Salary Growth in the Tech Industry

Ho Joon Jun

April 5, 2024

Introduction

The relationship between wages and workers has been a crucial topic in labor economics for decades. Having a quantitative relationship between these two can allow for workers to understand the value of their work and abilities and lead to more satisfactory job choices, and firms can create more accurate policies and compensation structures to reflect the benefit gained from the workers they have employed and want to employ in the future.

Several academic papers have discussed different approaches measuring the relationship between worker compensation and their tenure. Here we review some literature on similar topics, and the results they obtained.

- “Wages, Experience and Seniority” by Christian Dustmann and Coastas Meghir report findings on how a workers’ compensation changes depending on certain factors. Specifically the report mentioned creates a function based on explanatory variables found in their dataset, and gives a result of a numerical growth rate for each year of employment. (Dustmann,Meghir, 2005).
- “Do Wages Rise with Job Seniority?” by Joseph Altonji and Robert Shakotko discusses the topic by using a dataset of workers in the 60’s which include their ages and disabilities along with their tenure and salaries, and creates an equation to try and capture the effect the mentioned explanatory variables have on the workers wages. The paper reports that the effect of tenure on wages is small, and that the workers gain wage growth from their overall experience. (Altonji, Shakotko, 1985).
- “Job Duration, Seniority, and Earnings.” by Katharine Abraham and Henry Farber uses the same dataset as Altonji and Shakotko, but instead tries to determine how much the effect of reported salary growth actually comes from experience. The paper finds once again that experience doesn’t actually have much of an effect on salary growth, which suggests the hypothesis that wages and tenure are not related. (Abraham, Farber, 1987). Continuing off this idea, we look at a paper that talks about another factor that explains how worker’s salary is determined, the company said worker is employed.
- “The Wage Policy of a Firm” by George Baker, Michael Gibbs, and Bengt Holmstrom uses another imperial dataset from workers in Management. The paper uses a model that tries to attribute wage growth to the worker’s comparative advantage in their workplace. However, the paper finds that the learning model used doesn’t explain growth much, and that the method in which firms compensate their workers come from a predetermined salary increase that workers must compete for. (Baker, 1994)

The findings in the literature suggest that the relationship between worker salary and experience is much more complicated than expected, and that their isn’t much to report on in terms of new findings. However, since decades have passed since the initial publishing's of those papers, there is value in re-examining the topics discussed in a more modern viewpoint. In this new paper, a similar method will be used however a much more recent dataset will be used, and analysis on different explanatory variables will be done, as well as more economic intuition will be given to explain justification of the results found. The dataset created from the Hacker News salary survey in 2016 has thousands of relevant entries from over 1500 submissions. Using this dataset, results will be found in a variety of ways, each backed with historical economic theory, supported by similar literature, and will be presented in formats for both qualitative and quantitative examination.

In regards to the dataset, the survey collected data from those in the tech industry, a competitive field known for more service based work that has high entry level wages and a lucrative career path. The job market is known for being tough to enter and highly demanding, easily being considered a difficult field to work in. These characteristics support the fact that understanding how these workers are compensated is crucial for both the workers and firms, since workers need to be compensated fairly and have solid expectations from firms in such a demanding field, while firms must understand how to allocate their operating expenses to stay competitive.

The central question this paper will be asking is the following; What is the correlation between the longevity of employment and salary progression in the tech industry, and how does the relationship vary among job titles and locations? As mentioned above, numerous sources will be used as reference and inspiration for this paper, and will be recalled throughout to add more insight and explanation for choices in analysis made. Also, relevant papers will be used to justify principles made in this one. For example, running a linear regression will be done in section 4 of this paper, and many of the choices in feature selection and interpretations of the results require support from similar papers. A paper by Thomas Dohmen uses a dataset on Dutch aircraft workers and runs a linear regression to examine the relationship between seniority and earnings, and the steps made in their paper will be used as inspiration here (Dohmen, 2003).

An important category not included in this dataset is worker performance. Data like this can only be found from a company's perspective and is much more difficult to measure. However, not having this piece of information should not be too impactful in the results found in this paper, since merit based compensation is no longer the standard in today's employment market. The book "Reward Management" by Marc Thompson includes a chapter on salary progression and what systems can be found in a modern look of empirical data from the UK. The chapter finds the result in the UK dataset that "Workplaces using only merit payment systems accounted for 9 percent of the sample." (Thompson, 2009, pp 120–139). Drawing from these results, it should be interesting and justified in making the decision to ignore worker performance as it should not affect the results found in this paper.

The paper will have the following sections. Section 2 covers how the raw dataset was cleaned, manipulated for use in the other sections, and highlights important observations that can be made in just the tables and numbers. It will also highlight different sectors of firms and types of workers, along with graphs of the previously mentioned observations. Section 3 is focused on the geographical aspect of the dataset, and will showcase what impact location has on the research question. Section 4 runs several regressions on the dataset and gives an explicit equation, along with highlighting the causal effects of different variables. Finally, section 5 is the conclusion where major results are summarized and future steps are given.

1 Data and Data Interpretation

1.1 Data Cleaning / Loading

To begin to answer the main research question, we first need to define what specifically we are looking for in the data, and how we are going to acquire it. This leads to the following assignment of variables.

- Y: Salary. This variable is the main output of the question, and will be refereed to as the dependent variable. This will show the change in salary over a certain period of time.

The rest of the variables are for helping find Y, all of which will be relevant in determining the relationship

- x1: Current employment length at a given firm.
- x2: Total experience.

We treat x1 and x2 differently as to see what effect changing jobs in the middle parts of one's career will have on their salary growth. The relationship these two variables have and the affect it has on Y can be thought of as a loyalty factor. Traditionally, employee salaries often grow on a yearly basis, and one result of this paper will be what the marginal benefit of each additional year worked will lead to in salary growth.

- x3: Job title and job category. Wages differ from job to job, and the relationship between those wages to overall employment length can be found easier if we narrow down to different jobs. As a result, it will be helpful to distinguish salary growth patterns in different titles / categories, to see if there are any interesting results.

- x4: Geographic location. Just as how different tech fields have differing wages, so will different locations. Certain countries / cities will have a bigger focus on the tech sector compared to others, which could result in different salary growth paths in each location.

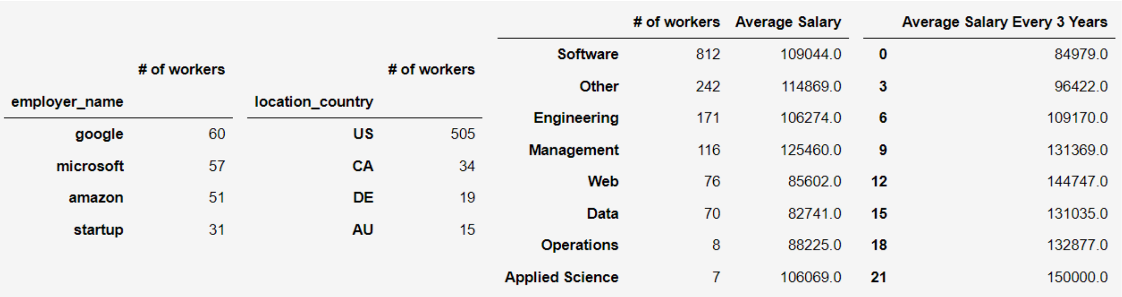
We can now start analysing the tech salaries dataset. This project will show every step of data harvesting, starting with cleaning the original dataset to streamline the research, to detailed summaries of important information and insightful graphs that help show trends that appear in the data. The original dataset comes with some unnecessary columns and a lot of missing values. As well, it includes row entries that aren't appropriate for the purposes of this research. All of these flaws will be removed into the new working dataset that will be studied.

Looking at the predefined variables in the dataset, we can define the desired explanatory variables as follows. Y will be 'annual_base_pay', and growth will be measured later on. Then, x1 and x2 are 'total_experience_years' and 'employer_experience_years' respectively. 'total_experience_years' will showcase how seasoned / mature a given worker is, regardless of their current position, and 'employer_experience_years' showcases the current tenure with their current salary. It's worth noting that the current salary might be found to be influenced by the total experience, and this result will be shown later on.

The variable x3 will be the 'job_title_category', and specific job titles in each category will be sorted out later. Finally, x4 will be 'location_country' and 'location_state', refering specfically to American states where this survey was taken. Also, the variables of 'signing_bonus' and 'annual_bonus' can help showcase why a worker may stay on with their current company, or if a company historically releases a higher signing bonus, it could explain why workers stay on longer.

1.2 Summary Statistics Tables and Interpretation

Now, we obtain relevant statistics of the cleaned data. First, it will be helpful to the average salaries for each variable to see general trends. Also, knowing the most frequent workplaces people are working at and the bonuses people receive will help in determining the loyalty factor previously mentioned.



