An Empirical Analysis of the Potential Factors Relating to Employees' Salary in US Tech Companies

Mengyuan Cui 1006955305

melissa.cui@mail.utoronto.ca

ECO225 Big Data Tools for Economists

University of Toronto

April 20, 2023

Introduction

As a conventional wisdom, tech companies are paying a relatively decent salary to their employees. These companies mainly focus on the innovation and manufacture of electronic, technology-intensive products and services, which create a more convenient and connected world only in a few decades. By innovating ideas and inventing products, tech companies earned huge profits, and working in the tech industry is also viewed as having a decent job with considerable income. For example, Amazon, Apple, Google, and Microsoft are some big names among tech companies based in the US. According to Wheelwright (2021), the average salary in the US tech industry ranges from 20% to 85% higher than workers in other industries. However, tech companies are facing a long-term shortage of skilled employees, making them harder to retain skilled workers and drives up tech company salaries (Harvey Nash Technology Survey, 2015). The survey found 77% of employees who changed jobs in 2015 considered a high salary as the main reason for their job change, more important than work/life balance (72%) as the main motivator in the previous year. Additionally, Ayala and Echeverri (2009) found that workers in specific geographical regions, such as the San Francisco Bay Area, earn approximately 17% more than workers in other places across the US. These researches and reports give valuable information about the tech company salary in the US in recent years; however, an overall review of the potential factors of tech salary is missing in the current field of research.

Contributing to the literature gap, this project focuses on the potential factors relating to US tech company salaries, such as the employee's total year of experience, education attainment ratio, different job categories, and companies' geographic locations. As a conventional wisdom, the longer time one work in a field, the person is more likely to get a higher salary, even a promotion. Also, skill-based positions (such as software engineers) usually make more money compared to non-skilled positions (such as HRs). Collected from

the online database Kaggle, an open-source dataset of tech company salaries in 2016 (Telle, 2017) is used in the project to test the conventional wisdom, with two main independent variables (total years of experience and college education attainment), and a group of geographic dummies and job category dummies.

The linear regression and the machine learning model showed that total experience and college education ratio are positively correlated to tech salary, although the results are not very significant, suggesting the conventional wisdom is not fully supported. Besides, employees working in different job categories earn different salaries, yet most of the differences are insignificant. As for geographic locations, employees in the Western states (e.g. Washington and California) tend to have higher salaries, and their incomes keep increasing even after they earn more than 100000 USD per year. Some limitations of the project and restrictions of the dataset were also discussed at the end of the project. Future studies are encouraged to use a difference in difference design to investigate the current trend in tech salary, or find a proper instrumental variable to examine whether there is a causal relationship between the potential factors and tech company salaries. In the following parts, I will explain how I prepare the data and analyze the summary statistics and regression results, as well as show visualizations that support the findings.

Data and Methodology

The main dataset in this project is from the result of the 2016 Hacker News Salary Survey, containing tech salaries data around the world and its potential factors such as the base and bonus received, employees' experience and job category, and geographical information of employers. The dataset is at an individual level, where each data point represents a tech company employee. Since the main focus of the project is the US tech industry, only the data incorporating US tech companies were selected, resulting in 547 observations in the dataset. Missing values were dropped from the data, resulting in 397

observations remaining. I also defined tech salary as the sum of employees' annual base, signing bonus, and annual bonus received. An outlier with an extremely high salary (larger than 1 million USD) was dropped from the data. The geographic locations of tech companies were converted into 3 state dummies: *Eastern*, *Western*, and *Central*. Now the data is ready to be analyzed.

The main dataset is lacking two aspects of information: one is the demographic information about our population of interest, such as gender and education attainment ratio among employees in tech salaries; another is the information on employees' annual salary in years other than 2016, which can change the structure of our data to a panel data. Knowing either aspect could provide more comprehensive information about tech company salaries, resulting in more comprehensive analyses and more valuable insights.

In this project, I added data on college education attainment (i.e. people having their first college degree) among the population of each state, as a proxy for the educational attainment among tech company employees. Specifically, I merged two datasets with the number of people having a college degree (named "Field of Bachelor's Degree for First Major") and the 2016 census data with state populations, then got the percentage of people having college degrees among states. Both datasets were retrieved from the American Community Survey listed on the website of the US Census Bureau. Since higher education often correlates with a higher salary in the labor force (as conventional wisdom), I hope to find a positive relationship between education attainment and salary. However, I need to acknowledge the dataset is focusing on the entire state population (i.e. having unemployed workers and people who are out of the labor force), instead of a specified group of tech company employees.

Before doing any data analysis and interpretation, it is important to acknowledge the nature of this dataset, which is a cross-sectional, voluntarily-filled survey data. In other

words, the explanatory variables in this data are not randomly assigned and the regressions can only generate observational, correlational findings from the data. This limitation must be addressed before doing a causal inference of the effects of the potential factors on tech salary.

Visualization and Summary Statistics

Summary Statistics

In this section, I will discuss some characteristics of the dataset used in this project. Table 1 shows the average salary earned among the sample is approximately 113286 USD in 2016, and the standard deviation of salary is about 61396 USD. Note some employees in the dataset are earning zero wages, which prevents me from applying a log transformation to salary. The sample has an average of 7 years of total experience (standard deviation = 5.77), indicating employees are young in the tech industry and usually don't stay in the same company for a long time. The education attainment ratio is on average 23% among the population (standard deviation = 0.03) since the state labor force data is used as a proxy of college degrees received by tech employees.

Table 1. Summary Statistics.

	No. Observations	Mean	Standard Deviation	Min	25%	50%	75%	Max
Salary	397.00	113285.63	61395.82	0.00	74500.00	108500.00	140000.00	420138.00
Total Years of Experience	397.00	7.10	2.04	2.00	6.14	6.70	7.60	25.00
Education Attainment Ratio	397.00	0.23	0.03	0.14	0.21	0.22	0.25	0.30
Eastern	397.00	0.35	0.48	0.00	0.00	0.00	1.00	1.00
Western	397.00	0.43	0.50	0.00	0.00	0.00	1.00	1.00
Central	397.00	0.22	0.42	0.00	0.00	0.00	0.00	1.00
Software	397.00	0.52	0.50	0.00	0.00	1.00	1.00	1.00
Engineering	397.00	0.13	0.34	0.00	0.00	0.00	0.00	1.00
Data	397.00	0.06	0.23	0.00	0.00	0.00	0.00	1.00
Web	397.00	0.05	0.21	0.00	0.00	0.00	0.00	1.00
Management	397.00	0.08	0.26	0.00	0.00	0.00	0.00	1.00
Applied Science	397.00	0.00	0.05	0.00	0.00	0.00	0.00	1.00
Other	397.00	0.17	0.38	0.00	0.00	0.00	0.00	1.00

In the Appendix, Figures 1-3 shows the distributions of these variables. For the total year of experience, most of the US tech employees have 3~6 years of working experience.

