STA141A Final Project

STEM Salary Analysis - Group 10

12/10/2021

Name	Contribution		E-mail
	Code	Report	
Xiquan Jiang	Model construction, model diagnostics, Statistical analysis	Model Set-ups, Models & Results, Limitations	xiqjiang@ucdavis.edu
Cindy Li	Data Clean-up & Preprocessing	Introduction, Questions of Interest, Outliers & Data Transformation, Statistical Analysis	whali@ucdavis.edu
Bo Zhang	Data Preprocessing	Data Background, Data Visualizations, Interpretation & Conclusion, Limitations	bzbzhang@ucdavis.edu

Emanuela Furfaro Instructor

STA 141A - Fundamentals of Statistical Data Science

University of California, Davis

Introduction

Almost all tech careers are inseparable from STEM basics. Recently, it is a noticeable fact that careers in STEM fields have bloomed at an incredible pace since 75% of the fastest-growing occupations today require a lot of science or math skills. The demand for highly skilled engineering professionals in those fields, such as data scientists, software development engineers, and product managers, has grown dramatically in order to accommodate the urgent needs for optimizations, automations, system integrations, as well as smarter decision making for a wide range of industry sectors. As a response, such demands have attracted an increasing number of students to pursue related educations so that they can be more competitive while stepping into those STEM job markets. Also, STEM majors almost dominate the top ten of the top ten high-paying majors for graduates in the United Kingdom and the United States. There is no doubt that promising job prospects and high salaries have attracted more and more students to choose STEM majors, which means the pressure of competition will become increasingly fierce. In fact, there is a large selection of roles within the STEM fields which are different from each other from many perspectives. So it may be interesting to take a closer look at the salaries for a variety of STEM field jobs and better understand the internal factors that influence them.

Data Background

This dataset was retrieved from levels.fyi and has already been cleaned by the creator of this dataset and uploaded on Kaggle. There are 62,000 salary records from 1,632 companies all over the world such as Amazon, Google and Microsoft in our dataset.
There are 28 columns in total:

timestamp: the time when this data was recorded
Company: the company where the observation is hired
Level: what level the observation is at
Title: the role title the observation is
total yearly compensation, base salary, stock grant value, bonus
location: the job location of each observation
years of experience
years at company: years of experience at the hired company
tag: the track this observation is taking (Full Stack, Distributed Systems, etc.)
gender: male, female, and other
other details: free form text field
city ID
dmaid
row number
Race
Education
Dummy variables:

Masters_Degree, Bachelors_degree, Doctorate_degree, High school, Some_College
Race_Asian, Race_White, Race_Two_or_More; Race_Black; Race_hispanic

Among these variables, we will set the total yearly compensation as our response variable, all other variables as the predictor variables.

Questions of Interest

We will train proper predictive models for predicting the STEM careers salaries based on given factors.

1) Does “title”, “yearsofexperience”, “yearsatcompany”, “gender”, “Race”, “Education” , “state” have significant effects on “total yearly compensation”?
2) Which factor among them has the most significant impact on the response variable?
3) Whether there are significant differences in salaries for different levels in variables “Education”, “Race”, and “Gender”?

In addition, backed up by statistical analysis, we will be able to provide insights about the employment orientations for job seekers for these job categories based on what we analyzed in this project.

Model Set-ups

Data Preprocessing & EDA

Missing Values & Variable Selection

First of all, we checked the missing values in our salary data set. Since there are over 40,000 data (almost ⅔ of the data) in our dataset with missing values, we decided to remove all of them in order to keep the most valuable data and to get a more accurate analysis over our data. After removing all the missing values, we have 21591 data in total.

Since we don’t need the dummy variables column for the ‘race’ and ‘education’ variables, we decided to remove them, keep ‘race’ and ‘education’ columns only. Besides, we decided to remove variables ‘other detail’, ‘cityID’, ‘dmaid’, and ‘row number’, since these are not related to our analysis.

The variable ‘tag’ is too concentrated on the job ‘software engineer’, which is too detailed that we will not focus on, so we decided to remove this column.

For the ‘level’ variable, there isn’t any consistent scale for people to fill it out. We could see that several people fill it by ‘L1, L2, ...’, and people from Microsoft will use the scale ‘59/60/…68/69’. There are over 20 types of level design for different companies, it’s tons of workload for us to process this variable reasonably. Since we still have another two variables ‘years of experience’ and ‘years at company’, we think this could relatively represent the ‘level’ variable. Therefore, we removed the ‘level’ variable after our careful consideration.

As for the ‘timestamp’ variable, we have data from 2017-2021. However, for the data from the time period 2017-2019, there exists too many missing values. Besides, the employment status every year has changed a lot recently because of the coronavirus, and we can't predict when this

epidemic will end, thus we decide to keep data from the epidemic period, which will have more reference value for our research.

For the location variable, we split that into 3 columns: ‘city’, ‘state’, ‘country’, and include only ‘state’ in our dataset, in order to analyze only employment circumstances inside the United States from a regional perspective based on different state locations.

Data Visualizations

Figure 1 indicates that, if we categorize the data by education, employers with bachelor’s degree and master's degree are the main components in the job market. Basically, it shows the trend for each education level vs. the numerical variables are similar but not the same.

Figure 1: Pair plots by education

According to figure 2, we could find that if we categorize the numerical data by race, Asian and White are the two largest race groups in the career market. There isn’t much difference for the distribution of every numerical variable categorized by race.

Based on Figure 3, it shows that based on the dataset we have, male have a larger percentage in the career market. There isn't a significant difference for the distribution by each gender.

