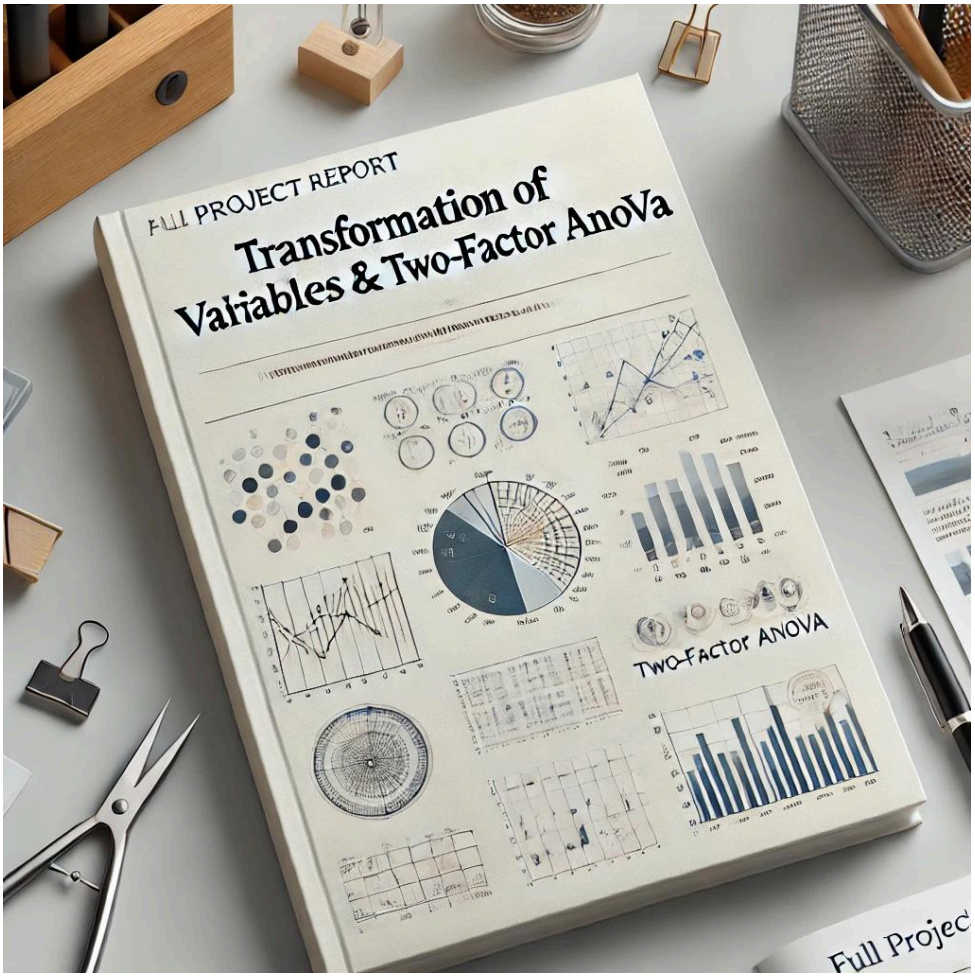
Transformation of Variables & Two-Factor ANOVA



Two Way ANOVA

Helicopter Frequency and Emergency Response Analysis Report

Part 1: Transformation of Variables

I. Introduction

The use of helicopters is essential in order to better assist law enforcement during times of emergency. The dataset 'Helicopter.csv' contains data from the sheriff's office on the times during the day helicopters were requested, by studying this dataset we can gain important understanding for the emergency response patterns present. By analyzing the frequency of dispatches across different shifts we can have better insight into better resource allocation and trends. Our data analysis focuses on transforming the data to make sure ANOVA assumptions are met using a significance threshold of 0.05.

II. Model Fit & Diagnostics

Our first goal was to fit an ANOVA model onto the dataset. The ANOVA model we used is $Y_{ij} = \mu_i + \epsilon_{ij}, i = 1, ..., 4, \epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. Through this model we are trying to determine if any shift has a different average number of times a helicopter was called compared to the other shifts. The μ_i represents the mean number of times the helicopter was called for the shift. The i represents each of the shifts, with I (between 2AM and 8AM), II (between 8AM and 2PM), III (between 2PM to 8PM), and IV (between 8PM to 2AM). In this model, ϵ_{ij} is the individual error for any given jth value in the ith group. Once we fit the model we obtained the following table

Source	DF	Sum Sq	Mean Sq	F-value	Pr(>F)
Shift	3	602.44	200.813	19.503	1.784e-09 (***)
Residuals	76	782.55	10.297		

Table 1.1 The ANOVA Table on the 'Helicopter.csv'

From the table it can be seen that since our p-value is quite low we reject the null hypothesis suggesting that there is significant evidence that at least one shift group is significantly different from others. However, we cannot be definitive about this conclusion as we have to ensure the assumptions of ANOVA are met.

In order to do so we tested the assumptions of independence of observations, groups of shifts are independent and the errors follow a normal distribution with homogeneity/constant variance. Since the dataset was given to us we will assume that the observations and groups are independent of each other. Therefore, we need to check the assumption for normality and constant variance. In order to test normality we created a QQ plot which using quantiles of the dataset compares to quantiles of a normal distribution. The QQ plot is as follows:

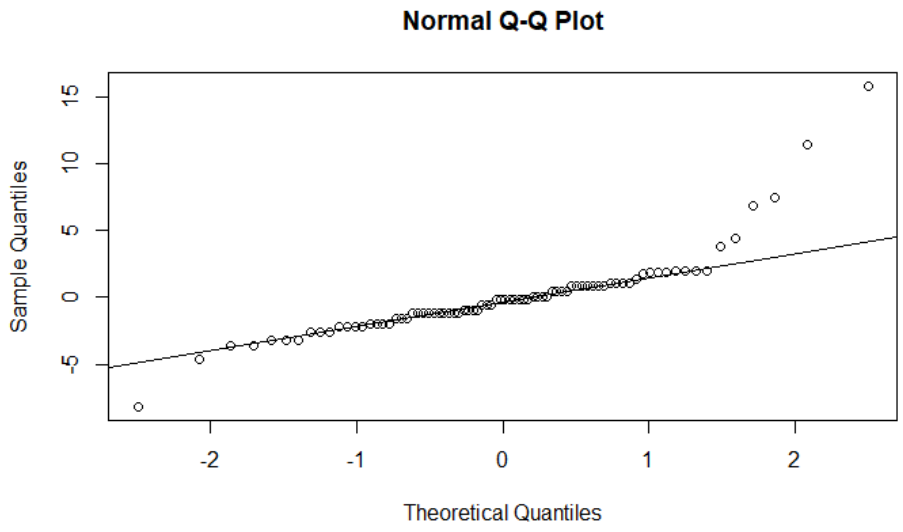


Figure 1.1 Normal QQ Plot of 'Helicopter.csv'

From the figure we can see that the data is mostly normal however there are points not following on the line especially towards the right. Since there is significant deviation it suggests that the data does not follow a normal distribution. We can additionally perform a Shapiro Wilk's test to

confirm that our data is not normally distributed. The null hypothesis would be that the error is normally distributed and the alternative hypothesis is the error is not normally distributed. Through R, we obtained a p-value of 3.945e-09 this means that we would reject the null since the p-value is so small, therefore the Shapiro Wilks test confirms that the data is not normally distributed and this ANOVA assumption is violated.

