Exploring Company Factors and Job Attributes for Salary Determination in the Data Industry

Tam Nguyen

Mercer University

May 2023

## 1. Introduction

The field of analytics in the United States has seen significant growth in recent years. The demand for data professionals has surged across diverse sectors, including technology, finance, healthcare, retail, marketing, and more. Understanding the factors that influence salaries in this industry is of great interest to job seekers, employers, and policymakers. Moreover, analyzing differences in salary based on company type and the presence of specific skills can provide insight into the value of these factors in the job market. This information can be useful for candidates seeking to identify areas of competitive advantage, and for job seekers looking to acquire the skills that are most in demand.

In this project, I will identify the key job characteristics that are most strongly associated with salary estimates for data related jobs. Furthermore, I will determine whether there are any significant differences in the relationship between company type and salary. Lastly, I will explore how the presence of specific skills (such as Python, Excel, or Machine Learning) affect salaries and use linear regression to predict salaries based on the presence of certain skills.

The "Data Related-Jobs in US" dataset from Kaggle will be used to conduct the analyses. By conducting a comprehensive analysis of this dataset, this project aims to provide valuable insights into the factors that influence data analyst salaries and shed light on the dynamics of the job market in this field.

## 2. Descriptive analysis

### *2.1. Dataset*

The data used in this study was sourced from Kaggle, which is broadly known for providing a great variety of datasets. This dataset was collected through web scraping of job postings on popular job websites such as Glassdoor. The data includes over 22,000 job postings for data-related jobs in the United States and was collected between October 2015 and February 2019. As we can see in the detailed account of the data sources table (table 2.1), there are 23 variables and 2084 observations in the data set: particularly, 11 categorical variables and 12 numerical variables in the data. This provided information on various data-related job positions in the United States, including job title, company name, location, job description, salary estimate, and more.

*2.2. Limitations*

One potential limitation of the data is that the data-related job postings are only posted on Glassdoor's platform. Therefore, it would not be representative of the entire job market for data-related roles in the United States. Additionally, the dataset may suffer from selection bias, as only those companies that choose to advertise their job postings on Glassdoor are included in the dataset. The salary estimates provided in the dataset may also be subject to bias because they are based on self-reported information and may not accurately reflect the true salaries of data-related roles.

*2.3. Understanding the data*

As we can see from the summary statistics table (table 2.2) for numerical variables below, it seems that min and max of each variable: salary estimate, description length, company age has a large range. For instance, the maximum company rating is 5.00, indicating that some companies have very high ratings. Moreover, the standard deviation of salary is large so we can conclude that these variables have large variability. Furthermore, the mean salary is $108,769, which may be considered a good salary range in the US.

*2.4. Exploring data by visualization*

We can see in the histogram (figure 2.1) that if we just use the salaries given to us in the data set, the salary distribution falls around a central range of values. This may suggest that there is a typical salary range for data-related jobs in the US.

The boxplot (figure 2.2) shows that the median and maximum salary of mid-level and senior level are similar, which means there is a salary limit that is being reached for these positions. However, the salary range of senior level is larger than mid-level.

Regarding the bar chart of the relationship between salary and job simplification (figure 2.3), data scientists have the highest salary compared to others. Besides, it also appears that data scientist jobs require the most skill set (figure 2.4).

The scatter plot (figure 2.5) shows a positive trend, indicating that there may be a positive relationship between salary and company rating. As company rating increases, there appears to be a higher average salary for data-related jobs. A significant of salary and company sector (figure 2.6) illustrates that a significantly higher salary can be seen in the Information Technology sector.

## 3. Hypothesis Test

### *3.1. Hypothesis Test*

Suppose we want to test whether the average salary for data analysts in our dataset is significantly different from $75,000. The null hypothesis value of $75,000 was chosen as the value for the population mean in the null hypothesis based on industry standards.

State the null and alternative hypotheses:

(H0): The average salary for data analysts in the dataset is equal to $75,000.

(H1): The average salary for data analysts in the dataset is not equal to $75,000.

t = (sample mean - hypothesized mean) / (sample standard deviation / sqrt(sample size))

t = 44.77

Since |t| = 44.77 > 1.96, we reject the null hypothesis. We conclude that the average salary for data analysts in the dataset is significantly different from $75,000.

