I Introduction

Question: How does the annual salary (USD) vary between different professions (Data Scientist, Software Engineer, Bioinformatics Engineer) across different regions (San Francisco and Seattle)?

Interest: Understanding the salary distribution across professions and regions can provide insightful economic. Insights into wage can help individuals make informed decisions about career paths and relocation opportunities. Additionally, businesses and governments can use this information to strategize their hiring plans and economic policies respectively.

Approach: Two Factor ANOVA (TFA)

II. Summary

It is important to understand and summarize the data to conduct an insightful analysis of the relationship between annual salaries among different professions and regions. By interpreting the sampled annual salary data, we can examine important measurements through statistical tables, and visualize the distribution of the data through plots. Overall allowing the discovery of trends that can not be seen just by looking at data alone.

In this section, we will summarize the randomly sampled and observed annual salary data according to its general distribution, distribution by category, and distribution among categories.

II-1 Overall Distribution of Sample Data

Looking at the overall distribution of the Annual salary sample data allows us to grasp a general trend of the data distribution.

To interpret the numerical values (Annual Salary), the data can be plotted into a histogram to visualize the variability of the overall sample population (Figure 2.1).

Figure 2.1

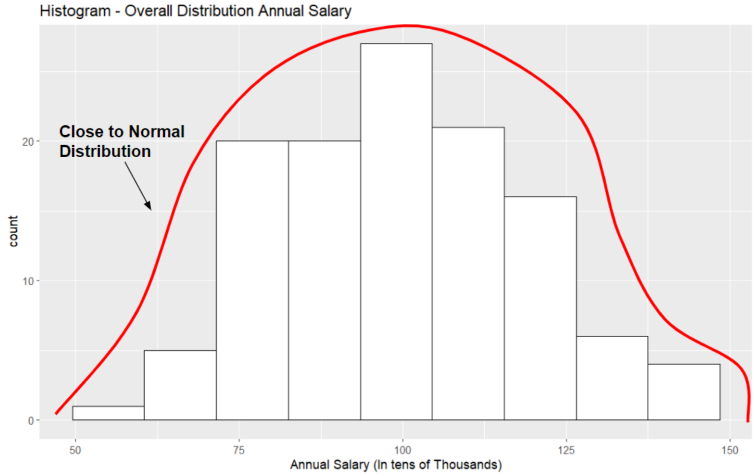


Figure 2.1 sorts the overall sample population (120 data points) by the frequency of the annual salary according to a range. Each bar has a bin size of 11, meaning that each bar represents a range of $11,000. Hence, the y-axis represents the frequency of a salary every $11,000. The highest frequency (the mode) of around 27 people tends to occur around the range of around $94,000-105,000, which also consists of the mean. The lowest frequency of around 1 person occurs at the smallest salary of around $50,000-$61,000. These values, along with the general symmetry of the data above allow us to assume that the distribution of Salary is almost normal, with a slight left tail. This could mean that the average person for all professions and regions in our sample tends to earn a bit more than $100,000.

While a Histogram gives us a general visualization of the salary distribution, it does not do well to give a more accurate representation of the symmetry if there are outlier's.

Analysis of the overall distribution of salary through a box plot helps clarify the symmetry of the data as it omits the outlier's and sees the spread of the data majority.

Figure 2.2

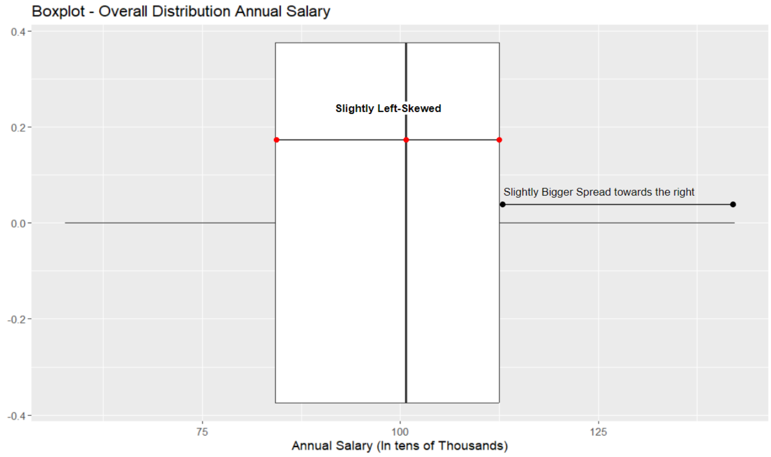


Figure 2.2 shows that the sample population does not consist of any outlier's (numerically distant points). Giving a further indication of its relatively normal distribution. The lower 25% whisker of ~$60-84 thousand is smaller than the upper 25% whisker of around $112.5-140 thousand, giving evidence of a slight left tail. The middle 50% (Interquartile range) has a slight left skew where the median is bigger than the mean. However, the interquartile range stays relatively in the middle. Meaning that the data has a relatively normal distribution.