As we look at the scatter plot, we could see that ‘stock bonus’ and ‘base salary’ have the highest relationship with our response variable ‘total compensation’. The scatter plot of ‘years of experience’ and ‘total compensation’, and that of ‘years of experience’ and ‘stock value’ both show bell shape distribution, which means that employers with 10-25 years of experiences would have higher compensation and stock. The scatter plot of ‘years at company’ and ‘stock grant value’

could tell us that employers with less years at company would potentially get more stock, and it decreased as the total year at company increased.

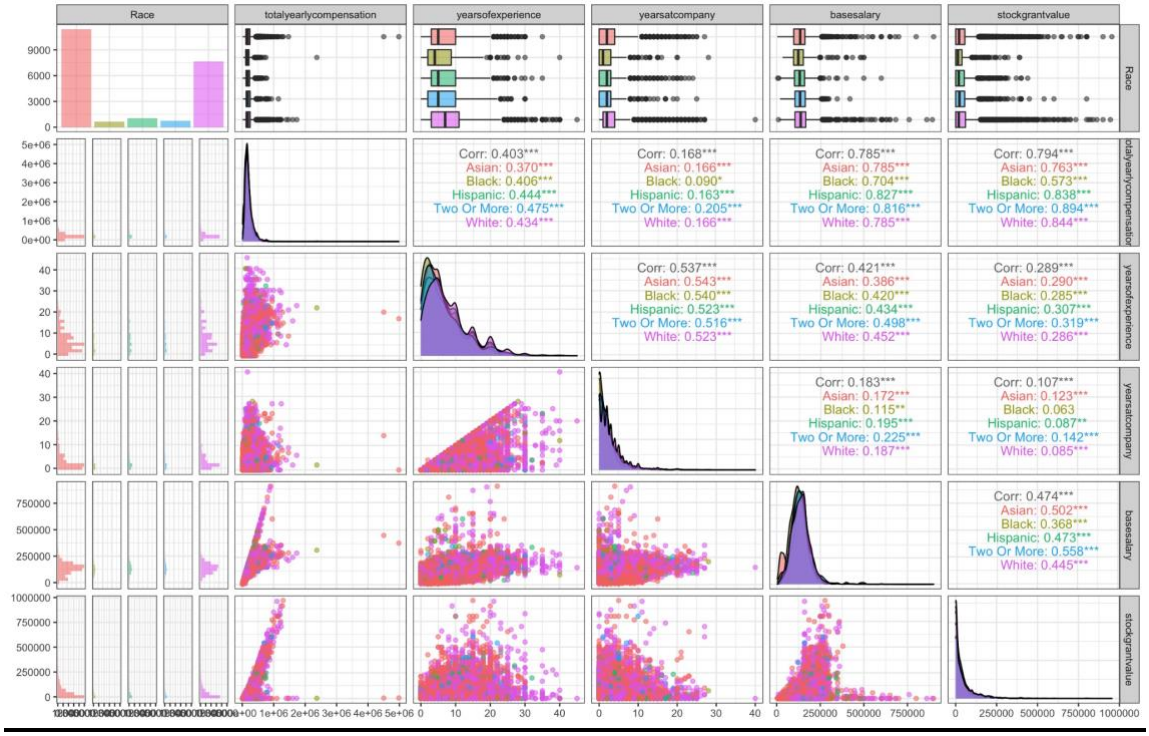

Figure 2: Paired plot by races

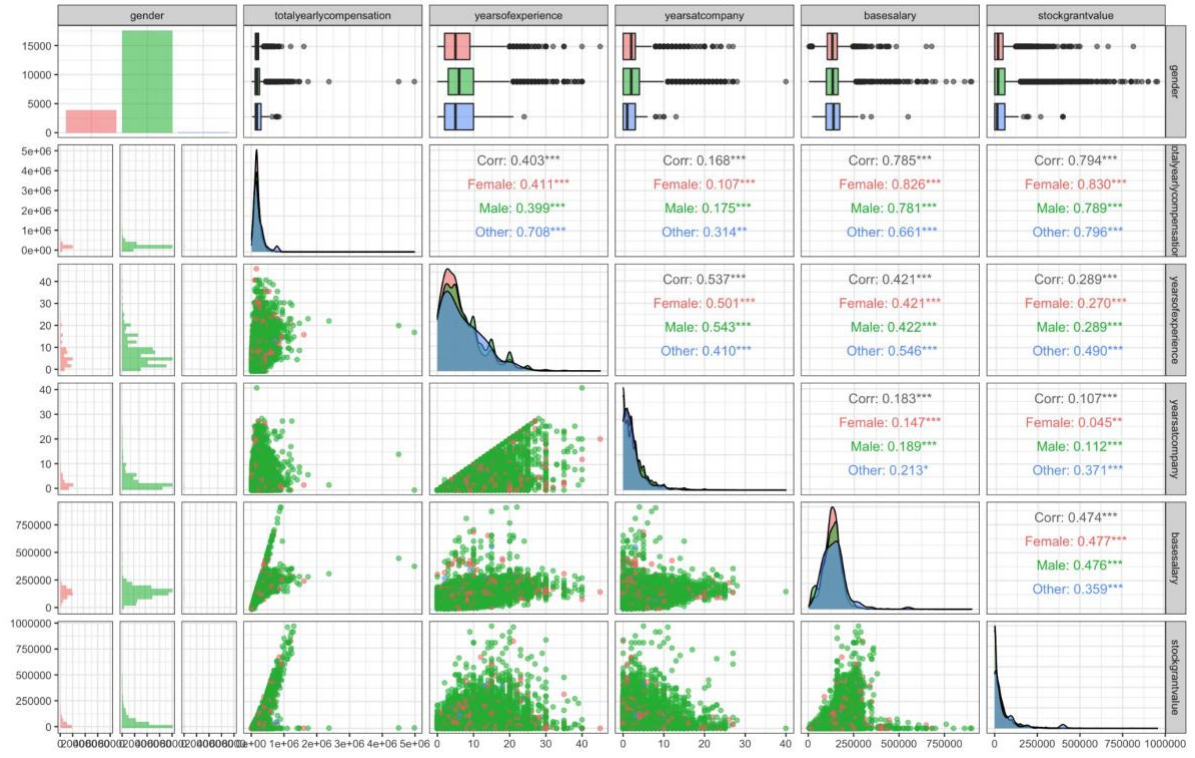

Figure 3: Paired plot by gender

Besides, we could clearly discover that all these histogram plots show strong right skewed distributions, which means for all these numeric variables, there exists outliers having extremely high value.

Only correlation between ‘base salary’ and ‘total yearly compensation’, and correlation between ‘stock value’ and ‘total yearly compensation’ are higher than 0.7. However, we decided to remove these two variables since these are components included in ‘total yearly compensation’, which is what you will get after people get an offer from a company, not the factor that would influence the total compensation. Thus, we will not include these variables in the regression model.

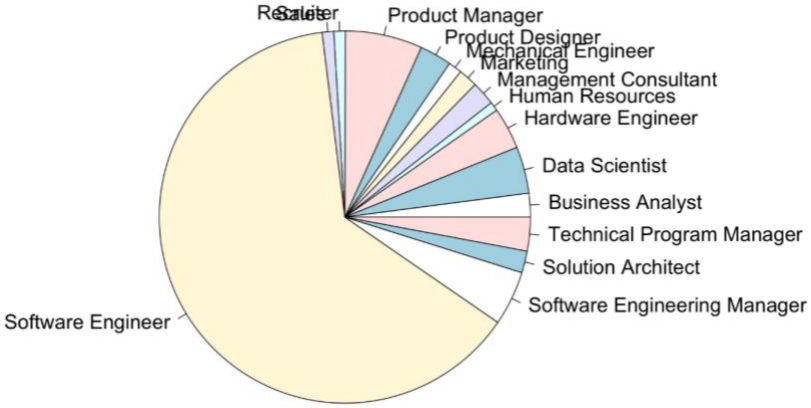

Figure 4: Pie Chart for Categorical Value ‘title’

