**Analysis of Technology Centered Career Salaries**

Matthew Carey

Ryan Kaplan

Daivik Nambiar

Hayden Ramirez Suarez

The University of North Carolina at Charlotte

Data Science 2302-002

Professor Ted Carmichael

Professor Iván Flores Martínez

April 29, 2024

Table of Contents

Introduction 3

1. Stakeholders 3
2. Background 4
3. Research Questions 6

Methodology 7

1. Data Exploration 7
2. Data Preprocessing 13

Model Selection 18

1. Multivariate Linear Regression 18
2. Decision Tree Regressor 21
3. Gradient Descent 22
4. Model Comparison 24

Results 26

Discussion 31

References 34

**Analysis of Technology Centered Career Salaries**

**Introduction**

In a world where a cheeseburger no longer costs a quarter, and less and less of the affluent can afford personal yachts, the value of the dollar has become less and more concurrently. Less valuable because the purchasing power of the dollar has diminished, but more valuable because it is more important than ever to squeeze each dollar out of income sources. In order to combat this reality, and add to the wealth of knowledge surrounding potentially increasing an individual’s income, we are analyzing the tech companies salary dataset to find potential relationships between salary and other variables.

The objective of our research is to determine which variables have an impact on tech centered career salaries, how strong of an impact those variables are, and to make a prediction as to what salary an individual can potentially expect based on the value of variables that have an impact on salary. The expectations of this analysis are that variables such as ‘Years of Experience’, *Market Cap Group*, and ‘Revenue per Employee’ have a strong influence on predicting the salary of an individual, with also a better understanding of how they affect the outcomes. These objectives and expectations will be formally addressed through the research questions and hypotheses in the upcoming *Research Questions* section.

*Stakeholders*

For stakeholders, this analysis presents an opportunity to be bought or funded by companies similar to LinkedIn, Indeed, and Monster, who have a large stake in the analysis and trends of the job market. Recruiters can use the research to better focus resources, increase success rate, and reduce time to successfully aid in finding employment for individuals. Social media companies with large reach could benefit by the traffic of posts and articles generated centering around the subject matter of this analysis. By gaining better insights into the driving factors of salaries, companies can implement initiatives to attract certain markets, ultimately contributing to the overall growth of the industry and the economy at large.

*Background*

Understanding the factors of influence on tech salaries is of the utmost importance to potential and current applicants, as well as the companies setting the salaries of said careers. One expected variable of influence on tech salaries is highlighted in research performed by the U.S. Bureau of Labor and Statistics. As reported it is clear that with more education, the median weekly wages increase, while at the same time unemployment rates decrease (Vilorio). Understanding the factors that could influence potential salaries equips individuals with the knowledge of what decisions to make when pursuing a career. This also aids companies in providing competitive compensation to individuals in relation to their education level, or information to advertise competitive compensation to attract individuals with a certain level of education.

Tech jobs are some of the most sought after jobs in the country, meaning competition is fierce. Speculation of jobs being eliminated with the arrival of generative artificial intelligence only adds to the pressure experienced in the marketplace. In layman's terms, demand is high for these positions and supply is expected to decrease, leading to a supply crunch driving down the salaries of careers affected. Evidence of a current supply crunch can be seen in the number of layoffs at tech companies increasing by nearly one hundred thousand from 2022 to 2023 (Fore). When it comes to salary, a high supply of workers and low demand of employees decreases the salary paid to an individual. These ideas contribute to our analysis by illustrating the need for a high level understanding of the variables that have influence or correlation with tech salaries. This knowledge aids individuals, looking to be competitive, to tailor career and educational decisions in order to maximize success in the technology centric career marketplace. With increasing competition any advantage an individual can gain will set them apart from the pack increasing the probability of success. As for the benefits gained by companies with the increase in competition, salaries decrease, but also, the quality of applicants increases.

There is another game changing factor that is coming with a high level of influence on the reduction of salaries in the tech space, such as the rapid rise of remote work. The founder of LinkedIn reported that remote work job postings have decreased by fifty percent while demand for these jobs is still extremely high accounting for 46% of all applications done through LinkedIn. This trend is also prevalent in hybrid positions (Naprys). It will be in the interest of these companies to reduce the salary paid based on these types of positions while demand, of work from home or hybrid jobs, is high for job seekers and low for employers.

With the vast amount of information available, and articles frequently contradicting each other, the ambiguity of the marketplace is at an all time high. Analyzing the data available, to demystify the prospects of the job market, is our answer to arm individuals with information to make the best career decisions. Companies can benefit by having the ability to offer competitive salaries, as well as hire higher quality employees. This enables both parties to make well informed decisions as they navigate their way through the marketplace. In this context, our research aims to bridge the gap between the industry and its impact on salary by identifying insights that enable individuals and organizations to better understand the field.

*Research Questions*

The main question we want to answer is which variables, of the variables of interest, can influence the salary of an individual with a technology centered career in the United States. Our variables of interest include the individual level variables years of experience, education level, title, sex and race. We will also look at company level data including the number of employees at the company, revenue per employee, sector, and market cap group. Finding an answer to this question will give an individual the ability to tailor career decisions in a way to increase the probability of receiving a higher level of compensation. Stakeholders can use this information to filter and attract applicants, as well as set competitive salaries. Gaining an enhanced insight into the determining components of salary can allow stakeholders to strategize unique initiatives to target specific populations. We expect that the most influential factors include the years of experience an individual has, the size of the company they are working in, and the revenue generated by the company. This will be analyzed through correlation statistics throughout the data analysis process.

We would like to also answer the question of how strong of a relationship do significant variables have in predicting one’s income within technology centric careers based in the United States. We expect years of experience to have the largest impact in predicting one’s income. This hypothesis was formulated based on the time spent honing skills, the propensity for yearly raises issued to employees, and many iterations of the data exploration process. In order to test this hypothesis, we will be looking at the correlation statistics between years of experience and base salary, as well as the value of the calculated coefficient of years of experience in a multivariate linear regression model.

**Methodology**

*Data Exploration*

Our analysis focuses on the tech companies salary dataset. This dataset consists of information corresponding to individuals with technology centered careers. The data includes general demographic information about the individual, such as ‘Years of Experience’, *Education* level, *Sex*, and *Race*. For income data, the dataset provides values for metrics such as ‘Base Salary’ and ‘Total Yearly Compensation’. Both being potential dependent variables. The success of the analysis predicated on choosing the correct dependent variable, or the variable we wanted to predict. ‘Total Yearly Compensation’ was an option as the dependent variable, but there are some issues with this value. The first being that ‘Total Yearly Compensation’ is a dynamic value. Many external factors can influence this value including, but not limited to, stock price, company performance, and employee yearly review. It is for this reason we are choosing to use the static value of ‘Base Salary’ as our dependent variable. Examples of company level information provided include the *Company Name* and *Title*. We will be evaluating the variables that have the most impact on an individual’s ‘Base Salary’.

