STA 106 Project II Group 14

By: Jiawei Zhu, Suzuran Schnyder, Jhaydine Bandola, Thomas Philip, Zahir Sabbah, Jiaqi Wang

Instructor: Maxime Guiffo Pouokam

2025-11-25

Topic I : Transformation of Variables

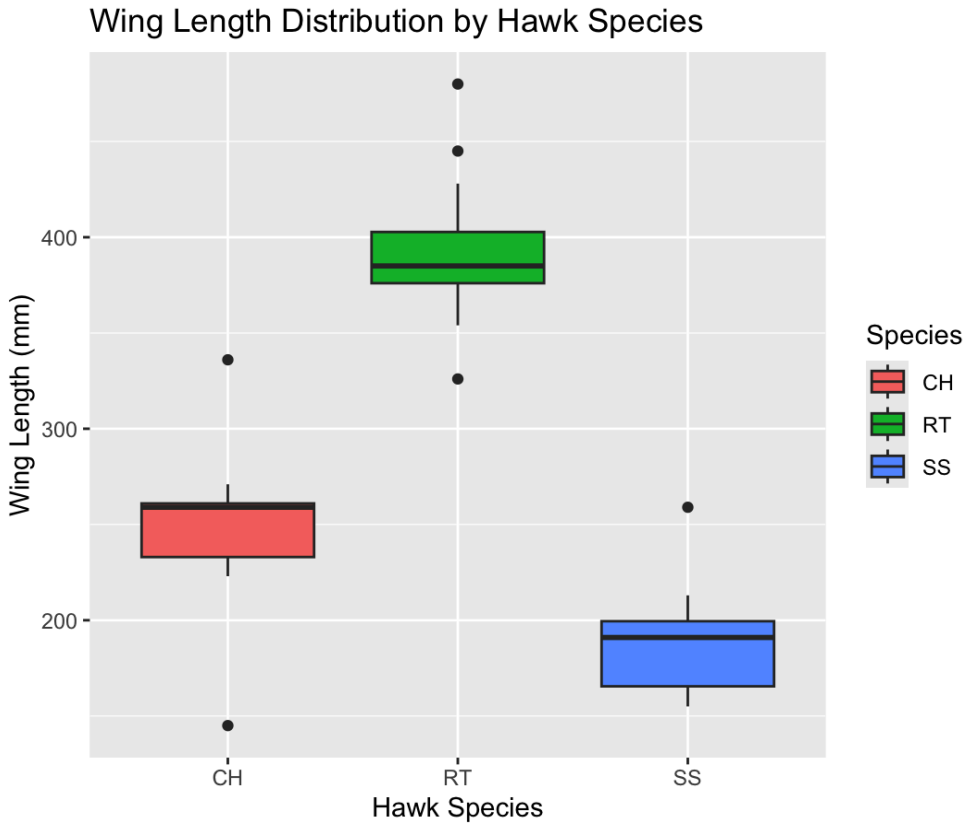
Question 2 - NewHawk.csv

A.

Hawk wing lengths differ across different species. In this report we will be analyzing the differences, if any, between the wing lengths of Cooper's, Red-tailed, and Sharp-Shinned hawks.

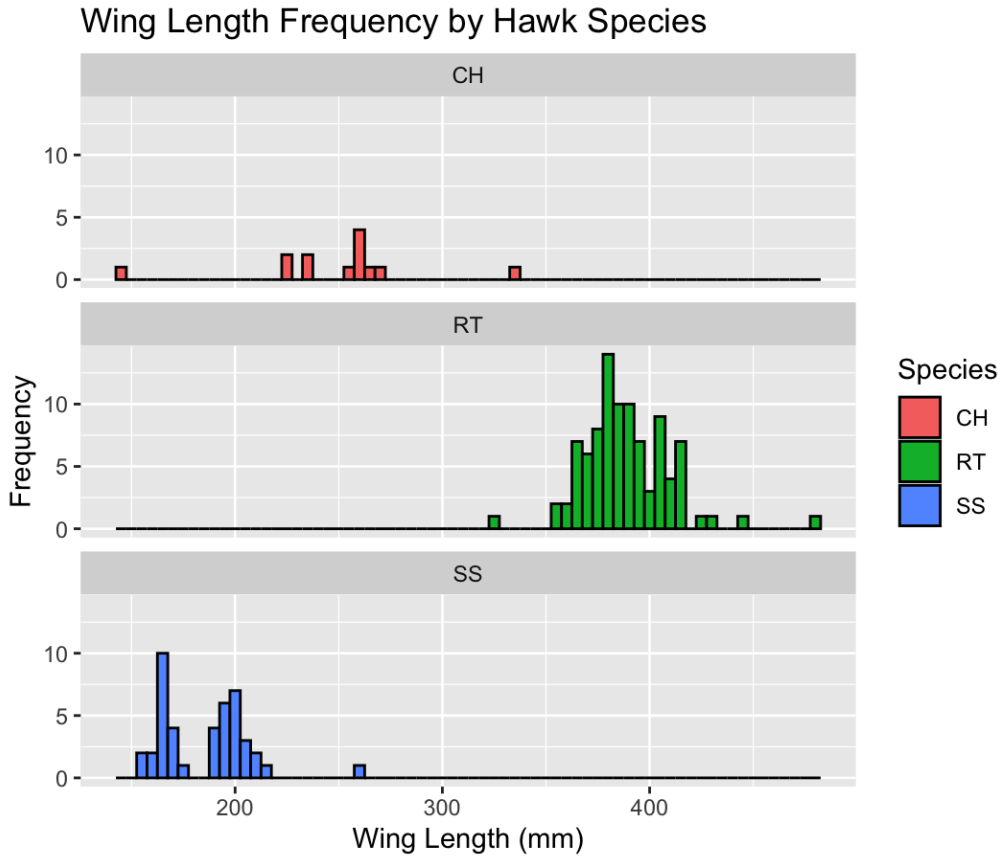
B.

Figure 1.2.1



These boxplots show a rough distribution of wing length (mm) by species of hawk. The species with the largest average wing length are Red-tailed (RT) hawks (~380mm), followed by Cooper's (CH) hawks (~260mm), and then Sharp-Shinned (SS) hawks (~190mm). All three species of hawks have outliers and similar spread, with 50% of the data for each species falling in a 30-35mm range (375mm-405mm for Red-tailed, 230mm-260mm for Cooper's, 165mm-200mm for Sharp-Shinned).

Figure 1.2.2



These grouped histograms show the distribution of wing length (in mm) by species of hawk. Cooper's and Red-tailed hawks have an approximately normal unimodal distribution,

although there are not many data points of Cooper's hawks. Sharp-shinned hawks have an approximately normal bimodal distribution. There are possible outliers for Cooper's Hawks at around 150mm and 340mm, for Red-tailed hawks at close to 500mm, and for Sharp-shinned hawks at 260mm. We can see a similar pattern in Figure 1.2.1 of Red-tailed hawks having the largest average wing length, followed by Cooper's hawks, with Sharp-shinned hawks having the smallest.