In summary, through a general analysis of the annual salaries without taking different categories into account, it can be concluded that the data has an almost normal distribution with no outlier's and a slight bias towards the right (higher salary).

II-2 Distribution of Annual Salary by Category

While the data above gave us a general idea of the distribution of the data, it does not tell us the distribution of data within each group. Conducting further observations of the data distribution according to each group per category allows us to analyze how each category has an impact on salary distribution.

The Salary Data Set is Distributed into 2 categories:

- **Profession:** With groups DS (Data-Scientist), SE (Software-Engineer), BE (Bioinformatics Engineer)
- **Region:** With groups SF (San Francisco), S (Seattle)

Figure 2.3

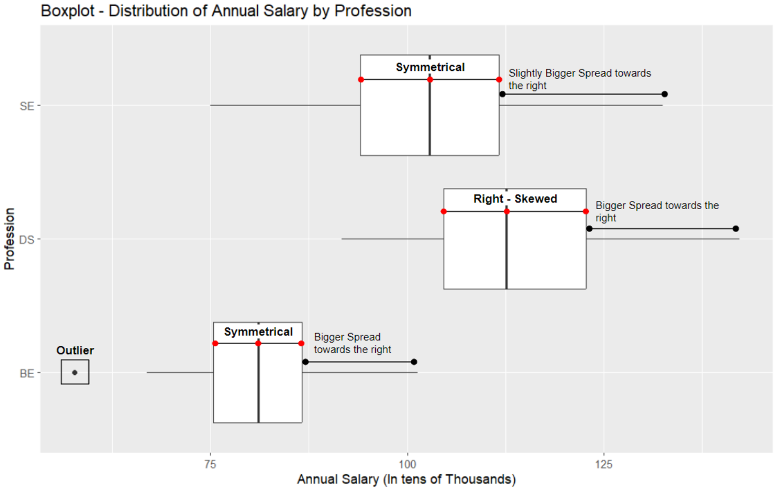


Figure 2.3 structures the Annual Salary of each worker by their profession. Based on the medians of each profession, Bioinformatics Engineers tend to earn the lowest and Data Scientists tend to earn the highest. The interquartile range further suggest the approximate normality of the data. However, it can be noted that majority of Data-Scientists tend to earn more than the median, and that one bioinformatic engineer tends to earn a lower salary than the sampled data of bioinformatic engineers. However, these points will not majorly affect our data.

The highest variability in salary occurs among Data Scientists and Bioinformatics engineers, in which the right tailed-bias (smaller left-whisker and bigger right-whisker) along with the symmetry in IQR suggests that the average salary per worker (mean) is higher than the median.

Figure 2.4

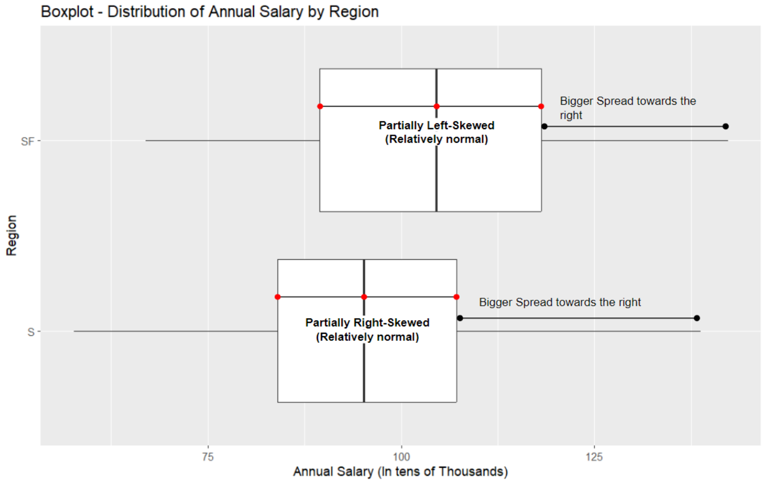


Figure 2.4 structures the annual salary per worker by the region they are in. While the median salary for workers in San Francisco appears to be higher than workers in Seattle, they are not beyond the bounds of the other regions IQR. Suggesting that it may be an insignificant difference. Therefore the interquartile range may further suggest the approximate normality of the data, which we will analyze further in the analysis section.

II-3 Mean distribution of Annual Salary among categories

II-3.1 Analysis of sample means

To further understand the interaction between the numerical values (annual salary), and the categorical values (Profession and Region), a interaction plot and a summary of mean values can be looked at.

To simplify the interpretation, the categories will be referred to as Factors.

