Introduction

The following paper will be analyzing data from Level.fyi. Level.fyi is a website that collects people's compensation information. Level.fyi also lets people upload their own compensation information anonymously with proof documents that can verify their compensation which makes the data reliable. The data we are going to use was collected from 2017 until now. The data contains 62643 rows of individual's compensation with 29 columns of variables such as city, level, and company. Since there are 28 parameters, we are going to use (1) company name, (2) title of the worker, (3) total yearly compensation, (4) location of where they worked, and (5) years of experience to study the relationship between each factor and the yearly compensation, studying observations recorded in the State of California.

Data Background and Questions of Interest:

As previously mentioned, the data from Level.fyi contains approximately 60,000 salary records from top companies. It measures the relationship between the yearly compensation and the five parameters mentioned above: company name, title of the worker, total yearly compensation, location of where they worked, and years of experience.

Here, the yearly compensation or the annual compensation, in the simplest terms, is the combination of your base salary and the value of any financial benefits your employer provides. We need to analyze how this annual compensation will be affected by the five factors.

Our goal, along with the combination of these variables is to test whether there is a statistically significant relationship between these potential predictors and the annual compensation. The methodology is basically to understand if the data maintains the normal distribution assumption, and if there is a need for data transformations or not. Our goal is to understand the applications of plots and correlation, linear regression, etc., amongst different factors to see its influence on Total Yearly Compensation.

Given this thought process, we have some questions we would like to answer:

1. Does company, title, location, and years of experience (YOE) have significant effects on Total Yearly Compensation?
2. Which of these factors have the greatest influence on Total Yearly Compensation.
3. When is Total Yearly Compensation highest?
4. Is there a relationship between Total Yearly Compensation and Title?

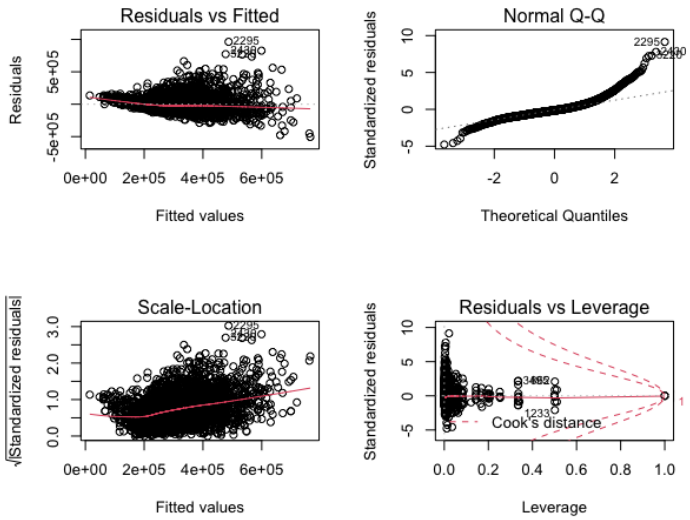
Through statistical analysis, we will be able to find answers to back up our claims and make decisions based on our findings. Having answered these questions will allow us to learn more about what influences the yearly compensation.

Set up for Model

1. Analysis of Full Model

(i) Assumption of Full Model

Model: lm(totalyearlycompensation ~ company + title + location + yearsofexperience)



From observing the Normal Q-Q plot, notice the points fall along a line in the middle of the graph, but curve off in the extremities. Normal Q-Q plots that exhibit this behavior usually means the data have more extreme values than would be expected if they truly came from a Normal distribution. Few outliers exist.

Looking at the Residuals vs Fitted plot, some dots do not bounce randomly around the 0 line and some and some dots stand out. In addition, the residuals do not form a horizontal band around the 0 line, which suggests that the variances of the error terms are not equal. Therefore, the assumption of equal variance may be violated.