We will also be merging the tech salaries dataset with a dataset of current stock market statistics. Utilization of stock market statistics adds variables to our dataset that would otherwise not be available. This will also focus our population on United States based public companies, which in turn will create a more accurate and useful model. This dataset will allow us to analyze variables like the *Market Cap Group*, a categorical grouping of companies based on the number of outstanding shares multiplied by the stock price. This can be interpreted as the economical impact a company has in the United States. Other metrics include the ‘Number of Employees’ corresponding to the number currently employed by the company, the ‘Revenue per Employee’ generated, the *Industry* of the company, the *Sector* of the company, and ‘Debt Growth Year Over Year’. By enhancing the dataset with metrics like these, we are able to further investigate variables that potentially influence or correlate to an individual’s salary.

The initial step of diving into our analysis, included loading our dataset into a Jupyter notebook to be explored. Upon loading the data, we found that the null values corresponded to categorical variables. Specifically, *Company* had five null values, *Sex* had 19,540 null values, *Race* had 40,215 null values, *Education* had 32,272 null values, *Tag* had 870 null values, and *Level* had 123 null values. With these being in the set of initial variables of interest in our analysis, the corresponding entries with null values in any of these categories were dropped from the dataset. The entries were dropped from the working dataset because replacing these values would adversely affect the categories corresponding to the replacement value. For instance, assigning the value of Bachelor’s Degree in *Education* to the null values in the *Education* variable could skew the mean and standard deviation of base salary for entries explicitly associated with the Bachelor’s Degree category. Also, replacing these values with a catchall label like Not Available, provides no insight unlike the meaningful categorical values. This type of category also has the issue that each of the people associated with these entries has a measurable level of education that would be expected to hold weight in this analysis as a predictor of base salary. Therefore the ambiguity could lower the model’s accuracy or capture noise in the data. With a large enough dataset resulting from the removal of these entries, 21,515 entries remaining, we had no reason to investigate further into methods of replacement. In the future, more accurate data can be gathered or existing entries can be traced back to the origin, in order to complete their data profile.

Values of categorical data were then checked for uniqueness. This process verified that categorical values were in a workable format and did not contain redundant values. For example, case sensitive issues like “Female” and “female”, or redundancy issues like “Apple” and “Apple Inc.” would be reassigned for uniformity.

A function was written to split the location column into a city and state column. This action served multiple purposes for our analysis. The first purpose was to have the ability to group entries by state in the future. The second, and more pertinent purpose, was to set values that contained more than two strings delimited by commas of entries to null. These entries correspond to locations outside of the United States and were dropped from the working dataset. Choosing to focus our analysis on entries in the United States refined the scope of our study, as well as eliminated many outliers in the dataset. Thus giving us the ability to create a more accurate and useful model of the data.

A second data set was then uploaded to the Jupyter notebook containing current stock information to be linked to the salaries dataset. The purpose of linking these datasets is to add more independent variables to research potential relationships between company related statistics and the dependent variable of base salary. In order to merge the datasets, the values of the company column, corresponding to company names, in both datasets were reformatted to uppercase strings. Also, the most recurring companies in the salaries dataset were renamed to the corresponding value in the stocks dataset. This was enough to create an initial merge of the datasets.

After the merge, there were numerous null values to process. Working in Excel,the dataset was cut down to only include variables from the original salaries dataset with the added category corresponding to the ticker symbol of the company. A ticker symbol is a unique identifier assigned to a publicly traded company for trading purposes and is oftentimes an abbreviation of the company name. For example, Amazon’s ticker symbol is ‘AMZN’. The null values remaining in the *Ticker Symbol* category were manually researched and the ticker symbol was added. Private companies, who are not traded on the stock market and therefore do not have corresponding ticker symbols, were assigned the *Ticker Symbol* label of “PRIVATE”. Since our stock dataset encompasses only U.S. based stock exchanges, companies that corresponded to public companies on foreign stock exchanges were assigned a *Ticker Symbol* label of “INTERNATIONAL”. Government operated companies, as well as companies who through the research process were still ambiguous, were left with null values. Subsidiaries were assigned the *Ticker Symbol* of their parent company. The column names were also changed to follow a uniform naming convention. This process was performed in Excel to minimize the number of hours spent labeling the data. With additional time, automating this process would be beneficial for expansion of the dataset. After this process was completed, the dataset was merged again with the stock dataset, this time on the *Ticker Symbol*. The new complete dataset was uploaded to the Jupyter notebook to be used and analyzed as the working dataset.

After uploading the dataset, the entries with null values in the *Ticker Symbol* category were dropped from the dataframe. The aggregated dataset was then subsetted into multiple datasets. The primary dataset contains entries where the company is publicly traded on United States based stock exchanges. The secondary datasets contain entries where the company is publicly traded on international stock exchanges and private companies, respectively. These secondary datasets have potential to be used for further comparisons in the future.

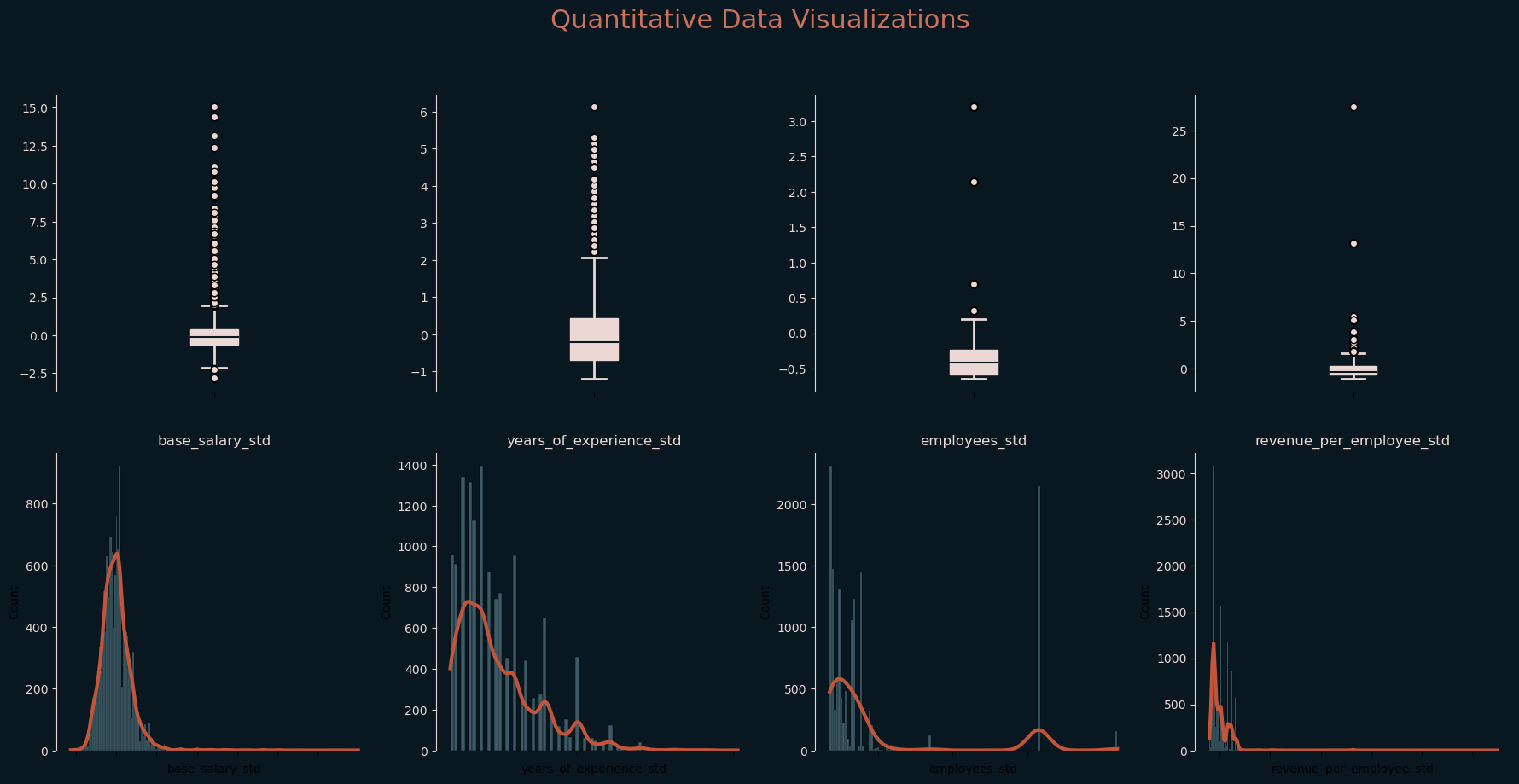
The public company dataset was narrowed further by the subsetting of variables to be investigated. The variables of interest included:

* ‘Base Salary’
* *Title*
* ‘Years of Experience’
* *Education Level*
* *Race*
* *Sex*
* *Market Cap Group*
* *Sector*
* ‘Number of Employees’
* ‘Revenue per Employee’

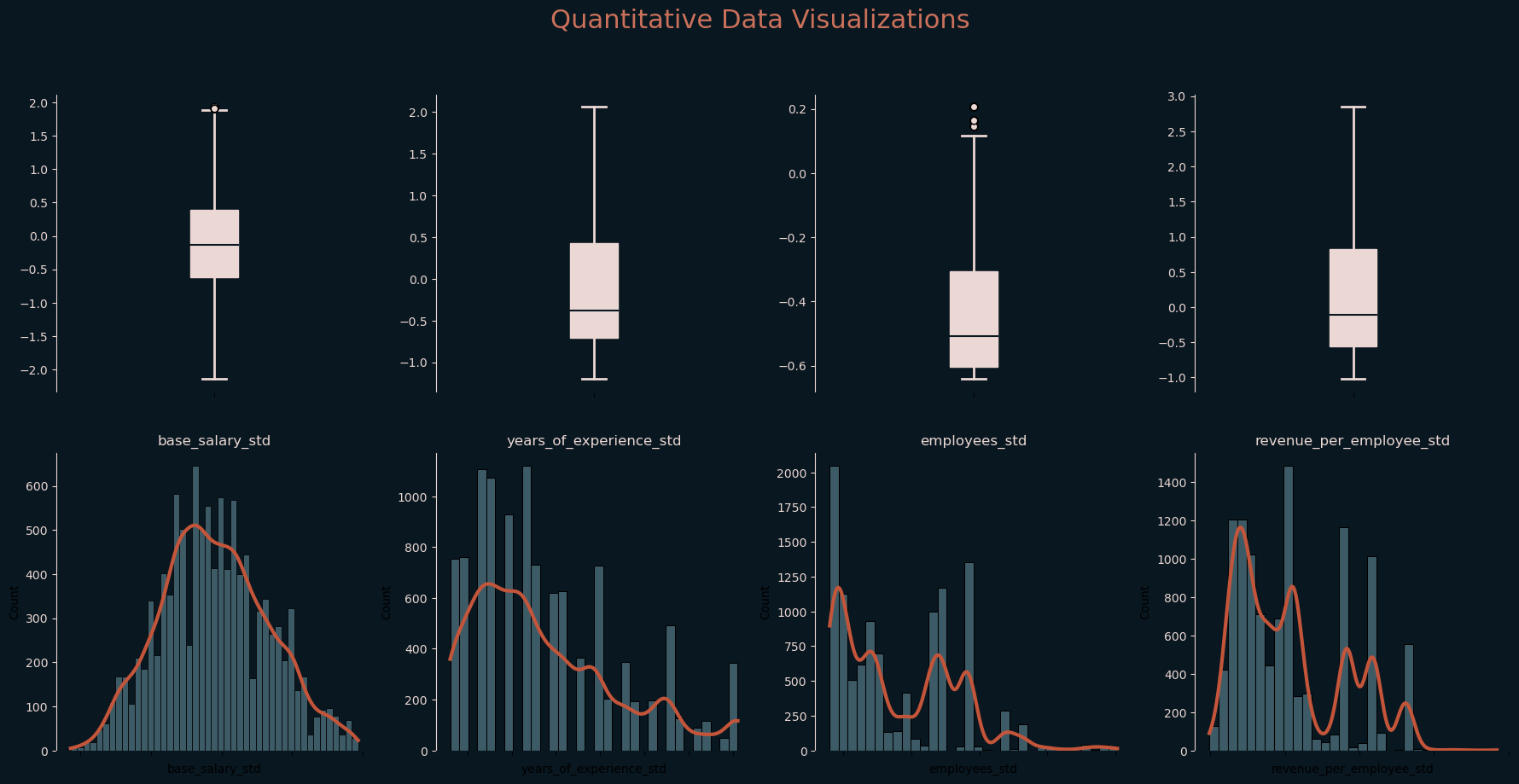
Originally, we had several other variables included in our variables of interest including ‘Profit per Employee’, ‘Years at Company’, ‘RSI’, ‘Cash Over Market Cap’, and ‘Debt Growth Year Over Year’. ‘Profit per Employee’ has a collinear relationship with ‘Revenue per Employee’. ‘Profit per Employee’ is a function of ‘Revenue per Employee’ and influenced by external factors like debt taken on by a company for the given year. This makes it a less reliable predictor, therefore we dropped it from the variables of interest. ‘Years at Company’ has a collinear relationship with ‘Years of Experience’, however ‘Years of Experience’ is a more reliable metric due to its holistic nature. For example an individual can have twenty years of experience and one year at the company simultaneously. It is for this reason we dropped ‘Years at Company’ from the variables of interest. ‘RSI’, or the relative strength index is a metric for evaluating if a stock is overbought or oversold. This is a function of stock price and volume over time. ‘RSI’ is a time sensitive metric and with little to no direct relationship in predicting ‘Base Salary’, we are dropping it from the dataset. ‘Cash Over Market Cap’ is a measure of the cash a company has on hand relative to the market cap of the stock. The market cap is the number of outstanding shares times the stock price. Because of the proportional property of this metric we are dropping it from the dataset, in lieu of more direct variables like ‘Revenue per Employee’ and *Market Cap Group*. ‘Debt Growth Year Over Year’ is also a time sensitive metric encompassing the last year of debt growth to overall debt. Given that our dataset has entries dating back to 2016 and our stock market data is current, we have eliminated this from being a variable of interest.