- **Factor A:** Profession, with values Data Scientist (DS), Software Engineer (SE), and Bioinformatics Engineer (BE)
- **Factor B:** Region, with values San Francisco (SF) and Seattle (S)

Figure 2.5

	BE	DS	SE	Group Means
S	79.75 (20)	112.53 (20)	95.55 (20)	95.94 (60)
SF	82.42(20)	117.77 (20)	110.26 (20)	103.48 (60)
Group Means	81.09(40)	115.15 (40)	102.91 (40)	99.71 (120)

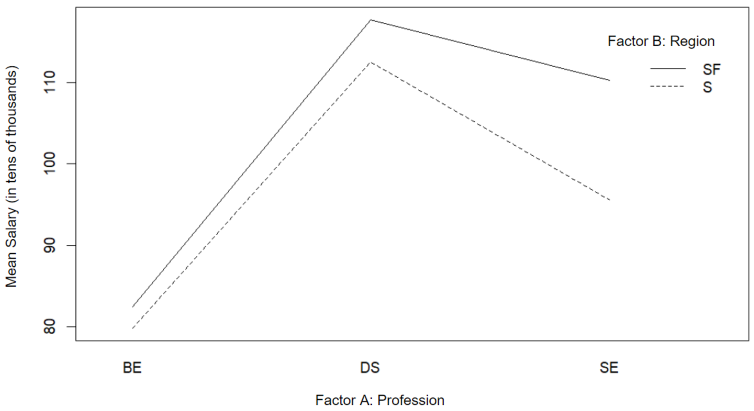
Figure 2.5 sorts each factor (A,B) by their average mean salary, and the value in parentheses shows the sample size of each interaction. Data Scientists in San Francisco tend to have the highest salary while Bioinformatics Engineers in Seattle tend to have the lowest. As all sample sizes are equal, it eliminates the possibility of a weightage bias in the overall factor mean.

When comparing the factor means to the overall sample mean, we can see that there is a bigger difference in worker salary when it comes to profession more than region. The difference ranges from \$3.2-18.62 thousand according to profession and ±\$3.77 thousand according to region.

Comparing pairwise differences of the individual Factor B (region) samples to the overall Factor B mean, we can see that there is not a significant difference in average salary for Bioinformatic Engineers and Data Scientists in a given region. With a difference ranging from \$1.33-1.34 thousand for Bioinformatic Engineers, and ±\$2.62 thousand for DS. However, software engineers have a \$7.35-7.36 thousand difference in Salary depending on what region they work in. Meaning that Factor B may not have a major impact on salary except for Software Engineers. As for comparing individual means of factor A (Profession) to the overall mean of factor A, we can see that there is a significant difference in average salary depending on the profession. With a range of \$0.39-16.59 thousand for workers in Seattle, and \$6.78-21.06 thousand for workers in San Francisco. This gives evidence of a possible interaction effect, with factor A having a stronger impact than factor B. However, further analysis testing through formal models is needed to conclude an interaction.

While the summary of mean values gives a numerical perspective of the sample salary distribution among workers according to profession and region, looking at the factors through an interaction plot is an informal method to gain a visual perspective on how profession and region interact to effect the average annual salary of a worker. It could also give further evidence of a possible interaction effect between both factors.

Figure 2.6



As seeen in Figure 2.6, the interaction plot further suggests a possible interaction effect between Profession and region. This can be seen as the slopes of SF and S do not run parallel to each other from BE to DS and DS to SE. Additionally, the steeper (bigger) slope between DS and SE in Seattle compared to San Francisco suggests that the difference in pay between data scientists and software engineers is a lot larger in Seattle than San Francisco, possibly meaning an interaction.

Comparing pairwise differences, it it seen that there is a larger difference in salaries (income gap) for software engineers in both regions as compared to the other 2 professions. This observation could mean that there is a higher demand for Software engineers in San Francisco than in Seattle, and that there is an interaction effect which we can further analyze in our analysis.

II-3.2 Analysis of group variances

Figure 2.7

Overall (Sample) Standard Deviation - 18.70

Profession	**BE**	**DS**	**SE**
Standard Deviation	9.67	13.67	13.24
Region	**S**	**SF**	
Standard Deviation	17.42	19.30	

An analysis of group standard deviations allows us to understand the spread of each group from its group mean. Where a lower spread (standard deviation) indicates the group values are denser near the mean. Overall, the salaries by group tend to vary from $9.67-$19.30 thousand per worker, which is around 10%-19% of the overall average salary per worker.

Through a comparison of the group standard deviations from the overall standard deviation, we can see that the standard deviation's for every group in the profession category is much lesser than the overall standard deviation compared to region. With a difference of 5.03-5.46 for profession and 0.6-1.28 for region from the overall standard deviation.