As we analyze the components of the variable ‘title’, according to the pie chart, there are 15 job titles in our dataset, among all of these, ‘software engineer’ has the most percentage in our dataset. Based on our question of interest, we only included ‘software engineer’, ‘data scientist’, ‘business analyst’ and ‘management consultant’ in our regression analysis, since these are the most popular employment directions for statistics students.

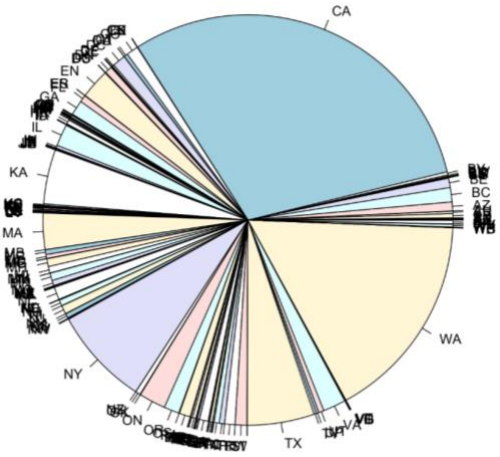

Figure 5: Pie Chart for Categorical Value ‘state’

For the ‘state’ variable, as we look into the figure 4 above, we sort the proportions of each state from high to bottom, and select 2 states from eastern side, 2 from the middle part, and 2 from the west coast of the United States from this rank. We eventually got six states that we are interested in, CA- California, WA - Washington, NY - New York, TX - Texas, VA - Virginia, IL - Illinois, which are the most popular and provide the most job opportunities’ state among the United states. After all the variable and data selections process, based on our question of interests, we have 9,414 data in total, with 7 predictors: ‘title’, ‘years of experience’, ‘years at company’, ‘gender’, ‘race’, ‘education’ and ‘state’, and ‘total yearly compensation’ as our response variable.

Outliers Analysis & Data Transformation

From both boxplots and histograms, there are outliers identified in these four quantitative variables. Their histograms also indicated that the distributions are right skewed with heavy tails, which suggested that data transformations were needed in order to obtain normal distributions (Figure 6).