Considering information technology is a profession with relatively young people, the results generated from the histograms are not very surprising. For the salary paid to these employees, I found most of them have an annual salary of approx. 50000 to 150000 USD, higher than the median income in the US, which is approx. 27419 USD in 2016 according to the United States Census Bureau. This finding aligns with our background information that people working in tech companies often have a decent income compared to other professions.

As for the dummy variables, I found there are 35% of employees in the sample working in Eastern states, 43% working in Western states, and 22% working in Central states. As for their job categories, most of the employees belong to the *Software* category, occupying 52% of the sample, and the least employees are from the *Applied Science* category, occupying only 0.25% of the sample. About 17% of the respondents are having jobs that are difficult to categorize, and thus were put in the *Other* category. Two bar charts (Figures 4 and 5) were generated to compare salaries in these geographic regions and different job categories. From the charts, I conclude employees in the Western states and working in the applied science category earn the highest salary compared to others, and the employees in Central states and working in the data category have the least wage than others.

Maps

Three maps were created: one for the outcome variable (salary), and two for our variables of interest (total years of experience and education attainment) in this project. Note some states are having white color since there is no company located in these states according to our dataset. Thanks to having the geographic information of US tech companies in our dataset, I labelled all companies with red dots in the maps.

The average tech company salaries by states are shown in the first map (Figure 6). Aligning with our previous findings, Washington and California are the two states having the highest salary at the state level, and both are Western states in the US. Some Eastern and

Central states (such as New York and Tennessee) also have a high state-average salary. In general, tech companies in the Eastern and Western states are paying on average a higher salary to their employees, compared to those in the Central states.

In the second map (Figure 7), total years of working experience are calculated by states. Similar to our previous findings, tech companies have many younger workers usually with less than 15 years of experience. The Western states recruit particularly young employees, with no more than 10 years of prior experience on average. Alabama stands out from the map, showing the companies are having the highest number of years of experience.

In the third map (Figure 8), I plot the percentage of people attaining college education among states. Generally speaking, Eastern states have a relatively high education attainment ratio, especially in the 13 founding states such as Massachusetts, New York, Maryland, and Virginia. The three west-coast states (Washington, Oregon, and California) also have relatively high education attainment compared to the central area. Among the Central states, Colorado stands out from the map with a higher percentage of people having college degrees. Linking to the results in the first map, this map may reveal that employers would like to choose states with a relatively more educated population as their companies' location. The Kulkarni et al. (2022) study brought out the same idea, contending tech unicorns (i.e. companies with a market value over $1 billion without being listed on the stock market) are actively expanding and looking for locations where they can absorb the most talented workers. My findings match the results from Kulkarni et al.'s research, suggesting locations may benefit tech companies by easier reaching out to and recruiting people with higher education.

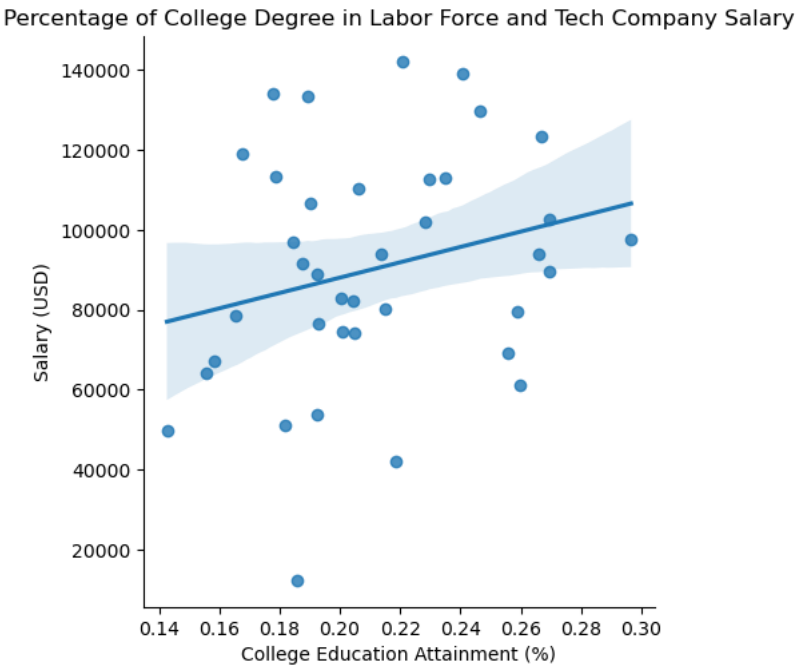
Initial Findings

With the preliminary results from summary statistics and visualizations, I drew some scatterplots to investigate the correlation between employees' total years of experience/

education attainment and tech salary, respectively. Here I created two bin scatterplots (Figures 9 and 10) to get the average value of the independent variables in each bin, showing the scatterplot more cleanly and nicely (where dots are not condensed).

Figure 9. The Relationship between Total Experience and Tech Salary.

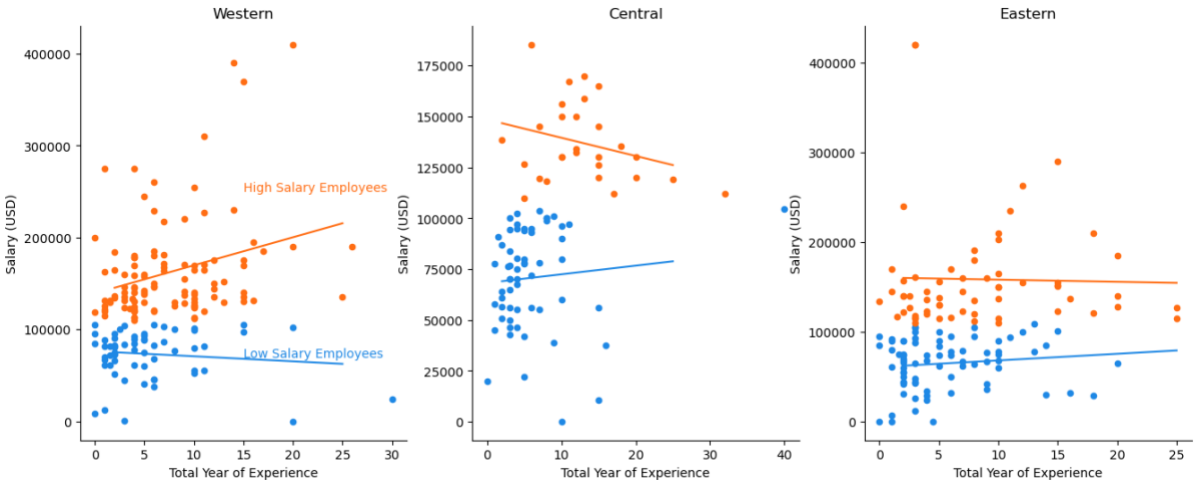
Figure 10. The Relationship between Educational Attainment Ratio and Tech Salary.