Quantitative variables were then standardized to reduce the scale of the numbers. Standardization reassigns values to z-score representation. This makes the mean of a variable equal to zero and the standard deviation equal to one. A function was then written to remove outliers from the dataset. This function implemented the standard formula for eliminating outliers, by eliminating values above the threshold of the upper quartile plus one and a half times the interquartile range and below the threshold of the lower quartile minus one and a half times the interquartile range.

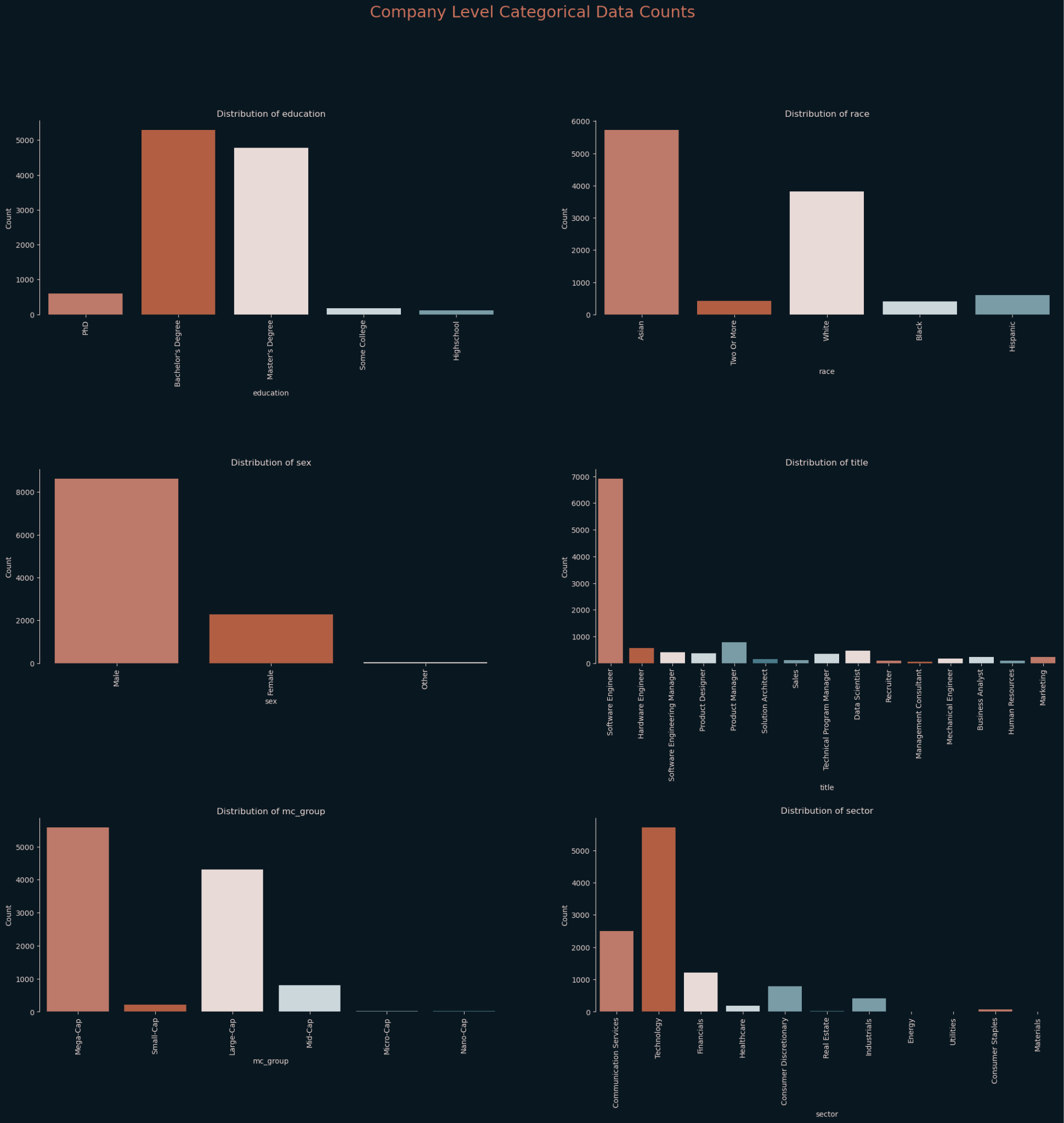
*Data Preprocessing*

Firstly, box plots and distribution plots were created to visualize quantitative variables. 

After visualization, it was evident that all quantitative variables needed to be processed for outliers. This would reduce the range of the variables and the skew. We applied the processOutliers function on all variables, and the results are visualized below.



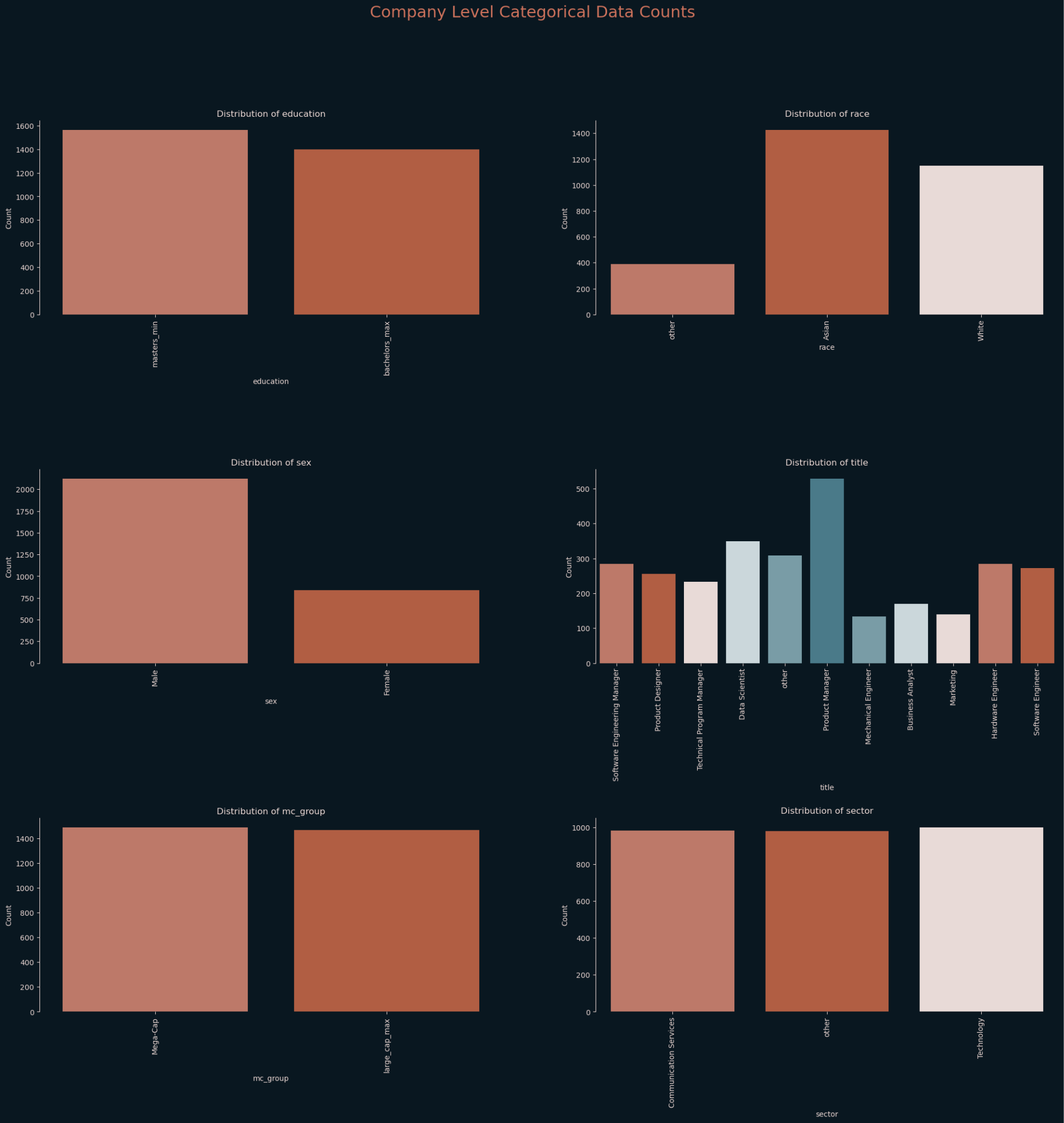
As illustrated, the range and skewness of each variable was significantly reduced. After processing outliers, we reviewed the summary statistics of the quantitative variables. Moving on to categorical data, these variables were visualized using count plots.