The standard deviation further allows us to interpret the difference in means. Since the average salary of a data scientist is more than 3 standard deviations (around 3.52) away from a Bioinformatics Engineer average salary, it suggests a difference in average salary between data scientists and bioinformatics engineers as the spread is high. In the same manner, the average salary of a worker in San Francisco is less than 1 standard deviation (around 0.43) away from the average salary of a worker in Seattle, suggesting a possibility of a lower difference is sample means.

In general, the difference in standard deviations among groups suggest that the sample have a roughly constant variance within each category. With a range of \$0.43-4 thousand among profession and \$1.88 thousand among region. Further diagnostic testing will be conducted to verify this.

II-4 Summary

In summary, the relatively normal distribution of normal salary and the relatively constant standard deviation suggests that our ANOVA assumptions are met. Through an analysis of the factor means, it appears that the profession (factor A) of the worker has a stronger individual impact on annual salary than the region (factor B) the worker is in. Even so, there appears to be an interaction effect between both factors on annual salary, where profession affects the salary of all regions and region impacts the salary of Software Engineers but not any other profession.

What could this mean.

While this section gave a relative understanding of the distribution of Annual salary per category, an analysis through formal models and methods is required to make a more accurate conclusion.

III Diagnostics

In this section, we will check whether our data satisfies the ANOVA assumptions.

The assumptions are: 1. All samples are independent 2. All groups in factor A are independent 3. All groups in factor B are independent 4. Errors are normally distributed and have constant variance where $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$

Due to limitations, we can only test whether the errors are normally distributed and have constant variance. We cannot determine whether assumptions one through three are satisfied as we did not sample the data. For simplicity, we will assume these hold.

III.1 Assessing Type I and Type II Errors

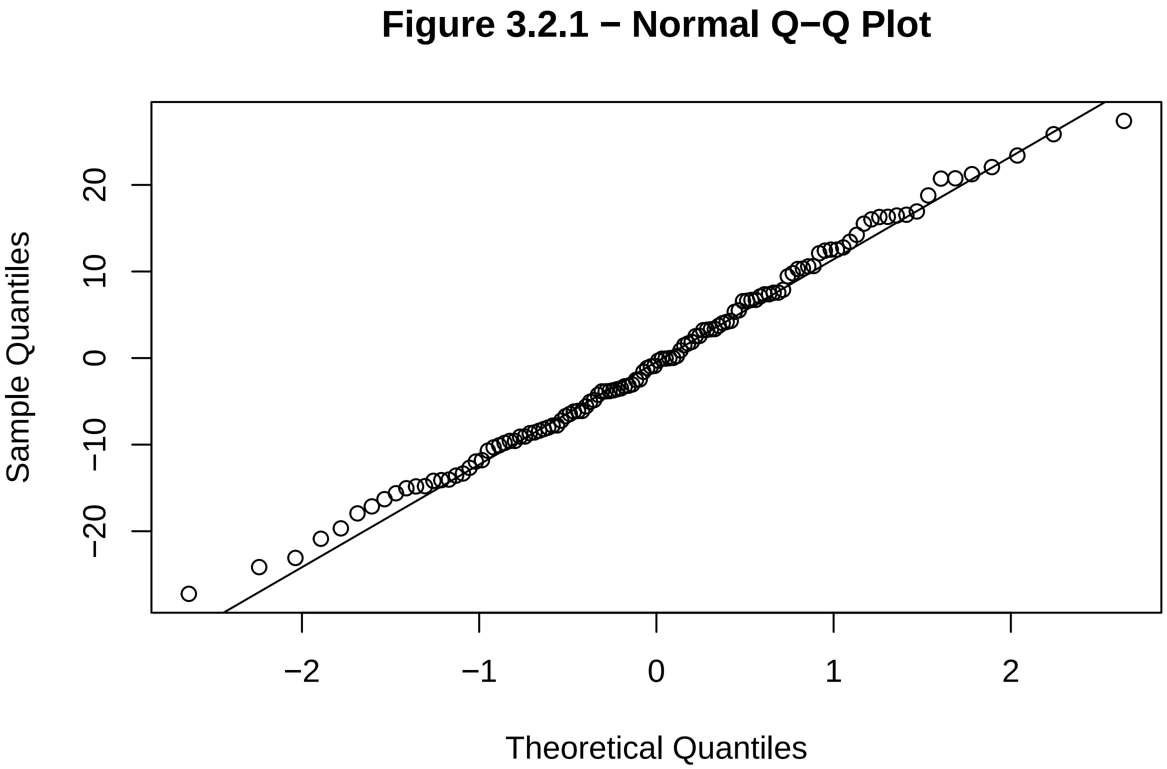
To determine which α to use for diagnostics, we need to assess whether we want to minimize the chance of a Type I Error or a Type II Error.