Both plots indicate a positive relationship between the independent variable and the salary in US tech companies. Especially for the education variable, the result matches the Harvey Nash Survey (2015), saying tech companies would like to attract and retain the best workers with required skills to catch up and even lead the change in the tech industry. Besides, the conventional wisdom of having higher education linked to a higher income is also validated in the context of the US tech industry. However, outliers in the data are messing with the correlation, making the result less significant. Another caveat is these two graphs only showed simple linear regression results with one x-variable and the y-variable, possibly having omitted variables that can cause inconclusive results.

Taking a step further, three geographic regions were incorporated in the visualization below, showing the relationship between total experience and salary in each region. Here I divided salary into two subgroups by 108500 USD (the median salary in the dataset) to see if any pattern can be found between the high- and low-salary groups.

Figure 11. Trends in High- and Low-salaries by Region.



From the graph above I found the high-salary group has increasing salaries with increasing years of working experience only in Western states, and has decreasing salaries in Eastern and Central states. However, this trend was reversed in the low-salary group: In Western states, salary is negatively correlated to the total experience, while in Eastern and

Central states, these two variables are positively associated. The underlying reason is currently unclear, yet we can still get the intuition that tech companies in the Western states are somehow different from those in other states. Moreover, some limitations in the graph need to be acknowledged, such as I failed to add more controls into this visualization and the significance of the results is unknown. To address these problems, a set of multiple linear regressions were run and discussed in the next section to further investigate the potential factors relating to tech salary.

Regression Results

Linear Regression

Continuing with the initial findings, there are two main x-variables in the regression models: one is the tech employee's total year of experience and the other is the percentage of people having their first college degree in each state (as a proxy of employee's educational attainment). Before making any model specifications, I assume there is a positive linear relationship between these factors and the outcome variable (salary). Then, the geographic state fixed effects and the job category fixed effects will be included in some models to better interpret the main variables of interest. Finally, I will use a non-linear specification with quadratic terms to compare the results with linear models and to see if the specification could lead to a higher predictability of the model.

Baseline Model

Starting with the baseline model, I use total years of experience and college educational attainment as two independent variables in a multiple regression model to study the relationship between these variables and salary. The baseline model has the following equation:

$$Salary_{ij} = \beta_0 + \beta_1 \times Total\ Experience_i + \beta_2 \times Education\ Attainment_j + \varepsilon_{ij}$$

Where β_1 and β_2 are the coefficients for the two independent variables, and ε_{ij} is the error term. The subscript i indicates the observation is on individual level and the subscript j indicates the observation is on state-level.

Table 2. Regression Outputs.

Table 2 - Regression Outputs.

	Dependent variable: Salary			
	Baseline (1)	Geographic FE (2)	Job Category FE (3)	Nonlinear (4)
Constant	60653.97**	63735.57*	55737.34	-436424.86***
	(29007.19)	(36458.02)	(38225.85)	(154717.59)
Total Year of Experience	652.91	1611.25	1356.17	2645.66
	(1559.43)	(1536.15)	(1542.57)	(4469.00)
Total Experience Squared				-24.35
				(217.38)
Educational Attainment	210186.59**	128001.46	91926.30	4254151.43***
	(105113.16)	(134288.23)	(134226.42)	(1292773.38)
Education Squared				-8831092.47***
				(2720996.90)
Central		-9813.60	-11339.36	-1804.32
		(10997.41)	(11043.23)	(11423.64)
Western		25831.45***	23003.33***	18519.97**
		(7271.88)	(7317.19)	(7407.28)
Applied Science			75790.72	75253.04
			(60999.38)	(60340.34)
Data			1781.47	3720.57
			(18684.34)	(18517.64)
Engineering			19460.59	22092.70
			(15967.78)	(15856.33)
Management			18936.17	21893.18
			(17431.65)	(17260.92)
Other			8699.95	10465.21
			(15540.11)	(15394.23)
Software			26671.62*	29123.85**
			(14278.53)	(14148.08)
Observations	397	397	397	397
R-squared	0.01	0.07	0.09	0.12
Adjusted R-squared	0.01	0.06	0.07	0.09
Residual Std. Error	61240.91	59591.27	59293.41	58627.36
Mean Squared Error	3750448983.67	3551119058.99	3515708482.16	3437167251.35
F Statistic	2	7.09***	3.86***	4.19***
Note:				*p<0.1; **p<0.05; ***p<0.01

In Table 2 Model (1) (see above), I found the annual salary of one with zero years of experience working in a US tech company is approximately 60654 USD. Note this is a

statistically and economically significant result: the coefficient has a p-value smaller than 0.05 (passing the 5% significant level) and has practical interpretations showing the entry salary in the US tech industry. After controlling education attainment, having one more year of total experience is associated with on average a 653 USD increase in salary. This result matches our conventional wisdom that the longer time stayed in an industry, the more income will earn. However, the coefficient is not significant. It is not statistically significant because the p-value failed to pass even the 10% significant level, and not economically significant once we considered the average salary shown in the summary statistic table: a one standard deviation difference in the x-variable is associated with only 6.14% of one standard deviation difference in salary.

As for another independent variable, I found a 1 percentage point increase in employee's college education attainment ratio is correlated with approximately 2102 USD increase in salary. Note this result is statistically significant at a 5% significance level (p-value less than 0.05) but economically insignificant due to only having a 10.27% difference in one standard deviation of salary. Overall, the coefficient of determination (R-square) in this model is close to zero (0.01), indicating there is almost no correlation and we should not be surprised by having an inconclusive result, also suggesting more variables and controls should be added to the regression to better explain the y-variable.