After the initial visualization, pictured above, of the categorical variables, we eliminated categories that had less than thirty observations. This led to the elimination of Micro-Cap and Nano-Cap categories in the *Market Cap Group*, as well as the Energy, Materials, and Utilities categories in *Sector* and the Other category in *Sex*.

Our categorical variables are unbalanced in the above visualization, so our next step was to try to balance out the categories. We first processed the categorical variables from the tech salaries dataset. *Education* categories were aggregated into two categories representing individuals with a maximum level of education equivalent to a Bachelor’s degree and a minimum level of education equivalent to a Master’s Degree. The lower represented *Race* categories were aggregated into a category named Other. This included individuals of Black, Hispanic, and multiple races. Lower represented *Title* categories were aggregated in a similar way with Management Consultant, Recruiter, Human Resources, Sales, and Solution Architect being renamed to Other. Software Engineer is an overly represented category in the *Title* variable, so we removed all entries associated with software engineers, took a sample of 400 of those entries and added the sample back into the dataset.

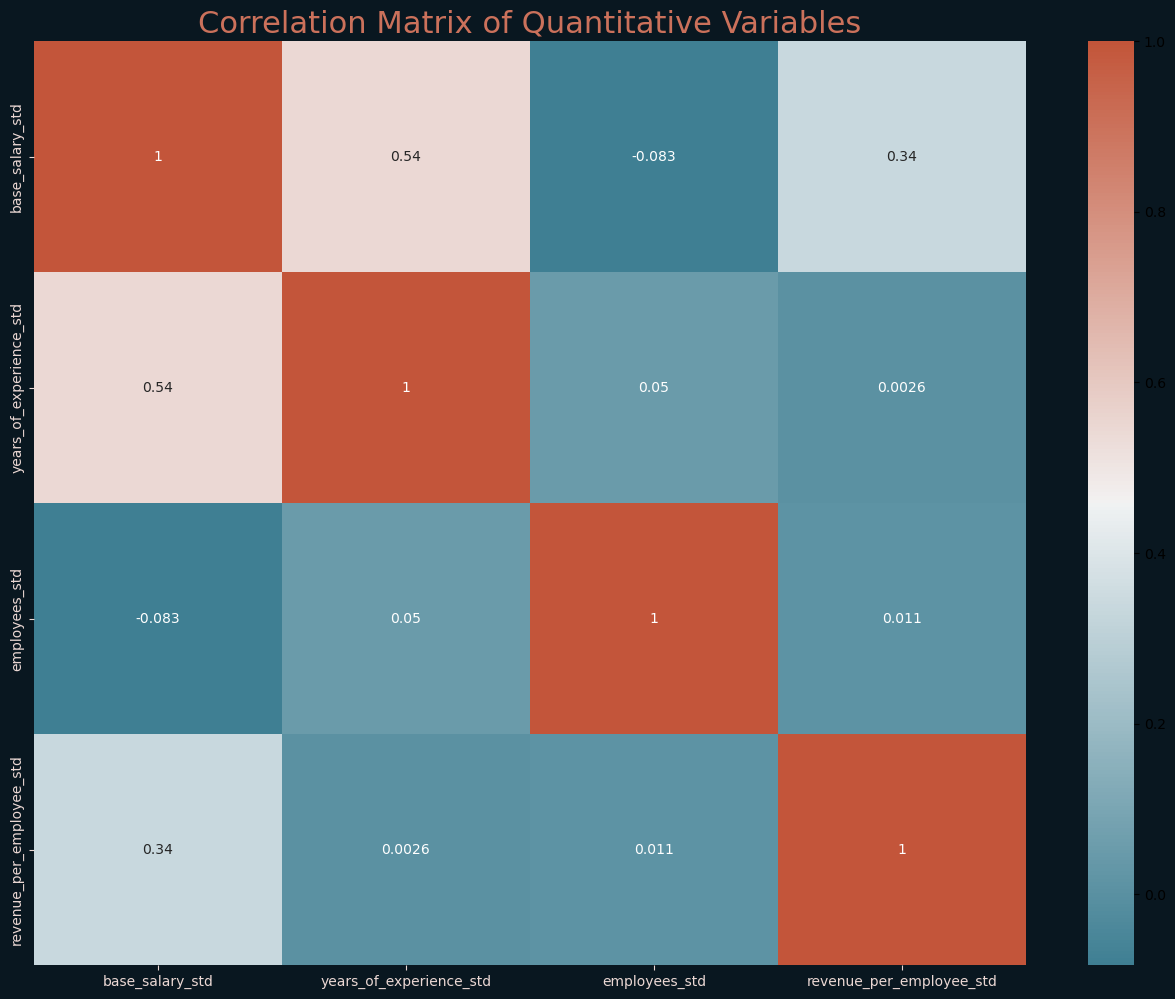
Next, we processed the categorical variables associated with the stocks dataset. Entries were removed corresponding to the Real Estate and Consumer Staples of *Sector* for having below thirty observations. We then aggregated the Healthcare, Consumer Discretionary and Financials categories of *Sector* to a category labeled Other. We also aggregated *Market Cap Group* by assigning Large Cap, Mid-Cap and Small Cap companies to a category labeled Large Cap Max. The Technology sector, as illustrated above, was overrepresented in the dataset, so we removed these from the dataset, and repopulated the dataset with a sample of one thousand entries associated with the Technology label of the *Sector* category. Below is the visualization of the change in distribution from these transformations.



As illustrated, the categorical variables are much more balanced and the categories with lower numbers of observations have a much more substantial representation than before, proportionally.

Columns with categorical data were then encoded in order to analyze these columns. Nominal categorical variables were dummy encoded. The ordinal categorical variables, including *Market Cap Group* and *Education*, were encoded with the lower values being assigned zero and the higher values being assigned one.

Upon the completion of processing quantitative and categorical variables, a correlation matrix was constructed to analyze the relationship between quantitative variables.