Next we had to test for homogeneity of variance, we did this by first creating a residual plot for the data.

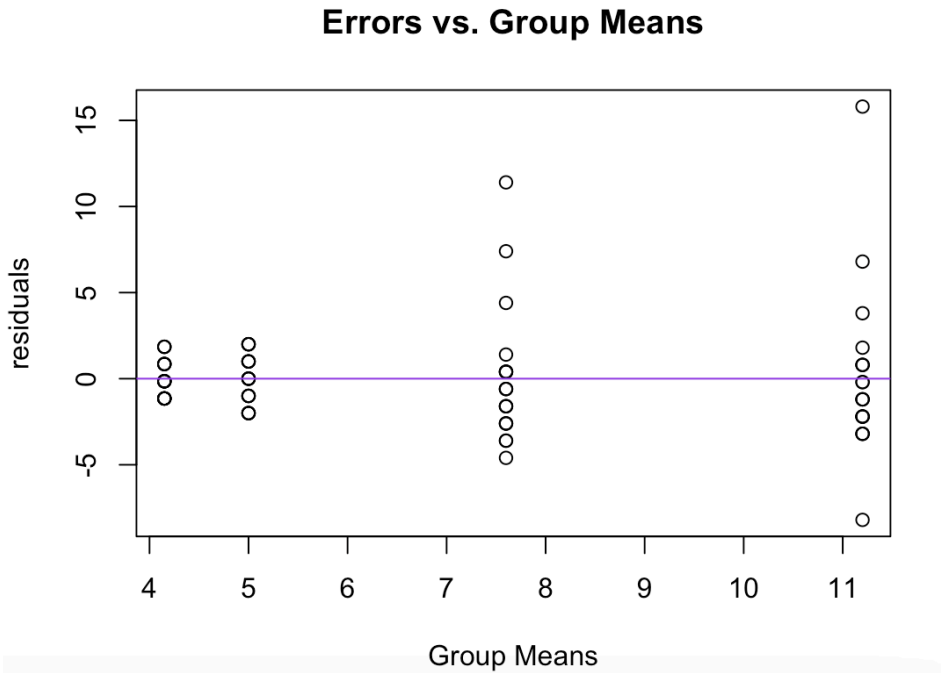


Figure 1.2 Residual Plot for Constant Variance

The plot helps us determine constant variance by plotting the errors vs each of the group means showing if there is equal spread of all the groups. Looking at the figure we can see that there seems to be an unequal distribution of the data's variability is different among all the group means; however, to confirm this we would need to perform a Brown Forsythe test. In order to do

so we defined our null that each of variances are equal whereas the alternative was that one or more of the variances is not equal. After computing this in R, we obtained that the p-value is 0.03185, therefore we would reject the null assuming a significance level of 0.05. This means that the variances are approximately not equal between the groups so this ANOVA assumption is also violated.

We next moved onto identifying the outliers so that we can do our best to transform the data to see if our assumptions are met. We determined the outliers by analyzing the standardized residuals. We calculated the residuals and used a threshold to determine which points were outliers. Through this we obtained that values at indices 15, and 67 are outliers, these points were (27, I) and (19, IV). Since these values were significantly different from the rest of the dataset we decided to create a new dataset with these outliers removed.

III. Data Transformations & Outlier Handling

Since our ANOVA assumptions were not met one way of making sure these assumptions are not violated is by performing transformations. There are a multitude of ways to do so such as log, square-root, or Box-Cox. We decided for our analysis we are going to perform a box-cox transformation on the dataset, in particular we utilized the Shapiro Wilks method.

We firstly performed the transformation on the dataset without the outliers removed in order to see if the transformation helped in improving the assumptions of our ANOVA model. We obtained a lambda value of -0.2322, fit our new ANOVA model, then performed our assumption checks. Through this we obtained that the p-value for the Shapiro Wilks test time was 0.109 since it is significantly higher than our threshold we fail to reject the null hypothesis that the

errors are normally distributed. This means there is significant evidence now that the data follows normality, this means our transformation improved our data. Next we checked if the variances are equal for this new model, to do so we again performed a Brown Forsythe test, the p-value was 0.654481 since we are assuming a threshold of 0.05 we can say we fail to reject the null hypothesis which is that the variances are approximately equal to one other. This means that there is significant evidence to suggest that the variances are equal and this ANOVA assumption is met.

The next dataset we analyzed was the one with our two outliers removed. We followed the same procedure of performing Shapiro Wilks box-cox method. We fit a new ANOVA model and performed our assumption checks once more. We obtained p-values of 0.04673 for the Shapiro Wilks test and 0.6589 for the Brown-Forsythe test. From this we can see that assuming a significance threshold value of 0.05 we reject the null hypothesis for Shaprio Wilks meaning that at least one of the errors is not normally distributed. Therefore this assumption of ANOVA is still violated. For the Brown-Forsythe test the null cannot be rejected therefore the variance is still equal.

IV. Discussion of Results

Our analysis focused on making sure that the dataset ‘Helicopter.csv’ met the assumptions of an ANOVA model. We found that our initial data failed to do so, therefore we performed a box cox transformation on the original dataset and one with the outliers removed. The Box-Cox transformation significantly improved the dataset without outlier removal, the Shapiro-Wilk test p-value increased allowing us to fail to reject the null hypothesis, suggesting that the transformed data was approximately normal. Additionally, the Brown-Forsythe test p-value was higher than

the significance level confirming that variances across shifts were now homogeneous. These results indicate that the transformation was successful in meeting ANOVA assumptions. For the dataset with outliers removed, the transformation had mixed effects; it did indeed raise the p-value to suggest equal variance; however, the normality was still violated. This result indicates that while outlier removal helped address the variance, it was not sufficient to fully normalize the dataset.

Overall these data transformations did help in the case with the outliers to satisfy ANOVA assumptions. The box-cox transformation did help in normalizing the data and equalizing the variance allowing us to fit an ANOVA model without the assumptions being violated. However, for the dataset with outliers removed, normality was still not fully achieved, suggesting that transformation alone was not enough in that case.