Geographic Region Fixed Effect

Continuing with the previous model, in Model (2) I added the geographic dummies (*Eastern*/*Western*/*Central*) into the regression, hoping to increase the explaining power of the overall model. Note the Eastern states dummy was left-out from the regression to make the comparison. Here is the regression formula:

$$Salary_{ij} = \beta_0 + \beta_1 \times Total\ Experience_i + \beta_2 \times Education\ Attainment_j + \delta_1 \times Central_i + \delta_2 \times Western_i + \varepsilon_{ij}$$

Where δ_1 and δ_2 are the coefficients for the two geographic dummies.

After controlling for geographic effects, the annual salary of one with zero experience and no college degree has slightly increased to approximately 63736 USD, which is statistically significant at a 10% significance level (p-value < 0.10). After controlling all other variables, having an extra year of total experience is now associated with on average a 1611 USD increase in salary, higher than the previous estimation in Model (1). Note the result is again statistically and economically insignificant. As for educational attainment, I surprisingly found an insignificant result, in which a 1 percentage point increase in college education attainment correlates with a 1280 USD salary increase, lower than in Model (1). The R-square in this model has increased to 0.07 (with adjusted R-square increased to 0.06), suggesting an improvement in model explanation power. Although the results are not very informative, at least I am on the right track explaining what relates to tech company salaries.

To look closer at the difference in geographic regions, I found that tech companies in Western states are paying way more salaries as employees work longer, with approximately 2831 USD higher than the average salary paid in Eastern states. The companies in the Central states pay the least to their employees, which is about 9814 USD less than the average salary in Eastern states. However, note only the result of the Western states is statistically significant (with a p-value smaller than 0.01). This notable difference is likely because Western states have some regions famous for developed tech industries and big-name companies, such as Meta in San Francisco Bay Area and Microsoft in Seattle. This result can be supported by Echeverri-Carroll & Ayala's (2009) study, in which they found a spatial concentration of tech companies in specific regions, such as Silicon Valley, where companies are competing for human resources and are willing to pay more salaries to their employees.

Job Category Fixed Effect

Similar to the previous step, in Model (3) I added a group of 7 job category dummies (such as engineering, software, management, etc.) to control for the job category fixed effects. Note the dummy *Web* is used as the omitted variable in the regression to compare the results. The regression formula can be found below:

$$Salary_{ij} = \beta_0 + \beta_1 \times Total\ Experience_i + \beta_2 \times Education\ Attainment_j + \delta_1 \times Central_i + \delta_2 \times Western_i + \delta_3 \times Applied\ Science_i + \delta_4 \times Data_i + \delta_5 \times Engineering_i + \delta_6 \times Management_i + \delta_7 \times Other_i + \delta_8 \times Software_i + \varepsilon_{ij}$$

Where δ_3 to δ_8 are the coefficients for the job category dummies.

After controlling all other variables, the annual salary of one with zero years of experience is now decreased to on average 55737 USD, and an extra year of total experience is now associated with an approximately 1356 USD increase in salary. Having a 1 percentage point increase in educational attainment is associated with an even lower, on average 919 USD increase in salary. However, all these coefficients are not significant. After controlling all other variables, employees working in the web category earn the least salary compared to those in other categories, and those who work as applied scientists have the highest salary, which is about 75791 USD more than *Web* employees. However, only the coefficient of the *Software* category is statistically significant (with p-values passing the 10% significant level), suggesting working as a software engineer may be an influential factor in tech salary. Although the overall model has an R-square further increased to 0.09 (adjusted R-square increased to 0.07), the F-statistic in this model is lower than Model (2), also suggesting an overall not very effective improvement in the model's predictability.

Non-linear Specification

Since Model (3) doesn't perform a substantial improvement in our regression model, I reconsidered the model and chose a nonlinear specification with quadratic terms (total

experience squared and education attainment squared) to further explore the results. Note this is still a linear regression model but has non-linear features (the quadratic terms).

$$Salary_{ij} = \beta_0 + \beta_1 \times Total\ Experience_i + \beta_2 \times Education\ Attainment_j + \gamma_1 \times Total\ Experience_i{}^2 + \gamma_2 \times Education\ Attainment_j{}^2 + \delta_1 \times Central_i + \cdots + \delta_8 \times Software_i + \varepsilon_{ij}$$

Where γ_1 and γ_2 are the coefficients of the quadratic terms.

In the final model (Model 4), the model intercept is uninterpretable because it changed to a negative value due to the model specification (salary should not be negative). After controlling all other variables, an extra year of total experience is now correlated with on average 2646 USD increase in salary, and a 1 percentage point increase in educational attainment is associated with a surprising 42541 USD increase in salary on average. Note the coefficients for total experience and its squared variable are both statistically and economically insignificant, and the coefficients for education attainment and its squared value are both significant (with p-value < 0.01; a one standard deviation difference in the education is associated with more than 2 standard deviation differences in salary). As for the controls, the Western state dummy and the software category dummy are still statistically significant as in the previous model. The overall model also shows an improvement in R-square (further increased to 0.12) and adjusted R-square (increased to 0.09), suggesting the nonlinear specification better fits the data and enhance the explaining power of the model.

Machine Learning

Methodology

In the final part of the project, I used machine learning to provide an accurate mapping from my independent variables to the output variable salary. The regression model is similar to the linear regression models in the previous part but with location longitude and latitude variables added. And since the decision tree is a non-parametric model, here I do not

need to drop the employer years of experience, *Eastern* states dummy, and *Web* category dummy to avoid collinearity. My goal is to minimize the mean squared errors (MSE) in the regression (i.e. the difference between the predicted value and the actual value). The objective function of the regression analysis is shown below:

$$\min_{j,s}\left[\sum_{i:x_{i,j}\leq s, x_i\in R1}(y_i-\hat{y}_{R1})^2+\sum_{i:x_{i,j}>s, x_i\in R2}(y_i-\hat{y}_{R2})^2\right]$$

In this function, R_1 and R_2 are both from a rectangular region R that contains all values of X. By iteration, a feature and location j and s is chosen to split R into two parts to minimize MSE, and this process is repeated with R_1 and R_2 and the smaller rectangles they produced. By iterating this process, finally I got a regression tree with all branches generated by machine learning. How does the model know when to stop iterating? This is because I chose a maximum depth of the tree, which is 3. The iteration would stop when the depth of the tree reached this value.