Using the correlation matrix pictured above and a threshold of .3, to indicate a strong positive linear relationship, and a threshold of -.3, to indicate a strong negative linear relationship, we can see that ‘Years of Experience’ and ‘Revenue per Employee’ have strong positive linear relationships with base salary.

**Model Selection**

*Multiple Linear Regression*

The primary model we used was a multivariate linear regression model. We chose multivariate linear regression because of the ability to create a formula for predicting ‘Base Salary’ based on its relationship to the independent variables, however, this method is susceptible to variables without linear relationships to the dependent variable. Another downside is that the model is sensitive to outliers. Even though outliers of quantitative variables were accounted for in this analysis, outliers of categorical variables were not accounted for in extreme detail. For example, if there are extreme values of ‘Base Salary’ within the distribution of ‘Base Salary’ among entries associated with Bachelor’s Degree educated individuals, these values would influence the coefficient corresponding to Bachelor’s degree.

The multivariate linear regression model uses the ordinary least squares method to reduce the sum of the squared residuals of the model. We also wrote a function to implement cross-validation of the statsmodels library. This technique divided our data into five folds. Four of the folds were used as training datasets and one fold was used as the test dataset. A model was created for each of the four training datasets. Each of the models’ statistics were aggregated into corresponding lists and the mean and standard deviation of each list was reported back. These statistics included the R2 value, the adjusted R2 value, the F-statistic, the p-value of the F-statistic, and the Durbin-Watson statistic. Coefficients of the models were also aggregated into a list with the mean being assigned as the final model’s coefficients. Using the mean of the coefficients instead of the coefficients of the model with the best fit increases the robustness of the final model, and reduces the impact of aforementioned remaining outliers. Standard deviations of the coefficients along with p-value were also reported in the summary. It was important to include the standard deviation with these statistics in order to be able to interpret the consistency between models. For instance, high standard deviations could mean that an independent variable may not be a good predictor of ‘Base Salary’ and would require further investigation.

From here we built many models using the cross validation function. The first model included all independent variables of interest. Our goal with this model was to evaluate the significance of the quantitative variables. Any quantitative variable with a p-value greater than 0.05 would be considered not significant and the variable with the highest p-value would be dropped from the model. The model would then be rebuilt and this process would repeat until all quantitative variables had a p-value below 0.05. None of our models’ quantitative independent variables needed to be dropped from the model based on the criteria laid out. For reference, here are the metrics of this model, R2 was 0.557, adjusted R2 was 0.553, the F-statistic was 147.614, the p-value of the F-statistic was 0.000, the Durbin-Watson statistic was 1.966. We will provide an interpretation of these values when evaluating the different models against each other.

We then analyzed the significance of categorical variables by dropping the entirety of the categorical variable from the dataset and constructing a multivariate linear regression model with this subset of features. Here we expected degradation of the model. Little to no degradation would imply that the categorical variable removed was not a good predictor of base salary and would subsequently be dropped from the final model. It was found that the removal of *Sex*, *Race*, *Sector* and *Market Cap Group* dropped R2 by 0.002, 0.001, 0.017 and 0.004 respectively. In efforts to simplify the model, these categories were dropped from the final model. The 0.017 drop of *Sector* is rather high compared to the other values, but the significance of the variable after being aggregated into three categories including Technology, Communication Services, and aggregated category Other had diminished the meaning of the overarching categorical variable and without a high level of impact on predicting ‘Base Salary’, it is hard to justify keeping this variable in the model. Further data collection to balance out all sectors could re-establish this as an impactful variable on predicting ‘Base Salary’ in the future.

For the final multivariate linear regression model, the dependent variable is ‘Base Salary’, the independent variables include ‘Years of Experience’, ‘Number of Employees’, ‘Revenue per Employee’, and categorical variables *Title* and *Education*. Summary statistics include an R2 value of 0.528, adjusted R2 value of 0.525, F-statistic of 188.165, p-value of the F-statistic of 0.000, and a Durbin-Watson statistic of 1.963.

With the final model in hand, we decided to create a secondary test dataset using the values originally removed from the dataset, for balancing purposes, corresponding with the Technology *Sector* entries and the Software Engineer *Title* entries. Using this dataset and the test dataset we calculated the mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R2. These values are listed in the table below and will be interpreted in the upcoming model comparison section.

|  | Test Data | Expanded Data |
| --- | --- | --- |
| MSE | 0.360 | 0.343 |
| RMSE | 0.600 | 0.585 |
| MAE | 0.471 | 0.343 |
| R2 | 0.469 | 0.412 |

*Decision Tree Regressor*

The secondary model we will be using will be a decision tree regressor. Using a decision tree regressor will also give us the ability to predict ‘Base Salary’ based on the independent variables. This method works by partitioning the training data and assigning a value to each partition or node. As the independent variables are analyzed while traversing the tree, these values are added together to create a prediction. A prominent issue to decision trees is the tendency to overfit the training data. This can lead to extremely high residuals and subsequently high MSE, RMSE, and MAE metrics.

In order to combat overfitting training data and creating a robust model, we used a grid search method with cross validation. This is a computationally expensive process and having a relatively small dataset with a non-extensive list of independent variables aids in keeping the time to process down. Also, with gridsearch, it is possible to implement parallelization allowing the algorithms to utilize multiple CPU cores when available. We were able to run a wide range of hyperparameters in one go at the grid search because of this. These hyperparameters included minimum samples to cause a split. For this we ran a range of integers from two to one hundred. We also specified the minimum samples per leaf in a range of integers from one to twenty. We specified the maximum depth of the tree from five to none. The next hyperparameter customized was the criterion which evaluates quality of the splits in the tree, the criteria tested included MSE, Friedman MSE, and MAE. The Maximum Features hyperparameter was also given a range of four to fourteen to be evaluated. Lastly, the Max Leaf Node hyperparameter was given a range of 10 to none. The none values in the hyperparameters are likely to cause many overfitting problems with decision trees and was one of the driving factors in creating the aforementioned expanded dataset to test against the decision tree model in order to evaluate performance. Many restrictions were placed on the ranges of the criterion, but ultimately led to a model that did not perform as well as the multivariate linear regression model so the best parameters of the grid search are what is illustrated in the Jupyter Notebook, for simplicity. These include the criteria of MSE, no max depth, 9 max features, no limit to number of leaf nodes, five minimum samples per leaf, and fifty minimum to cause a split.

Again, we calculated the MSE, RMSE, MAE, and R2 for the model against the test dataset and the expanded dataset. MSE for the test dataset was equal to 0.371 and 0.385 for the expanded dataset. RMSE for the test dataset was equal to 0.609 and 0.621 for the expanded dataset. MAE for the test dataset was equal to 0.473 and 0.480 for the expanded dataset. R2 was equal to 0.451 for the test dataset and 0.340 for the expanded dataset.

*Gradient Descent*

Because the decision tree model did not perform as well as the multiple linear regression model, we also created a model using gradient descent. Gradient descent traverses a gradient finding the local minima and maxima based on the derivatives of the slope of the gradient. A downfall to gradient descent modeling is that it can converge with multiple linear regression modeling, however the idea behind trying to create a gradient descent model was to find coefficients that captured more of the relationship between independent variables and base salary. We can do this by fine tuning the number of iterations of the traversal, as well as the learning rate of the algorithm. With a high number of iterations and a low learning rate, we can expect similar results to the multiple linear regression model, but by decreasing iterations and increasing the learning rate, we may be able to capture more of the data.