*3.2. Confidence Interval*

To compute a 95% confidence interval for the population mean salary, we can use the formula:

CI = sample mean ± (t-value * sample standard deviation / sqrt(sample size))

CI = (107290.23, 110247.05)

The 95% confidence interval for the population mean salary of data analysts in the United States is (107290.23, 110247.05).

**4. Analysis of Variance**

*4.1. The importance of testing*

In my study, I would like to compare the mean salaries of analyst jobs in different locations. It is crucial to test for significant differences between them because it can help us understand the dynamics of the job market in this field. Firstly, it can help job seekers to determine which locations are more likely to find higher salary jobs. This can influence their job search and relocation decisions. For example, “tech” cities like New York, Seattle, and San Francisco often offer decent salaries. Additionally, comparing mean salaries across locations can allow researchers to understand the dynamics of the job market and identify potential areas for improvement. Last but not least, it can provide valuable insights for employers who are looking to attract and retain top talent. If employers in a particular location are offering lower salaries compared to other locations, they may struggle to compete for talent.

*4.2. Hypothesis test*

State the null and alternative hypotheses:

H0: There is no significant difference in mean salaries among the different locations.

H1: There is a significant difference in mean salaries among the different locations.

The output of summary model shows that t-statistic = 6.344 and p-value = 2e-16

Since p-value < 0.05, we can reject the null hypothesis. We can conclude that there are significant differences in mean salaries between different locations.

*4.3. Discussion*

The significant differences in mean salaries across locations suggest that employers need to be aware of the regional differences in salaries when negotiating compensation packages or making hiring decisions. This can affect decisions on where to live or which job offers to accept.

Additionally, the findings may have broader implications for policies aimed at addressing economic inequality and promoting equal pay. By highlighting the disparities in salaries across different locations, policymakers can more effectively target their actions to guarantee that individuals in all regions have access to fair compensation for their skills and contributions. Moreover, the results may suggest that there are structural differences in the demand and supply of data-related jobs across locations, which could inform decisions on where to invest in education and training programs to meet the needs of the labor market.

## 5. Regression Analysis

*5.1. Research Question*

- What are the key job characteristics (rating company, company age, description length) that are most strongly associated with salary estimates for data related jobs?
- What is the relationship between company type and salary?
- How do specific skills (such as Python, Excel, or Machine Learning) affect salaries?
- Can we use linear regression to predict salaries based on the presence of certain skills?

These questions are important in the context of the field of study because they provide insights into the factors that influence salaries in the data industry. This information can be useful for job seekers, employers, and policymakers in making informed decisions about job offers. Additionally, understanding the relationship between job characteristics and salaries can help identify potential areas for career growth and advancement within the industry. Finally, being able to predict salaries based on specific skills and job characteristics can provide a valuable tool for employers in making hiring and compensation decisions. Besides, it also helps job seekers figure out what skills are required the most to prepare for their career.

*5.2. Dependent Variable*

The dependent variable in this study is the salary of job-related data in the United States. It is of interest to understand its variation because it can provide insights into the factors that influence compensation in the field of data professionals and can help employers and job seekers make informed decisions about salary negotiations.

*5.3. Factors influencing dependent variable*

- Rating company: Higher-rated companies may be more likely to offer higher salaries to attract and retain top talent, so it could be a predictor of higher salaries.

- Description length: Jobs with longer descriptions may require more skills or experience, or may be more complex, and therefore could be associated with higher salaries.

- Company age: Older companies may have more established reputations and financial stability, and may be more likely to offer higher salaries compared to newer companies.

- Company type: Different types of companies may have different salary structures and compensation policies, and this could result in variations in salaries for data-related jobs.

- Specific skills: The presence of specific skills such as Python, Excel, or Machine Learning may impact a candidate's salary. These skills are in high demand and may command higher salaries.

*5.4. Regression Analysis*

*5.4.1. Model 1*

Using regression model to determine which job characteristics are most strongly associated with salary and explore the difference between salary and company type. As we can see table 5.1, p-value of rating < 0.05 indicates that the rating company and description length are statistically significant and have a significant effect on the salary. For the rating company variable, a coefficient of 10700 indicates that on average, each additional rating is associated with an increase of 10700 units on the salary, given other variables constant. This means that working for a company with a higher rating is associated with a higher salary. Meanwhile, the coefficient for company age is -1.02, this means that on average, for each unit increase in rating, the salary is 1.02 units lower holding other variables constant.