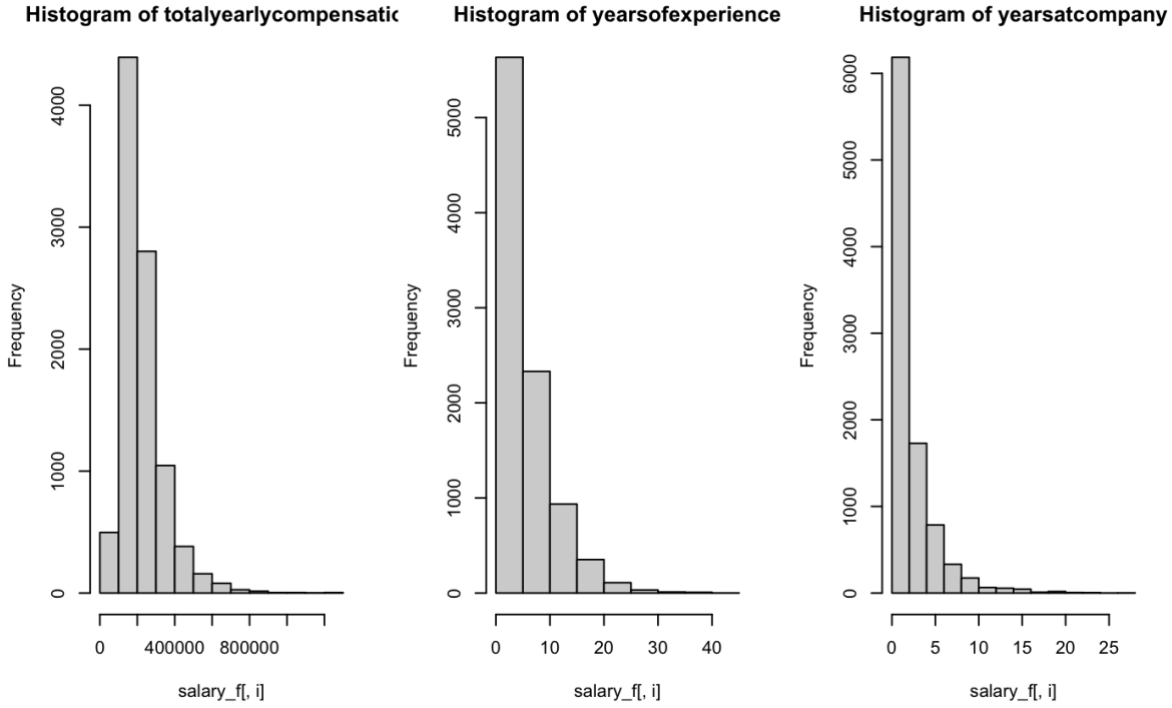

Figure 6: Histograms for Quantitative Variables before transformations

After trying different transformations on every variable, we decided to use log transformations for ‘total yearly compensation’ and square root transformation for ‘years of experience’. The resulting distributions obtained after transformations displayed in Figure 7 are roughly normal. No transformation is selected for ‘years at company’, since the distributions after all transformations were worse compared to the original data.

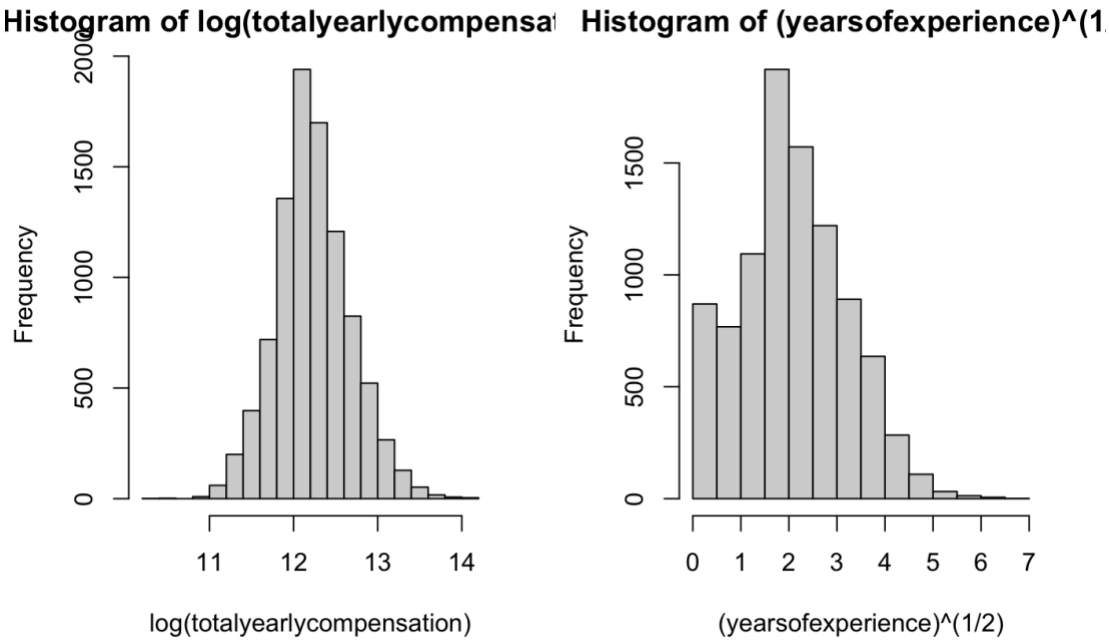

Figure 7: Histograms for transformed Quantitative Variables

After transformations, we noticed that the outlier issues were largely resolved, so there is no need to drop the outliers.