For this we used modules provided by Dr. Andrew Ng of Stanford University in his Machine Learning by DeepLearning, AI Coursera specialization. To use the tools provided we transformed our training and testing datasets into numpy arrays. We passed the training arrays to the gradient descent algorithm with iterations set to three thousand and the learning rate set as low as 1e-9. From here, we increased the learning rate by a factor of ten on each run, to ultimately have a learning rate value of 0.1 and number of iterations was cut to one thousand. Evaluation metrics of the model were close, but not as good as the multiple linear regression model. We then decreased the number of iterations to four hundred and increased the learning rate to one. Final evaluation metrics for all three models are in the table below.

| Statistic | Method | Test Data | Expanded Data |
| --- | --- | --- | --- |
| MSE | MLR | 0.360 | 0.343 |
|  | DT | 0.371 | 0.385 |
|  | GD | 0.359 | 0.343 |
| RMSE | MLR | 0.600 | 0.585 |
|  | DT | 0.609 | 0.621 |
|  | GD | 0.599 | 0.585 |
| MAE | MLR | 0.471 | 0.343 |
|  | DT | 0.473 | 0.480 |
|  | GD | 0.470 | 0.457 |
| R2 | MLR | 0.469 | 0.412 |
|  | DT | 0.451 | 0.340 |
|  | GD | 0.469 | 0.412 |

*Model Comparison*

Based on the testing data, the R2 value of the multivariate linear regression model and the gradient descent model are equal with a value of 0.469. This means that about 46.9% of the variance in the dependent variable is explained by the model when tested against the testing data. This is higher than the R2 value of the decision tree model with a value 0.451. The same happens when tested against the expanded dataset where R2 is 0.412 for the multivariate linear regression model and the gradient descent model. The decision tree has an R2 value of 0.340 when tested against the expanded dataset. This value could signify overfitting of the training data, but also that because Software Engineers corresponded to a large number of entries reintroduced to the dataset, that the decision tree does not make an accurate distinction in this title.

When evaluating the RMSE of the models, we see a similar pattern. RMSE is calculated by taking the square root of MSE, which further emphasizes our discrepancy in residuals. Against the test data, the RMSE of the multivariate linear regression model is 0.600 and for the gradient descent model the value is 0.599. The decision tree model is slightly higher with a value of 0.609. These values indicate how far the predicted values are from the actual values, so the lower values are more accurate. This disparity widens when the expanded dataset is tested against the models. Here we see the RMSE of the multivariate linear regression and gradient descent models equal to 0.585. The decision tree RMSE value is 0.621.

Based on the metrics above, the multiple linear regression model and the gradient descent model perform about equally when looking at R2, MSE, and RMSE. However, looking at MAE, the multiple linear regression model has a significantly lower MAE score of 0.310 on the expanded dataset when compared to the gradient descent model 0.432. This is likely due to the attempt to capture extra noise in the gradient descent model, compared with the strict calculation of the multivariate linear regression model. It is for this reason we believe the multiple linear regression model is the best model and will be the model used for interpretation and make predictions by the function written at the end of the Jupyter Notebook.

**Results**

For the final multivariate linear regression model, there were only three variables with a p-value above zero. These included *Title -* Mechanical Engineer (0.001), *Title* - Product Designer (0.006), and *Education* (0.011). This signifies that all independent variables used in the model are statistically significant. The standard error of the model was equal to 0.055. In order to calculate the hypothesis test statistics, we divided the coefficients of the model by the standard error. After examining the absolute values of the test statistics, the lowest calculated value belonged to the Other category of *Title*. Calculated to be 3.018, this is higher than the typical threshold of 3.0 that indicates strong evidence to reject the null hypotheses, meaning the coefficient can be considered statistically significant. This being the lowest value, of the absolute values of the coefficients hypothesis test statistic, means all of our coefficients can be considered statistically significant.

As for the confidence intervals of the coefficients, when tested using a ninety-five percent confidence interval, none of our intervals included zero. This provides evidence that the effect of every variable is statistically significant. The variable with the lowest range in confidence interval was ‘Years of Experience’ with a range of 0.6 and a value of 0.480. The variable with the widest range in confidence interval was the Mechanical Engineer category of the *Title* variable. The range of the interval was 0.285 and the value of the coefficient was 0.206. The results of the fine-tuning process and tradeoffs made are illustrated in the *Multivariate Linear Regression* section. The coefficients for our multivariate linear regression model are depicted in the table below.

| Independent Variable | Coefficient |
| --- | --- |
| ‘Years of Experience’ (Standardized) | .480 |
| ‘Number of Employees’ (Standardized) | -.472 |
| ‘Revenue per Employee’ (Standardized) | .382 |
| *Title* - Data Scientist | .657 |
| *Title* - Hardware Engineer | .370 |
| *Title* - Marketing | .263 |
| *Title* - Mechanical Engineer | .206 |
| *Title* - Product Designer | .553 |
| *Title* - Product Manager | .658 |
| *Title* - Software Engineer | .616 |
| *Title* - Software Engineering Manager | .910 |
| *Title* - Technical Program Manager | .460 |
| *Title* - Other | .166 |
| *Education* | .180 |

There is only one variable with a negative relationship to ‘Base Salary’ and that is the ‘Number of Employees’. Every other variable has a positive relationship to ‘Base Salary’. The magnitude of these coefficients can be evaluated after destandardization. For the quantitative variable coefficients, the coefficients were multiplied by the unstandardized standard deviation of ‘Base Salary’, and then divided by the unstandardized standard deviation of the quantitative variable. The ‘Number of Employees’ and ‘Revenue per Employee’ coefficients were multiplied by a factor of ten thousand for interpretation. For categorical variables, the coefficient was multiplied by the unstandardized standard deviation of ‘Base Salary’. The approximate unstandardized coefficients are in the table below.

| Independent Variable | Coefficient |
| --- | --- |
| ‘Years of experience’ | 3,852 |
| ‘Number of employees’ | -424 |
| ‘Revenue per employee’ | 208 |
| *Title* - Data Scientist | 32,405 |
| *Title* - Hardware Engineer | 18,253 |
| *Title* - Marketing | 12,965 |
| *Title* - Mechanical Engineer | 10,175 |
| *Title* - Product Designer | 27,276 |
| *Title* - Product Manager | 32,448 |
| *Title* - Software Engineer | 30,370 |
| *Title* - Software Engineering Manager | 44,893 |
| *Title* - Technical Program Manager | 22,670 |
| *Title* - Other | 8,177 |
| Education | 8,867 |