While transformations can be great to help fit data there are also some key downsides one must consider. One of the most prevalent downsides is the lack and change in interpretability, since the transformed data is no longer in the same units as the original there can be some real practical issues in interpreting what certain numbers mean and stand for, additionally it is not possible to simply undo the transformation such as squaring the number when a square root transformation is performed. Another limitation is that there are multiple different transformations present and to pick the best one additional testing on its particular effects may be needed. Lastly, a transformation may not have full effectiveness as seen by our dataset, by removing the outliers the transformation we used did not work therefore sometimes a certain transformation may not be enough. If a client was using our dataset we would have some suggestions, firstly to consider

performing a box-cox transformation on a dataset without outliers removed, second to perform the assumption checks before and after the transformation to ensure ANOVA can be applied, and finally to consider different transformations or different methods to perform ANOVA if a certain kind does not work.

In conclusion, while transformations can help improve a dataset to match ANOVA assumptions, there must be careful consideration of the different kinds of transformations and the limitations that follow suit. For the 'Helicopter.csv' we found that the best approach was to fit an ANOVA model on the transformed original data containing outliers.

Part 2: Two-Factor ANOVA

I. Introduction

Salary disparities within the technology sector play a crucial role in shaping career decisions, hiring practices, and economic policy development. Understanding how salaries vary based on profession and geographic region is essential for job seekers negotiating wages, employers designing competitive compensation packages, and policymakers addressing wage inequalities. This study examines whether profession and region significantly impact salary levels and investigates the potential interaction effect between these factors.

Using the dataset Salary.csv, we analyze salary data for technology professionals in San Francisco (SF) and Seattle (S) across three job titles: Data Scientist (DS), Software Engineer (SE), and Bioinformatics Engineer (BE). The primary objective is to determine whether salaries

vary based on profession, location, or their combination. A profession-based salary difference would suggest that compensation is influenced by specialized skill sets and market demand for certain roles. Regional differences in salary may reflect cost-of-living adjustments, economic conditions, or location-specific industry demand. The presence of an interaction effect would indicate that salary variations across professions are not uniform across locations, revealing broader industry and economic patterns.

To examine these relationships, this study applies Two-Factor ANOVA (Analysis of Variance), a statistical method used to evaluate the main effects of profession and region on salary while also testing for an interaction effect, all using an alpha of 0.05. This approach determines whether observed salary differences are statistically significant and whether profession and region independently or jointly influence earnings. By analyzing salary as the dependent variable, this study evaluates whether disparities arise primarily from job title, geographic location, or the combined influence of both factors.

This analysis provides valuable insights into salary structures within the technology industry. The findings contribute to a deeper understanding of wage distribution patterns, offering meaningful implications for professionals seeking equitable compensation, employers striving to remain competitive in the job market, and policymakers working toward fair labor practices. The following sections include data exploration, statistical modeling, hypothesis testing, and interpretation of results to assess the significance of these factors and their broader impact.

II. Data Summary

A good summary analysis is essential for our data analysis as it provides insights into the dataset structure and may lead to the orientation of the analysis and thus inferences. I will start our data summary by summarizing the mean, standard deviation, and group length of each group.

II - 1 Data Statistics

Prof (A) / Region(B)	San Francisco (SF)	Settle (S)
Data Scientist (DS)	117.769 (SD = 14.289, n = 20)	112.527 (SD = 12.839, n = 20)
Software Engineer (SE)	110.264 (SD = 10.552, n = 20)	95.549 (SD = 11.599, n = 20)
Bioinformatics Engineer (BE)	82.419 (SD = 10.521, n = 20)	79.755 (SD = 8.787, n = 20)

Table 2.1 : Summary statistics of the data: mean, standard deviation, and sample size for each treatment.

Table 1 shows the summary of data. The first column is Profession titles, which are Data Scientist (DS), Software Engineer (SE), and Bioinformatics Engineer (BE). The First row is Region titles, which are San Francisco (SF) and Settle (S). The Prof is Factor A, and the Region is Factor B. The statistics are represented in the table. For example, the mean annual salary for people who are Data Scientists and work in San Francisco is 117.769 thousand dollars. In this group, the standard deviation is 14.289 thousand dollars. Besides, all the groups are equally weighted as every group has the same sample size, which is 20.

The statistics from the mean, standard deviation, and group length show that the mean from the sample who are Data Scientists and work in San Francisco have the largest annual salary. Oppositely, those subjects who are Bioinformatics Engineer and work in Settle have the least amount of salary by comparing the mean. From the table, there is a trend that the mean annual salary decreases by the order from Data Scientist to Software Engineer to Bioinformatics Engineer, no matter the Region. Also, no matter their Profession titles, the annual salary mean from San Francisco is higher than that from Seattle. This suggests the idea that there is factor A and factor B effects. It also seems to suggest that there may be a slight interaction effect since when you look at the mean of Software Engineer, the average annual salary decreased by almost 15 when looking at San Francisco compared to Seattle, however, the other two professions decreased by 5 or less when looking at San Francisco compared to Seattle. To confirm this would require further testing.

II - 2 Boxplot of Annual Salary Index by Region

Salaries vary significantly based on profession and location. To better understand these variations, we present a boxplot that illustrates the annual salary distribution across different regions for three professional roles. This visualization helps identify salary trends, regional disparities, and potential outliers within these professions.

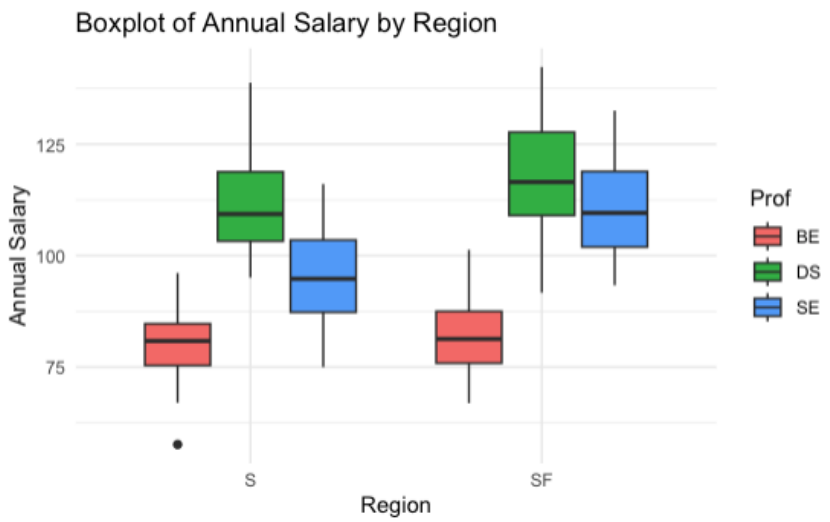


Figure 2.1 : Boxplot of Annual Salary by Professions and Regions

This boxplot shows the annual salary by region. The y-axis is Annual Salary and the x-axis is Region. In the figure, Red represents Bioinformatics Scientist (BS), green represents Data Scientist (DS), and blue represents Software Engineer. We observe that no matter the Region, Bioinformatics Engineer has the lowest annual salary, and the Data Scientist has the highest annual salary. Also, the boxplot shows that the annual salary for a Data Scientist and Software Engineer is higher in San Francisco than in Settle. Besides that, there is an outlier observed by the point at the bottom of the first red box, so in the group Bioinformatics Engineer and Seattle.