Models & Results

Multiple Linear Regression

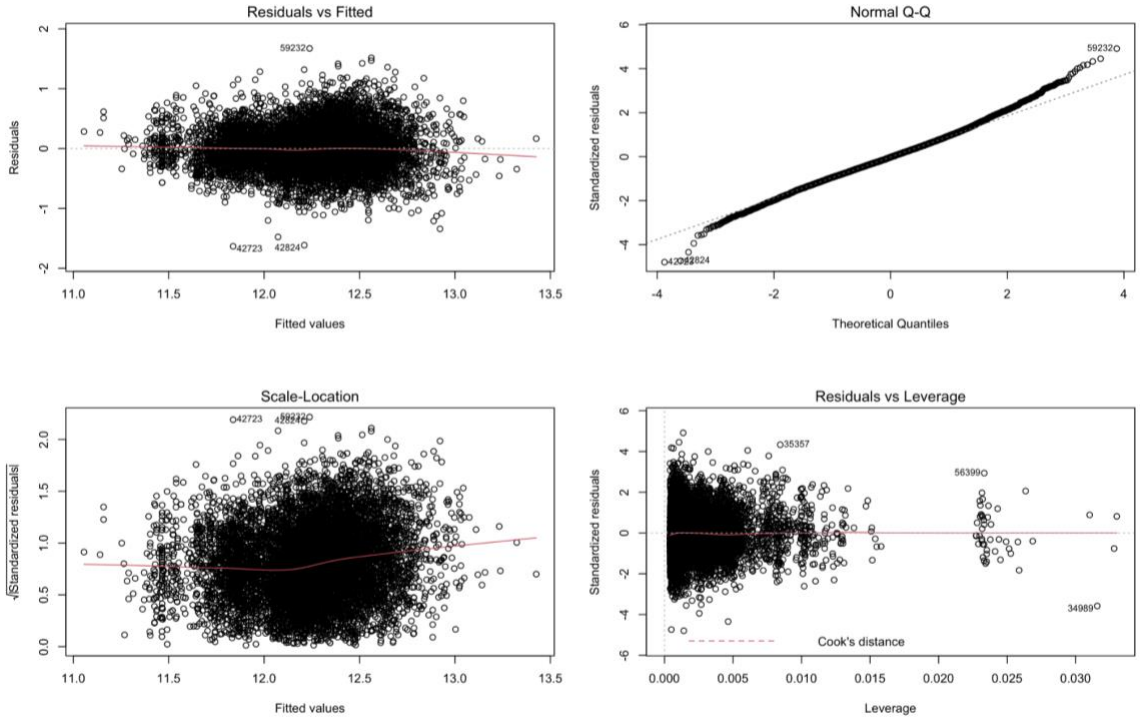

Figure 8: Diagnostic Plots for MLR

Full model:*totalyearlycompensation ~ title + yearsofexperience + yearsatcompany + gender + Race + Education + state*

From the Residuals vs Fitted Plot (Figure 8), there is no systematic correlation between residuals and fitted values, most residuals are distributed around the 0.

From the Normal Q-Q Plot, most points fall on the straight line. So, the normality assumption for residuals also holds. However, there are some outliers (#59232, #42723 and #42824).

From the Scale-Location Plot, the points around the horizontal line are randomly distributed, the assumption that the variance in the predicted model is a fixed value holds.

From the Residuals vs Leverage Plot, #35357, #56399 and #34989 can be seen as outliers.

Stepwise Regression

Null Model: A basic model with no X variable.

Full Model: A model with all first order X variables terms.

total yearly compensation ~ yearsofexperience + state + title + Education + yearsatcompany + gender + Race

Stepwise regression is to introduce independent variables into the model one by one and do an F test after introducing an independent variable to judge whether the introduction of the variable makes a significant change in the model. If a significant change occurs, then the variable is introduced into the model, otherwise ignore the variable until all variables have been considered. Since the stepwise regression arranges the variables in descending order of contribution and introduces them in order, we can get the order of importance of each variable.

Random Forest

Random forest uses bootstrap and bagging to train the data, which are algorithms for improving accuracy of machine learning models. Through Stepwise regression we have already know the order of importance of each variables, however, we are also curious about how exactly their ability of this model to contribute to the further prediction of the target variable are.

Table 1: Random forest importance

```
Call:
 randomForest(formula = totalyearlycompensation ~ ., data = salary_r,      ntree = 150, importance = TRUE, proximity =
TRUE)
               Type of random forest: regression
                     Number of trees: 150
No. of variables tried at each split: 2

          Mean of squared residuals: 0.1120224
                    % Var explained: 43.29
                    %IncMSE
title              42.99083
yearsofexperience 86.84319
yearsatcompany     32.30841
gender             15.85066
Race               20.18217
Education          47.62043
state              69.40663
```

The idea of random forest to evaluate the importance of variables, in general, is to see how much contribution each feature has made to each tree in the random forest. Then, take the average value of the contribution and compare the magnitude of the contribution between the variables. The

variable significance level is ranked as: Years of Experience > state > title > Education > years at company > gender > Race.

Statistical Analysis

F-test of Regression Effect

We state the hypothesis as following:

$$H_0: \beta_0 = \beta_1 = \beta_2 = \ldots = \beta_{20} = 0$$
$$H_\alpha: At\ least\ one\ of\ \beta_k \neq 0$$

Table 2: ANOVA table for MLR

```
Analysis of Variance Table

Response: totalyearlycompensation
                   Df  Sum Sq Mean Sq   F value    Pr(>F)
yearsofexperience   1  410.90  410.90 3540.2550 < 2.2e-16 ***
state               5  235.96   47.19  406.5952 < 2.2e-16 ***
title               3   57.92   19.31  166.3545 < 2.2e-16 ***
Education           4   47.86   11.96  103.0833 < 2.2e-16 ***
yearsatcompany      1   10.23   10.23   88.1752 < 2.2e-16 ***
gender              2    5.06    2.53   21.7865 3.632e-10 ***
Race                4    1.45    0.36    3.1143    0.0143 *
Residuals        9393 1090.21    0.12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to the ANOVA table from R for the multiple linear regression model, we could tell that the P-value is very small (<2.2e-10), which is smaller than the critical value 0.05. Thus, the null hypothesis is rejected, and the regression effect is significant.

T-test of Standardized Regression

We then performed standardized regression to identify the variable having the most significant impact on the regression model.

For every single variable, the hypothesis is stated as following:

$H_0: \beta_k = 0\ ;\ \ H_\alpha: \beta_k \neq 0$

From the ANOVA table of the forward stepwise regression displayed above, we could obtain the t-test p-values for all variables. Since all p-values are smaller than the critical p=0.05, all the variables are statistically significant to the model. However, “yearsofexperience”, “state”, “title”, “Education”, and “yearsatcompany” are relatively more significant ones. The importance level of these four variables is decreasing in order due to the fact that stepwise regression selects factors based on their importance. So it is fair to state that “yearsofexperience” has the most significant effect on the “totalyearlycompensation”.

The p-value results from regression summary table also agree to that in the ANOVA table, concluding that “title”, “yearsofexperience”, and “state” are the most important factors overly.

Table 3: Regression Summary Table for MLR