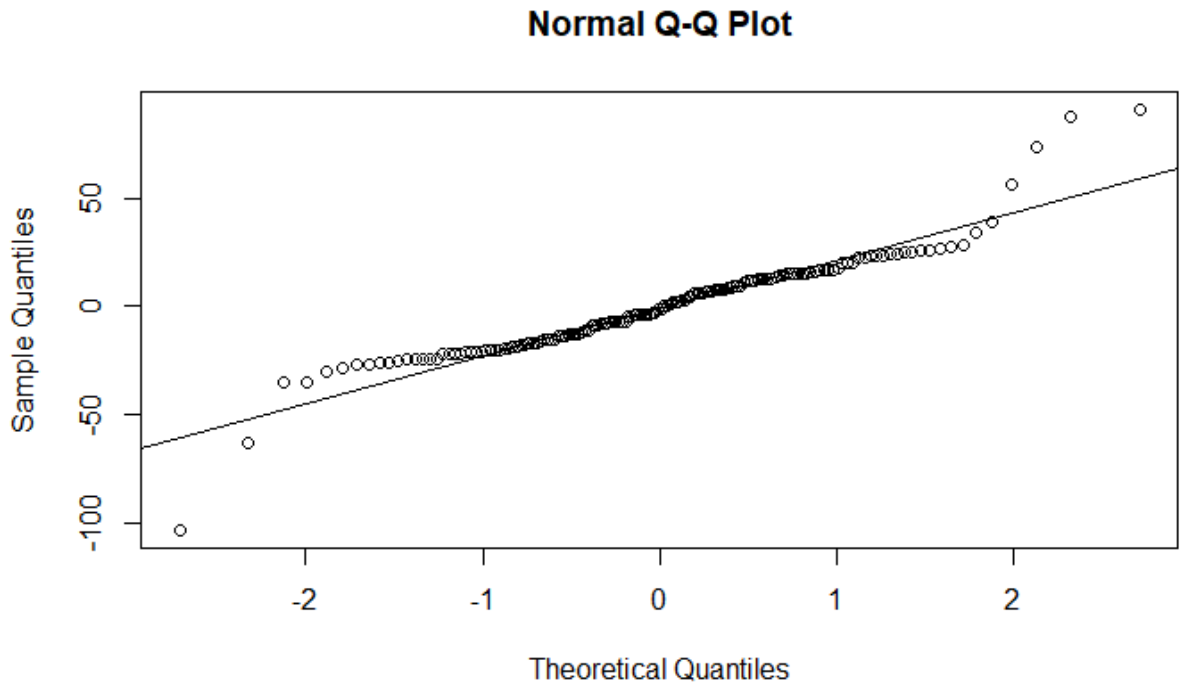
Table 1.2.1

Species	Mean Wing Length (mm)	Standard Deviation	Sample Size
CH	248	42.1	13
RT	389	20.8	94
SS	185	21.6	43

This table summarizes our observed data by hawk species: Cooper's, Red-tailed, and Sharp-Shinned. We can see that Red-tailed hawks have the largest mean wing length (389mm), followed by Cooper's hawks (248mm), then by Sharp-Shinned (185mm) with the lowest. We can also see that Cooper's hawks have significantly higher variance than Red-tailed hawks (42.1mm standard deviation to 20.8mm), with more than double the standard deviation, and nearly double that of the Sharp-Shinned hawks (21.6mm). There is also a clear disparity in sample size, as there were only 13 Cooper's hawks observed in this dataset, while the other two species had more than 40 subjects observed.

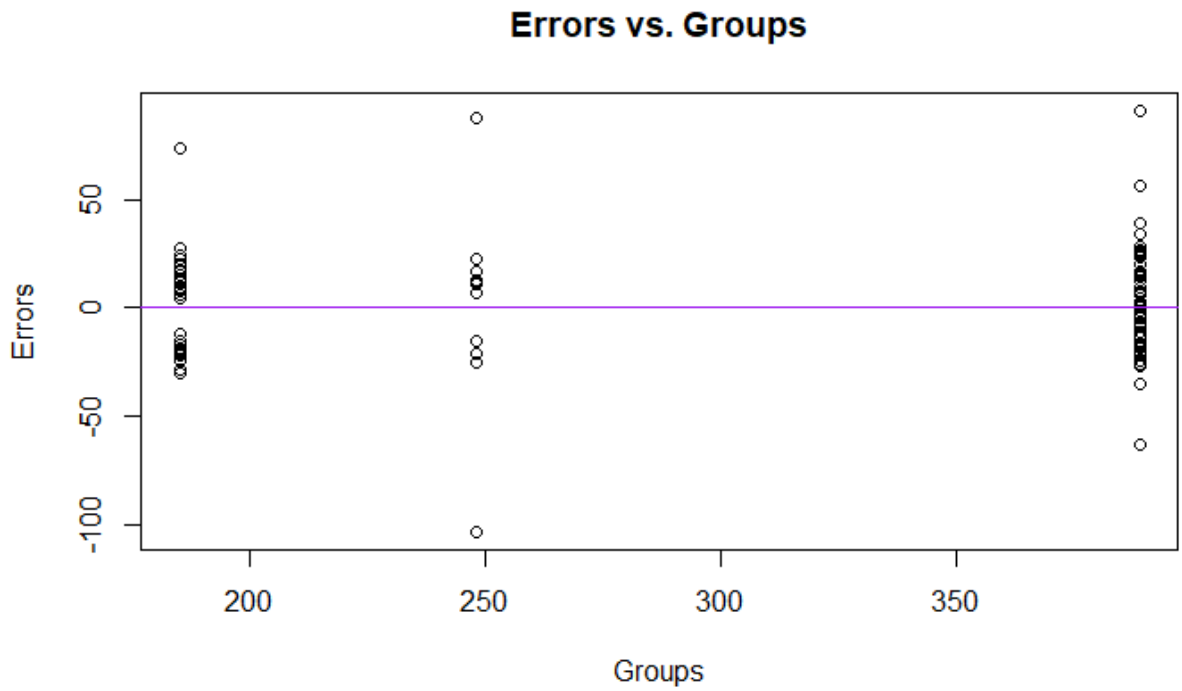
Before we can fit an ANOVA model, we must first check to see if the data meet the following core assumptions: normality of errors and homogeneity of errors amongst groups. We first plot the QQ-plot and the Errors vs. Fitted Values plots to subjectively assess if the data fails to meet the two core assumptions.