	# of workers
employer_name	
google	60
microsoft	57
amazon	51
startup	31

	# of workers
location_country	
US	505
CA	34
DE	19
AU	15

	# of workers	Average Salary
Software	812	109044.0
Other	242	114869.0
Engineering	171	106274.0
Management	116	125460.0
Web	76	85602.0
Data	70	82741.0
Operations	8	88225.0
Applied Science	7	106069.0

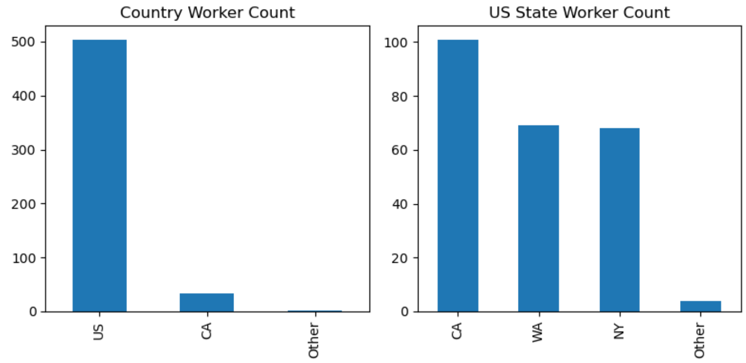
	Average Salary Every 3 Years
0	84979.0
3	96422.0
6	109170.0
9	131369.0
12	144747.0
15	131035.0
18	132877.0
21	150000.0

As expected, Google, Microsoft and Amazon are the most popular companies, with startups also taking a large portion of the working population, and since the survey was mostly conducted in America, it makes sense that basically all of the workers are located in America as well. A more interesting observation is that more than half of the workers in the survey were in the job category of Software. This may indicate that the tech industry is taking a bigger focus on software and services and is currently the most in demand field. It also could be true that it is over saturated, but that has yet to be proven in this data. More relevant to the research question, we can see that the salary per year of experience follows a good growth path, with jumps of around $10,000 happening every 3-4 years after 0.

A relevant question that will be examined later in this paper is how much influence the choice of firm has on the salary growth. The article “High Wage Workers and High Wage Firms” by John Abowd cites several sources that debate how correlated a workers returns based on their experience is to their firm. (Abowd, 1999) The generally accepted theory in the literature is that should be correlation, and that allowing for it results in more diversified data that leads to more refined results.

1.3 Plots, Histograms, Figures

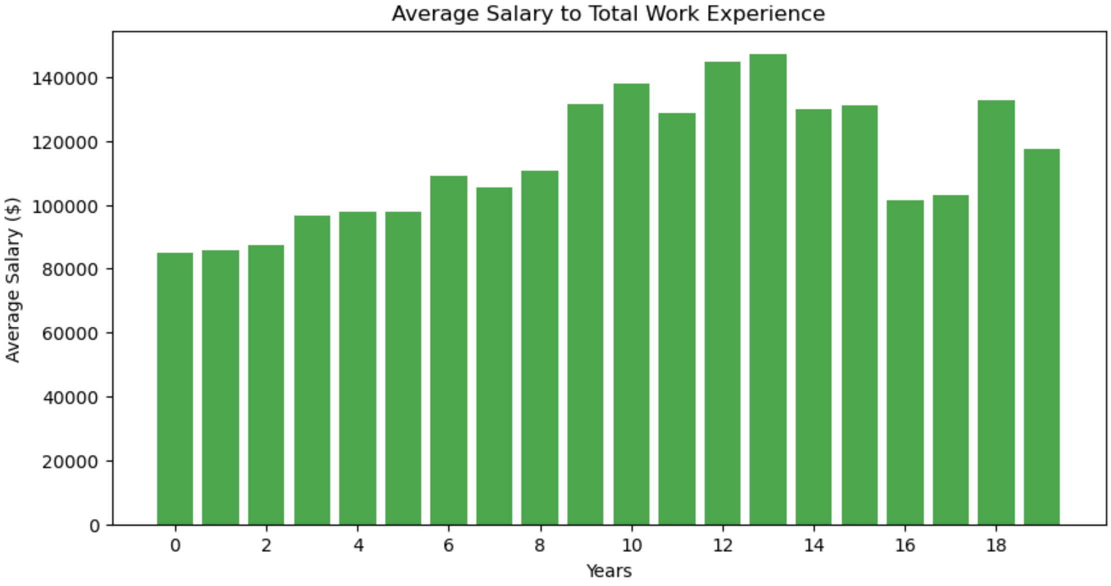
With the dataset and relevant statistics, graphs can be created that help showcase the results found, and interesting observations that serve as inspiration for the topics covered in the rest of this paper.



Almost the entire dataset comes from North America, majority being the US, meaning all the salary data should be assumed to be in USD, and following American economic trends. Had the dataset been more diverse internationally, the growth trends could be even more interesting since different countries could offer very different growth paths in salary for the tech industry.

Looking only at the US, we can see the most popular states to work in are California, New York, and Washington. California has a little over 10% more of the total workers, with New York and Washington being tied. The rest of the results found in this report will mostly be coming from these three states in one major country, meaning salary differences across locations will only be state to state differences.

On average, the highest paying job category is in Management, although it is worth noting that the average salary won't be as diluted as the most popular category Software.



This graph is the most important so far, as it acts as a benchmark answer for the main question of this paper. The effect years worked have on salary appears to be upwards trending, which is to be expected. However one interesting observation is that the growth does not stay consistent at around 10 years. From this point onwards, salary growth appears to be stagnant and even decreasing. One possible explanation for this is due to the nature of the survey that collected this data. Individuals had to write down their current total experience, meaning the entries we see after the 10 year mark are people who have been in the field for that long. This along with their lower average salary could imply that the jobs the mid to senior works are currently at are underpaying them, and the workers may not have realized.

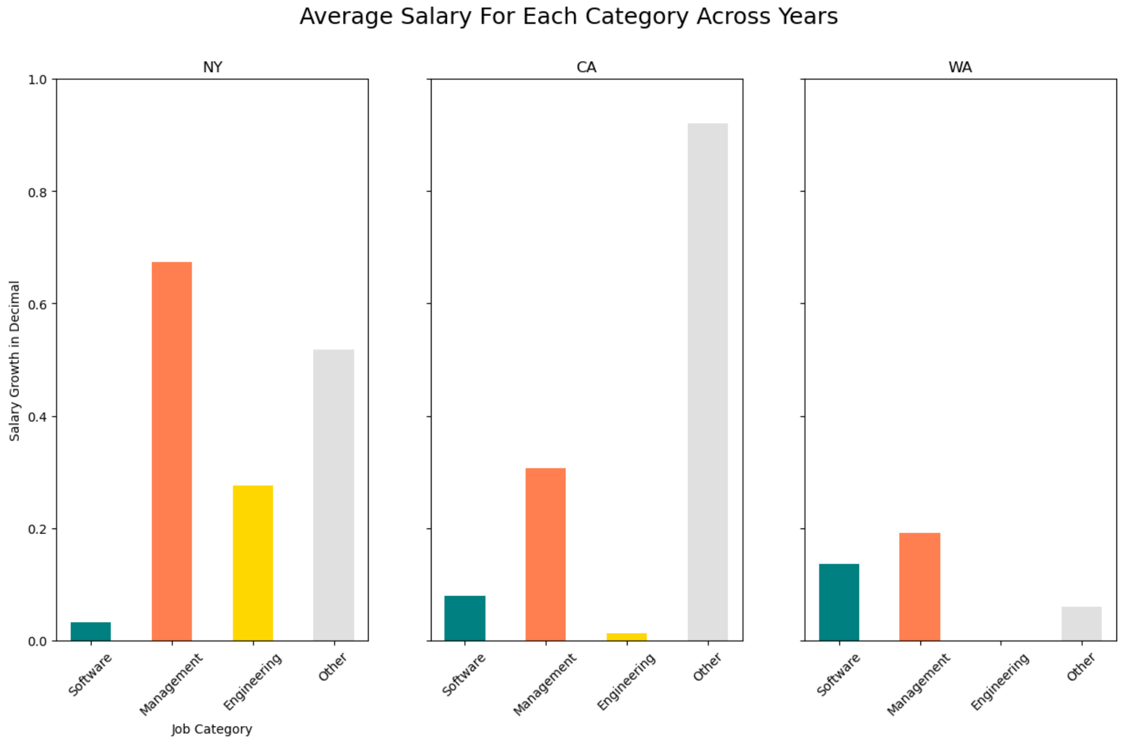
Another observation in this graph is the pattern in salary change every three years. Starting from year 0, we can see that growth is small, but every 3 years there is a big jump in average salary. This pattern really only holds in the first 10 years, but it is worth noting a potential reason for this. As mentioned before, the ability to switch jobs every few years is a practice that has become increasingly more popular in the modern world. Once a worker gains experience, they can start demanding better compensation than what they were at previously, and they may look to another firm to hire them for that higher compensation. Literature such as "The importance of firms in wage determination" by Max Gruetter uses the same hypothesis to support their data manipulation. Gruetter takes a similar dataset and creates a more refined dataset by splitting the data into 8 different tenure categories. (Gruetter,2009) Looking at workers in this interval perspective may be more beneficial than just a per year increase, since the data suggests that workers feel comfortable changing every 3 years.

One question that stems from this result is whether or not the firms perform well due to their workers being more productive from being compensated better, or if all the workers in the field have relatively equal productivity but the higher salary comes from the massive firms they work for? This chicken and egg problem isn't necessarily the focus of this research paper, but it is definitely worth studying later on.