II - 3 Interaction Plot

To get a further glimpse into the relationship between Profession and Region we will create an interaction plot. This plot will allow us to see factor A effect, factor B effect, as well as interaction effects, if they are present in the data.

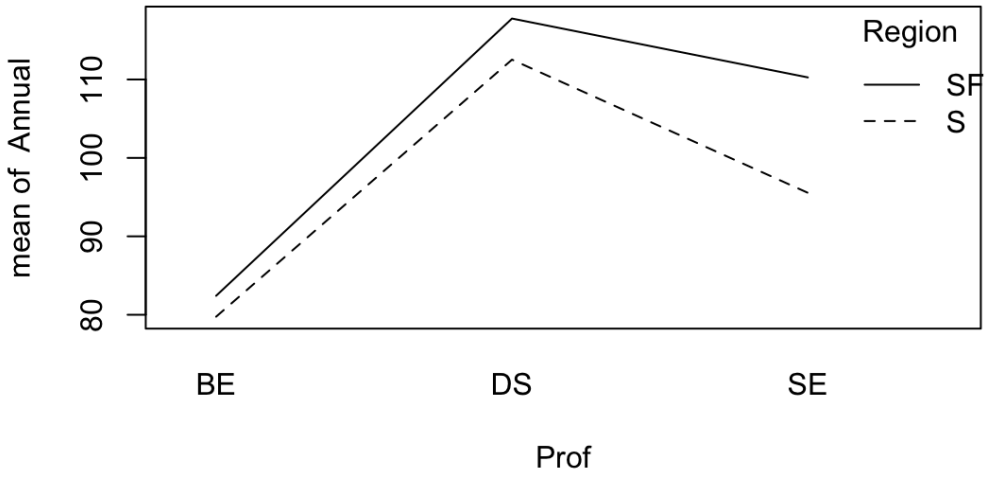


Figure 2.2 : Interaction Plot of Profession and Region

The interaction plot shows the relationship between Profession (BE, DS, SE) and estimated mean annual salary, with the Region (S vs. SF) as a grouping variable. Salaries increase significantly from BE to DS and then decline for SE in both regions, with SF consistently offering higher salaries than S. The non-parallel lines suggest a potential interaction effect. The decrease in average annual salary from DS to SE is much larger for Seattle compared to San Francisco. Overall, profession strongly influences salary, and region SF offers a salary advantage, particularly for DS professionals.

II - 4 Bar Plot of Means and Standard Errors

Finally, we will create a bar plot of the treatment means.

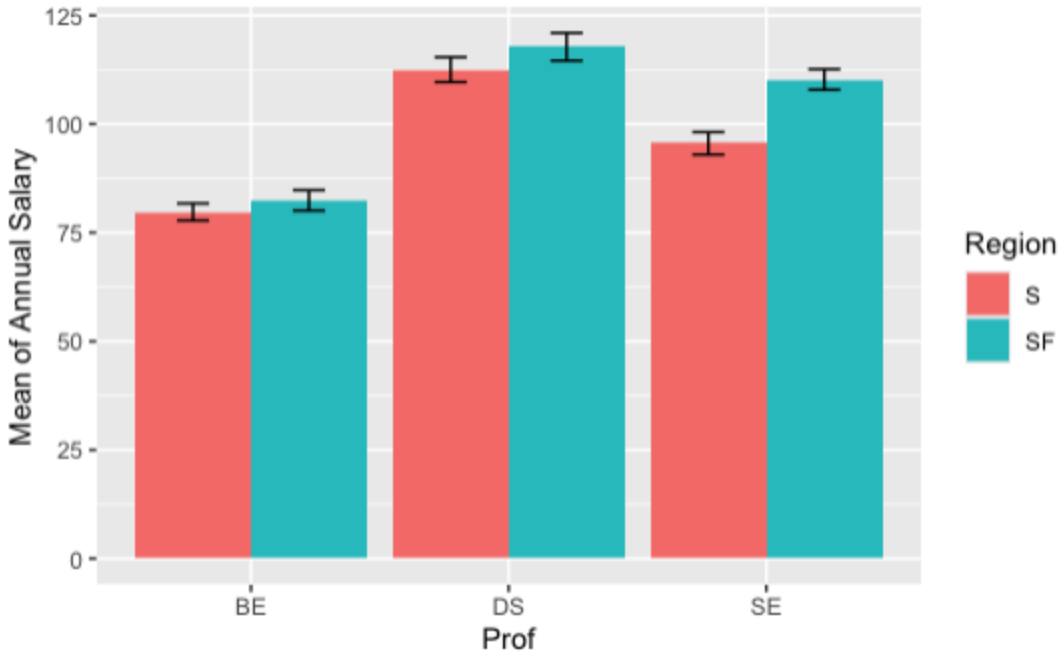


Figure 2.3 : Bar Plot of Means of treatment means

The bar plot illustrates the mean annual salary across different Professions (BE, DS, SE), categorized by Region (S and SF). Each bar represents the mean salary, with error bars indicating the standard error of the mean (SEM), providing insight into salary variability. In all three professions, Region SF (blue bars) generally has higher mean salaries compared to Region S (red bars). The salary difference is most pronounced in SE, where SF has a noticeably higher mean salary than S. The error bars suggest relatively low variability, implying consistent salary estimates within each group. This visualization highlights a regional disparity in salaries across professions, particularly in SE and DS, where SF offers a clear salary advantage.

III. Diagnostics

In this section, we will be finding our final model and checking the assumptions that come with that model.