```
Call:
lm(formula = totalyearlycompensation ~ ., data = salary_r)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63378 -0.22133 -0.00795  0.20836  1.67230

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                11.490813   0.023941 479.962  < 2e-16 ***
titleData Scientist         0.337483   0.026087  12.937  < 2e-16 ***
titleManagement Consultant  0.132040   0.030386   4.345 1.40e-05 ***
titleSoftware Engineer      0.380293   0.021820  17.429  < 2e-16 ***
yearsofexperience           0.196834   0.003794  51.873  < 2e-16 ***
yearsatcompany             -0.013500   0.001433  -9.421  < 2e-16 ***
genderMale                  0.062284   0.009573   6.506 8.10e-11 ***
genderOther                -0.037911   0.051637  -0.734  0.46286
RaceBlack                  -0.029588   0.021861  -1.353  0.17595
RaceHispanic               -0.040488   0.016424  -2.465  0.01372 *
RaceTwo Or More             0.036995   0.019331   1.914  0.05568 .
RaceWhite                  -0.001013   0.008487  -0.119  0.90499
EducationHighschool        -0.032824   0.034113  -0.962  0.33597
EducationMaster's Degree    0.025859   0.007943   3.255  0.00114 **
EducationPhD                0.329013   0.016703  19.698  < 2e-16 ***
EducationSome College      -0.065095   0.030842  -2.111  0.03484 *
stateIL                    -0.394167   0.021620 -18.231  < 2e-16 ***
stateNY                    -0.096015   0.011186  -8.584  < 2e-16 ***
stateTX                    -0.467596   0.013435 -34.805  < 2e-16 ***
stateVA                    -0.411778   0.020849 -19.750  < 2e-16 ***
stateWA                    -0.066214   0.008703  -7.608 3.05e-14 ***
```

Pairwise Two Sample T-tests for Different Levels of Education

We then conducted two sample t-test for different levels in categorical factor “Education” to see whether there are significant differences among them on influencing the salaries. From the t-test outputs in R attached below, we could conclude that there are significant differences between all three pairs: Bachelor’s and Master’s, Bachelor’s and PhD, as well as Master’s and PhD based on the fact that all their p-values are extremely small (<0.05). Therefore, it is reasonable to reject all the null hypotheses and make the above conclusion.

Table 4: T-test Results Table for Education vs Salary

	bachelor vs master	bachelor vs phd	master vs phd
P-value	9.414519e-37	1.421995e-80	2.887902e-49
Decision	Significant Difference	Significant Difference	Significant Difference

Pairwise Two Sample T-tests for Different Races

Likewise, two sample t-tests for different levels in categorical factor “Race” are performed to see whether there are significant differences among them. The results from R indicated that the p-values for Asian vs Hispanic and Asian vs Black are extremely small (<0.05). Therefore, we reject the null hypotheses for them and conclude they different significantly. However, we can’t reject the null hypothesis for the rest pairs, since the p-values for them are larger than 0.05.

Table 5: T-test Results Table for Race vs Salary

	Asian vs White	Asian vs Two Or More	Asian vs Hispanic	Asian vs Black	Hispanic vs Black
P-value	0.767709	0.1525306	8.395671e-07	1.684251e-10	0.01458304
Decision	No Significant Difference	No Significant Difference	Significant Difference	Significant Difference	No Significant Difference
	White vs Two Or More	White vs Hispanic	White vs Black	Two Or More vs Hispanic	Two Or More vs Black
P-value	0.1347465	1.511649e-06	2.507882e-10	0.0339581	5.122671e-05
Decision	No Significant Difference	Significant Difference	Significant Difference	No Significant Difference	Significant Difference

Pairwise Two Sample T-tests for Different Genders

Similarly, another two sample t-tests for different levels in categorical factor “gender” are generated. The results from R indicated that only the p-value for male and female is smaller than 0.05. As a result, we reject the null hypothesis and conclude that salaries for male and female different significantly. There aren’t significant differences between other level pairs.

Table 6: T-test Results Table for Gender vs Salary

	male vs female	male vs other	female vs other
P-value	5.081358e-30	0.5465673	0.3274828
Decision	Significant Difference	No Significant Difference	No Significant Difference

Interpretation and Conclusion

Based on our model, we could conclude that “title”, “yearsofexperience”, “yearsatcompany”, “gender”, “Race”, “Education”, “state” do have significant effects on the total yearly compensation, and the influence level of each variable for an employer could be ranked as: Years Of Experience > state > title > Education > years at company > gender > Race.

According to our two sample T-tests, education level is a significant variable in the career market. How much you will get as compensation will depend on the professional ability you have. More education a person pursues, the more knowledge he/she will obtain, better handling ability he/she will have when facing the work, more salary he/she will deserve.

Besides, this is not a very fair career market in the United States for minorities. There exists a major difference within gender and race minorities. According to the research data from the United States Census Bureau, Women made up 27% of STEM workers in 2019. This also got proved in our two sample t-tests. There exists a significant difference between males and females. And based on the data provided by Pew Research Center, the components of STEM workers in the U.S. are white (69%), Asians (13%), blacks (9%) and Hispanics (7%). Blacks and Hispanics are underrepresented in the STEM workforce. By our two sample t-tests, there are significant differences between white and minorities. Although this has been improved a lot over the last 50 years and it already shows diversity in the career market, there is still a long path for eliminating the stereotype and building a market without the influence of employers’ identification or background.

From the scatter plot, we conclude that people with 10-20 years of experience would get the most stock, and those who work less years at the company would potentially get more stock. These may have four major reasons:

1. The proportions of Employers having more than 20 years of working experience which have a lot of stock may retire or quit office already, since they’ve earned a lot from previous career life which is enough for their later expenses.
2. Employers with 10-20 years of experience are basically at the age of 30-40 years old, which is a group of people that have not only more experiences than the newer members, but also better energy and passion at work, many of them are at the mid-level in the company. Even if they are job-hopping to another company, companies would be willing to give more stock to this group of employers.
3. People with 0-5 years at the company are those who graduated or changed jobs recently. For those people with working experience, when they are job-hopping, companies would give more stock in order to attract more excellent employers to the company, thus the stock value giving to them will increase, and those people who changed their job to a specific company more than 5 years ago would probably change to another company already, thus their stock value at this company may not appeared on the plot, which will explain why the stock value got decreased as the years at company increased.

4. Those who stay at the company for a long time may not have a significant increase on their salary, including base salary or the stock value, if they failed to gain promotion over their career life. Thus, the trend for people at a company for more than 10 years may be lower than for new people to the company.

From the outliers of the total salaries, we could see that people may earn extremely high salaries if they are at a very high level, such as vice president or manager. There is a very significant gap between the minimum and maximum salary, which means that the salary each company provides to those who are extremely excellent and have significant contributions to the company will be extremely high.

Limitation

During our data cleaning process, we found that almost ⅔ of the data in our dataset has missing values, in order to get a more accurate analysis we decided to remove them. It could be better when setting up the questionnaire, those important variables could be set as mandatory questions to prevent too many missing values.

Also, there are several columns such as ‘level’ and ‘company’, it appears hundreds of different level naming ways and doesn’t have a uniform standard, which will lead to a significant workload to combine the same level and manipulate them. It will be easier to process if those variables could be set to multiple choice questions such as to choose levels from ‘entry-level’, ‘mid-level’, ‘senior’, ‘principle’, ‘vice presidents’, etc. This would need the questionnaire designer to have a certain degree of specialized knowledge on the direction this questionnaire wishes to be explored, in order to make the data better processed and get more accurate results.

Meanwhile, the size of the sample of male and female is inconsistent and has a huge difference, which may lead to some bias. This also happens to other categorical variables. It would be more precise and somewhat eliminate the bias to get the conclusion if we could have a relatively equal size of the data for each level of the categorical variables.

References

Martinez, A., & Christnacht, C. (2021, October 8). *Women are nearly half of U.S. workforce but only 27% of STEM workers*. Census.gov. Retrieved December 10, 2021, from https://www.census.gov/library/stories/2021/01/women-making-gains-in-stem-occupations-but-still-underrepresented.html.

Funk, C., & Parker, K. (2018, January 9). *Diversity in the stem workforce varies widely across jobs*. Pew Research Center's Social & Demographic Trends Project. Retrieved December 10, 2021, from https://www.pewresearch.org/social-trends/2018/01/09/diversity-in-the-stem-workforce-varies-widely-across-jobs/.

Appendix Code