2 Data Visualization

2.1 The Message

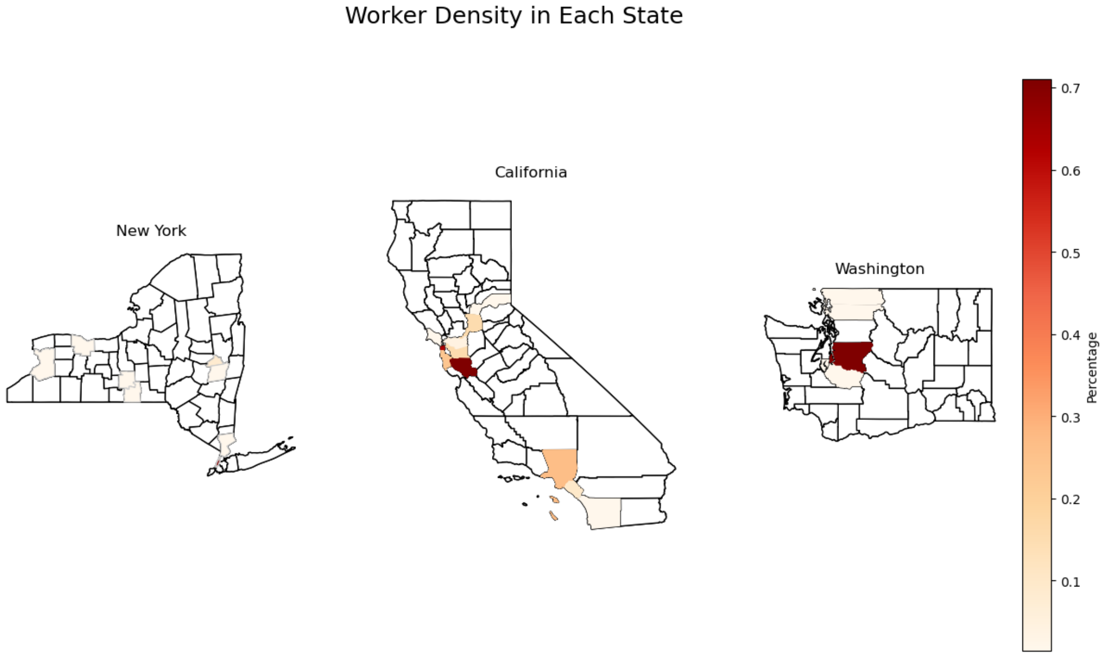
The main message of this paper is to describe quantitatively what the relationship between a workers experience and salary growth is. However, there are many external factors that also impact salary growth that are related to the amount of experience a worker has, such as where they are located or their skill / ability. These external factors can be shown to be related to experience though, and connecting these elements together is what the following sections will explain. For example, the salary growth to experience of a Management worker at Amazon in Washington will be different from a Google Software Engineer in California, and part of the message of this paper will be determining the estimated amount from datasets and economic intuition for the explanation. As a baseline however, the main graph that this paper should show is how a workers salary changes throughout years and across different states and categories.



We can see in these plots that the salary progression of workers are extremely varied across years. In Software and Engineering, we can see there isn't very consistent growth across states. The average salary growth has large variance across all three states, and the graphs don't suggest a consistent growth rate, which was unexpected. Just based on preconceived understandings of salary, people should expect a steady increase in their salary as they gain more work experience. However the data suggests that even the growth in the yearly 1 to 3 years is very volatile even in the same state and job category, implying that the firm's compensation to their workers doesn't follow any consistent trends. The message of this paper is to show that the growth rate in salary is much more complicated than expected, and will try to decompose what factors make it so complicated to create a better understanding for the reader.

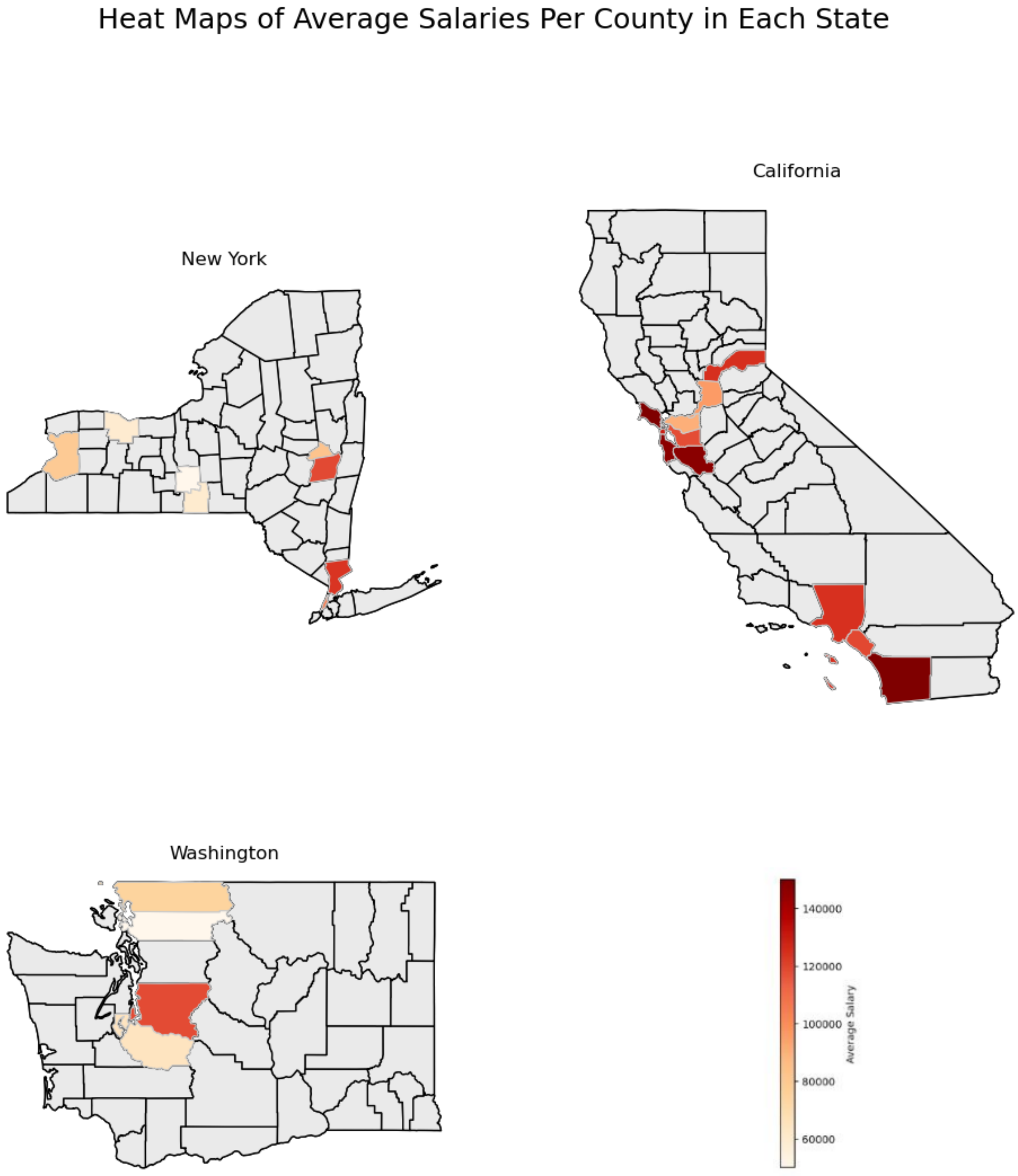
2.2 Maps and Interpretations

Now, we focus on looking at the geographical data and come to try and understand how worker location affects their salary. The dataset will be merged with US Census data since the survey was mostly conducted in America, and furthermore the focus will be primarily on the 3 most popular states in the US found in the dataset. New York, California, and Washington were found in section 1.4. to be the most common locations for workers in the dataset by an overwhelming amount. While more investigation could be conducted on some of the other locations in the survey, there will be a lot less data to work with to create any meaningful results. First we look at where within the 3 states the workers are located at. A density plot of the workers will help narrow down more accurately the workers real working locations, and data from this can be used to adjust for the workers actual cost of living and more.



As we can see in the mapS, California workers are primarily located in San Francisco and Los

Angeles, which are locations historically known for white-collar, tech based work. We also can find out which cities are the most popular, and examine them more closely for different trends. In New York, the worker locations are very spread out, but we can observe that New York city is where the biggest clump of workers are. Similarly, California has two main regions of workers on this map, San Francisco and San Diego. The city more interesting to observe will be San Francisco, the city regarded as more of the tech central of California. Finally, Washington's map shows dispersion throughout a couple counties, but the most populated will be Seattle. Having shown this, we can examine the two cities and learn more about how the geographical location affects salary growth.


Heat Maps of Average Salaries Per County in Each State

New York

California

Washington

The heat maps shown give interesting insights into salary patterns in the three major states. California has the highest average salary counties out of the three, with San Diego and San Francisco being the highest. Also, New York has a very dispersed map of worker locations, although the east appears to have a higher average salary overall compared to the west. Finally, Washington's dispersion appears to be very centralized with the highest average salary being in King county. As we can see, there are large differences in salary depending on a workers location, even down to different counties within the same state.

The main observation to take away from these graphs is how workers choose to location depending on experience. In all three major states, the worker population can be seen focusing in the central cities mentioned above, NYC, San Francisco, and Seattle. It can be hypothesised in these graphs that the workers with more experience choose to move to these cities will also experiencing higher salaries as found above in section 1.4.. This results in an interesting but expected correlation between a worker's location to salary, with more experienced workers choosing to move to where their could be more money to be made.

Another research paper that discusses this idea is "Are Older Workers Overpaid?" by Daniel Vuuren, 2011. Vuuren talks about the effect of returns to tenure changing with the mobility of the worker, with more mobile workers getting a higher pay. This is due to workers being able to find a better paying job easier if they can move easily, and applying this idea to the maps above we can see this in action. Many tech workers don't appear to migrate across the country, limiting their mobility. A possible explanation could be the workers being unable to find jobs in a certain city due to the high cost of living, a factor that will be examined later on in this paper. Another possible explanation is that less experienced workers are more scattered geographically as they are trying to work at newer start-up's or less competitive areas.