Figure 1.2.3



Clearly, towards the tail ends of the graph (Figure 1.2.3), the points begin to deviate from the qq-line. This indicates to us that errors are not normally distributed.

Figure 1.2.4



The errors vs fitted values of the groups plot (Figure 1.2.4) appear to display slightly differing spreads of error data. This tells us that the data does not display constant variance amongst group errors. Although both these diagnostic graphs indicate to us that the data fail to meet two core assumptions for ANOVA, we will need to perform a more objective analysis on whether or not the data meet these criteria. Therefore, we will conduct a Shapiro-Wilk test and the Brown-Forsythe test, setting alpha to 0.05.

Shapiro-Wilk Test:

Again, we will use the Shapiro-Wilk test to see if the errors of our data follow a normal distribution using the hypothesis:

H_0: The errors are normally distributed

H_a: The errors are not normally distributed

Table 1.2.5: Shapiro-Wilk Test for Errors of Wing Length and Species

W-Statistic	0.91732
P-Value	1.431e-07

At alpha = 0.05, we strongly reject the null hypothesis and conclude that the errors are not normally distributed.

Brown-Forsythe Test:

Again, the Brown-Forsythe test will determine if the errors follow a constant variance among the Species groups. Here are the following null and alternative hypotheses.

H_0: The variances among the species groups are equal

H_a: At least one of the group variances is not equal to the others

Table 1.2.6: Brown-Forsythe Test for Constant Variance

F-Statistic	2.3481
P-Value	0.09913

At alpha = 0.05, we fail to reject the null hypothesis and conclude that the errors follow constant variance amongst species groups.

On the one hand, our Brown Forsythe test provides results that support one assumption of ANOVA. However, the p-value tells us that the probability of observing our data, given we assume constant errors among groups, is still relatively low. On the other hand, our SW-test

provides extremely convincing results that our data fails to meet the ANOVA assumption of normally distributed errors. This tells us that outliers may need to be removed, transformations need to be made, or a combination of both. These modifications will hopefully allow the data to meet the necessary ANOVA assumptions.

C.

As established before, the original data fails to meet the assumptions of ANOVA. Therefore, our first step is to check for outliers. As we can see from Figure 1.2.1, there appears to be points that deviate from most of the data. However, we want to use a more systematic approach to detect outliers. We will detect outliers using semi-studentized and standardized residuals.

Semi-Studentized Residuals

By using semi-studentized residuals approach, we assume equal variance among species groups. This is a fair assumption for us to make because of our Brown-Forsythe test. We estimate the z-score of the errors:

$$e_{ij}^{*} = (e_{ij} - 0)/\sqrt{MSE} \sim t_{df=n_t-a}$$

Using $t_{1-a/2n}$, we test which errors are too 'unusual'. We find that rows 58, 68, 122 of our data are outliers. After removing these outliers from the data, we again test the data to see if it meets the ANOVA assumptions of constant variance of group errors and normality of error distribution.

Shapiro-Wilk Test:

H_0: The errors are normally distributed

H_a: The errors are not normally distributed

Table 1.2.6: Shapiro-Wilk Test for Errors of Wing Length and Species

W-Statistic	0.97182
P-Value	0.004021

At alpha = 0.05, we strongly reject the null hypothesis and conclude that the errors are not normally distributed.

Brown-Forsythe Test:

H_0: The variances among the species groups are equal

H_a: At least one of the group variances is not equal to the others

Table 1.2.7: Brown-Forsythe Test for Constant Variance

F-Statistic	0.9823
P-Value	0.3769

At alpha = 0.05, we fail to reject the null hypothesis and conclude that the errors follow constant variance amongst species groups.

Although the Shapiro-Wilk test on outlier-removed data gives the same conclusion as before, the new test's p-value increased from our first SW-test. This indicates the probability of observing our data, assuming normality of error, is higher after removing outliers. The Brown-Forsythe test for the outlier-removed data also resulted in a higher p-value, increasing the probability that our data has constant variance of errors among groups.

Our next steps include the following: transforming the outlier-removed data and transforming the original data. We will conduct Shapiro-Wilk and Brown-Forsythe tests on both transformed data. Whichever data meet the two needed error assumptions and have the highest p-values will be the final data we use.

Box-Cox Transformation on Original Data:

The following is the equation of the Box-Cox transformation. The Box-Cox transformation utilizes a grid search to find the best lambda for the best transformation of our data.

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda}, & \lambda \neq 0, \\ \ln(Y), & \lambda = 0. \end{cases}$$

The three ways of defining the "best" lambda include the following: the lambda that maximizes correlation of data to the QQ-line, the lambda that maximizes p-value of the Shapiro-Wilk test, and the log-likelihood method. Once this lambda is found, the y-data is modified using the above equation. We will calculate the lambda using all three methods and conduct the SW-tests and BF-Tests on all three transformations. Here are the results for alpha = 0.05.

Box-Cox Transformation "QQ-line" Method:

Lambda Value	1.46001
Shapiro-Wilk P-Value	3.473e-07
Brown-Forsythe P-Value	0.2001

Box-Cox Transformation "Shapiro-Wilk" Method:

Lambda Value	1.449575
Shapiro-Wilk P-Value	3.476e-07
Brown-Forsythe P-Value	0.2035

Box-Cox Transformation "Log-Likelihood" Method:

Lambda Value	2
Shapiro-Wilk P-Value	8.275e-08
Brown-Forsythe P-Value	0.02109

Box-Cox Transformation on Outlier-removed Data:

We again do the Box-Cox transformation, this time on the data after the outliers are removed.

Box-Cox Transformation "QQ-line" Method:

Lambda Value	1.612634
Shapiro-Wilk P-Value	0.01156
Brown-Forsythe P-Value	0.1778

Box-Cox Transformation "Shapiro-Wilk" Method:

Lambda Value	1.701168
Shapiro-Wilk P-Value	0.01179
Brown-Forsythe P-Value	0.0893

Box-Cox Transformation "Log-Likelihood" Method:

Lambda Value	2
Shapiro-Wilk P-Value	8.275e-08
Brown-Forsythe P-Value	0.02109

From all the possible lambda transformations, "QQ-Line" Box-Cox transformation with outlier-removed data provides the best combination of Shapiro-Wilk and Brown-Forsythe p-values.

D.

From our analysis, we found that transforming the data did improve our Shapiro-Wilk and Brown-Forsythe p-values, but we had to do so through a Box-Cox Transformation, with the "QQ-line" Method, after removing outliers. A downside to this transformation of the data is that we removed 3 of 150 observations, which may reduce representativeness if those observations reflect true biological variability. The other downside is that using the Box-Cox transformation makes it more difficult to interpret the data. While we do believe that the transformation of the data creates a better fit, I would caution a client to use these transformations (removing outliers, then doing a Box-Cox transformation with the "QQ-line" Method) with the knowledge that the

Shapiro-Wilk test still rejects normality at $\alpha = 0.1$ and 0.5. . This means that, while it is our best transformation, at a 0.1 or 0.05 significance level we would reject the null hypothesis that the residuals have a normal distribution.