A related concept of the stopping rule is regularization and pruning. The regularization process penalizes a regression tree if it has extra nodes and branches, but does not increase the model predictability (i.e. it's overfitting the data). Pruning is a way to regularize a regression and avoid model overfitting. A penalty function often includes regularization parameters such as the minimum leaf size, maximum depth of the tree, and α in the pruning process. Below is the penalty function used in my regression:

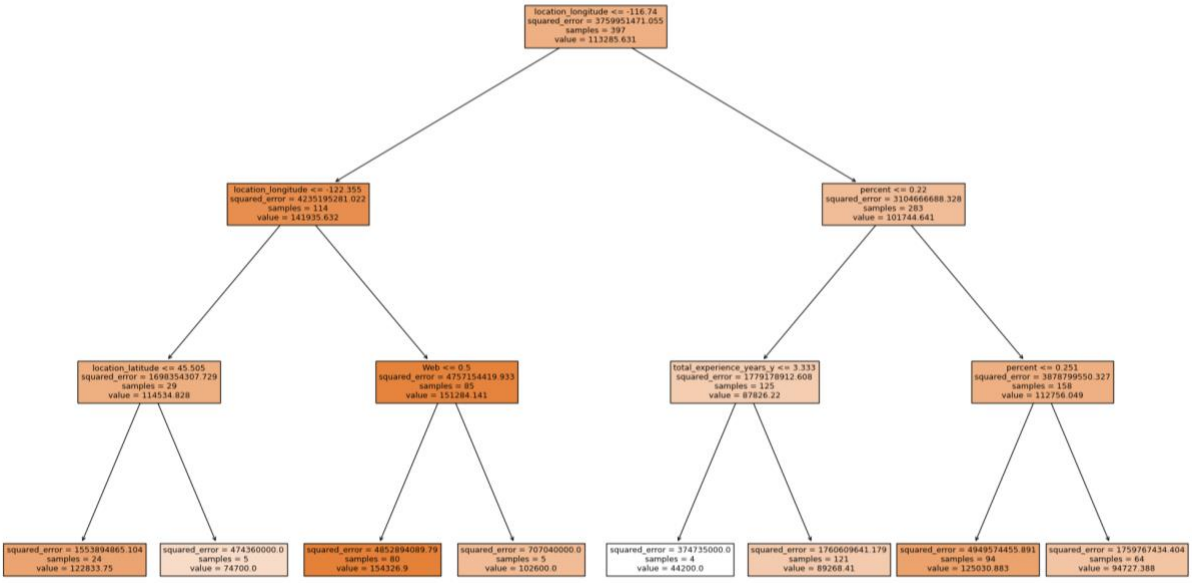
$$\min_{tree\subset T}\sum(\hat{f}(x)-y)^2+\alpha|\text{terminal nodes in tree}|$$

Where α is the pruning parameter, usually between 0 and 1. Increasing the value of α would lead to a smaller tree and vice versa. As for another important regularization parameter, the maximum tree depth in our regression is set to be 3, which limits the number of terminal nodes in the tree. Increasing the maximum tree depth would lead to a larger tree, often with smaller MSE, but may be overfitting the data. With the regression tree created, I

found the mean squared error (MSE) of the tree with a depth of 3 is 3082745152, which is better (smaller) than the MSE of linear regression models. Although I can further decrease the MSE in regression by increasing the tree depth, I chose not to do so to avoid overfitting because I don't want the model to have a low prediction power.

Regression Tree

Figure 12. Regression Tree.



In the regression tree with the max tree depth of 3 (see above), I found the most contributed variable in the regression is the location longitude of US tech companies. The data is separated into two groups, one with a longitude less than or equal to -116.74 (i.e. US West Coast including Washington, Oregon, and California), and the other group with a longitude more than -116.74. This result matches the previous finding in linear regression, suggesting a 40191 USD difference in salary between Western and other states (for example, the Bay Area is in the West). Note the longitude variable is not included in the linear regressions because its numerical value has no meaning while interpreting, and I changed the company locations into 3 categorical dummy variables. However, since regression trees do not impose linearity or monotonicity, the longitude and latitude variables can be used as a feature to separate data.

According to the regression tree, educational attainment further separates the sample at the East of longitude -116.74. Employees in states with a college education attainment ratio greater than 22% are earning a 24930 USD higher salary on average. This finding also matches the result in the linear regression models, suggesting the importance of having a higher education/ living in a state with more educated people. As for the sample at the West of longitude -116.74, the longitude -122.36 further divides the sample into two groups. This line of longitude separates the well-famous high-tech regions such as Seattle and Silicon Valley from the East of the states. However, due to the small sample size (one group with 85 observations, and the other group has only 29 observations), I could not conclude any insightful results.

Speaking of the number of observations in each subgroup, a regression tree can show the divisions of data by showing the size of the subgroups. For example, the longitude variable separates the sample into two groups: one has 114 observations and the other contains 283 observations. However, the number of observations at each terminal node is no greater than 121. Having such a small sample size at each node is not good for the data analysis, making me difficult to determine if a particular result is true and unbiased.

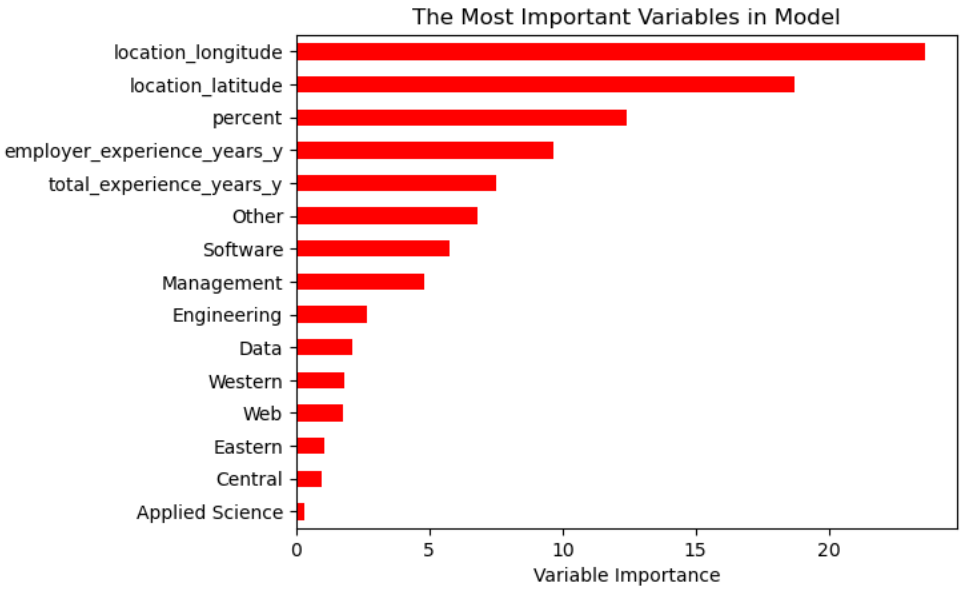
Random Forest

Finally, I used a random forest to find the best regression that minimizes the mean squared error. Random forest is a machine learning algorithm combining the output of several decision trees to attain a single result. A random sample drawn from the set of independent variables is selected at each node, then the tree chooses the best-split variable from the sample. By having different groups of x-variables, the random forest provides a solution of having similar/ correlated trees, which may result in little improvement of the regression.