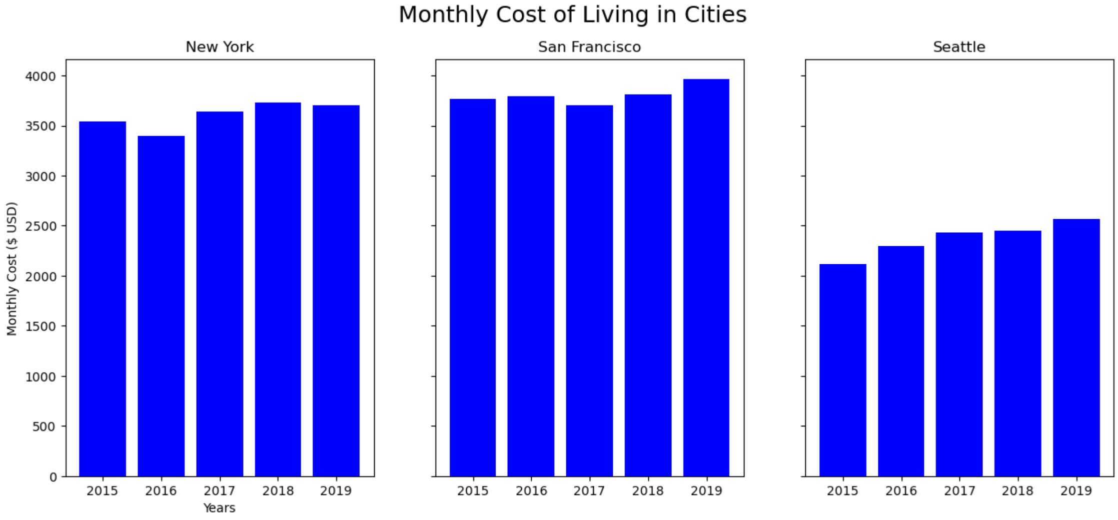
3 Merging New Data

Data that will complement the existing dataset of tech salaries will be information that can help further distinguish the difference between nominal and real salary growth rates. An immediate issue when trying to observe a workers true salary growth to their experience is whether or not they are gaining any increase in salary. Comparing different worker's salary growth is only significant if their unique costs of living are held constant. For example, if the cost of living goes up more than the annual salary growth rate, the workers are really making a negative growth rate over all. Therefore, finding some measure of cost of living in the different locations will be economically important.

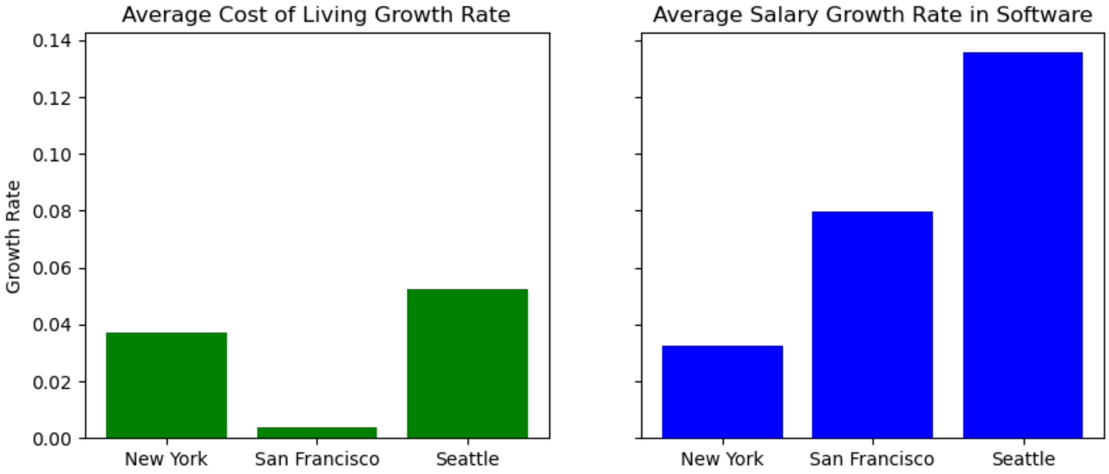
The website "https://numbeo.com/" is the ideal place to find cost of living will be a website that has historical data and several different categories such as average rent prices and average transportation costs. While this dataset is ideal, there are some technical problems that stop it from being perfect. One possible concern is the accuracy of the reported data, since it uses user inputted data, however since the original tech salaries dataset was also surveyed and user input, the accuracy concern isn't a big concern. Another potential issue would be in the accuracy of different parts of the cities compared to overall city. Even though the scope of the search will be the primary cities found earlier, the website still doesn't have much to distinguish what costs are like across different parts of the same city.

We scrape this data and analysis it to determine what the real salary gains a worker achieves depending on where they are located.

3.1 New Dataset Visualization



As we can see, the monthly cost of living in New York city and San Francisco are roughly the same, but New York's growth rate seems to be higher across the years compared to San Francisco. More interestingly, Washington DC is by far the cheapest, being almost $1000 cheaper each year, and a much lower growth in the cost of living. When paired with the heat map of the different counties found in section 2.2, an important result is that New York and Washington had similar average salaries, although we can now see the cost of living in Washington is a lot lower, which means Washington workers are experiencing more real salary growth overall.

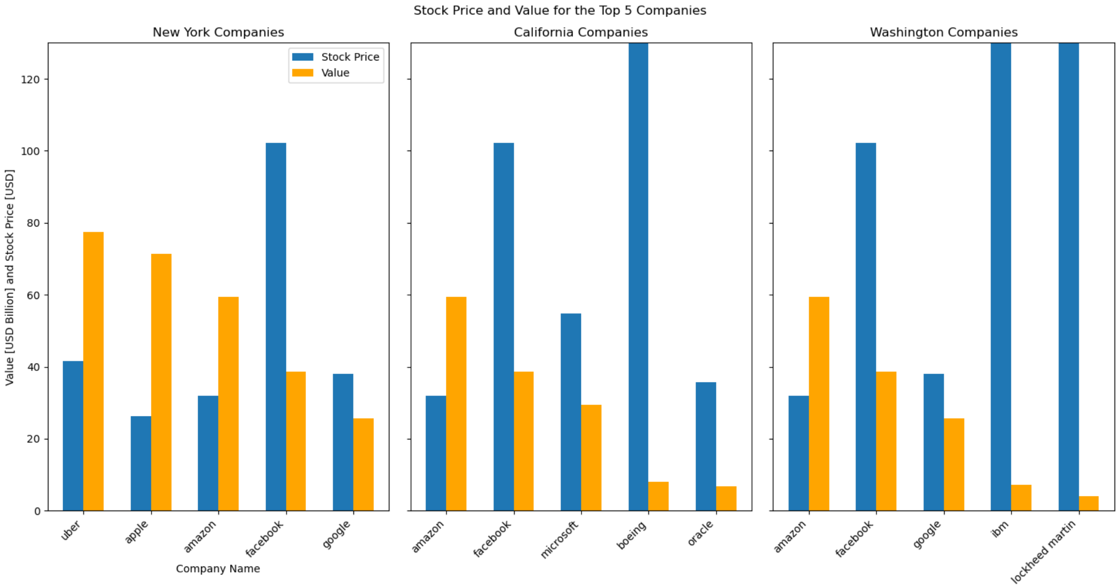


As the graph suggests, a New York worker in Software has a similar salary growth rate to their cost of living growth rate. While this doesn't mean the worker is seeing any actually growth in annual income, it does help explain a workers incentive to stay in New York if their salary growth is matching the cities cost of living growth. The worker may choose to live in New York longer with the knowledge that the cost of living won't exceed their jobs salary if they are able to already maintain living in New York. Another perspective though is that if a New York worker isn't currently able to afford living comfortably in New York, they now know their job's annual salary increase won't be enough to support them in the coming years. Applying this idea to the other two cities, San Francisco sees an extremely high salary increase to relative to their cost of living increase, which suggests the idea that if a worker is able to begin living in San Francisco to work, they should be able to expect a comfortable salary after a few years of working. Most interestingly is Seattle, with the highest salary growth out of the three cities, but with also the highest cost of living growth rate. However, since Seattle was shown to have a significantly lower cost of living already, it suggests that workers are already realizing that Seattle is the city to try and break into next. Workers are seeing the low cost of living and the high salary growth rates, and potentially beginning work in Seattle, which causes Seattle's apartment prices and other cost of living goods to go up with demand.

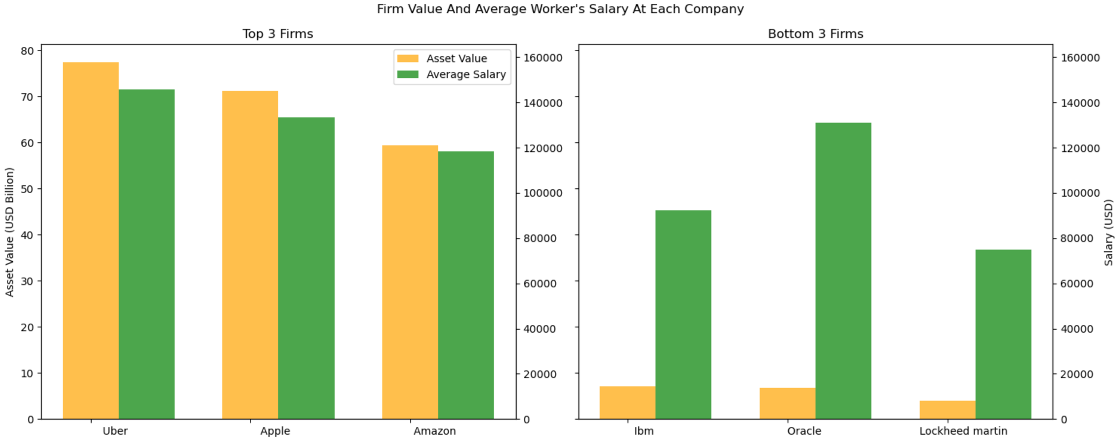
3.2 Merging a New Dataset

Another new dataset that will help compliment the existing tech salary data will be information that better explains the salary amount for each worker. While the results so far have been created from the existing dataset, more variables can be found from other sources and can be combined to create a bigger dataset with more explanatory power. Also, an interesting observation will be measuring the firm's financial performance and seeing what effect the performance has on a workers salary growth. Workers should know if their firm experiences high growth, and receives the same or less salary growth then another worker whose firm didn't see high growth, to understand if their value is better reflected at a different firm. The place the new dataset will come from will be Yahoo Finance, as it is a great source for financial data given a firm's name.