Topic II (30pts): Two Factor ANOVA

Question 1 - Salary.csv

I. Introduction

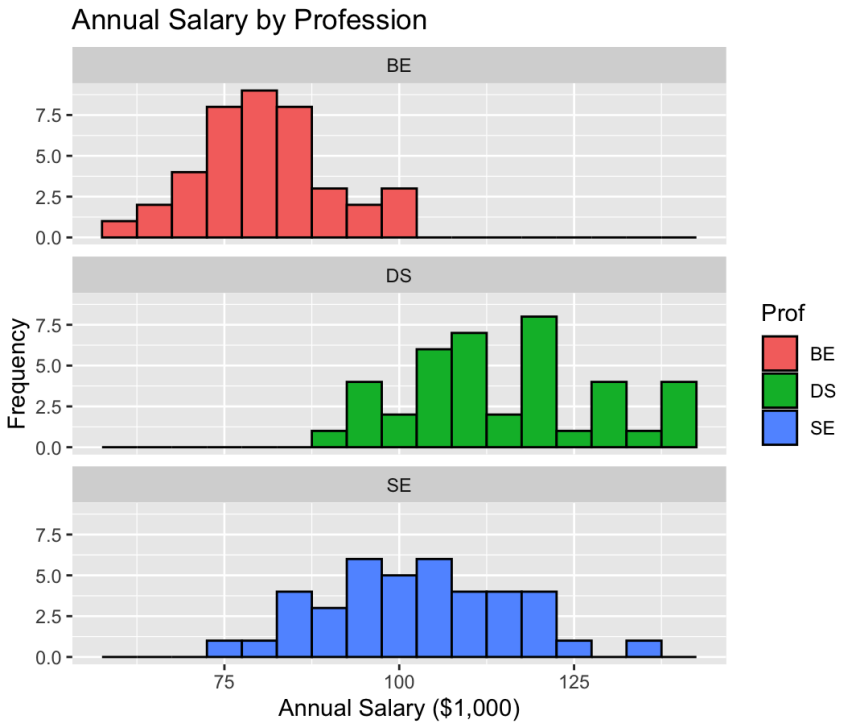
Place and type of work are both predictors for annual salary, making them important variables to consider when moving and starting a career. In this report we examine annual salary data in three occupations: Data Scientist, Software Engineer, and Bioinformatics Engineer. The dataset includes data from tech-hubs in the United States: Seattle and San Francisco. This allows us to analyze differences in annual salary by geography and occupation. The primary objective of this analysis is to determine whether there is a statistically significant effect on annual salary from location, occupation, or the interaction between location and occupation. To address this objective, we conduct a two-factor ANOVA to evaluate main effects and a possible interaction effect. Prior to conducting the two-factor ANOVA we visually analyzed the distribution of annual salaries by city and occupation. This was done in order to help clarify if city and occupation were good predictors, and if a combination of the two factors had a different effect that couldn't be explained by either one individually.

II. Summary of Data

Figure 2.2.1

This grouped histogram shows the distribution of salaries by city, Seattle (S) and San Francisco (SF). For the distribution of annual salaries in Seattle, the data appears to be normally distributed, with a large majority of the data being in the $75,000 to $125,000 range. For the distribution of annual salaries in San Francisco, the data also appears to be approximately normal. The majority of the annual salaries in San Francisco is similar to Seattle's, with a large majority being in the $75,000 to $125,000 range. Both cities appear to have similar spread in annual salaries, with neither having clear outliers.

Figure 2.2.2



This grouped histogram shows the distribution of annual salaries by occupation, specifically Bioinformatics Engineer (BE), Software Engineer (SE), and Data Scientist (DS). All three occupations appear to have an approximately normal distribution, but the distribution of salaries for data scientists also appears to be multimodal, as there are multiple peaks in frequency of certain salary intervals ($107,500 to $112,500 and $117,500 to $122,500). A large majority of the salaries for bioinformatics engineers are in the range of $70,000 to $90,000, $90,000 to $135,000 for data scientists, and $80,000 to $120,000 for software engineers. The spread of the salaries for bioinformatics engineers appears to be the lowest, followed by salaries of data scientists, and salaries of software engineers having the largest spread. It isn't clear if there are outliers without a formal test, but there may be some bioinformatics engineers with a salary of $55,000 or $100,000 labeled as such because of the generally low variance of the data.

Figure 2.2.3

This grouped boxplot shows salaries per occupation by city that the subjects work in. The salaries of bioinformatics engineers are very similar between those who work in Seattle and San Francisco, the only difference is that there is an outlier in Seattle making around $55,000, while the median is around $82,500. The median salary of data scientists in San Francisco ($115,000) is slightly higher than those who work in Seattle ($107,500), but a test would have to be done to see if there is a statistically significant difference for the overall population. For software engineers, there is a more clear difference, as those who work in San Francisco have a significantly higher median annual salary ($110,000) than the median of those who work in Seattle ($92,500).

Table 2.2.1: Mean Annual Salary by City and Position

Prof/Region	S	SF
BE	79754.85	82419.14
SE	95548.75	110264.12
DS	112527.15	117768.83

This table shows the mean annual salary for every combination of profession and city. The highest average salary among this combination of factors are software engineers who work in San Francisco. We also see that jobs in San Francisco pay better than in Seattle for these occupations, especially for software engineers, where the difference in annual salary is nearly $15,000. We also note that data scientists have the highest mean salary, followed by software engineers, and bioinformatics engineers with the lowest.