III - 1 Model Fitting

The goal of this project is to see if profession and region affect salary. To find the best model to illustrate the relation between these three variables we conducted multiple hypothesis tests with an $\alpha = 0.05$, between a bigger model and a smaller model which is a subset of the bigger model. Before this we computed the partial R^2 values using the general formula:

$$R^2 = (M_F \mid M_R) = \frac{SSE(Reduced\ Model) - SSE(Full\ Model)}{SSE(Reduced\ Model}$$

The results of these calculations are in the table below:

R^2 (AB | A + B)	0.0501551
R^2 (A + B | B)	0.5972622
R^2 (A + B | A)	0.09602243

Table 2.2 : Partial R^2 values

The first row is the partial R^2 value of adding an interaction term to a model that already contains factor A, profession, and factor B, region. The result of this computation is the value 0.0501551, which tells us the proportion of reduction in error when adding an interaction term to our model with factor A and B separately. With such a low R^2 value, it suggests that it may not be as important in our model and is not necessary to add. In the second row, we can see how adding the effect of profession to our model that only contains regions reduces overall error by 59.73%, a decent amount of reduction. This suggests that factor A may be an important factor to have in our model. The last row then tells us that adding the region effect to a model already containing

profession reduces the overall error by 9.60%. This implies that factor B may not be as important in our model as factor A in reducing error but may still be beneficial to have.

Next, we used multiple hypothesis tests to confirm our thoughts and select the best model. We started off with fitting the biggest model, a model that contains both variables and an interaction term, and comparing it to the model that only contains the two variables separately. The null and alternative hypotheses and their corresponding models are as follows:

H_O: all $(\gamma\delta)_{ij} = 0$ Reduced Model: $Y_{ijk} = \mu.. + \gamma_i + \delta_j + \varepsilon_{ijk}$

H_A: at least one $(\gamma\delta)_{ij} \neq 0$ Full Model: $Y_{ijk} = \mu.. + \gamma_i + \delta_j + (\gamma\delta)_{ij} + \varepsilon_{ijk}$

γ represents factor A effect in this case, the effect of profession, and δ represents factor B effects which is region. The levels of profession are represented by "i" and run through DS, SE, and BE. The levels of the region are represented by "j" and run through SF and S.

Using the Anova function in R, we get this table listing all the p-values for hypothesis tests:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Prof	2	23814.6004	11907.3002	88.994837	0.0000000
Region	1	1705.7513	1705.7513	12.748738	0.0005228
Prof:Region	2	805.4077	402.7039	3.009798	0.0532358
Residuals	114	15252.9323	133.7977	NA	NA

Table 2.3 : ANOVA table

The F statistic for our test for interactions is 3.01, with a corresponding p-value of 0.0532358. The p-value tells us the probability of observing our data or more extreme under the null, which is that there is no interaction effect. Since the p-value is greater than our α of 0.05, we fail to reject the null hypothesis and conclude that all interaction terms equal 0; in other words, there is no interaction effect between profession and salary. This tells us the reduced model of this test is

a better fit. Next, we want to check for factor A and factor B effects separately. Starting of with testing for the effect of profession, the null and alternatives are:

H_O: all $\gamma_i = 0$ Reduced Model: $Y_{ijk} = \mu.. + \delta_j + \varepsilon_{ijk}$

H_A: at least one $\gamma_i \neq 0$ Full Model: $Y_{ijk} = \mu.. + \gamma_i + \delta_j + \varepsilon_{ijk}$

Looking at the table, the F-statistic for testing for factor A effects, profession, is 88.994837, with a p-value of 2.2e-16. This means the probability of observing our data or more extreme under the null that there is no factor A effect is 2.2e-16. With a p-value that is lower than our α, we reject the null hypothesis and conclude that there is a significant factor A effect. This corresponds with the full model being a better fit.

The final hypothesis test we will do is testing for factor B, region, and effects to see if it is necessary to include it in our model. The new nulls and alternative hypotheses are as follows:

H_O: all $\delta_j = 0$ Reduced Model: $Y_{ijk} = \mu.. + \gamma_i + \varepsilon_{ijk}$

H_A: at least one $\delta_j \neq 0$ Full Model: $Y_{ijk} = \mu.. + \gamma_i + \delta_j + \varepsilon_{ijk}$

Looking at that same table, we now get an F statistic of 12.7487 with a p-value of 0.0005228. Since the p-value for this test is also less than our α of 0.05, we reject the null hypothesis and conclude that there is a factor B effect and that it should be included in our model.

This leads to our final model being:

$$Y_{ijk} = \mu.. + \gamma_i + \delta_j + \varepsilon_{ijk}$$

Constraints:

$$\sum_i \gamma_i = 0 \qquad \sum_j \delta_j = 0$$

Where $\mu..$ represents the overall mean of the data regardless of group and γ_i represents factor A, profession, with "i" running through levels Data Scientist (DS), Software Engineer (SE), and Bioinformatics Engineer (BE). Factor B, region, is represented by δ_j, with "j" running through

levels San Francisco (SF) and Seattle (S). The index "k" represents the kth subject that is in the i-th and jth-group. "ε_{ijk}" represents the error that is associated with each individual subject in our data. The assumptions of this model are as follows: Y_{ijk} are random, levels of profession are independent, levels of region are independent, and ε_{ijk} follows a normal distribution with mean 0 and constant variance σ^2_{ε}. The first three assumptions of this model cannot be checked with tests so we will assume they have been satisfied and will now be moving on to check the assumptions associated with the errors in our model.

III - 2 Outliers

Having outliers in our data may cause issues with satisfying our assumptions of normality and constant variance in the errors so, we must first check for and remove outliers. We will be using the studentized method to determine outliers using an α of 0.05 in the formula for the cutoff, which is listed below:

$$\text{t.cutoff} = 1\text{-alpha}/(2*nt) \qquad df = nt - a - b + 1$$

```
named integer(0)
```

Figure 2.4 : Outlier Result

Running these calculations in R, we observe that no outliers exist, so we do not need to remove outliers. We can continue to check normality and equal variance of residuals.

III - 2 Checking Normality

We begin with a QQ plot to roughly observe the normality of our data.