However, there are several potential problems when trying to scrape Yahoo Finance is the lack of privately traded firms data. Yahoo Finance will carry the majority of the publicly traded companies in the dataset, but won't be able to provide any information on privately traded ones. This severely impacts the ability to get data, and it will be shown later on that a good portion of the dataset doesn't have any Yahoo Finance data. Also, the ability to find historical income sheets make it hard to find how much money the firm made during the time of the survey. Instead, another metric in determining firm performance will be done. The dataset will find a firm's stock price and value during 2016, as both help explain what financial state the company was in during the time of the survey. Also, Yahoo Finance has stock price data and the number of shares in a given year readily available, even during 2016. Obtaining the stock names from the company titles can be very tricky, since a lot of the company names aren't publicly traded or have a ticker name that can be found from the company name. Also, finding the data from just 2016, which is the same time as the survey, can be difficult since many of the firms don't have their financial data public at during that time. To make up for that, the code will attempt to find the most recent year after 2016's data, only up until 2020. If Yahoo! Finance had historical income sheets that went back to 2016, this section would instead be trying to find the firms operating costs and total revenue for that year to deepen the connection between firm performance and their compensation to their workers.



In these graphs we can see that the values of the firms vary greatly. The stock prices seem to be a little more consistent in the data, with Amazon, Google, and Microsoft all having relatively similar stock prices. However it is worth noting how high Facebook's stock price is. Overall, the most popular companies stay consistent with the results found in section 1.4., and now we can combine these results with the existing dataset to see how the companies performance is related to their worker's compensation. The main companies to examine will be amazon, google, and apple.



This graph shows the average worker's salary at the 3 highest market value firms compared to the salaries at the 3 lowest out of the 15 (not unique) firms found above. The immediate result being shown in this graph is that the worker's at the bottom 3 firms do appear to have a lower average salary relative to the top 3 firms workers. Ibm and Lockheed have significantly lower salaries compare to the workers at Uber, Apple, and Amazon, which could be the result of the firm's having a lower market value. The only exception is Oracle, with a very similar average salary to the top 3 firm's workers, so the effect of firm value on worker salary doesn't appear to have strong correlation. Nonetheless, the effect does appear to be present, which suggests that firm performance does have a significant effect on worker compensation.

4 Regressions

4.1 OLS Regression

The chosen output Variable for our regression was Annual Salary, and the main explanatory variable was Experience. Under economic theory, the relationship should be linear as there should be relatively constant changes in salary each year. The gain in salary for an inexperienced worker might be less than the gain a more experienced worker will get, but at the same time, the more experienced worker isn't taking on a marginally high extra workload to compensate for the higher salary, so returns should even out. If the question was a worker's market value with experience, then I would believe the relationship to be nonlinear and probably quadratic with age, but at the same job position and title the relationship should be linear. Some further explanatory variables are taken into consideration and will be inclued in different regressions below.

- Total work experience will support the theory that a worker's overall compensation should change to their overall experience. Note, we will be using a owrkers prior experience, which is their total experience minus their current experience. If we don't make this distinction, there will be correlation between current experience and total experience, which ruins the regression.
- Their geographical location with longitude and latitude, along with dummy variables outlining the 3 major states that were analyised above. Dummy variables for California, New York, and Washington will be created.
- The signing and annual bonuses will help see if a worker's salary growth is negatively impacted due to a higher bonus, which is to be expected economically.
- The job category will be in dummy variables too, with the categories of Software, Management, and Engineering being chosen for their popularity.
- Finally, the stock price and firm value will be regressed to see how firm performance impacts salary.

The first regressions that will be created will be as follows. - Regression 1: Current Experience and Prior Experience. This will be the simplest regression and should give a very basic answer to the research question.

- Regression 2,3: Building off Regression 1, we include coordinate locations, along with the signing and annual bonuses. Regression 3 will add on stock price and firm value variables, to see how firm performance affects a workers salary.

These three regressions will explain the main question of this paper, showing how salaries are affected based on experience, location, and firm choice.

The next 5 regressions will be implementing binary variables, which characterize different subsets of workers. Each regression will include current experience and prior experience, but the other variables will be ignored for these models to see what the effect of these binary variables are on their own.

- Regression 4,5,6: Model 4 will have 3 binary variables on job category. Model 5 will have 3 on US state location. Model 6 will have 3 on Job title. The binary variables will represent the 3 most popular of each category, all of which have been discussed previously in this paper. It was a complete coincidence that each subset just happened to have 3 top popular options in the dataset, with the options afterwords being significantly less popular.

- Regression 7,8: These two models will be much larger and include both the explanatory variables found in the first 3 models, and the binary variable models swell. Model 7 will include all 9 binary variables, to see how each subset interact with each other, and Model 8 will take the most popular binary variable in each subset, along with the most important explanatory variables in Models 2 and 3. The "importance" coming from both how the R^2 and P-value of the coefficients and models increase in each model

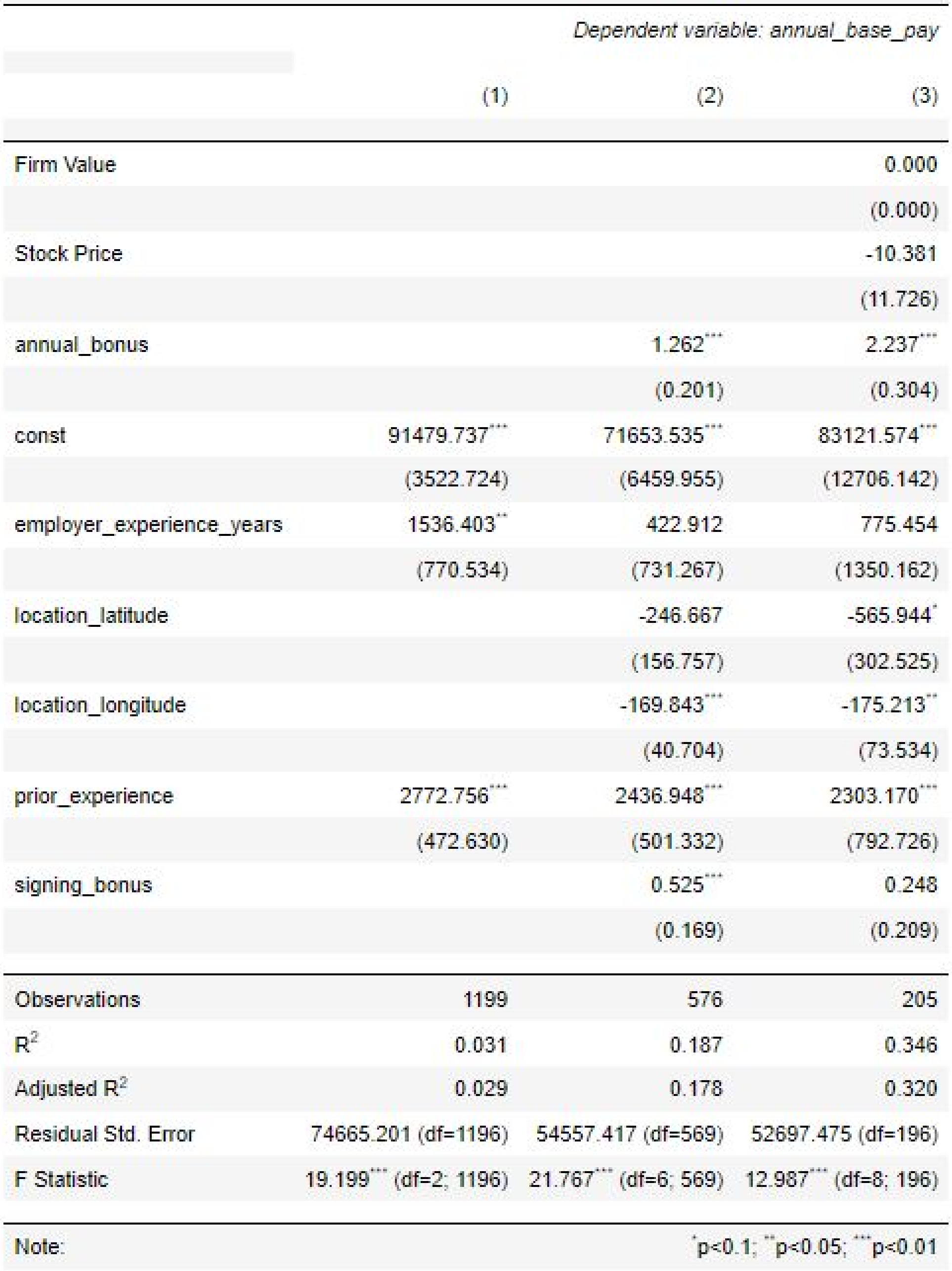
Inspiration for this model selection is drawn from a paper by Robert Gibbons, 2005, in which they regress the log of workers hourly wage on years of education, dummy variables, experience, industry and occupation. The results they found due differ from what will be found later on in this paper as the dataset also included gender and education, information missing in the Tech Salaries data, but it is still worth noticing that the regression models created for this paper do follow a similar structure as those found in other literature on the topic of wage to experience.