- Type I Error: When you reject H_0 when in reality H_0 is true. In this case, this represents the chance we conclude the data violates our ANOVA assumptions when in reality it satisfies our ANOVA assumptions.
- Type II Error: When you accept H_0 when in reality H_0 is false. In this case, this represents the chance we conclude the data satisfies our ANOVA assumptions when in reality it violates our ANOVA assumptions

For determining normality and constant variance, we want to minimize our probability of incorrect assumptions, so a Type II error is worse than a Type I error. As a result, we want to maximize α, so we will use $\alpha = 0.1$ as our threshold.

III.2 Determine Normality

III.2.1 QQ Plot



The QQ plot shows our original data plotted against a theoretical normal distribution. From the plot, the majority of the data points converge to the normal line, suggesting that our data is most likely normal. We formalize this plot by running a Shapiro-Wilk Test next.

III.2.2 Shapiro Wilk Test

H_0 : Our data is normal.

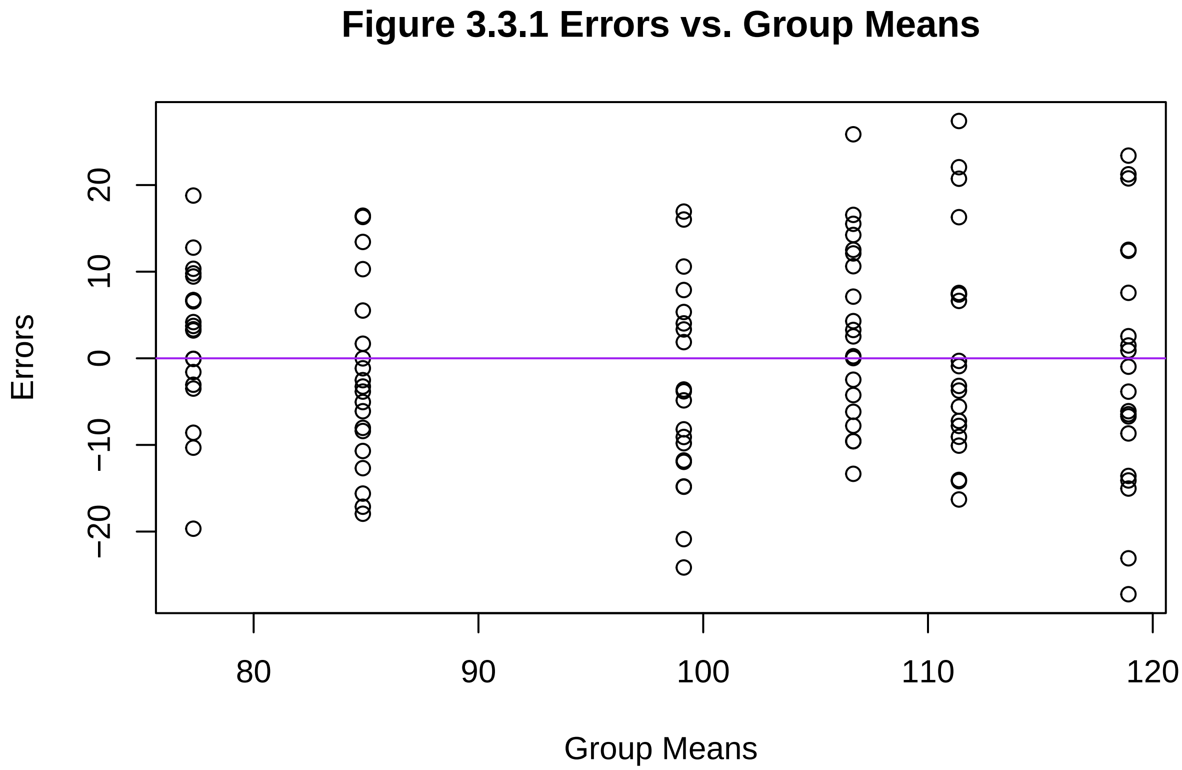
H_a : Our data is not normal.

$p = 0.6698$

Since $p > \alpha$, we accept H_0. Therefore, our data is normal.

III.3 Assessing Constant Variance

III.3.1 Plot on Errors vs Groups Means



From figure 3.3.1, the dots for each group mean seem to have approximately the same spread, suggesting that there is constant variance between groups. To formalize this, we will run the BF-test next.

III.3.2 BF-Test

H_0 : The data have constant variances.

H_a : The data does not have constant variances.

Since $p = 0.3048 > \alpha$, we accept H_0. Therefore, the data has constant variances.

III.4 Final Verdict

We can conclude that the errors are normally distributed and have constant variances, satisfying one of our ANOVA assumptions. No transformation nor outlier removal is needed.