Table 2.2.2: Standard Deviation of Annual Salary by City and Position

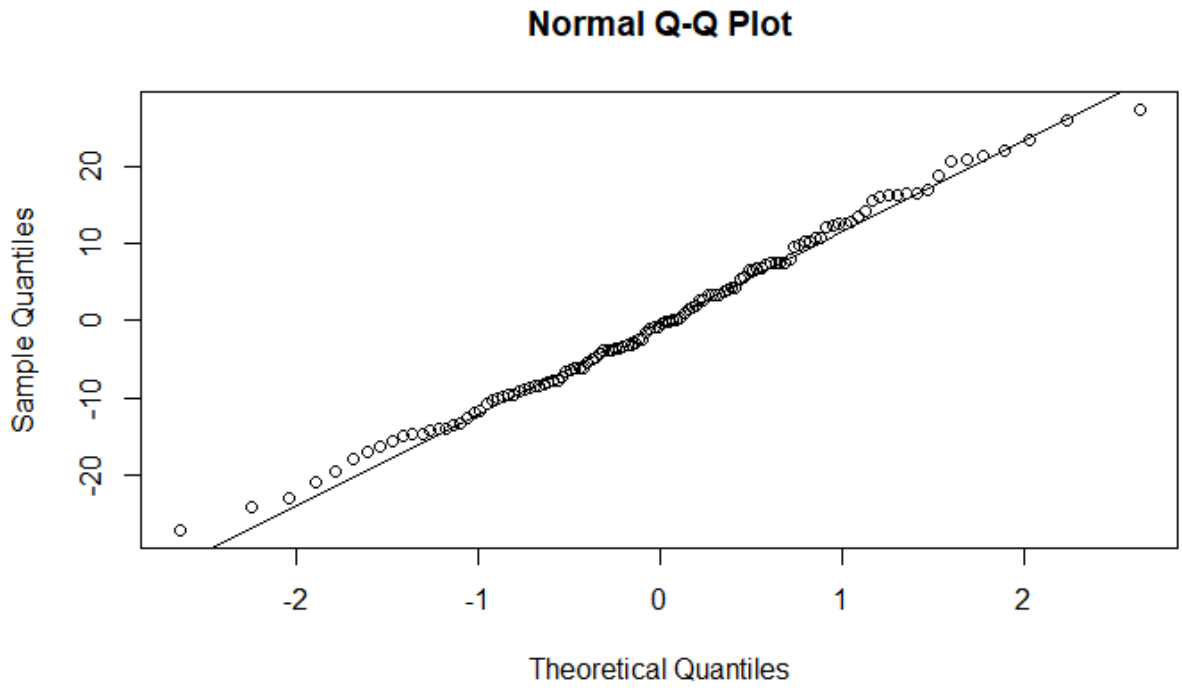
Prof/Region	S	SF
BE	8786.63	10521.48
SE	12838.57	14289.23
DS	11598.72	10551.71

This table shows the standard deviation of annual salary for every combination of profession and city. For bioinformatics and software engineers, there is more variance in annual salary for those who work in San Francisco, while for data scientists there's more variance for those who work in Seattle. By profession, software engineers have the highest variance in annual salary, followed by data scientists, and then bioinformatics engineers with the lowest variance.

III. Diagnostics

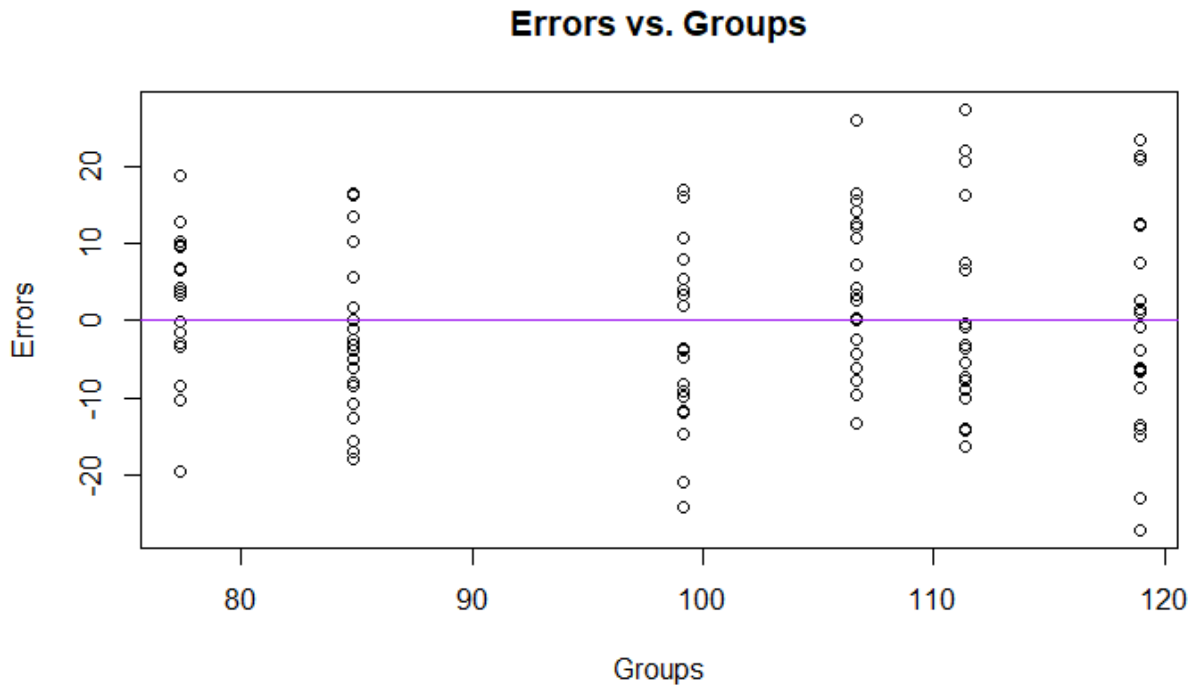
Just like Single Factor ANOVA, Two Factor ANOVA requires the three core assumptions for a model: (1) independence of Yij, (2) constant variance among group errors, and (3) normal distribution for errors. We will test to see if the data follows assumptions (2) and (3). Before we conduct the SW-test and FB-test, we will graph the QQ and Errors vs Fitted Value plots.

Figure 2.2.3:



According to the QQ-plot (Figure 2.2.3), the data points appear to adhere to the qq line. This suggests that the errors of the data follow an approximately normal distribution.

Figure 2.2.4:



According to the Errors vs Fitted Values plot (Figure 2.2.4), the data points appear to follow similar distribution among groups. This indicates that errors appear to follow constant variance amongst groups.

Though both plots provide promising results, we will use the Shapiro-Wilk and Brown-Forsythe Tests to investigate if our ANOVA model meets the two crucial assumptions. We set our alpha to 0.05 for both tests.

Shapiro-Wilk Test:

H_0: The errors are normally distributed

H_a: The errors are not normally distributed

Table 2.2.5: Shapiro-Wilk Test for Normally Distributed Errors.

W-Statistic	0.99146
P-Value	0.6698

With alpha = 0.05, we fail to reject the null hypothesis at p-val = 0.6698 and conclude that the normality assumption of TFA holds.

Brown-Forsythe Test:

H_0: The variances among the species groups are equal

H_a: At least one of the group variances is not equal to the others

Table 2.2.6: Brown-Forsythe Test for Constant Variance

F-Statistic	1.2188
P-Value	0.3048

With alpha = 0.05, we fail to reject the null hypothesis at p-val = 0.3048 and conclude that the Homoscedasticity assumption of TFA holds.

Both these conclusions tell us the test does not provide evidence against these two core assumptions of TFA. Therefore, it will not be necessary to remove any outliers or transform the data.