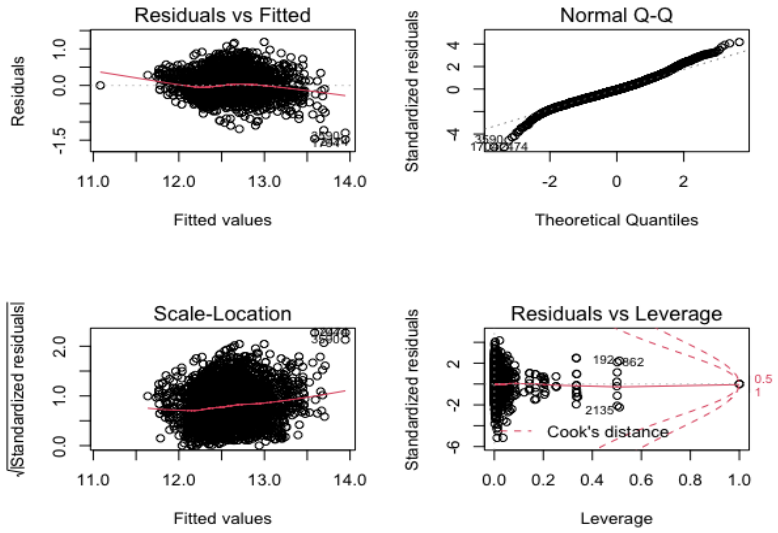
Multiple R-squared: 0.4741, the predictor variables explain 47.41% of the changes in the response variable (total yearly compensation). The Residual standard error: 106400, is extremely high. When running an F test for a multiple regression model with an intercept, the small p-value(<2.2e-16) is obtained. In other words, there is a relationship between these predictor variables and response variable Y.

2. Transformations and diagnostics

Since there are outliers and assumption of equal variance is violated, transformations are needed.

(i) Log Transformations

Log Transformations make sense due to the presence of high variance. Multiple R-squared: 0.5205, which is higher than before. The residual standard error greatly reduces to 0.2877. Therefore, this log transformation model is better than the original model.

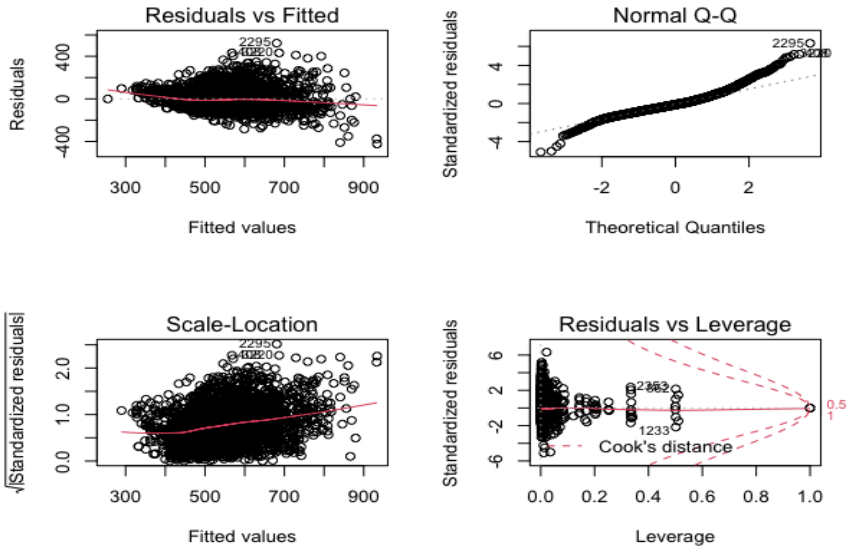


From the Normal Q-Q Plot, the assumption of normality holds because most dots are close to the line or on the line. However, some outliers still exist.

From the Residuals vs Fitted Plot, many dots bounce randomly around the 0 line but with fewer dots standing out from the basic random pattern of residuals. In addition, the residuals roughly form a horizontal band around the 0 line. Therefore, the assumption of equal variance may hold. (outliers: 2474, 3590, 1704).

(ii) Square Root Transformation

After square root transformations, Residual standard error: 83.61, which is higher than before and Multiple R-squared: 0.5091, which is less than the log transformation…



From the Normal Q-Q Plot, the assumption of normality holds because most dots are close to the line or on the line. However, some outliers still exist.