4.1.1 Data Imputation

To remove the high amount of missing in some of the important columns, we can implement Data Imputation techniques to fill in the missing values. Literature on the topic suggests to only use Median or Mean data imputation when only a few entrees are missing, but for example, the bonuses columns are missing almost 20% of their entrees each. (Zhang, 2016). The same paper discusses using regressions to produce "smarter" values, but talks about the issue with this method. By the nature of imputating with regression, the imputated data will cause the bonuses columns to be highly correlated with the other columns. However, this shouldn't be a big problem as there will still be low enough correlation in the data to safely run an accurate regression.

The technique used here will be regressing the missing column on the other variables, to then predict the missing column, and then imputate the missing values with the predicted value. We can see below a table of the columns missing the most entrees. Knowing this, the only data that will be imputated for now will be the bonuses.

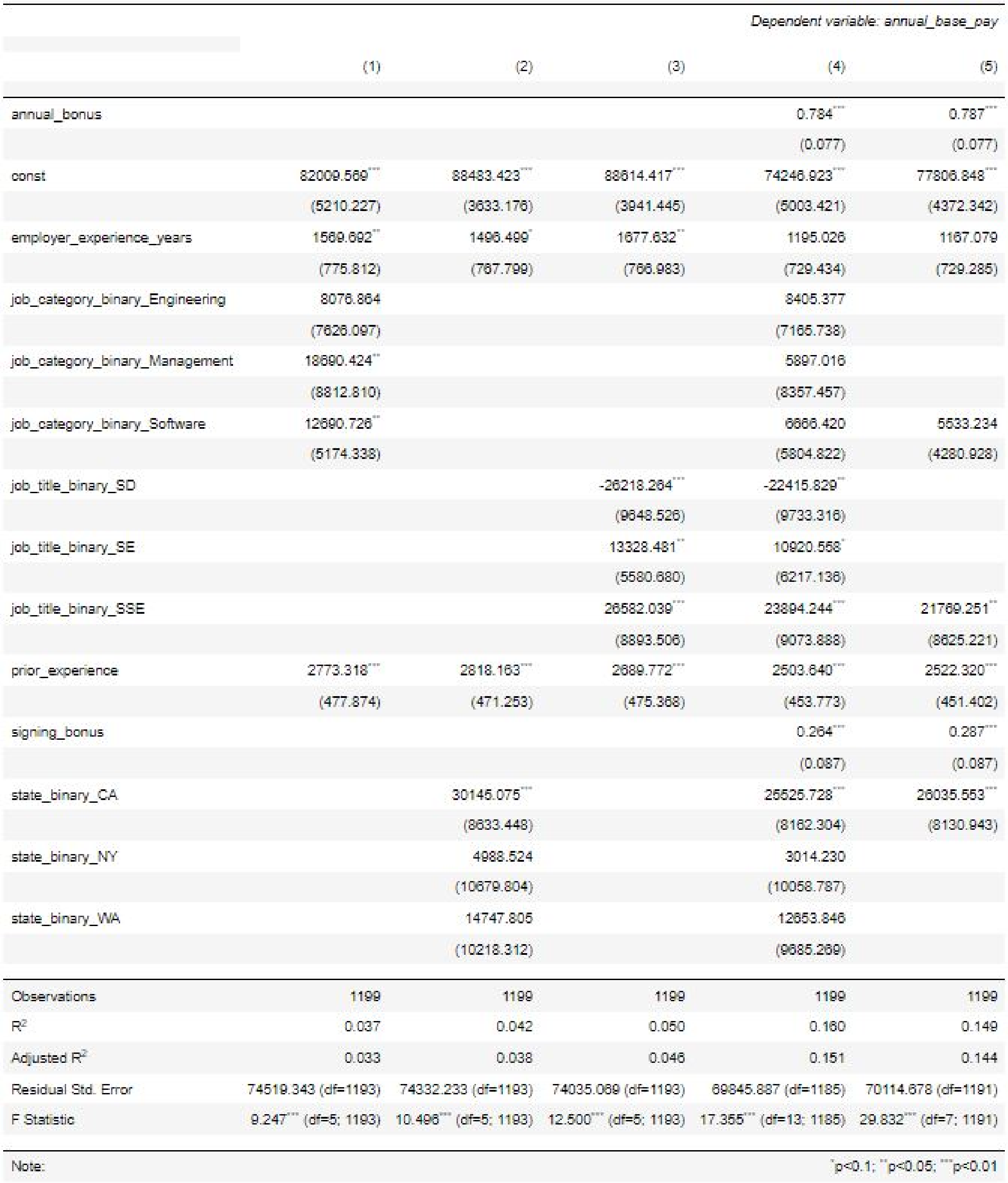
4.1.2 Basic Regressions



	Dependent variable: annual_base_pay		
	(1)	(2)	(3)
Firm Value			0.000
			(0.000)
Stock Price			-10.381
			(11.726)
annual_bonus		1.262***	2.237***
		(0.201)	(0.304)
const	91479.737***	71653.535***	83121.574***
	(3522.724)	(6459.955)	(12706.142)
employer_experience_years	1536.403**	422.912	775.454
	(770.534)	(731.267)	(1350.162)
location_latitude		-246.667	-565.944*
		(156.757)	(302.525)
location_longitude		-169.843***	-175.213**
		(40.704)	(73.534)
prior_experience	2772.756***	2436.948***	2303.170***
	(472.630)	(501.332)	(792.726)
signing_bonus		0.525***	0.248
		(0.169)	(0.209)
Observations	1199	576	205
R^2	0.031	0.187	0.346
Adjusted R^2	0.029	0.178	0.320
Residual Std. Error	74665.201 (df=1196)	54557.417 (df=569)	52697.475 (df=196)
F Statistic	19.199*** (df=2; 1196)	21.767*** (df=6; 569)	12.987*** (df=8; 196)
Note:			*p<0.1; **p<0.05; ***p<0.01

As we can see in the table, all the models show that the effect of an additional year of working results in an increase in $$500 $ to $1500. We can see that the effect of prior experience is always larger than the effect of a workers current experience, which supports the idea that a workers salary is determined more on employee market value rather than in firm experience. Another important result to take away from Model 3 is to look at the coefficient of Stock price and Firm value. The values show a very small beta, which considering the size of the stock prices (the five largest in each state were found above, and the max was still small), have almost very little influence in salary. Overall, the regression suggests the idea that firm performance has little to no impact on how their workers are paid, and is more reflected in how the worker's market is at the time.

4.1.3 Regressions With Binary Variables



	Dependent variable: annual_base_pay				
	(1)	(2)	(3)	(4)	(5)
annual_bonus				0.784***	0.787***
				(0.077)	(0.077)
const	82009.569***	88483.423***	88614.417***	74246.923***	77806.848***
	(5210.227)	(3633.176)	(3941.445)	(5003.421)	(4372.342)
employer_experience_years	1569.692**	1496.499*	1677.632**	1195.026	1167.079
	(775.812)	(767.799)	(766.983)	(729.434)	(729.285)
job_category_binary_Engineering	8076.864			8405.377	
	(7626.097)			(7165.738)	
job_category_binary_Management	18690.424**			5897.016	
	(8812.810)			(8357.457)	
job_category_binary_Software	12690.726**			6666.420	5533.234
	(5174.338)			(5804.822)	(4280.928)
job_title_binary_SD			-26218.264***	-22415.829**	
			(9648.526)	(9733.316)	
job_title_binary_SE			13328.481**	10920.558*	
			(5580.680)	(6217.136)	
job_title_binary_SSE			26582.039***	23894.244***	21769.251**
			(8893.506)	(9073.888)	(8625.221)
prior_experience	2773.318***	2818.163***	2689.772***	2503.640***	2522.320***
	(477.874)	(471.253)	(475.368)	(453.773)	(451.402)
signing_bonus				0.264***	0.287***
				(0.087)	(0.087)
state_binary_CA		30145.075***		25525.728***	26035.553***
		(8633.448)		(8162.304)	(8130.943)
state_binary_NY		4988.524		3014.230	
		(10679.804)		(10058.787)	
state_binary_WA		14747.805		12653.846	
		(10218.312)		(9685.269)	
Observations	1199	1199	1199	1199	1199
R^2	0.037	0.042	0.050	0.160	0.149
Adjusted R^2	0.033	0.038	0.046	0.151	0.144
Residual Std. Error	74519.343 (df=1193)	74332.233 (df=1193)	74035.069 (df=1193)	69845.887 (df=1185)	70114.678 (df=1191)
F Statistic	9.247*** (df=5; 1193)	10.496*** (df=5; 1193)	12.500*** (df=5; 1193)	17.355*** (df=13; 1185)	29.832*** (df=7; 1191)
Note:					*p<0.1; **p<0.05; ***p<0.01

Looking at Model 4 and Model 5, we can look at the coefficients of the binary variables, and we can see how the effect of working in different categories and states change a workers salary. Model 4 suggests that Management workers get paid the most from being in that category, getting around $6000 more than Software workers, and around $11000 more than Engineering workers, showing how important the workers field is in determining their salary. Doing a similar examination on the effect of the workers state, we can see that California workers gain $25000 more than New York workers, and $15000 more than Washington workers, which show a clear disparity in geographic location. Combining this fact with the cost of living data, it shows that New York workers really don't benefit from their location, having a lower salary, yet equal cost of living to California workers, and being paid even less than Washington workers even though Washington has the lowest cost of living out of all of them.