For the company type variable, the p-value of College/University, Public Company, Private Company, Government, Hospital, Private Practice / Firm, Subsidiary or Business Segment are less than 0.05, it suggests that there is a statistically significant relationship between these types of company and salary. Noticeably, the coefficient estimates on Private Practice / Firm states that, on average, salaries on Private Practice / Firm are about 83134 USD higher than in the College/University company types. However, we can see that the regression coefficient for Franchise is -3861.3. This means that, on average, salaries on Franchise are 3861.3 units lower compared to salaries on College/University. Similarly, salaries on remaining company types are higher than salaries on College/University company types.

Model evaluation: The R-squared for this model is 0.12, which is not a good fit for the data. In our case, the standard error of the estimate is 31472.86. This means that, on average, the predicted salaries from the model will be within $31472.86 of the actual salaries. With a low R2 and large standard error of the estimate, we can conclude that this model is not strong. To strengthen the model and improve its accuracy, additional variables or data that could be considered is company sector, which salaries can vary significantly depending on the industry. For example, salaries in the tech industry tend to be higher than salaries in the

non-profit sector. Another variable is that company size because larger companies may have more resources to offer higher salaries, while smaller companies may have more flexibility in terms of non-monetary benefits. Furthermore, location is one of the additional variables that should be examined as salaries can vary significantly depending on the cost of living and the job market in different regions.

### *5.4.2. Model 2*

Using regression model to explore how the presence of specific skills) affect salaries use linear regression to predict salaries based on the presence of certain skills. Based on table 5.2, a job requiring Python, Spark, Azure, AWS, and Machine Learning skills would have a salary equal to the intercept increased by $18643, $9335, $1122, $1980, $2977 respectively compared to jobs not requiring these skills. The regression analysis revealed that Python has a significant impact on salaries for data-related jobs, highlighting the importance of acquiring these skills for job seekers in this field. However, a job requiring Excel skill would have a lower salary by 7940 than a job not requiring, which means that a job listing that requires knowledge of Excel is associated with a lower salary.

Using a linear model to predict the salary based on the presence of a certain combination of skills, an individual has strong skills in Python, Spark, and Machine Learning, the model predicts that their expected salary will be 125714.4 units higher compared to an individual with no skills in those areas.

## *5.5. Discussion*

One surprising result is the lack of significance of the company age variable, as its coefficient is negative and not statistically significant. This suggests that the age of the company may not be a strong predictor of salary for this dataset. Another surprising result is that the coefficient for the description length variable is positive. Some people think that shorter job descriptions are typically connected with greater salaries because they may signal a more specialized position, but longer job descriptions might give more thorough information about the position and its responsibilities, which might result in higher salaries. Besides, the coefficient for the "Franchise" in the company type variable is negative and not statistically significant.

This means that being a franchise may not have a significant impact on the salary of the data jobs, compared to being employed by a college/university company type. This result may be surprising because franchises are often linked to profitable businesses. However, it is possible that in the field of data, being a franchise does not offer a salary advantage compared to other company types. In terms of skillset, the coefficient of "excel_yn" variable is negative. This result may be unexpected, especially given that Excel is a widely used software in many industries.

## 6. Conclusion

This work focuses on finding the key job characteristics that influence salaries, identifying the correlation between salary and company type, and predicting the salary based on the presence of certain skill sets. The result shows that rating company and description length are most strongly associated with salary estimates for data related jobs, while the effect of company type is mixed. Moreover, the inclusion of programming skills such as Python, Spark, Azure, AWS, and Excel, has shown that having expertise in these skills can significantly affect job salaries. The salary prediction model illustrates that the presence of some skill set can remarkably increase job salaries. Overall, these results provide valuable insights into the job market for data-related jobs in the United States. The findings of this project can inform discussions about the importance of acquiring certain skills for job seekers in this field, and the factors that influence salaries in this industry.