* The coefficient for ‘Years of Experience’ indicates that for every year of experience, ‘Base Salary’ is expected to increase at a rate of $3,852. This makes ‘Years of Experience’ the most influential quantitative predictor of ‘Base Salary’, aligning with our hypothesis detailed in *Research Questions*.
* The coefficient for ‘Number of Employees’ indicates that for every increase of 10,000 employees employed at a company, ‘Base Salary’ is expected to decrease at a rate of $424. This suggests that the larger the company, the less will be paid out in salaries.
* The coefficient for ‘Revenue per Employee’ indicates that for every increase of $10,000 of ‘Revenue per Employee’, ‘Base Salary’ is expected to increase at a rate of $208.
* The coefficient for *Title* - Data Scientist indicates that data scientists are expected to make about $32,405 more than the base case of Business Analyst.
* The coefficient for *Title* - Hardware Engineer indicates that hardware engineers are expected to make about $18,253 more than the base case of Business Analyst.
* The coefficient for *Title* - Marketing indicates that marketing titles are expected to make about $12,965 more than the base case of Business Analyst.
* The coefficient for *Title* - Mechanical Engineer indicates that mechanical engineers are expected to make about $10,175 more than the base case of Business Analyst.
* The coefficient for *Title* - Product Designer indicates that product designers are expected to make about $27,276 more than the base case of Business Analyst.
* The coefficient for *Title* - Product Manager indicates that product managers are expected to make about $32,448 more than the base case of Business Analyst.
* The coefficient for *Title* - Software Engineer indicates that software engineers are expected to make about $30,370 more than the base case of Business Analyst.
* The coefficient for *Title* - Software Engineering Manager indicates that software engineering managers are expected to make about $44,893 more than the base case of Business Analyst. This means that software engineering managers are paid more than any other title evaluated by the model.
* The coefficient for *Title -* Technical Program Manager indicates that technical program managers are expected to make about $22,670 more than the base case of Business Analyst.
* The coefficient for *Title* - Other indicates that titles not explicitly outlined here are expected to make about $8,177 more than the base case of Business Analyst.
* The absence of a coefficient for Business Analyst indicates that this is the base case and because all *Title* categories have positive coefficients, Business Analyst is the lowest paid *Title* in the dataset.
* The coefficient for *Education* indicates that those with at least a Master's Degree are expected to make about $8,867 more than those with less education.

**Discussion**

To begin a review of this analysis, we are going to interpret the results within the context of the research questions outlined in the *Introduction*. Our first question was which variables, of the variables of interest, can influence the salary of an individual with a technology centered career in the United States. Throughout our analysis we refined the list of variables of interest and ultimately ended with ‘Years of Experience’, ‘Revenue per Employee’, ‘Number of Employees’, *Title*, and *Education* being the variables used for the final model. As stated earlier, we believe that there is potential for expansion of variables like *Sector* in the future. Further data collection, or extensive sampling of the dataset, to balance out distributions of categorical variables to avoid aggregation of categories, could lead to variables like *Market Cap Group* and *Education* to have higher significance when building a model.

Our initial hypothesis was that ‘Years of Experience’, ‘Number of Employees’, and ‘Revenue per Employee’ were going to be the most influential factors. ‘Years of Experience’ proved to be the most influential quantitative variable, however ‘Number of Employees’ and ‘Revenue per Employee’ did not have much influence in the multivariate linear regression model. Even more surprisingly, ‘Number of Employees’ had a negative relationship with ‘Base Salary’ illustrating that with more employees a company employs, salaries tend to decrease.

Our second research question focused on the strength of the relationship between the independent variables and ‘Base Salary’. With ‘Years of Experience’ being the most influential quantitative variables, we will discuss the impact of the categorical variables. The final model had *Education* and *Title* as the remaining categorical variables. It was unexpected to see such a vast range in coefficients of the *Title* category. With the base case of Business Analyst being the lowest paid *Title*, we saw four categories with coefficients representing a pay gap of over $30,000. These figures depending on the number of ‘Years of Experience’ could be the most influential factor in predicting ‘Base Salary’. The coefficient for *Education* indicated that those with at least a Master’s Degree made nearly $9,000 more than their counterparts, roughly equivalent to the same as three years of experience. This did not meet expectations, but we expect with more data, or intense sampling methods, we could deaggregate the category and obtain more illustrative results of the relationship with a wider range of categories. Depending on the results, this could potentially increase the demand for higher level education.

For users, our analysis can provide a roadmap to increasing or obtaining higher salaries for individuals currently, or potentially, in technology centric positions. For example with the nearly $9,000 increase in salary for individuals with a Master’s Degree or higher, individuals can have a better informed analysis of the cost-benefit relationship of pursuing a career leveraged against entering the industry earlier to increase years of experience. They can also use the information obtained about *Title* to make career decisions aligned with successfully entering a position correlated with a higher salary.

As for our stakeholders, recruiting companies or organizations can use our analysis as a way to connect potential employees and employers more effectively, based on metrics of position requirements and individual’s qualifications. They can also use it to generate articles and posts to drive traffic to the company. Companies, in general, can use the analysis to better attract the best candidates for the requirements of the position based on metrics desired with competitive salary ranges.

Consequences could lead to fewer individuals pursuing higher education because the approximate $9000 salary increase, gained by individual’s with a Master’s Degree or higher, may be interpreted as being less valuable than other factors like years of experience. This could produce less demand for educational programs, less revenue for these programs, and therefore fewer opportunities. This would also lower the average quality of employees in these industries. Another potential consequence is the avoidance of applying to large companies due to the negative relationship between ‘Number of Employees’ and ‘Base Salary’. These companies employ a massive amount of individuals with tech centered careers, amassing to many employment opportunities. This would also cause competition to increase for positions in smaller companies, in turn driving down the salaries of these companies. With fewer positions available, this effect would be realized quicker than the supply decrease of individual’s applying to positions in smaller companies’ larger counterparts.

In the future, we would like to create a larger dataset, with specific focuses on gathering information based on current variables like Education Level and Sector for reasons specified throughout this analysis. Another current variable to focus on would be ‘Number of Employees’, there was a large gap separating companies with the largest number of employees from the distribution leading to these entries being dropped by outlier elimination. Filling in this gap to make this information relevant could lead to this variable telling a different, or more accurate, story than currently explored. Other limitations we have could be mitigated through adding more variables of interest to analyze. Potential variables could include variables like ‘Debt Growth Year Over Year’ whose time sensitive nature was a bad fit for our dataset. Backdating the data, to include accurate statistics linking entries to relevant stock market data, would make a more robust model and could explain more of the variance between ‘Base Salary’ and the independent variables analyzed. We explored many other ideas to improve the analysis in the future throughout the various sections of this report.

**References**

Fore, Preston. “Average Tech Salaries Decreased in 2023, According to a New Dice Report. The Company’s CEO Says You Shouldn’t Worry.” *Fortune Education*, Fortune Media IP Limited, 30 Jan. 2024, fortune.com/education/articles/dice-tech-salary-trends-report-2024/.

Naprys, Ernestas. “Remote Work Reducing Salaries, Tech Pros Also Affected.” *Cybernews*, Cybernews, 28 Mar. 2024, cybernews.com/tech/remote-work-reducing-salaries-tech-pros-also-affected/.

Vilorio, Dennis. “Education Matters.” *U.S. Bureau of Labor Statistics*, U.S. Bureau of Labor Statistics, Mar. 2016, www.bls.gov/careeroutlook/2016/data-on-display/education-matters.htm.