From the Residuals vs Fitted Plot, many dots bounce randomly around the 0 line but there are some dots observations standing out from the basic random pattern of residuals. In addition, the residuals roughly form a horizontal band around the 0 line. Since some dots are still far away from the 0 line, the assumption of equal variance may not hold.

(iii) Transformation Decision

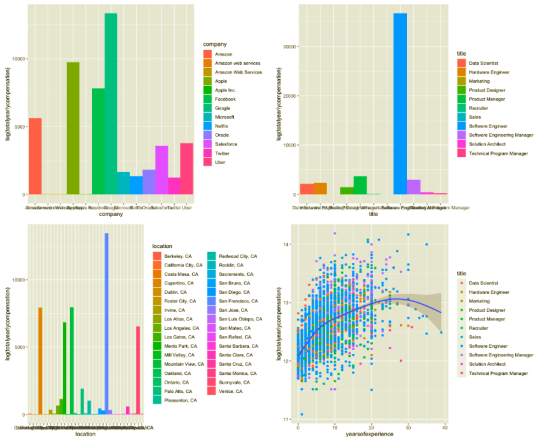
Since the logarithmic transformation holds the assumption of normality and equal variance and has a higher R-squared value, we will use the resulting model in our analysis. However, it is important to note that the resulting transformation still has some heteroskedasticity from residuals, but is much improved from the original model.

3. Outlier Analysis

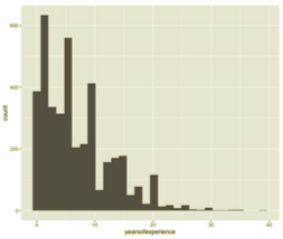
According to Residuals vs Fitted and Normal Q-Q plots of the square-root transformation, we see that 2474, 3590, and 1704 are outliers. These observations had low total yearly compensation than the mean Total Yearly Compensation for California, along with having higher total Years Of Experience than the mean. These observations were removed from the data set.

Visualization and Statistical Analysis

1. Plots



Plot	Correlation
Company vs Total Yearly Comp.	Kruskal-Wallis: 317.27
Title vs Total Yearly Comp.	Kruskal-Wallis: 308.28
Location vs Total Yearly Comp.	Kruskal-Wallis: 286.75
YOE vs Total Yearly Comp.	Kruskal-Wallis/Cor: 1889.2/0.578371



As we can, from the plots and correlation values, Years of Experience (YOE) and Total Yearly Compensation (TYC) are very strongly correlated. We can assume that Location and TYC do not have much of a linear relationship because the variables have a weak correlation.

We have also modeled a histogram of YOE, filled with TYC, and saw that observations with less than 20 years of experience have higher total yearly compensation on average, mostly Software Engineers. In contrast, having more than 10 years of experience seems to decrease average yearly compensation.

2. Statistical Analysis

(i) F-Test of Multiple Regression Analysis

Hypothesis:

$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4$

H_a : At least one of $\beta_k \neq 0$

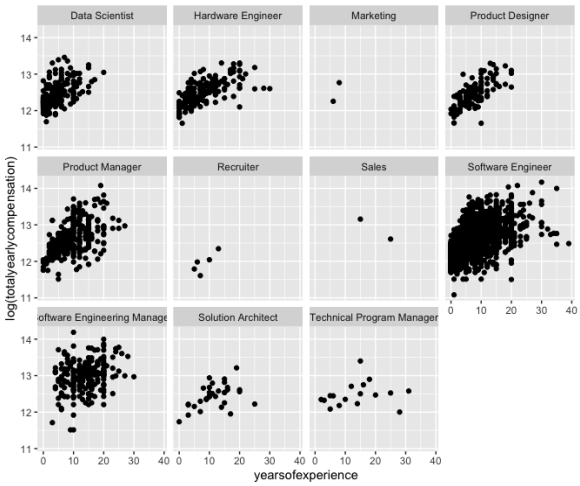
Since the F-stat's p-value is small (< 2.2e-16), we would reject the null hypothesis and conclude that these predictor variables and total years of experience have significant relationship.

(ii) T-test and Standard Regression

In order to find which factor has the most significant influence on Total Yearly Compensation, standardized regression analysis is needed because each predictor variable has different units.