In this model, a group of 5 x-variables was randomly selected at each node. By iteration, I found the best model with MSE of 1917663837, which is only 2/3 of the error in the previous tree! This substantial decrease in MSE shows a great benefit of having a random forest, which forces the trees to split on different variables and avoid having common nodes.

From the random forest result, the algorithm records the amount of averaged reduction in mean squared error due to splits, among all regression trees. If a splitting variable chosen from the set of x-variables leads to a large reduction in MSE, we conclude this variable is important. Thus, I plotted a variable importance figure (see below) to show which variables are important in reducing the MSE in the regression.

Figure 13. Variable Importance.



From this graph, I found the location variables with longitude and latitude data are most important in explaining tech company salary. Following the geographic variables, the third important variable is the college education attainment ratio, similar to the linear regression results. The experience-related variables (total year of experience and the year of experience at current employer) are not the most important factors relating to salary, ranking at 4th and 5th on the variable importance graph. Although the linear regression also generates insignificant results for the two variables, this result is still surprising because it contradicts

our conventional wisdom: it's not always the case that the longer one stayed in an industry/ a position, the more money one earned. Following the two experience variables, two job category dummies (*Other* and *Software*) are ranking 6th and 7th on variable importance, partially matched to the linear regression results (where the *Software* category has statistically significant coefficients).

Limitation

This project provides an overview of the potential factors such as working experience, education attainment, job types, and geographic locations relating to the salary earned in the technology industry in the US. However, this project has some caveats and many results are inconclusive. One possible reason for the insignificant results can be having a small sample size (with only 397 observations in the dataset), leading to large standard errors. A further study can increase the sample size by adding more observations to the data. For example, one can find the tech salary data in different years to create a panel data, which would improve the significance of regression results and show a trend over time.

The second limitation is lacking control variables (such as age, gender, race, and ethnicity among employees) in the regressions, and the regressions may suffer from omitted variable bias. For example, according to Lee (2013), Asian immigrants working in the tech industry are experiencing some patterns between ethnic niches (i.e. companies are disproportionately owned and/or staffed by ethnic minorities) and their salaries. To address the omitted variable bias, future studies can add related information and run regressions with more comprehensive data. A potential dataset, State and Metro Area Employment from the US Bureau of Labor Statistics, has state-level demographic data of employees' age, gender, race, and employment rate in US labor force that can be used to scrape and analyze. By adding these variables into the regression model, the multivariate model should provide a stronger prediction of salary in the US tech companies. However, this data has a population

of the entire US labor force, which cannot perfectly match the population of employees in tech companies. Future researchers can also use a dataset with a population that better match the topic.

The third limitation is also about the data itself. Since the data only has tech salaries in the year 2016, it could not give us the most recent trend in the field, resulting in less valuable insights. For example, what are the changes in salary during the pandemic? In this project, I found a relatively competitive labor market in the tech industry, where companies are paying high salaries to more educated people. Is this trend become more prevalent during Covid, when more people are working from home? According to Normandin Beaudry (2022), IT companies in Canada are planning to increase average salaries to a 5.8% increase in 2023, as evidence of adapting to the competitive labor market. Would US tech companies do the same and would these firms continue to increase their salary in the future? Future research could source a similar dataset containing tech salaries in the post-Covid era, and use the panel data to implement a difference in difference design to study the effect of the pandemic on tech salary and answer the above questions.

Last but most importantly, since the data does not randomly assign the independent variables to salary (the survey respondents are not a random sample), this project only captured some correlations between the potential factors and salary but failed to investigate any causal effects. To solve this problem, future studies can collect the data from a random sample with individual-level education data to investigate if the potential factors are affecting tech salary. For example, education attainment may have a causal effect on tech salary, but it could not be randomly assigned. Besides, motivation is possibly a confounding variable affecting both education attainment and salary: people with higher education tend to be highly motivated, yet this motivation can also lead them to more competitive positions and thus higher salaries. If researchers can find a proper instrumental variable correlated with

educational attainment, the instrumental variable can control for the confounding motivation and thus can estimate the true treatment effect of education on salary.

Conclusion

This project investigates the potential factors (such as total years of experience, education attainment, company locations, and job categories) of US tech company salaries, aiming to find any meaningful relationship between these factors and salary. With the two independent variables and a group of geographic and job category dummies, the linear regression and machine learning results showed that geographic locations and education attainment are the two main factors of tech company salary, suggesting companies are willing to pay a higher salary in Western states and/or to employees with a college degree. Doing a software job is also found to have a higher income. To my surprise, no substantial results are showing that experience is correlated to salary, which contradicted the conventional wisdom that the longer one worked, the more money one earned. Future studies are encouraged to use a difference in difference design to investigate the effect of the pandemic on salary, or find a proper instrumental variable to examine whether there is a causal relationship between education attainment and tech company salaries.

References:

Echeverri-Carroll, E., & Ayala, S. G. (2009). Wage Differentials and the Spatial Concentration of High-technology Industries. Papers in Regional Science, 88(3), 623–641. https://doi.org/10.1111/j.1435-5957.2008.00199.x

IT Skills Crisis Drives Tech Salaries Up, Companies Struggle to Retain Talent, Says Harvey Nash Survey. (2015). PR Newswire Association LLC.

Kulkarni, A. U., Murkute, P. A., Kamble, H., & Sinha, K. (2022). An Exploratory Study on Salary Bubble & its repercussions on employee retention on techies of Indian Edutech and Tech Companies. International Journal of Early Childhood Special Education, 14(5), 2281–2286. https://doi.org/10.9756/INTJECSE/V14I5.238

Lee, J. C. (2013). Employment and Earnings in High-Tech Ethnic Niches. Social Forces, 91(3), 747–784. https://doi.org/10.1093/sf/sos199

Normandin Beaudry. (2022). Salary increases: Organizations are constantly adapting. Cision Canada. Retrieved from https://www.newswire.ca/news-releases/salary-increases-organizations-are-constantly-adapting-868298820.html

Telle, B. (2017). 2016 Hacker News Salary Survey Results. [Data set]. Retrieved 2023, from https://data.world/brandon-telle/2016-hacker-news-salary-survey-results.