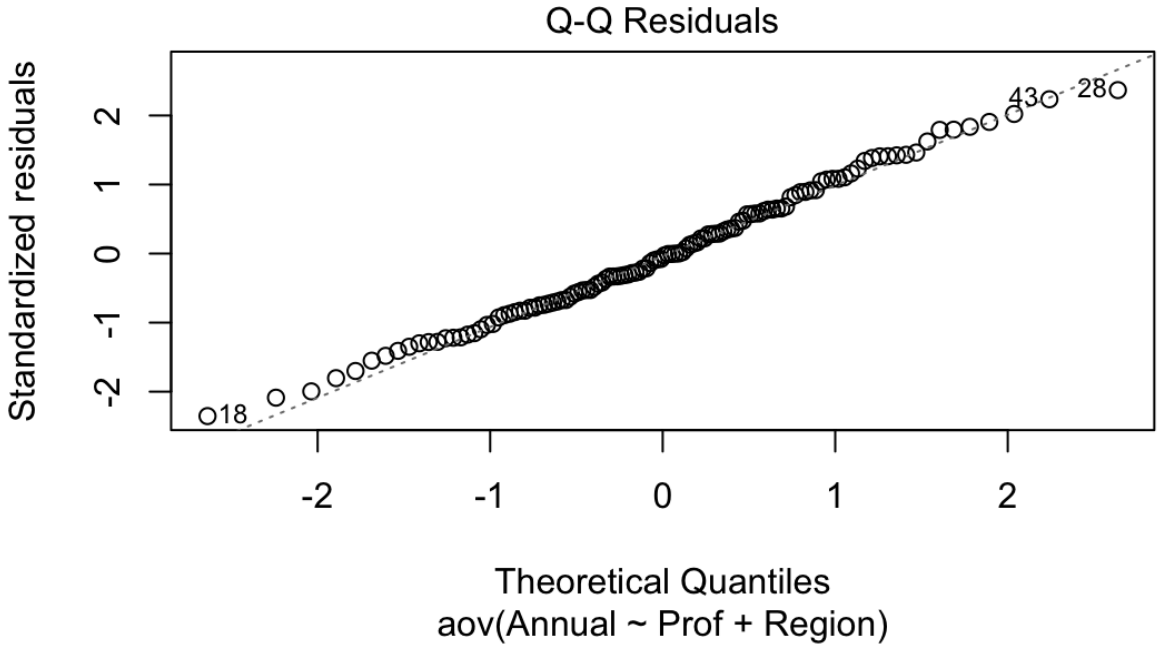


Figure 2.5 : QQ Plot

A QQ plot plots our data against what our data would be if it were a normal distribution. The QQ plot shows a pattern that suggests it satisfies normal distribution because nearly all the points are concentrated on one line.

We will now use a Shapiro-Wilk test to check and come to a conclusion on the normality of residuals.The hypothesis for a Shapiro-Wilk test are as follows:

H_O: the errors are normally distributed.

H_A: the errors are not normally distributed.

```
        Shapiro-Wilk normality test

data:  model.fit$residuals
W = .99146, p-value = 0.6698
```

Figure 2.6 : Shapiro Wilks Result

From the figure, we can see that the p-value for the Shapiro-Wilk test is 0.6698, which is greater than the significance level (alpha = 0.05), so we would fail to reject the null hypothesis and conclude that the residuals of our data follow the normal distribution. The normality of errors assumptions is met.

III - 3 Homogeneity of Variance

To check whether the residuals of our data follow equal variance, we can use a Residuals vs. Fitted Values plot to see the general trend.

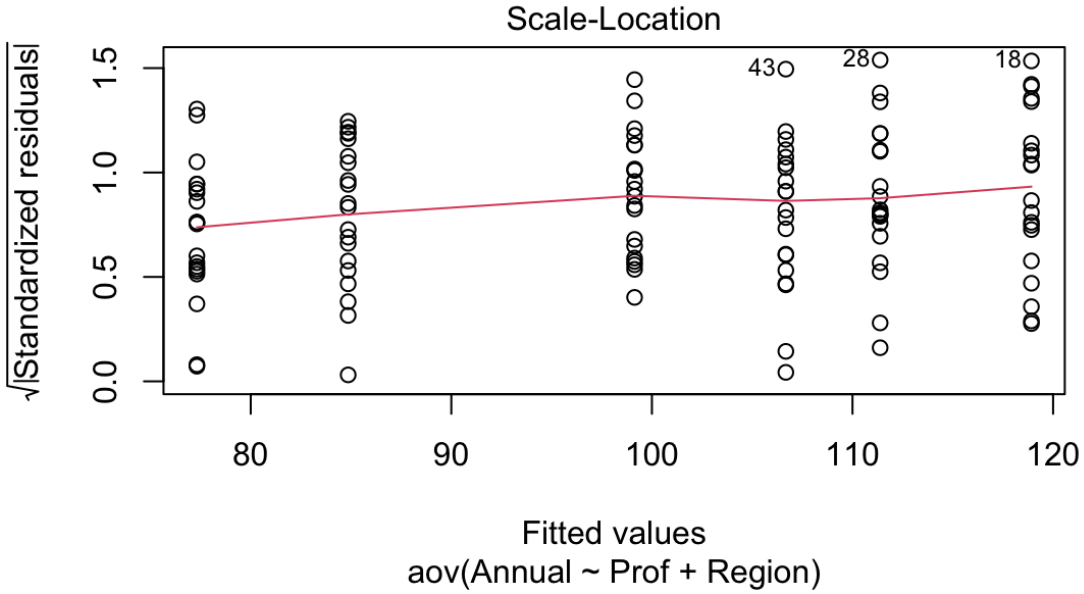


Figure 2.7 : Residuals vs. Fitted values plot

We observe that the red line in the middle is not straight from this plot, so it's hard to say the residuals of our data follow an equal variance pattern. We need to use hypothesis testing by Levene's Test to conclude.

Here is our hypothesis test:

H_O: $\sigma^2_{DS,\,SF} = \sigma^2_{SE,\,SF} = \sigma^2_{BE,\,SF} = \sigma^2_{DS,\,S} = \sigma^2_{SE,\,S} = \sigma^2_{BE,\,S}$

H_A: at least one = σ^2_{ij} not equal to the rest

Then, let's see the result of Levene's Test.