```
install.packages("plotrix")
install.packages("tidyverse")
install.packages("dplyr")

#Read data
data<-read.csv("Levels_Fyi_Salary_Data.csv")
head(data)

#Data preprocessing
#Delete encoded columns
keeps <- c('timestamp', 'company', 'level', 'title', 'totalyearlycompensation', 'location',
'yearsofexperience', 'yearsatcompany', 'basesalary', 'stockgrantvalue', 'bonus', 'gender', 'Race',
'Education')
data_new <- data[keeps]
head(data_new)

#checking sum of missing values for every columns
sapply(data_new, function(x) sum(is.na(x)))

#since we have 60,000+ observations, we can omnit all the NAs and still have enough
observations
salary <- na.omit(data_new)

#check whether there is still any nas
sapply(salary, function(x) sum(is.na(x)))

#parse location data into city, state, country
func_parse <- function(dataset){
n <- nrow(dataset)
city <- c()
state <- c()
country <- c()
for (i in (1:n)){
  l <- stringr::str_split(dataset[i, 6], ", ")
  city <- append(city, l[[1]][1])
  state <- append(state, l[[1]][2])
  if (is.na(l[[1]][3]) == T){
    country <- append(country, 'USA')
  }else{
```

```
    country <- append(country, l[[1]][3])
  }
}
}
func_parse(salary)
drop <- c('location',"timestamp","level")
df <- salary[,!(names(salary) %in% drop)]
df$country <- country
df$state <- state
df$city <- city

#ggpairs
library(GGally)
ggpairs(df[,c(9,3,4,5,6,7)],aes(color=gender,alpha=0.8))+theme_grey(base_size=9)+theme_bw()
ggpairs(df[,c(10,3,4,5,6,7)],aes(color=Race,alpha=0.8))+theme_grey(base_size=9)+theme_bw()
ggpairs(df[,c(11,3,4,5,6,7)],aes(color=Education,alpha=0.8))+theme_grey(base_size=9)+theme_
bw()

#distribution of title
library(plotrix)
title<-table(df$title)
pie(title)
lbl<-c("Business Analyst ","Data Scientist ","Hardware Engineer","Human
Resources","Management Consultant","Marketing","Mechanical Engineer","Product
Designer","Product Manager","Recruiter","Sales","Software Engineer","Software Engineering
Manager","Solution Architect","Solution Architect","Technical Program Manager")
pie3D(title, labels=lbl, main="Pie Chart of Titles")