United States Census Bureau. (2022). American Community survey 5-year data (2009-2021). Retrieved 2023, from https://www.census.gov/data/developers/data-sets/acs-5year.html.

U.S. Bureau of Labor Statistics. (2023). State and Metro Area Employment, Hours & Earnings. [Data set]. Featured SAE Searchable Databases. Retrieved 2023, from https://www.bls.gov/sae/data/.

U.S. Census Bureau. (2023). American Community Survey (ACS) Demographic and

Housing Estimates. [Data set]. Retrieved 2023, from https://data.census.gov/table?q=population%2Bin%2B2016&g=010XX00US%240400000.

U.S. Census Bureau. (2023). Field of Bachelor's Degree for First Major. [Data set]. Retrieved 2023, from https://data.census.gov/table?q=s1502&g=010XX00US$0400000.

Wheelwright, T. (2021). Top tech salaries in the US. Business.org. Retrieved from https://www.business.org/hr/benefits/highest-tech-salaries/.

Appendix

Figure 1. Histogram of Employee's Total Years of Experience.

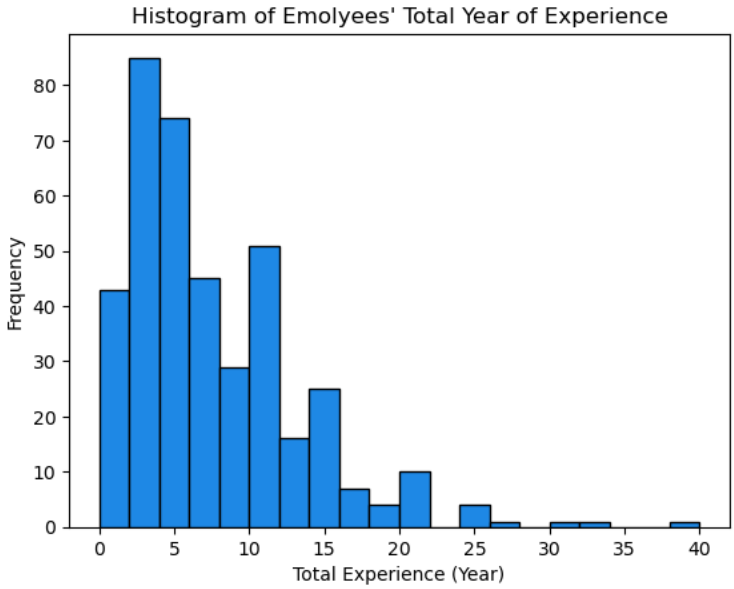


Figure 2. Histogram of Employees' College Educational Attainment Ratio.

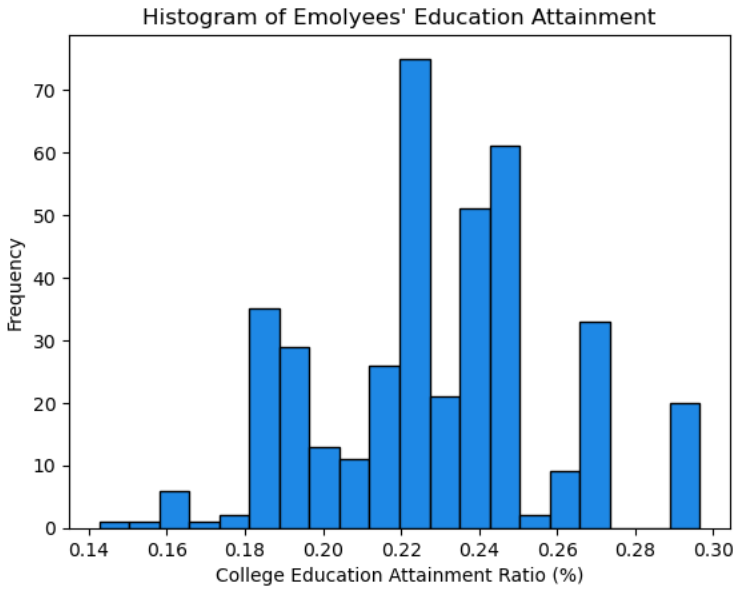


Figure 3. Histogram of Employee's Salary in US Tech Companies.

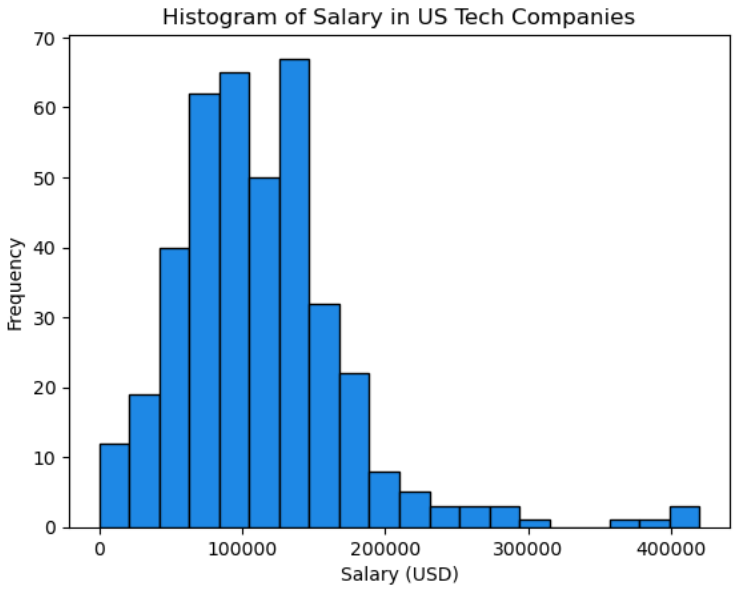


Figure 4. Average Salaries in Different Geographic Regions.

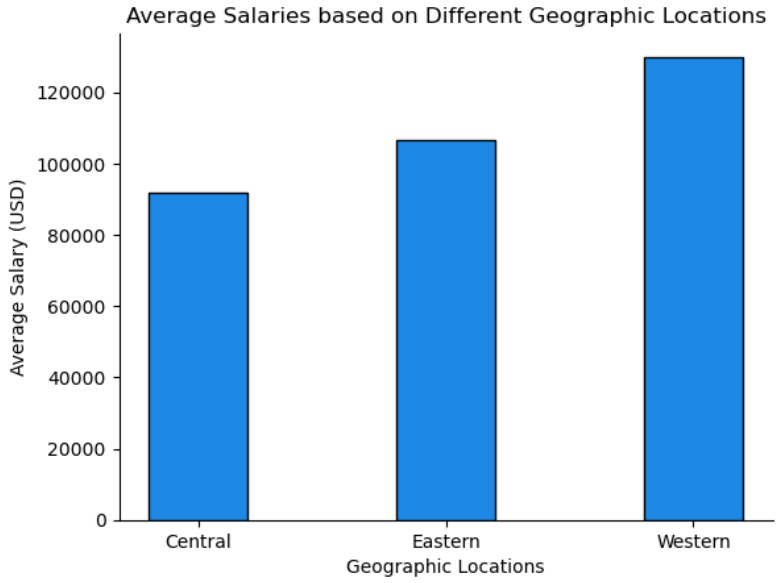


Figure 5. Average Salaries in Different Job Categories.