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   5  1.2188 0.3048
      114
```

Figure 2.8 Levene Test Result

From the result, we found that the p-value is 0.3048, which is larger than the significance level (alpha = 0.05), so we would fail to reject the null and conclude that all the residuals follow equal variance so, the equal variance assumption of ANOVA is met.

IV. Analysis and Interpretation

With our final model chosen and assumptions satisfied, we now want to get more information about how the true averages compare between groups. To do this, we will create six 95% confidence intervals, $\mu_{BE,S} - \mu_{BE,SF}$, $\mu_{DS,S} - \mu_{DS,SF}$, $\mu_{SE,S} - \mu_{SE,SF}$, $\mu_{.S} - \mu_{.SF}$, $\mu_{SE,S} - (\mu_{BE,S} + \mu_{DS,S})/2$, and $\mu_{SE,SF} - (\mu_{BE,SF} + \mu_{DS,SF})/2$, using the assumption of equal weights.

Creating multiple confidence intervals will increase our α or type I error. To combat this issue, we will be using a correction multiplier rather than the t value normally used when finding confidence intervals. There are a total of three multipliers, Bonferroni, Tukey, and Scheffe, but because we are creating group mean confidence intervals as well as treatment mean confidence intervals, we can only use a Bonferroni multiplier. The formula for the Bonferroni multiplier with degrees of freedom from our final model is as follows:

$$B = t_{1-\alpha/2g} \qquad \text{with } df = df(SSE)$$

The multiplier we get is 2.684, which we will be using for all six of the confidence intervals we make. We then calculated the three pairwise confidence intervals of treatment means using the following formula:

$$(\bar{y}_{ij.} - \bar{y}_{i'j'.}) \pm B\sqrt{MSE(\frac{1}{nij} + \frac{1}{ni'j'})}$$

After implementing this formula, we get:

Pairwise	Confidence Interval
$\mu_{BE, S} - \mu_{BE, SF}$	(-12.650575, 7.321990)
$\mu_{DS, S} - \mu_{DS, SF}$	(-15.227965, 4.744600)
$\mu_{SE, S} - \mu_{SE, SF}$	(-24.701656, -4.729091)

Table 2.4 : Pairwise Confidence Intervals for treatment means

The 95% simultaneous confidence interval for the true average difference in annual salaries between bioinformatics engineers in Seattle compared to San Francisco is between -12.650575 and 7.321990 thousand dollars. The 95% simultaneous confidence interval for the true average difference in annual salaries between data scientists in Seattle compared to San Francisco is between -15.227965 and 4.744600 thousand dollars. A 95% confidence interval tells us if we used multiple samples and created many intervals, 95% of the intervals will contain the true population mean. With these two intervals containing 0, it tells us that there is no significant difference in average salaries for Bioinformatics Engineers and Data Scientists who are in Seattle compared to San Francisco. The 95% simultaneous confidence interval for the true average difference in annual salaries for Software Engineers in Seattle compared to San Francisco is between -24.701656 and -4.729091 thousand dollars. Since this interval only contains negative

values, it tells us that the average salaries of Software Engineers in San Francisco are higher than those working in Seattle.

We will now compute the pairwise confidence interval for factor B using the formula listed below:

$$\sum_j c_j \hat{\mu}_{.j} \pm B \sqrt{\frac{MSE}{a^2} \sum_j cj^2 \left(\sum_i \frac{1}{nij}\right)} \qquad \sum_i c_i = 0 \qquad \text{with } \hat{\mu}_{.j} = \frac{1}{a} \sum_i \bar{y}_{ij.}$$

Plugging the values into the formula, we get the interval:

$\mu_{.S} - \mu_{.SF}$	(-13.306032, -1.774866)

Table 2.5 : Pairwise Confidence Interval for factor B means

The 95% simultaneous confidence interval for the true average difference in annual salaries between Seattle and San Francisco is between -13.306032 and -1.774866 thousand dollars. With this interval being negative, it tells us that the average annual salary of someone in San Francisco is higher than someone working in Seattle.

To compute the two contrast confidence intervals, we use the formula:

$$\sum_i \sum_j c_{ij} \bar{y}_{ij.} \pm B \sqrt{MSE \sum_i \sum_j \left(\frac{cij^2}{nij}\right)} \qquad \sum_i c_i = 0$$

Using this formula the result of the confidence intervals is listed in the table below:

Contrast	Confidence Interval
$\mu_{SE,\,S} - (\mu_{BE,S} + \mu_{DS,\,S})/2$	(-9.240624, 8.056124)
$\mu_{SE,\,SF} - (\mu_{BE,SF} + \mu_{DS,\,SF})/2$	(1.521762, 18.818510)

Table 2.6 : Contrast Confidence Intervals of treatment means

We are simultaneously 95% confident that the true average difference in annual salary of Software Engineers compared to the average annual salary of Bioinformatics Engineer and Data Scientist all from Seattle is between -9.240624 and 8.056124 thousand dollars. Since this interval contains 0, we can conclude that this contrast contains no statistically significant difference. We are simultaneously 95% confident that the true average annual salary of Software Engineers is higher than the average annual salary of Bioinformatics Engineer and Data Scientist in San Francisco by between 1.521762 and 18.818510 thousand dollars.

V. Conclusion

This study explored the impact of profession and geographic region on salary levels within the technology sector, analyzing whether these factors independently or jointly influence earnings. The results indicate that both profession and region significantly affect salary, but there is no statistically significant interaction effect between the two. This suggests that while salaries vary by job title and location, the relationship between profession and earnings remains consistent across regions.

Our findings emphasize that profession is the strongest predictor of salary, highlighting the role of specialized skills and market demand in determining compensation. Additionally, regional differences persist, likely reflecting cost-of-living variations and local industry conditions. These insights provide valuable guidance for job seekers in evaluating career prospects, employers in structuring competitive compensation packages, and policymakers in addressing wage disparities.

Despite these meaningful conclusions, this study has certain limitations. The dataset focuses on only two cities, which may limit its generalizability to broader geographic areas. Additionally,

factors such as experience, education, company size, and industry-specific demand were not included in the analysis but could provide further insights into salary disparities. Future research could benefit from a larger, more diverse dataset that incorporates these additional variables to develop a more comprehensive understanding of salary determinants in the technology industry.

Overall, this study enhances our understanding of salary structures, confirming the independent effects of profession and location on earnings. By identifying key trends, this research contributes to informed decision-making for professionals, businesses, and policymakers, ultimately fostering greater transparency and fairness in compensation practices.

Appendix

Part 1: Transformation of Variables

Model Fit:

```
#Part I: Transformation of Variables

#III Model Fit and Diagnostics

#ANOVA model fit
helicopter <- read.csv("C:/Users/vs24/Downloads/Helicopter.csv")
the.model = lm(Count ~ Shift, data = helicopter)
anova.table = anova(the.model)
anova.table

#QQ Plot & Shapiro Wilks
qqnorm(the.model$residuals)
qqline(the.model$residuals)

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

#Errors vs Groups Plot & Brown-Forsythe Test
plot(the.model$fitted.values, the.model$residuals, main = "Errors vs. Group Means",xlab =
"Group Means")
abline(h = 0,col = "purple")

boxplot(ei ~ Shift, data = helicopter)