After the standardized regression is conducted, R^2 and p-values do not change, and the difference exists on the estimator of each variable. According to the t-test's p-values, the same results are those obtained

from previous linear regressions. The results include that Company and Years of Experience have statistically significant effects on mean yearly compensation because of high correlation values. For the Years of Experience, it results in the largest absolute value of the estimator (0.578), which means that the mean yearly compensation increases 0.578 as years of experience increases one unit.




```
        Welch Two Sample t-test

data:  data1$yearsofexperience and data1$totalyearlycompensation
t = -130.9, df = 3981, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -306904.1 -297846.3
sample estimates:
   mean of x    mean of y
7.035259e+00 3.023822e+05
```

From the graph and the result, it shows that most workers in California have less than 20 years of experience, with most having an average of 7 years of experience. Results show that most workers earn between $297,846 and $306,904. The majority of workers are Software Engineers, but along with other workers as well, the mean years of experience for each title is 7 years, with an average yearly compensation of $302,382.

Interpretation and Reporting

Does Company, Title, Location, and Years of Experience (YOE) have significant effects on Total Yearly Compensation?

The predictor variables affecting total yearly compensation are company, title, location and total years of experience. With the F test of multiple linear regression performed, the terms' coefficients are not all equal to zero, meaning that there is sufficient evidence to support a significant association between the predictor variables and the response variable, total yearly compensation. As we can, from the plots and relationship esteems, Years of Experience and Total Yearly Compensation are emphatically corresponded. We can expect that Location and TYC don't have a very remarkable direct relationship in light of the fact that the factors have a frail connection. We have likewise demonstrated a histogram of YOE, loaded up with TYC, and saw that perception with under 20 years of involvement have higher complete yearly remuneration by and large, generally Software Engineers. Interestingly, having over 10 years of involvement appears to diminish normal yearly compensation.

Which of these factors have the greatest influence on Total Yearly Compensation.

The factor with the greatest influence on total early compensation is the company, where it has the largest absolute value, largest absolute value of the estimator (0.578), which means that the mean yearly compensation increases 0.578 as years of experience increases one unit.

When is Total Yearly Compensation highest?

As we can, from the plots and connection esteems, Years of Experience and Total Yearly Compensation are firmly associated. We can accept that Location and TYC don't have a very remarkable direct relationship in light of the fact that the factors have a feeble connection.
We have likewise demonstrated a histogram of YOE, loaded up with TYC, and saw that perceptions with under 20 years of involvement have higher absolute yearly pay by and large, for the most part Software Engineers. Interestingly, having over 10 years of involvement appears to diminish normal yearly pay. Specifically, highest when working at Google as Software Engineer in San Francisco, with about average, 7 years of experience.

Is there a relationship between Total Yearly Compensation and Title?

There is a relationship between these two variables, where title has higher correlation with total yearly compensation than location, but less than Company and YOE. It is used in the final regression model as it holds a strong correlation value, but since the majority of the observations recorded are Software Engineers, it's only possible to calculate variance amongst that group only, rather than a wide scale of careers.

Conclusion

Based on the data from Level.fyi, we can conclude that Total Yearly Compensation increases as YOE increases. This conclusion makes sense because the more experience a person has, the better the more knowledgeable they'll be, which makes them more valuable to the company. Salary seems highest at Google in San Francisco which, again, makes sense since San Francisco is a relatively expensive city to live in [2]. Google is also one of the top 5 richest companies, which would account for how they're able to pay their employees well [4].

In order to find the factors affecting TYC, the reduced model: *TotalYearlyCompensation ~ Company + Title + Location + YearsOfExperience.* The logarithmic transformation holds the assumption of normality and equal variance, and it has a higher R-squared value, so we used the resulting model in our analysis. With the scatter plots we can draw that TYC and YOE are positively correlated, and based off of t-tests, though Company also has an effect on TYC, YOE has the most significant difference. However, TYC seems to decrease when YOE is over 20. This may be due to the fact that there aren't many data point, whether it's because few people work over 20 years or not many people who work over 20 years submitted their data to be collected.