IV Analysis and Interpretation

IV.1 Finding Best Model

We will first observe the conditional R^2 and differences between mean values to see what to expect. Then we will use F-statistic test to find out which model to use. When conducting our test, we will first test for interaction effect. If there is interaction effect, we use the model with interaction effect and stop the testing.

Otherwise, we will continue testing for factor A and factor B. We will only use these factors if there are significant effects.

IV.1.1 Conditional R^2

Figure 4.1.1

	AB	X.A.B.	A	B	Empty.Null
SSE	15252.93	16058.34	17764.09	39872.94	41578.69

Figure 4.1.2

$R^2(AB \mid (A+B))$	$R^2((A+B) \mid A)$	$R^2((A+B) \mid B)$	$R^2(A \mid Empty)$	$R^2(B \mid Empty)$
0.0502	0.096	0.5973	0.5728	0.041

From figure 4.1.2, there seems to be a significant factor A effect since its R^2 value is large when adding it both to the empty model and the model with factor B. While the conditional R^2 may hint which factors may be significant, this is not a conclusive test.

IV.1.2 Assessing Type I and Type II Errors

To determine which α to use for diagnostics, we need to assess whether we want to minimize the chance of a Type I Error or a Type II Error.

Type I Error: The chance we reject H_0 when in reality H_0 is true. In this case, it is the chance we conclude that there is a significant effect when in reality there is no significant effect.

Type II Error: The chance that we accept H_0 when in reality H_0 is false. In this case, it is the chance we conclude that there is no significant effect when in reality there is a significant effect.

In our analysis, we want to capture the factor and interaction effects if it exists, and thus want to minimize unexplained error. The lower the Type II Error, the less chance of falsely concluding no significant effect, increasing the chance of less unexplained error. Since a type II error is worse, we want to minimize it, so we will choose a higher α value. As a result, we will choose $\alpha = 0.1$.

IV.1.3 Testing for Interaction Effect

Our full model is $Y_{ijk} = \mu_{ij} + \gamma_i + \delta_j + (\gamma\delta)_{ij} + \epsilon_{ijk}$

With constraints: - $\sum \gamma_i = 0$ - $\sum \delta_j = 0$ - $\sum \sum (\gamma_i \delta_j) = 0$

Our reduced model is $Y_{ijk} = \mu_{ij} + \gamma_i + \delta_j + \epsilon_{ijk}$

With constraints: - $\sum \gamma_i = 0$ - $\sum \delta_j = 0$

Hypothesis: - H_0 : There is no significant interaction effect (i.e. all $(\gamma\delta)_{ij} = 0$). Do not use the full model. - H_a : There is significant interaction effect (i.e. at least one $(\gamma\delta)_{ij} \neq 0$). Use the full model.

Test results: - $F_s = 3.0098$ - $p = 0.0532$

Since $p \leq \alpha$, we reject H_0. Therefore, we will use the full model.

IV.1.4 Model Choice

From our testing, our final model choice is $Y_{ijk} = \mu_{ij} + \gamma_i + \delta_j + (\gamma\delta)_{ij} + \epsilon ijk$

With constraints: - $\sum \gamma_i = 0$ - $\sum \delta_j = 0$ - $\sum \sum (\gamma_i \delta_j) = 0$

IV.2 Comparisons of Different Factors

First, we will create pairwise confidence intervals to test how much professions affects average annual salary. Then, we will create another pairwise interval to test how much region affects average annual salary. Lastly, we will create non-pairwise intervals to test for more complex differences.

IV.2.1 Accuracy

We want to minimize the error of our confidence interval for stronger interpretation. As a result, we want to minimize the probability that the value does not lie within our confidence interval by minimizing α. Therefore, we will choose $\alpha = 0.001$.

IV.2.2 Multiplier

We will compute 3 pairwise confidence intervals for group A and 1 pairwise confidence interval for group B to determine the difference in annual salary given the difference in profession or region. For these intervals, we can use either the Bonferonni, Tukey, or Scheffe multipliers. We will pick the smallest multiplier for higher precision.

We will also compute 2 non-pairwise confidence intervals. For these intervals, we cannot use Tukey since that is for pairwise comparisons only. We will pick either Bonferonni or Scheffe, whichever one is smaller.

IV.2.3 Effect of Profession on Average Annual Salary

We will analyze the difference in average annual salary between all professions.

Our 99.9% confidence interval for the difference of average annual salary between BE (bioinformatics engineer) and DS (data scientist) is $[-43.6051, -24.5169]$. This means that we are 99.9% confident that on average BE makes around 24.5169 to 43.6051 less annually compared to DS. Since 0 is not in our confidence interval, we are 99.9% certain that there is a difference in average annual salary between BE and DS (i.e. BS makes less than DS).