IV: Analysis

For this dataset, we considered 6 confidence intervals total: 4 pairwise confidence intervals comparing mean salaries between groups and 2 contrast intervals comparing average San Francisco (SF)- Seattle (S) differentials across professions. We will first use Tukey HSD for the 4 pairwise confidence intervals since this test compares pairwise group means. and then

linear contrast testing for the 2 contrast intervals as Tukey HSD cannot do these complex comparisons.

Table of Tukey post-hoc test:

Table 2.2.2: Standard Deviation of Annual Salary by City and Position

	diff	lwr	upper	P adj
DS:S-BE:S	32.772300	22.169050	43.375550	0.0000000
SE:S-BE:S	15.793900	5.190650	26.397150	0.0004752
BE:SF-BE:S	2.664292	-7.938958	13.267542	0.9780507
DS:SF-BE:S	38.013982	27.410732	48.617233	0.0000000
SE:SF-BE:S	30.509273	19.906023	41.112523	0.0000000
SE:S-DS:S	-16.978400	-27.581650	-6.375150	0.0001335
BE:SF-DS:S	-30.108008	-40.711258	-19.504758	0.0000000
DS:SF-DS:S	5.241682	-5.361568	15.844933	0.7069137
SE:SF-DS:S	-2.263027	-12.866277	8.340223	0.9894616
BE:SF-SE:S	-13.129608	-23.732858	-2.526358	0.0063648
DS:SF-SE:S	22.220082	11.616832	32.823333	0.0000002
SE:SF-SE:S	14.715373	4.112123	25.318623	0.0014227
DS:SF-BE:SF	35.349690	24.746440	45.952940	0.0000000
SE:SF-BE:SF	27.844981	17.241731	38.448231	0.0000000
SE:SF-DS:SF	-7.504709	-18.107959	3.098541	0.3202317

We decided to use the Tukey HSD test for the four pairwise confidence intervals at a 95% confidence because observations are independent within and across groups, the residuals are normally distributed, and there are equal variances. The four pairwise confidence intervals are

between Data Scientists in Seattle vs. Bioinformatics Engineer in Seattle (DS:S-BE:S) with a difference = 32.77, CI = (22.17, 43.38), Software Engineer in Seattle vs Bioinformatics Engineer in Seattle (SE:S-BE:S) with a difference = 15.79, CI = (5.19, 26.40), Data Scientist in San Francisco vs Bioinformatics Engineer in Seattle (DS:SF-BE:S) with a difference = 38.01, CI = (27.41, 48.62), and Software Engineer in San Francisco vs Bioinformatics Engineer in Seattle (SE:SF - BE:S) with a difference = 30.51, CI = (19.91, 41.11). These are the CIs I identified as most relevant as they compare salaries across different professions in the same region (S) or between the BE group and other groups across regions.

Contrast	Estimate	SE	df	Lower CL	Upper CL
SF premium DS vs SE	-45.90	5.17	114	-56.15	-35.7
SF premium SE vs BE	5.62	5.17	114	-56.15	15.9

The two contrast intervals using a 95% confidence level are for SF premium Data Scientist vs. Software Engineer with a CI of (-56.15, -35.7) and estimate of -45.90 and for SF premium Software Engineer vs. Bioinformatics Engineer with a CI of (-4.62, 15.9) and estimate of 5.62.

Firstly, we want to see if there are interaction effects between profession and region. Profession is set as factor A and region is set as factor B. The full model, which contains the effect of factor A, the effect of factor B, and the interaction effect would be:

$$Y_{i\square\square} = \mu + \gamma_i + \delta_{\square} + (\gamma\delta)_{i\square} + \varepsilon_{i\square\square}$$

Degrees of freedom is of SSE is $n\cdot - ab$

Our reduced model which only contains factor A and factor B effects would be:

$$Y_{i\square\square} = \mu + \gamma_i + \delta_{\square} + \varepsilon_{i\square\square}$$

Degrees of freedom is of SSE is $n\cdot - a - b + 1$

The goal of the hypothesis test would be to determine if the average annual salary value differs across different levels of professions and regions, and if there is an interaction effect between profession and region.

Our null hypothesis is:

H_0: All $(\gamma\delta)_{i\square} = 0$, which means that there are no interaction effects.

Our alternative hypothesis is:

H_1: at least one $(\gamma\delta)_{i\square}$ is not equal to 0, which means that interaction effect exists.

Now we perform a two-way ANOVA test to see whether the interaction effect between profession and region on annual salary is there. The test compares variation explained by the interaction term (SSE_F) with variation remaining in groups (SSE_R). We compare the variations using the formula: $F_S = \{\{SSE_R - SSE_F\} / \{df(SSE_R) - df(SSE_F)\}\} \div MSE_F$. If the variation explained by the interaction is large relative to the random error variation (F_S is large enough), we would conclude that the interaction effect is statistically significant. Below is the ANOVA table:

Model	Res. Df	RSS	DF	Sq Sum	F-Value	P-Value
Y ~ A + B	116	16058	NA	NA	NA	NA
Y ~ A * B	114	15253	2	805.41	3.0098	0.05324

As shown in the table, F value is 3.0098. F value shows that variation caused by interactions terms is three times of variation caused by random error. The P value suggests that if null hypothesis is true, we would observe our data or more extreme at a probability of 0.05324. We then compare the p-value with a significance level of 0.05. Since p-value is larger than 0.05, we fail to reject our null hypothesis. There are no interaction effects.

Then we calculate the partial R^2 for interaction effect. This can help us measure how important the interaction effect is to the overall effect. Partial R^2 is calculated using the formula:

$$R^2\{AB|(A+B)\} = (SSE(A+B) - SSE(AB)) / SSE(A+B)$$

The result we got is 0.0501551. This shows that when the interaction effect is added to a model with factor A effect and factor B effect, the overall error is reduced by 0.0501551. This number is far from 1 and shows that the contribution of interaction effect is limited.

Since there are no interactions, we move on to testing for factor A and B effects.

Now we want to see if factor A effect exists by using hypothesis tests. The full model, which contains both factor A effect and factor B effect is:

$$Y_{i\square\square} = \mu_{..} + \gamma_i + \delta_{\square} + \varepsilon_{i\square\square}$$

Degrees of freedom is of SSE is $n_{\cdot} - a - b + 1$

The reduced model, which contains only factor B effect is:

$$Y_{i\square\square} = \mu_{..} + \delta_{\square} + \varepsilon_{i\square\square}$$

Degrees of freedom is of SSE is $n_{\cdot} - b$

The goal of the factor A effect test is to see whether the mean annual salary differs across different levels of profession after accounting for the effect of region.

The null hypothesis is:

H_0: $\gamma_i = 0$ for all i, which means there are no factor A effects.

The alternative hypothesis is:

H_1: at least one $\gamma_i \neq 0$, which means that factor A effect exists.