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

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

```
alpha = 0.05
t.cutoff= qt(1-alpha/(2*nt), nt-a)
CO.eij = which(abs(eij.star) > t.cutoff)
CO.eij

rij = rstandard(the.model)
CO.rij = which(abs(rij) > t.cutoff)
CO.rij

#remove outliers
outliers = CO.rij
new.data = helicopter[-outliers,]
new.model = lm(Count ~ Shift,data = new.data)

#Outlier Data Transformation
#Shapiro-Wilks
boxcox(the.model ,objective.name = "Shapiro-Wilk")

L2 = boxcox(the.model ,objective.name = "Shapiro-Wilk",optimize = TRUE)$lambda


#Outlier Data with Shapiro Wilk
Y_SW_O = (helicopter$Count^(L2)-1)/L2
o_sw.data = data.frame(Count = Y_SW_O, Shift = helicopter$Shift)
o_sw.model = lm(Count ~ Shift,data = o_sw.data)

anova.table = anova(o_sw.model)
anova.table

qqnorm(o_sw.model$residuals)
qqline(o_sw.model$residuals)

eisw = o_sw.model$residuals
the.SWtest = shapiro.test(eisw)
the.SWtest

plot(o_sw.model$fitted.values, o_sw.model$residuals, main = "Errors vs. Group Means",xlab =
"Group Means")
abline(h = 0,col = "purple")
```

```
library(car)
the.BFtest = leveneTest(eiqq~ Shift, data=o_sw.data, center=median)
p.val = the.BFtest[[3]][1]
p.val
```

Diagnostics:

```
#IV Data Transformations & Outlier Handling

library(EnvStats)
#Shapiro-Wilks
boxcox(new.model ,objective.name = "Shapiro-Wilk")
L2 = boxcox(new.model ,objective.name = "Shapiro-Wilk",optimize = TRUE)$lambda

#Transformed Data with Shapiro Wilk
Y_SW = (new.data$Count^(L2)-1)/L2
t_sw.data = data.frame(Count = Y_SW, Shift = new.data$Shift)
t_sw.model = lm(Count ~ Shift,data = t_sw.data)

anova.table = anova(t_sw.model)
anova.table

qqnorm(t_sw.model$residuals)
qqline(t_sw.model$residuals)

eisw = t_sw.model$residuals
the.SWtest = shapiro.test(eisw)
the.SWtest

plot(t_sw.model$fitted.values, t_sw.model$residuals, main = "Errors vs. Group Means",xlab =
"Group Means")
abline(h = 0,col = "purple")

library(car)
the.BFtest = leveneTest(eiqq~ Shift, data=t_sw.data, center=median)
p.val = the.BFtest[[3]][1]
p.val
```

Part 2: Two way ANOVA

Summary:

```
# mean
aggregate(Annual~Prof+Region,
        data = salary,
        FUN = mean
        )
# SD
aggregate(Annual~Prof+Region,
        data = salary,
        FUN = sd
        )
# Group Length
aggregate(Annual~Prof+Region,
        data = salary,
        FUN = length
        )

# Boxplot
ggplot(salary, aes(x = Region , y = Annual, fill = Prof)) +
  geom_boxplot() +
  labs(title = "Boxplot of Annual Salary by P",
       x = "Region",
       y = "Annual Salary") +
  theme_minimal()

# Interaction Plot
with(
  salary,
  interaction.plot(Prof, Region, Annual)
)

# Barplots
library(Rmisc)
# compute mean and standard error of the mean by subgroup
summary_stat <- summarySE(salary,
  measurevar = "Annual",
  groupvars = c("Prof", "Region")
)
ggplot(
```

```
  subset(summary_stat, !is.na(Region)), # remove NA level for sex
  aes(x = Prof, y = Annual, fill = Region)
) +
  geom_bar(position = position_dodge(), stat = "identity") +
  geom_errorbar(aes(ymin = Annual - se, ymax = Annual + se), # add error bars
    width = 0.25, # width of error bars
    position = position_dodge(.9)
  ) +
  labs(y = "Mean of Annual Salary")
```

Diagnostics:

```
AB = lm(Annual ~ Prof * Region, data = Salary)
A.B = lm(Annual ~ Prof + Region, data = Salary)
A = lm(Annual ~ Prof, data = Salary)
B = lm(Annual ~ Region, data = Salary)

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

anova(A.B, AB)
anova(B,AB)
anova(A,AB)

anova.results = anova(AB)

knitr::kable(anova.results, caption = "Anova Results")

the.model = A.B

a = length(unique(Salary$Prof))
b = length(unique(Salary$Region))
alpha = 0.05
```

```
nt = length(Salary$Annual)

t.cutoff = qt(1-alpha/(2*nt), nt - a - b + 1)
rij = rstandard(the.model)
CO.rij = which(abs(rij) > t.cutoff)
outliers = CO.rij
outliers

model.fit <- aov(Annual ~ Prof + Region,
  data = salary
)

library(car)

plot(model.fit,which = 2)

shapiro.test(model.fit$residuals)

plot(model.fit, which = 3)

leveneTest(Annual~Prof*Region, data = salary)
```

Analysis:

```
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.multi = find.mult(alpha = 0.05, a = 3, b = 2, dfSSE = nt - 3 - 2 + 1, g = 6, group = "AB")
Bon = all.multi[1]

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)
}

the.means = find.means(Salary)
SSE = sum(the.model$residuals^2)
MSE = SSE/(nt - a - b + 1)

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)
}
}

# Pairwise between Bioinformatics in Seattle vs. San Francisco
A.B.cs.1 = matrix(0, nrow = a, ncol = b)
the.means$AB
A.B.cs.1[1,1] = 1
A.B.cs.1[1,2] = -1
scary.CI(Salary, MSE, equal.weights = TRUE, Bon, "AB", A.B.cs.1)

# Pairwise between Data Science in Seattle vs. San Francisco
A.B.cs.2 = matrix(0, nrow = a, ncol = b)
A.B.cs.2[2,1] = 1
A.B.cs.2[2,2] = -1
scary.CI(Salary, MSE, equal.weights = TRUE, Bon, "AB", A.B.cs.2)

# Pairwise between Software in Seattle vs. San Francisco
A.B.cs.3 = matrix(0, nrow = a, ncol = b)
A.B.cs.3[3,1] = 1
A.B.cs.3[3,2] = -1
scary.CI(Salary, MSE, equal.weights = TRUE, Bon, "AB", A.B.cs.3)

# Pairwise between Seattle and San Francisco
A.B.cs.4 = c(1, -1)
scary.CI(Salary, MSE, equal.weights = TRUE, Bon, "B", A.B.cs.4)

# Contrast between DS Seattle and average of BE and SE Seattle
A.B.cs.5 = matrix(0, nrow = a, ncol = b)
A.B.cs.5[2,1] = -1/2
A.B.cs.5[1,1] = -1/2
A.B.cs.5[3,1] = 1
scary.CI(Salary, MSE, equal.weights = TRUE, Bon, "AB", A.B.cs.5)

# Contrast between SD SF and average between BE and SE SF
A.B.cs.6 = matrix(0, nrow = a, ncol = b)
```

```
A.B.cs.6[2,2] = -1/2
A.B.cs.6[1,2] = -1/2
A.B.cs.6[3,2] = 1
scary.CI(Salary, MSE, equal.weights = TRUE, Bon, "AB", A.B.cs.6)
```