Our 99.9% confidence interval for the difference of average annual salary between DS and SE (software engineer) is $[2.6974, 21.7856]$. This means that we are 99.9% confident that on average DS makes around 2.6974 and 21.7856 more annually compared to SE. Since 0 is not in our confidence interval, we are 99.9% certain that there is a difference in average annual salary between DS and SE (i.e. DS makes more than SE).

Our 99.9% confidence interval for the average difference of annual salary between BE and SE is $[-31.3635, -12.2753]$. This means that we are 99.9% confident that on average BE makes around 12.2753 and 31.3635 less annually compared to SE. Since 0 is not in our confidence interval, we are 99.9% certain that there is a difference in average annual salary between BE and SE (i.e. BE makes more than SE).

IV.2.4 Effect of Region on Average Annual Salary

We will analyze the difference of average annual salary between S and SF.

Our 99.9% confidence interval for the average difference of annual salary between S and SF is $[-14.6743, -0.4066]$. This means that we are 99.9% confident that on average people in S makes around 0.4066 and 14.6743 less annually compared to SF. Since 0 is not in our confidence interval, we are 99.9% certain that there is a difference in average annual salary between people in S and SF (i.e. S makes less than SF).

IV.2.5 Addressing Large Salary Difference Between Regions for SE

From the summary, we see there is an interaction effect where SE gets much higher pay in SF than S compared to other professions. This means that regional differences may not apply to professions other than SE. We will use pairwise confidence interval where we compare the effect of different regions on average annual salary for average professional other than SE.

Our 99.9% confidence interval for the average difference in annual salary between S and SF is $[-12.6901, 4.7841]$ for the average professional other than SE. Since 0 is in our confidence interval, we cannot conclude that there is a difference between regions for the average professional other than SE. Therefore we believe that the result from IV.2.4, the difference of average annual salary on region, mainly applies to SE.

IV.2.6 Addressing Wage Inequality in SF

From the summary, we noticed a huge gap between the average annual salary for the lowest paying profession compared to the other professions. This gap may be a sign of wage inequality. Lets test how big this gap is.

Our 99.9% confidence interval for the difference in average annual salary between BE and the average profession other than BE is $[-42.2981, -20.8966]$ for SF. This means that we are 99.9% confident that in SF the profession BE pays an average annual salary of around 20.8966 to 42.2981 lower than the average other professions. Since 0 is not in our confidence interval, we conclude that there is a difference between BE and the average other professions in SF.

V Conclusion

In conclusion, we find that the profession does affect annual salary. We also found that region only affects salary for SE (software engineers). Lastly, we found a significant wage gap between the lowest paying job, BE (bioinfograhics enginering), and other professions in SF. We are confident in our results as our data does not violate normality or constant variance ANOVA assumptions, and in our diagnostics and we chose conservative α values for each test yielding more accurate results.

One limitation is the fact that we have very few groups and not enough data. For example, there are other technology roles like Machine Learning and hardware engineering. We could have also separated entry level roles from senior level roles. These ideas can present more insighful results.