There are some limitations to this study, however, as people had to voluntarily give their salaries and other information. Other than Google, there aren't many people with common companies amongst themselves. Another limitation is that the range of when each data point was entered is quite large. The earliest point is in 2017 and the latest in 2021, which may give an inconsistency to the data, however slight. The inflation rate from 2017 to 2021 is 12.84%. Therefore, future research with more employees from the same company, and people with more YOE can ensure a more accurate and persuasive analysis.

Function

```
> fun = function(x) {
+ a = data1 %>% filter(grepl(x, title))
+ b = max(a$totalyearlycompensation)
+ return(b)
+ c = mean(a$totalyearlycompensation)
+ return(c)
+ d = max(a$company)
+ return(d)
+ e = mean(a$yearsofexperience)
+ return(e)
+ }
```

This function takes one of the 11 unique titles for tech workers and returns the max salary, mean salary, the company with the most of a certain title, and average years of experience for the specified title.

This function allows the user to input their choice of career and lets them see the most important parts of what they desire in their career, similar to the output of what levels.fyi displays for workers at certain companies.

References

(1) "STA 141A- Project (example 1)", June 2021

(2) Brinklow, Adam, "What it really costs to live in the Bay Area", September 2019. https://sf.curbed.com/2019/9/20/20876203/24-7-wall-street-san-francisco-most-expensive-cost-of-living-epi

(3) CPI Inflation Calculator, "Value of 2017 Dollars Today", November 2021. https://www.in2013dollars.com/us/inflation/2017

(4) Kabra, Archana, "20 Richest Companies in the World by Market Cap 2021", August 2021. https://www.thetealmango.com/featured/richest-companies-in-the-world/

(5) Ogozaly, Jack, "Data Science and STEM Salaries", October 2021. https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries/version/1

Code

```
> data1 <- Levels_Fyi_Salary_Data[ -c(1,3, 8:29)]

> View(data1)

> data1 <- data1[-c(10001:26476),]

> data1 <- Levels_Fyi_Salary_Data[ -c(1,3, 8:29)]

library(dplyr)

data1 %>%
filter(grepl("Apple|Microsoft|Amazon|Google|Facebook|Netflix|Oracle|Twitter|Uber|Salesforce",
company))

> data1 = data1 %>%
filter(grepl("Apple|Microsoft|Amazon|Google|Facebook|Netflix|Oracle|Twitter|Uber|Salesforce",
company))

data1 = data1 %>% filter(grepl("CA", location))

full_model = lm(totalyearlycompensation ~ company + title + location + yearsofexperience, data = data1)

> full_model1 = lm(log(totalyearlycompensation) ~ company + title + location + yearsofexperience, data
= data1)

> plot(full_model1)

full_model2 = lm(sqrt(totalyearlycompensation) ~ company + title + location + yearsofexperience, data =
data1)

plot(full_model2)

data1 = na.omit(data1)

> ggplot(data1, aes(x=company, y=log(totalyearlycompensation), fill=company))

  +     geom_bar(width = 1, stat = "identity")

kruskal.test(data1$totalyearlycompensation ~ data1$location)

kruskal.test(data1$totalyearlycompensation ~ data1$company)

kruskal.test(data1$totalyearlycompensation ~ data1$title)

kruskal.test(data1$totalyearlycompensation ~ data1$yearsofexperience)
```

```
cor(data1$totalyearlycompensation ~ data1$yearsofexperience)

ggplot(data = data1, mapping = aes(x = yearsofexperience, y = totalyearlycompensation))
  + geom_point(mapping = aes(colour = title))
  + geom_smooth()

ggplot(data1, aes(x=location, y=log(totalyearlycompensation), fill=location))
  + geom_bar(width = 1, stat = "identity")

ggplot(data1, aes(x=title, y=log(totalyearlycompensation), fill=title))+
  + geom_bar(width = 1, stat = "identity")

> ggplot(data = data1)
  + geom_point(mapping = aes(x = yearsofexperience, y = log(totalyearlycompensation)))
  + facet_wrap(~title)

t.test(data1$yearsofexperience, data1$totalyearlycompensation, alternative = "two.sided")
```