Finally, we can see in Model 6 that Senior Software Engineers benefit the most out of the three most popular job titles, but what an unexpected result is the negative effect of being a Software Developer, with a decrease of $26000 in salary, with no real change in the effect of tenure or constant as seen in Models 4 and 5.

Model 7, which essentially combines Model 4, 5, and 6 shows very similar coefficients in all 6 binary variables, supporting the results found. Overall, California workers in Management appear to have the highest market value for a given year of work, and New York workers appear to have a less generous worker market. These results support commonly accepted ideas about workers salary in the tech industry, where the reputation of California Silicon Valley workers get paid the most, especially when in a Management position that carries more responsibilities and tasks.

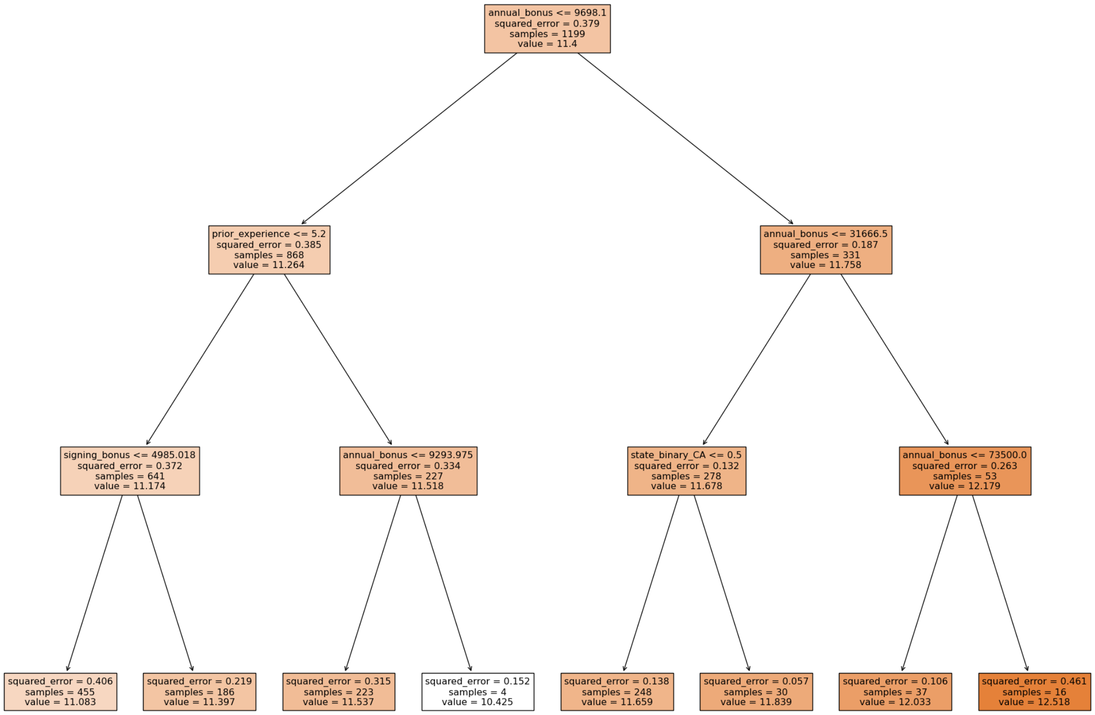
4.1.4 Preferred specification

As expected, the regression with the most variables has the highest R-squared and most explanatory power. However the low number of observations relative to the others make it seem unfavourable. Instead, the preferred specification will be the regression that contains the most important variables economically. In this perspective, regression Model 8 will be the preferred specification, as it carries the most significant dummy variables along with the bonuses. These variables where chosen to show the impact of the most popular case of worker, as these factors are what are mentioned most when discussing worker value in the tech industry. So the preferred regression model is as follows:

annual_base_pay = 72885.475 + 1682.218employer_experience_years + 2356.869prior_experience + 0.18signing_bonus +1.425annual_bonus + 5447.069job_category_binary_Software + 24123.525state_binary_CA + 21690.802job_title_binary_SSE

4.2 Machine Learning

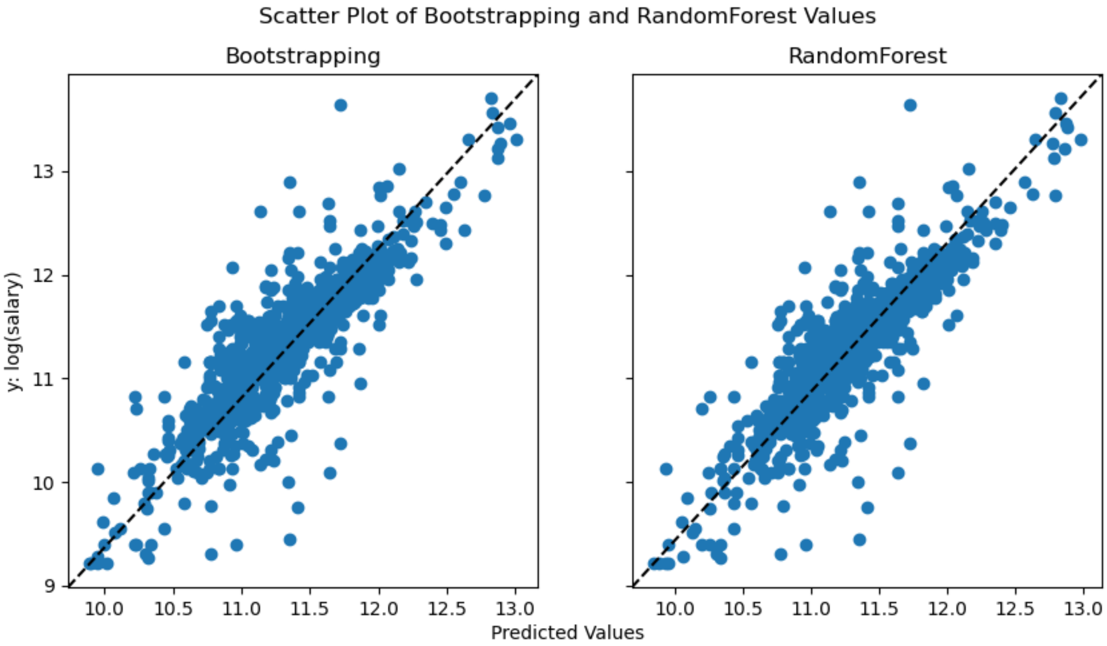
Since the preferred specification is regression 8, we can reuse the explanatory variables and fit them into a decision tree model. The following image is what the tree looks like after being fed the dataset, and it where the main results will be drawn from.



The tree showcases the most important features are the two bonuses and experience, with the splits happening to contain bonuses and experience first, and then splitting on prior experience. The right branch shows that differentiating different levels of annual bonus are best for identifying different salaries, and the left branch splits on prior experience, and then focuses on signing bonuses. We can also see how the squared errors vary across each leaf and apart from the farthest, all the squared errors are comfortably low, showing the accuracy of the decisions to be high. Overall, we can see the log salary to be predicted to roughly be between 11.1 to 12.5. Looking at the MSE, we can see that the error is fairly high, with roughly 29%. However, with the lack of explanatory variables that aren't binary, it is difficult to decrease the MSE with more depth in the tree. So, using the ML model will include more error than the OLS regression done above.

4.2.1 Bootstrapping and RandomForest

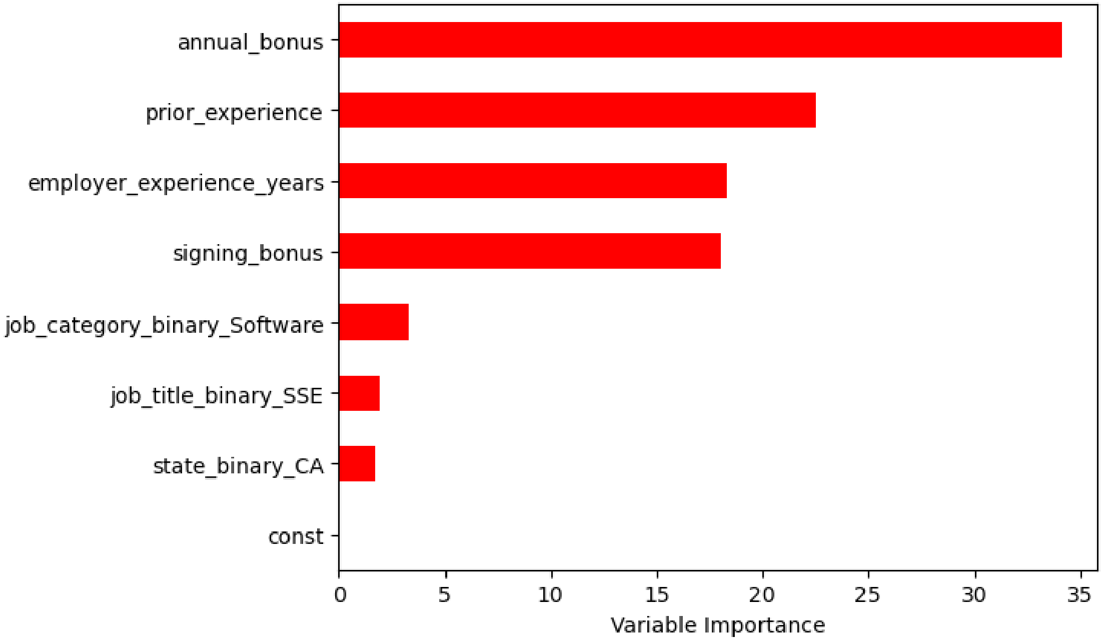
The next two graphs will use two ensemble models, which will combine different models to create a tree that can predict better. The two models will be Bootstrapping, which samples the data and then replaces said data with the newly predicted value, and RandomForest, which does the same thing as Bootstrapping, but takes a smaller number of random features in each split to create a less correlated model. The total number of features is 7, and the RandomForest model will use only 3. Below is a scatter plot graph that shows each models predicted values of the logged salary, and can show the difference in running the Bootstrapping model and RandomForest model.



```
The MSE of the Bootstrapping Model is: 0.09563382923433217
The MSE of the RandomForest Model is: 0.09421448833526744
```