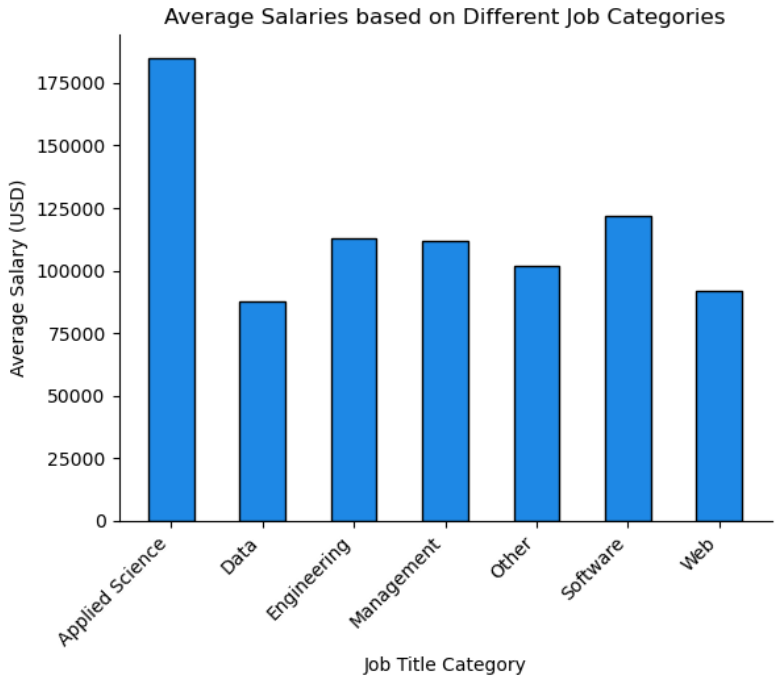


Figure 6. State-level Average Salary in US Tech Companies.

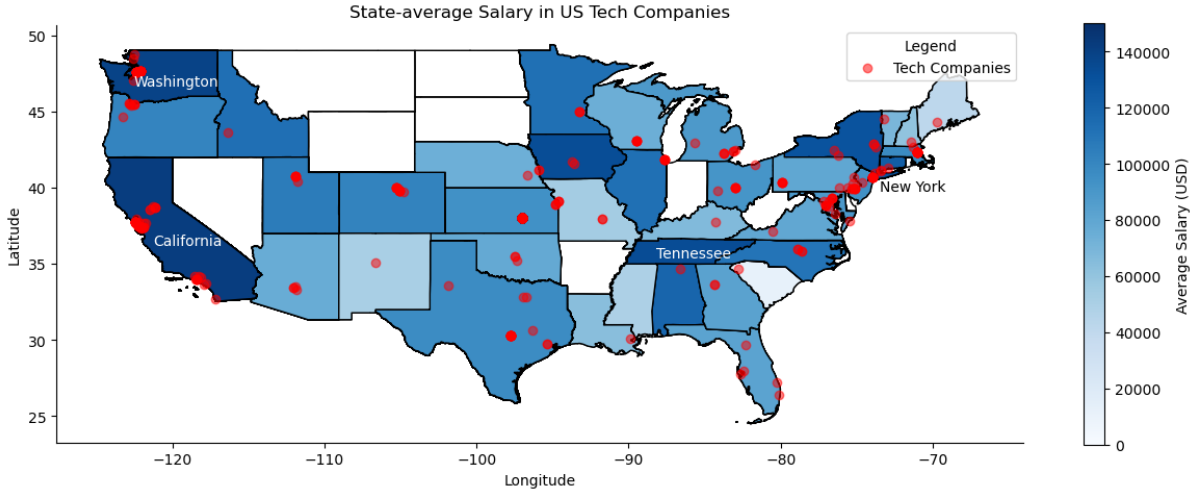


Figure 7. State-level Average Total Years of Experience in US Tech Companies.

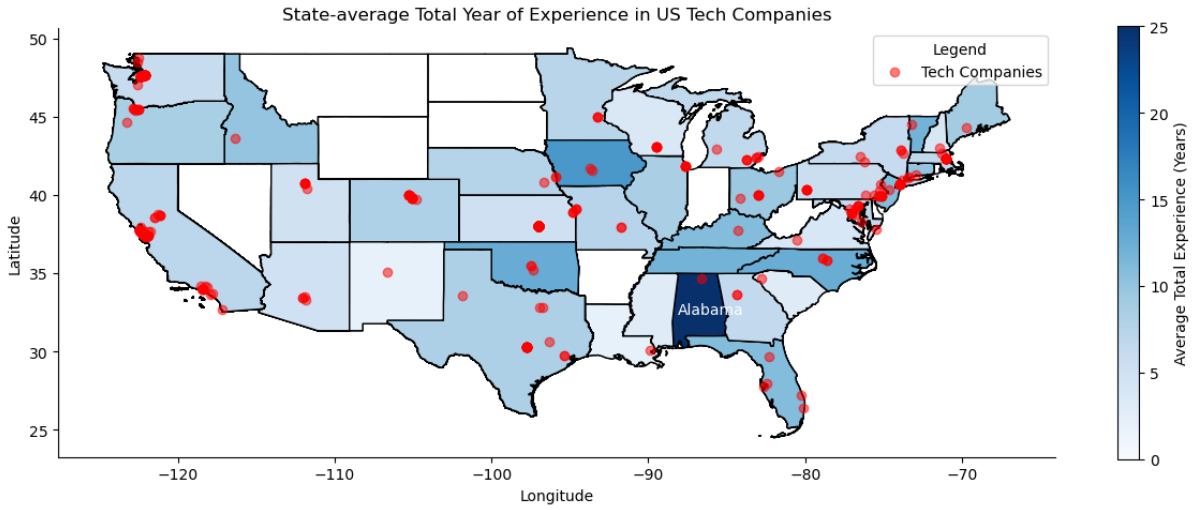


Figure 8. State-level Average Educational Attainment Ratio.