Appendix

```r
# Read the data from the downloaded file
file_name <- "Salary.csv"
data <- read.csv(file_name)
```

```r
# Give me means
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)
}

# Get all means for context
# all.means <- find.means(the.data)
# the.means <- data.frame(all.means$AB)
# the.means$Avg <- all.means$A
# the.means <- rbind(the.means, Avg = c(unlist(all.means$B), "NA"))
```

```r
# Get model
the.model <- lm(Annual ~ Prof + Region, data = data)
the.residuals <- residuals(the.model)
```

```r
# Do QQ Plot
qqnorm(the.residuals, main = "Figure 3.2.1 - Normal Q-Q Plot")
qqline(the.residuals)
```

```r
the.SWtest <- shapiro.test(the.residuals)
# the.SWtest
```

```r
# Do plot
plot(the.model$fitted.values, the.residuals,
     main = "Figure 3.3.1 Errors vs. Group Means",
     xlab = "Group Means",
     ylab = "Errors")

abline(h = 0,col = "purple")
```

```r
# Do BF-test
the.BFtest <- car::leveneTest(the.residuals ~ paste(Prof, Region), data=data,
                              center=median)
the.p.val <- the.BFtest[[3]][1]
# the.p.val
# Rename for more efficient typing
the.data <- data
names(the.data) <- c("Y", "A", "B")

# Fit the models
AB <- lm(Y ~ A * B, the.data)
A.B = lm(Y ~ A + B,the.data)
A = lm(Y ~ A,the.data)
B = lm(Y ~ B,the.data)
N = lm(Y ~ 1, the.data)

# Find the SSE values
all.models = list(AB,A.B,A,B,N)
SSE = t(as.matrix(sapply(all.models,function(M) sum(M$residuals^2))))
colnames(SSE) = c("AB","(A+B)","A","B","Empty/Null")
rownames(SSE) = "SSE"
knitr::kable(round(data.frame(SSE), 4))
```

```r
# Get partial R^2
get.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)
}

the.partial.R2 <- data.frame(get.Partial.R2(A.B, AB), get.Partial.R2(A, A.B),
                               get.Partial.R2(B, A.B), get.Partial.R2(N, A),
                               get.Partial.R2(N, B))
colnames(the.partial.R2) <- c("$R^2(AB\\mid (A+B))$", "$R^2((A+B)\\mid A)$",
                              "$R^2((A+B)\\mid B)$", "$R^2(A\\mid Empty)$",
                              "$R^2(B\\mid Empty)$")
knitr::kable(round(the.partial.R2, 4))
```

```r
# Interaction effect test
# anova(A.B, AB)
```

```r
# Get relavent values for CI based on model choice
n_T <- nrow(the.data)
a <- length(unique(the.data$A))
b <- length(unique(the.data$B))
sse <- SSE["SSE", "AB"]
df_sse <- n_T - a * b
mse <- sse / df_sse
alpha <- 0.001

# Give me multipliers
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)
}

# Give me CI
give.me.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)
  }
}
```

```
# Give me multiplier for pairwise comparisons between different professions
all.mult <- find.mult(alpha = alpha, a = a, b = b, dfSSE = df_sse, g = 3,
                      group = "A")
the.mult <- min(all.mult)
```

```
# Give me CI for BE vs DS
the.CI <- give.me.CI(the.data, mse, equal.weights = TRUE, the.mult, "A",
                     c(1, -1, 0))
names(the.CI) <- c("$\\mu_{1.} - \\mu_{2.}$", "lower bound", "upper bound")
# data.frame(the.CI)
```

```
# Give me CI for DS vs SE
the.CI <- give.me.CI(the.data, mse, equal.weights = TRUE, the.mult, "A",
                     c(0, 1, -1))
names(the.CI) <- c("$\\mu_{2.} - \\mu_{3.}$", "lower bound", "upper bound")
# data.frame(the.CI)
```

```
# Give me CI for BE vs SE
the.CI <- give.me.CI(the.data, mse, equal.weights = TRUE, the.mult, "A",
                     c(1, 0, -1))
names(the.CI) <- c("$\\mu_{1.} - \\mu_{3.}$", "lower bound", "upper bound")
# data.frame(the.CI)
```

```
# Give me multiplier for pairwise comparisons between different regions
all.mult <- find.mult(alpha = alpha, a = a, b = b, dfSSE = df_sse, g = 1,
                      group = "B")
the.mult <- min(all.mult)
```

```
# Give me CI for S vs SF
the.CI <- give.me.CI(the.data, mse, equal.weights = TRUE, the.mult, "B",
                     c(1, -1))
names(the.CI) <- c("$\\mu_{.1} - \\mu_{.2}$", "lower bound", "upper bound")
# data.frame(the.CI)
```

```
# Give me multiplier for pairwise comparisons between different regions
# g is one here because the two non-pairwise confidence intervals
# that we will calculate are unrelated
```

```
all.mult <- find.mult(alpha = alpha, a = a, b = b, dfSSE = df_sse, g = 1,
                      group = "AB")
bon <- all.mult[1]
sch <- all.mult[3]
the.mult <- min(bon, sch)
```

```
# Give me CI on S vs SF for avg(BE, DS)
AB.cs = matrix(0,nrow = a, ncol = b)
AB.cs[1, 1] = 0.5
AB.cs[2, 1] = 0.5
AB.cs[1, 2] = -0.5
AB.cs[2, 2] = -0.5
the.CI <- give.me.CI(the.data, mse, equal.weights = TRUE, the.mult, "AB", AB.cs)
names(the.CI) <- c(
  "$0.5 * \\mu_{11} + 0.5 * \\mu_{21} - 0.5 * \\mu_{12} - 0.5 * \\mu_{22}$",
  "lower bound", "upper bound"
)
# data.frame(the.CI)
```

```
# Give me CI on S vs SF for avg(BE, DS)
AB.cs = matrix(0,nrow = a, ncol = b)
AB.cs[1, 2] = 1
AB.cs[2, 2] = -0.5
AB.cs[3, 2] = -0.5
the.CI <- give.me.CI(the.data, mse, equal.weights = TRUE, the.mult, "AB", AB.cs)
names(the.CI) <- c("$\\mu_{12} - 0.5 * \\mu_{22} - 0.5 * \\mu_{32}$",
                   "lower bound", "upper bound")
# data.frame(the.CI)
```