c(0, 1, -1)
```

```
B.cs1 = c(1, -1)
AB.cs = matrix(0, nrow = a, ncol = b)
AB.cs[2,1] = 1
AB.cs[2,2] = -1
AB.cs2 = matrix(0,nrow = a, ncol = b)
AB.cs2[3,1] = 1
AB.cs2[3,2] = -1
SSE = sum(A.B$residuals^2)
MSER = SSE/ (nT - a - b + 1)

#pairwise confidence intervals
scary.CI(Salary, MSE, equal.weights = TRUE,Tuk,"A",A.cs1) # data
scientists vs software engineers
scary.CI(Salary,MSE,equal.weights = TRUE,Tuk,"A",A.cs2) # data
scientists vs bioinformatics engineers
scary.CI(Salary,MSE,equal.weights = TRUE,Tuk,"A",A.cs3) # software
engineers vs bioinformatics engineers
scary.CI(Salary,MSE,equal.weights = TRUE,Tuk,"B",B.cs1) # san
francisco vs seattle
# CI
scary.CI(Salary,MSE,equal.weights = TRUE,Bon,"AB",AB.cs) #software
engineer S vs SF
scary.CI(Salary,MSE,equal.weights = TRUE,Bon,"AB",AB.cs2)
#bioinformatics engineer S vs SF

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(B, A.B)
Partial.R2(A, A.B)
```