#trimed data set for regression model
#not include; company, basesalary, bonus, city
drop2 <- c('company', 'basesalary', 'bonus', 'city')
temp <- df[,!(names(df) %in% drop2)]
t1 <- temp[which(temp$title=='Software Engineer'), ]
t2 <- temp[which(temp$title=='Data Scientist'), ]
t3 <- temp[which(temp$title=='Business Analyst'), ]
t4 <- temp[which(temp$title=='Management Consultant'), ]
temp2 <- rbind(t1, t2, t3, t4)
#country: select observations in USA only
temp3 <- temp2[which(temp2$country=='USA'), ]
#state: CA, WA, NY, TX, VA, IL
```

```
s1 <- temp3[which(temp3$state=='CA'), ]
s2 <- temp3[which(temp3$state=='WA'), ]
s3 <- temp3[which(temp3$state=='NY'), ]
s4 <- temp3[which(temp3$state=='TX'), ]
s5 <- temp3[which(temp3$state=='VA'), ]
s6 <- temp3[which(temp3$state=='IL'), ]
salary_reduced <- rbind(s1, s2, s3, s4, s5, s6)
head(salary_reduced)

barplot(table(salary_reduced$state),main="State Bar Plot",xlab="State",ylab="Frequency")
state<-table(df$state)
pie(state)

par(mfrow=c(1,3))
salary_f<-salary_reduced[,c(2:4)]
for(i in 1:3) {
hist(salary_f[, i], main=paste("Histogram of", names(salary_f)[i]))}

par(mfrow=c(1,2))
with(salary_f,{
  hist(log(`totalyearlycompensation`))
  hist((`yearsofexperience`)^(1/2))
})

salary_r<-salary_reduced[,c(1:4,6:8,10)]
salary_r$`totalyearlycompensation`<-
log(salary_reduced$`totalyearlycompensation`)#log(`totalyearlycompensation`)
salary_r$`yearsofexperience`<-
(salary_reduced$`yearsofexperience`)^(1/2)#(`totalyearlycompensation`)^(1/2)

#Linear Model
df_lm <- lm(totalyearlycompensation~.,data=salary_r)
summary(df_lm)
anova(df_lm)
par(mfrow=c(2,2))
plot(df_lm)

#AIC stepforward
none_mod <- lm(totalyearlycompensation ~ 1, data = salary_r)
full_mod <- lm(totalyearlycompensation ~., data = salary_r)
```

```
library(MASS)
df_fs= stepAIC(none_mod, scope = list(upper = full_mod, lower = ~1), direction = "forward", k = 2, trace = FALSE)
anova(df_fs)
summary(df_fs)

library(randomForest)
set.seed(100)
df_rf=randomForest(totalyearlycompensation~.,data=salary_r,ntree=150,importance=TRUE,proximity=TRUE)
print(df_rf)
importance(df_rf,type=1)

#t-test
head(salary_r)
#education & compensation
bachelor <- salary_r[which(salary_r$Education=="Bachelor's Degree"),]
master <- salary_r[which(salary_r$Education=="Master's Degree"),]
phd <- salary_r[which(salary_r$Education=="PhD"),]
t.test(bachelor$totalyearlycompensation,master$totalyearlycompensation, alternative = "two.sided")[3]
t.test(bachelor$totalyearlycompensation,phd$totalyearlycompensation, alternative = "two.sided")[3]
t.test(master$totalyearlycompensation,phd$totalyearlycompensation, alternative = "two.sided")[3]
#race & compensation
asian <- salary_r[which(salary_r$Race=="Asian"),]
white <- salary_r[which(salary_r$Race=="White"),]
twoone<- salary_r[which(salary_r$Race=="Two Or More"),]
hisp<- salary_r[which(salary_r$Race=="Hispanic"),]
black<-salary_r[which(salary_r$Race=="Black"),]

t.test(asian$totalyearlycompensation,white$totalyearlycompensation, alternative = "two.sided")[3]
t.test(asian$totalyearlycompensation,twoone$totalyearlycompensation, alternative = "two.sided")[3]
t.test(asian$totalyearlycompensation,hisp$totalyearlycompensation, alternative = "two.sided")[3]
t.test(asian$totalyearlycompensation,black$totalyearlycompensation, alternative = "two.sided")[3]
```

```
t.test(white$totalyearlycompensation,twoone$totalyearlycompensation, alternative =
"two.sided")[3]
t.test(white$totalyearlycompensation,hisp$totalyearlycompensation, alternative =
"two.sided")[3]
t.test(white$totalyearlycompensation,black$totalyearlycompensation, alternative =
"two.sided")[3]
t.test(twoone$totalyearlycompensation,hisp$totalyearlycompensation, alternative =
"two.sided")[3]
t.test(twoone$totalyearlycompensation,black$totalyearlycompensation, alternative =
"two.sided")[3]
t.test(hisp$totalyearlycompensation,black$totalyearlycompensation, alternative = "two.sided")[3]

#gender & compensation
male <- salary_r[which(salary_r$gender=="Male"),]
female <- salary_r[which(salary_r$gender=="Female"),]
other<- salary_r[which(salary_r$gender=="Other"),]

t.test(male$totalyearlycompensation,female$totalyearlycompensation, alternative =
"two.sided")[3]
t.test(male$totalyearlycompensation,other$totalyearlycompensation, alternative =
"two.sided")[3]
t.test(female$totalyearlycompensation,other$totalyearlycompensation, alternative =
"two.sided")[3]
```