We can see that running a Bootstrapping model with all 14 features, and using Random Forest with only 5 features results in virtually the same graph. The data is scattered in virtually the same way, with only minor variances in the outlier plots. Also, the MSE of both models is roughly the same at 10%. Since the RandomForest model gives a smaller MSE, we will use this in the following section, finding the Importance Matrix of the variables.

4.2.2 Importance Matrix



The importance matrix of the RandomForest model shows the most important explanatory variable in predicted salary growth is the annual_bonus, which makes sense economically. While the annual_bonus isn't reflected in a repeated paycheck, it defineitly is a growth in the earnings made by a worker in a year, with the firm handing them out each extra year the worker stays in a firm. The next most important variables fall in line with results found previously, with prior expereience being more important than current worker experience, but both being the next two most important variables in explaning salary growth. We can also see that the binary variables defined for the preferred specification are quite insignificant in explaining salary growth, which contrasts the regression coefficents found in the previous section. However this could come from the models not being well equipped to preform splits on a binary variable.

4.3 OLS and ML comparison

Since both the OLS and ML models both attempt to predict salary with the same explanatory variables, it seems intuitive to compare the two directly in predictive power. We will use the remaining 20% sample that wasn't used in the regression creation or fed to the Machine Learning model to test the predictive power of each model. We will run each model on the explanatory variables in the sample, and compare the predicted salary values with the real values found in the sample set. Note, since the ML models log the salary to keep the values smaller, we will have to log the salary again for both the OLS regression and ML models. However, since the log function is bijective, it won't change the results at all. The following results show the different Mean Squared Errors, and total Explanatory power of each model.

```
OLS Regression Mean Squared Error: 0.3547375337024843
OLS Regression R-squared: 0.1257939729094122
```

```
RandomForest Regression Mean Squared Error: 0.2896512131409575
RandomForest Regression R-squared: 0.2861910223058186
```

Looking at each models error of prediction, we can see that the RandomForest model is more accurate and more explanatory. The MSE is lower, meaning the errors the decision trees make are lower than the errors of the OLS model, and the R-squared of the RandomForest model is higher, so the trees have more explanatory power than the OLS model aswell. The main reason for the better performance of the ML model is that the trees can capture more of the non-linear relationships the dataset could have. The OLS models created don't have any non-linear elements, mostly because half the variables are binary already, but the RandomForest model isn't constrained by the form the regression and finds patterns regardless. There could be a hidden economic relationship between terms such as a workers prior experience and the signing bonus they receive, or their job title depending on their field, and the RandomForest model is able to caputre that. Another reason could be the OLS model's lack of outlier removal, since all the data is taken in regardless of how far out that point is to the rest of the dataset. By the nature of collecting a country wide survey, data points will vary greatly while still being important in determining economic effects, but the OLS model will become less accurate because of that. Proof of this comes from looking at the standard errors of the OLS coefficents found. The binary variables specifically have massive variances, and this shows that the OLS model isn't able to accurate capture all the datapoints in the dataset, while the RandomForest model can. Overall, the findings from the RandomForest model are more accurate and economically useful than the OLS model. These results could change if more robust OLS models were used, such as more non-linear terms or finding an Instrument Variable to include, and so further work could be done in these areas.

5 Conclusion

We have found a variety of results in each of the major sections of this paper, and the most important results are as follows. Section 2 was a basic examination of the dataset along with some graphs and patterns that were identified. We examined what salary trends were like in each of the major companies, categories, job titles, and found an interesting pattern in salary growth in 3 year intervals of tenure. The workers were located most frequently in California, New York, and Washington, and the highest paying job category was Management, while Software was the most popular. The 3 year interval trend is theorized to be related to a job hopping interval, where workers are given job offers every 3 years to a new job that has a significant increase in salary. Further testing of this hypothesis could be done with more data on firm entry and exit data, and would be worth examining to draw more conclusions about a workers salary growth coupled with their mobility. Furthermore, section 3 looked at the geographical data, and we can find the impact in a workers location to their salary being quite significant. While it is heuristically obvious that the city where Silicon Valley is located has the highest average salary, it is still interesting to note that the workers in California still tend to migrate towards San Francisco, a pattern that isn't observed as strongly in the other two major states. We also saw how the different counties in each state changes in their average salary, with New York's being very spread apart and Washington very concentrated. It is also worth noting that the dataset did not include enough geographical data to draw more refined conclusions from, and more interesting results could be found if the workers included zip-code location, or at least the dataset being bigger so that there was a bigger spread of worker location. Section 4 web-scrapped the locations cost of living data, and switched the examination to the firms in finding the firms performance, and how it relates to workers salary. The cost of living data found implies that the higher salary in California doesn't actually translate into a higher compensation, as San Francisco has a cost of living significantly higher than Seattle, even though the average salary isn't too different between them. New York's has it the worst, having a similar cost of living to California, while having less salary growth. Looking at the firm performance dataset, we saw that the company's stock price and firm value actually has little to no effect on how they pay their workers. This result supports the hypothesis that workers are paid based on the workers market value, and has very little to do with the firms. The reason the tech industry has the high salaries to begin with is how valuable that sector is in the overall market, both within America and internationally. The results could be far more useful if we were able to access all the firms data, including the privately traded ones, along with all the firms actual revenue and operating costs found in their financial history. However due to the sensitivity of that data, and the hypothesized lack of impact of that data anyways, future research on this topic doesn't necessarily need firm performance, and instead examination of the workers market. Finally, section 5 uses linear regressions and machine learning models to predict a workers salary. The models emphasized looking at the workers current experience and prior experience, and the results found suggest that both models also place heavy importance on those measures. Also, we found that a workers signing bonus and annual bonus influence the workers salary a high amount, and can be related back to the cost of living data found earlier, as these bonuses can help influence a workers decision in job location and how long they spend at a given firm, i.e. if they are deciding between two firms with differing salaries at differing locations, looking at their offered signing bonus and the firms historical annual bonus will be the most influential factor. In conclusion, this paper was able to take a close examination on how the tech industry's worker salary is affected by their length of employment, both at their current firm and their overall experience. We examined how results change across location, field, and level of experience, and found some expected results and surprising results. Overall, future steps in this topic would involve looking at each worker more closely, finding perhaps a time series dataset of each worker, and a much larger sample of workers with more diversity. Also, future work should include a workers age, gender, race, education, and more unique data to examine how the results differ among each of those categories, and possibly examine topics such as a worker's barrier to entry, how salary differs among gender, how older workers current compensation differ to a new worker, and how different levels of education affect a workers prospective salary growth.

6 References

Abowd, J. M., Kramarz, F., & Margolis, D. N. (1994, November 1). High Wage Workers and High Wage Firms. NBER. https://www.nber.org/papers/w4917

Abraham, Katharine G., and Henry S. Farber. (1987) Job Duration, Seniority, and Earnings. The American Economic Review 77, no. 3: 278–97. http://www.jstor.org/stable/1804095.

Altonji, Joseph G. and Shakotko, Robert A. (1985). Do Wages Rise with Job Seniority? NBER Working Paper No. w1616, Available at SSRN: https://ssrn.com/abstract=227194

Dohmen, T. (2015). Performance, seniority, and wages: formal salary systems and individual earnings profiles. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0927537104000132

Dustmann, C., & Meghir, C. (2005). Wages, experience and seniority. OUP Academic. https://academic.oup.com/restud/article/72/1/77/1585793

Gibbons, R., Katz, L. F., Lemieux, T., & Parent, D. (2002, April 11). Comparative advantage, learning, and sectoral wage determination. NBER. https://www.nber.org/papers/w8889

Gibbs, Michael & Baker, George & Holmström, Bengt. (1994). The Wage Policy of a Firm. The Quarterly Journal of Economics. 109. 921-55. 10.2307/2118352.

Gruetter, M. (2009). The importance of firms in wage determination. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0927537108001012

Hek, P. D., & Vuuren, D. V. (2011, January 15). Are older workers overpaid? A literature review. SpringerLink. https://link.springer.com/article/10.1007/s10797-011-9162-3

Thompson, M. (2009). Salary progression systems. In Reward Management (2nd ed., pp. 120–139). essay, Routledge. Wooldridge, J. (2019). Introductory econometrics: a modern approach (7th ed). Cengage Learning

Zhang, Z. (2016, January). Missing data imputation: Focusing on single imputation. Annals of translational medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4716933/