Now we perform a two-way ANOVA test to see whether profession has an effect on annual salary. The test functions the same as the test for interaction effects, as explained above. We compare the variations using the formula: $F_S = \{\{SSE_R - SSE_F\} / \{df(SSE_R) - df(SSE_F)\}\} \div MSE_F$. If the variation explained by factor A effect is large relative to the random error variation (F_S is large enough), we would conclude that factor A effect is statistically significant. Below is the ANOVA table:

Model	Res. Df	RSS	DF	Sq Sum	F-Value	P-Value
Y ~ B	118	39873	NA	NA	NA	NA
Y ~ A + B	116	16058	1	23815	86.014	2.2e-16

As shown in the table, F value is 86.014. F value shows that variation caused by factor A terms is more than 86 times of variation caused by random error. The p-value suggests that if the null hypothesis is true, we would observe our data or more extreme at a probability of 2.2e-16.

We then compare the p-value with a significance level of 0.05. Since p-value is smaller than 0.05, we reject our null hypothesis. Factor A effect is statistically significant.

Then we calculate the partial R^2 for interaction effect. This can help us measure how important the factor A effect is to the overall effect. Partial R^2 is calculated using the formula:

$$R^2\{A + B|B\} = \{SSE(B) - SSE(A + B)\}\{SSE(B)\}$$

The result we got is 0.5972622. This shows that when factor A effect is added to a model with factor B effect, the overall error is reduced by 0.5972622. This number is close to 1 and shows that the contribution of factor A effect is significant.

We then want to see if factor B effect is significant by using a hypothesis test. The full model, which contains factor A effect and factor B effect is:

$$Y_{i\square\square} = \mu_{..} + \gamma_i + \delta_{\square} + \varepsilon_{i\square\square}$$

The degree of freedom is $n_T - a - b + 1$

The reduced model, which contains factor A effect is:

$$Y_{i\square\square} = \mu_{..} + \gamma_i + \varepsilon_{i\square\square}$$

The degree of freedom is $n_T - a$

The goal of the factor B effect test is to see whether the mean annual salary differs across different levels of region after accounting for the effect of profession.

The null hypothesis is:

H_0: $\delta_{\square} = 0$ for all j, which means there are no factor B effects.

The alternative hypothesis is:

H_1: at least one $\delta_{\square} \neq 0$, which means that factor B effect exists.

Now we perform a two-way ANOVA test to see whether region has an effect on annual salary. The test functions the same as the test for interaction effects, as explained above. We compare the variations using the formula: $F_S = \{\{SSE_R - SSE_F\} / \{df(SSE_R) - df(SSE_F)\}\} \div MSE_F$. If the variation explained by factor B effect is large relative to the random error variation (F_S is large enough), we would conclude that factor B effect is statistically significant. Below is the ANOVA table:

Model	Res. Df	RSS	DF	Sq Sum	F-Value	P-Value
Y ~ B	117	17764	NA	NA	NA	NA
Y ~ A + B	116	16058	1	1705.8	12.322	0.0006385

As shown in the table, F value is 12.322. F value shows that variation caused by factor B is more than 12 times of variation caused by random error. The P value suggests that if the null hypothesis is true, we would observe our data or more extreme at a probability of 0.0006385. We then compare the p-value with a significance level of 0.05. Since p-value is smaller than 0.05, we reject our null hypothesis. Factor B effect is statistically significant.

Then we calculate the partial R^2 for interaction effect. This can help us measure how important the factor B effect is to the overall effect. Partial R^2 is calculated using the formula:

$$R^2\{A + B|A\} = \{SSE(A) - SSE(A + B)\}\{SSE(A)\}$$

The result we got is 0.09602243. This shows that when factor B effect is added to a model with factor A effect, the overall error is reduced by 0.09602243. This number is not as close to 1 and shows that the contribution of factor B effect is modest.

Since interaction does not exist, while factor A effect and factor B effect both exist, and that the contribution of factor A effect and factor B effect is either large or modest, we use the two factor ANOVA model without interactions:

$$Y_{i\square\square} = \mu_{\cdot\cdot} + \gamma_i + \delta_{\square} + \varepsilon_{i\square\square}$$

V. Interpretation

In our results of data analysis for the Two-Factor ANOVA, show P-value 0.05324 for the interaction effect between Prof (The subject's title) and Region (San Francisco and Seattle). Because the p-value is greater than the alpha, we fail to reject the null hypothesis. So there is no interaction effect between Prof (The subject's title) and Region (San Francisco and Seattle). This means that the effect of people's subject title does not depend on the region they work in.

From the pairwise comparison, the confidence intervals between Data Scientists in Seattle vs. the Bioinformatics Engineers in Seattle is (22.17, 43.38). Because the confidence interval doesn't contain 0, we are 95% confident that there is a significant difference in mean salary between Data Scientists in Seattle and the Bioinformatics Engineer in Seattle; the Data Scientists in Seattle earn more than the Bioinformatics Engineer in Seattle. Also, the confidence

interval between Software Engineer in Seattle and Bioinformatics Engineer in Seattle is (5.19, 26.40). Because the confidence interval doesn't contain 0, we are 95% confident that there is a significant relationship between Software Engineer in Seattle and Bioinformatics Engineer in Seattle. A Software Engineer in Seattle earns more than a Bioinformatics Engineer in Seattle.

The contrast comparison for San Francisco salaries premium Data Scientist vs. Software Engineer with a confidence interval of (-56.15, -35.7). Because the confidence interval doesn't contain 0, we are 95% confident that there is a significant relationship between San Francisco salaries premium for Data Scientist and Software Engineer; the salary increase for working in San Francisco is significantly lower for Data Scientists than for Software Engineers. Also, San Francisco salaries premium for Software Engineer vs. Bioinformatics Engineer with a confidence interval of (-4.62, 15.9). Because the confidence interval contains 0, we are 95% confident that there is no significant relationship between San Francisco salaries premium Software Engineer and Bioinformatics Engineer.

VI. Conclusion

From our analysis, we concluded that there is no interaction effect between profession and region, meaning the difference in salaries across professions does not depend on whether someone works in Seattle or San Francisco. However, we did find significant mean differences in salaries across professions and across regions when considered separately. Data scientists and software engineers tend to earn more than bioinformatics engineers, and salaries in San Francisco are generally higher than in Seattle. Our pairwise comparisons and contrast intervals

support these findings by identifying which specific groups differ from one another. Overall, both profession and region matter for annual salary, with profession having a larger effect.

Code Appendix:

```
## Pt. B
newhawk = read.csv("NewHawk.csv")
library(car)
ggplot(newhawk, aes(x = Species, y = Wing)) +
  geom_boxplot()
the.model = lm(Wing ~ Species,data = newhawk)
newhawk$ei = the.model$residuals
qqnorm(the.model$residuals)
qqline(the.model$residuals)
plot(the.model$fitted.values, the.model$residuals, main = "Errors vs. Groups",xlab =
"Groups",ylab = "Errors") + abline(h = 0,col = "purple")
ei = sal.model$residuals
the.SWtest = shapiro.test(ei)
the.SWtest

the.BFtest = leveneTest(ei~ Prof*Region, data=salary, center=median)
the.BFtest

## Pt C
nt = nrow(newhawk)
a = length(unique(newhawk$Species))
SSE = sum(newhawk$ei^2)
MSE = SSE/(nt-a)
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

outliers <- c(58, 68, 122)
newhawk_clean <- newhawk[-outliers, ]
newhawk_clean

new.model = lm(Wing ~ Species,data = newhawk_clean)
ei2 = new.model$residuals
the.SWtest2 = shapiro.test(ei2)
the.SWtest2

the.BFtest = leveneTest(ei2~ Species, data=newhawk_clean, center=median)
the.BFtest

library(EnvStats)

L1 =boxcox(the.model ,objective.name = "PPCC",optimize = TRUE)$lambda
L2 = boxcox(the.model ,objective.name = "Shapiro-Wilk",optimize = TRUE)$lambda
L3 = boxcox(newhawk$Wing,objective.name = "Log-Likelihood",optimize = TRUE)$lambda
L1
L2
L3
```

```
YT1 = (newhawk$Wing^(L1)-1)/L1
t.data1 = data.frame(Wing = YT1, Species = newhawk$Species)
t.model1 = lm(Wing ~ Species,data = t.data1)

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

the.BFtest = leveneTest(ei~ Species, data=t.data1, center=median)
the.BFtest
###

YT2 = (newhawk$Wing^(L2)-1)/L2
t.data2 = data.frame(Wing = YT2, Species = newhawk$Species)
t.model2 = lm(Wing ~ Species,data = t.data2)

ei = t.model2$residuals
the.SWtest = shapiro.test(ei)
the.SWtest

the.BFtest = leveneTest(ei~ Species, data=t.data2, center=median)
the.BFtest
###
YT3 = (newhawk$Wing^(L3)-1)/L3
t.data3 = data.frame(Wing = YT3, Species = newhawk$Species)
t.model3 = lm(Wing ~ Species,data = t.data3)
```

```
ei = t.model3$residuals

the.SWtest = shapiro.test(ei)

the.SWtest


the.BFtest = leveneTest(ei~ Species, data=t.data3, center=median)

the.BFtest
```


```{r}
####
L1 =boxcox(new.model ,objective.name = "PPCC",optimize = TRUE)$lambda

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

L3 = boxcox(newhawk_clean$Wing,objective.name = "Log-Likelihood",optimize =

TRUE)$lambda

L1

L2

L3


YT1 = (newhawk_clean$Wing^(L1)-1)/L1

t.data1 = data.frame(Wing = YT1, Species = newhawk_clean$Species)

t.model1 = lm(Wing ~ Species,data = t.data1)


ei = t.model1$residuals

the.SWtest = shapiro.test(ei)

the.SWtest
```

```
the.BFtest = leveneTest(ei~ Species, data=t.data1, center=median)
the.BFtest
###

YT2 = (newhawk_clean$Wing^(L2)-1)/L2
t.data2 = data.frame(Wing = YT2, Species = newhawk_clean$Species)
t.model2 = lm(Wing ~ Species,data = t.data2)

ei = t.model2$residuals
the.SWtest = shapiro.test(ei)
the.SWtest

the.BFtest = leveneTest(ei~ Species, data=t.data2, center=median)
the.BFtest
###
YT3 = (newhawk_clean$Wing^(L3)-1)/L3
t.data13 = data.frame(Wing = YT3, Species = newhawk_clean$Species)
t.model3 = lm(Wing ~ Species,data = t.data3)

ei = t.model3$residuals
the.SWtest = shapiro.test(ei)
the.SWtest

the.BFtest = leveneTest(ei~ Species, data=t.data3, center=median)
the.BFtest
```

```

## Diagnostic Pt.2

salary = read.csv("Salaryt.csv")

salary

sal.model = lm(Annual ~ Prof + Region,data = salary)

salary$ei = sal.model$residuals

qqnorm(sal.model$residuals)

qqline(sal.model$residuals)

plot(sal.model$fitted.values, sal.model$residuals, main = "Errors vs. Groups",xlab =

"Groups",ylab = "Errors") + abline(h = 0,col = "purple")

```

```{r}

ei = sal.model$residuals

the.SWtest = shapiro.test(ei)

the.SWtest

the.BFtest = leveneTest(ei~ Prof*Region, data=salary, center=median)
```

```
the.BFtest

ggplot(NewHawk, aes(x = Species, y = Wing, fill = Species)) +
geom_boxplot() +
labs(title = "Wing Length Distribution by Hawk Species",
x = "Hawk Species", y = "Wing Length (mm)", fill = "Species")

ggplot(NewHawk, aes(x = Wing, fill = Species)) +
geom_histogram(binwidth = 5, color = "black", position = "identity") +
facet_wrap(~ Species, ncol = 1) + # Separates into individual plots
labs(title = "Wing Length Frequency by Hawk Species",
x = "Wing Length (mm)", y = "Frequency")

NewHawk %>%
group_by(Species) %>%
summarise(n = n(), mean_wing = mean(Wing),  sd_wing = sd(Wing))

ggplot(Salary_2, aes(x = Prof, y = Annual, fill = Prof)) +
geom_boxplot(alpha = 0.8, color = "black") +
labs(title = "2. Annual Salary Distribution by Profession",
x = "Profession", y = "Annual Salary (USD)" )

ggplot(Salary_2, aes(x = Annual, fill = Prof)) +
geom_histogram(binwidth = 5, color = "black", position = "identity") +
facet_wrap(~ Prof, ncol = 1) +
labs(title = "Annual Salary by Profession",
```

```
x = "Annual Salary ($1,000)", y = "Frequency")


ggplot(Salary_2, aes(x = Annual, fill = Region)) +
geom_histogram(binwidth = 5, color = "black", position = "identity") +
facet_wrap(~ Region, ncol = 1) +
labs(title = "Annual Salary by Location",
x = "Annual Salary ($1,000)", y = "Frequency")


ggplot(Salary_2, aes(x = Prof, y = Annual, fill = Region)) +
geom_boxplot(alpha = 0.8, color = "black") +
labs(title = "Annual Salary Distribution by City and Location",
x = "City", y = "Annual Salary (USD)")
tapply(Salary_2$Annual, list(Salary_2$Prof, Salary_2$Region), mean)
tapply(Salary_2$Annual, list(Salary_2$Prof, Salary_2$Region), sd)
```