However, there are several limitations in this project. One is that the dataset used was the relationship between a limited set of variables and salary, and there may be other important factors that were not considered. Second, the data was self-reported by individuals, which may introduce inaccuracies. Therefore, in further analysis, it would be essential to examine the impact of additional factors on salaries in the data-related field, such as education level, years of experience. The research question may be "How do salaries for data professionals compare to salaries for professionals in other fields with similar education and experience levels?". It is useful for job seekers to understand the competitiveness of the data field deeply.

# Appendix

Table 2.1. The detailed account of dataset

| Variable | Description |
|---|---|
| Company name | The name of company |
| Job Title | Job title |
| Location | Location in the US |
| Job description | Job description |
| Salary Estimate | Salary estimate |
| Company Size | The size of company |
| Company Type | Company type (Private, Public, Franchise, etc.) |
| Company sector | Company sector (Healthcare, Financial Services, Education, etc.) |
| Company industry | Company industry (Accounting, HR, Construction, etc.) |
| Company founded | The year of company founded |
| Company Revenue | The revenue of company |
| Hourly | Whether they pay salary by hour or not |
| Rating | Rating of company |
| python_yn, spark_yn, azure_yn, aws_yn, excel_yn, machine_learning_yn | Whether these skills are required or not |
| Job_simple | Types of job, including data scientist, data analyst, data engineer, machine learning engineer or other |

| | |
|---|---|
| Seniority | Level of job: junior, mid, senior |
| Description length | The length of description job |
| Company age | Age of Company |

Table 2.2. Summary Statistics of all numerical variables

| Variable | Min | Max | Mean | Std. Dev. |
|---|---|---|---|---|
| Salary Estimate | 3760 | 297000 | 108769 | 34435 |
| Company Rating | 1 | 5 | 4.1 | 0.49 |
| Description Length | 26 | 2781 | 409 | 222 |
| Company Age | 1 | 397 | 57 | 63 |

Figure 2.1. The distribution of salary

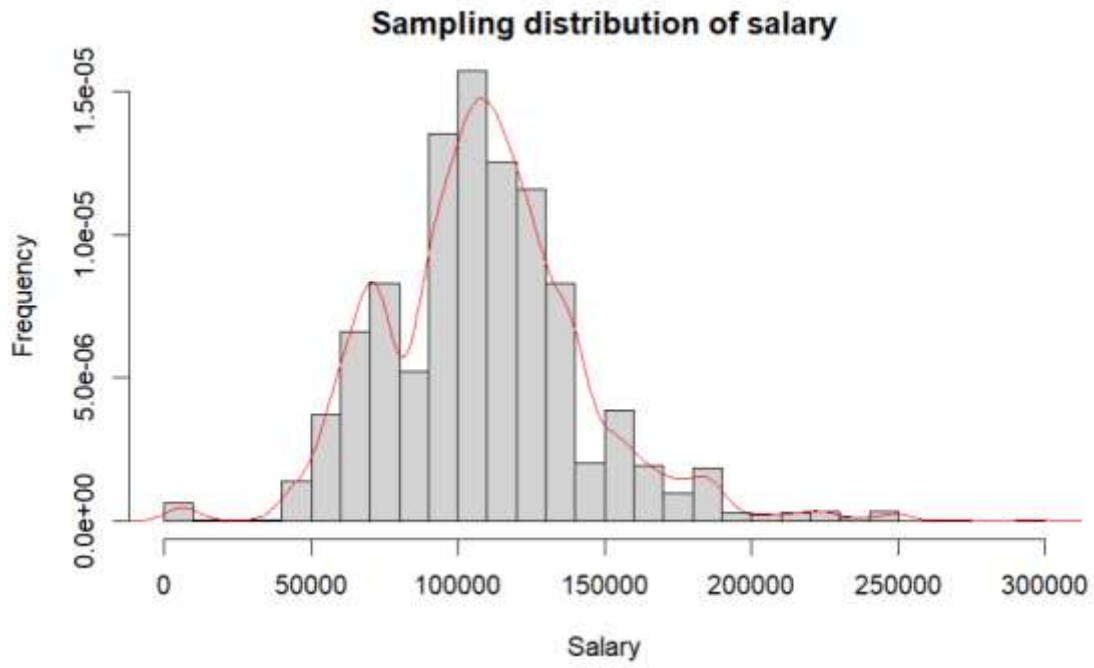

Figure 2.2. Salary distribution based on level

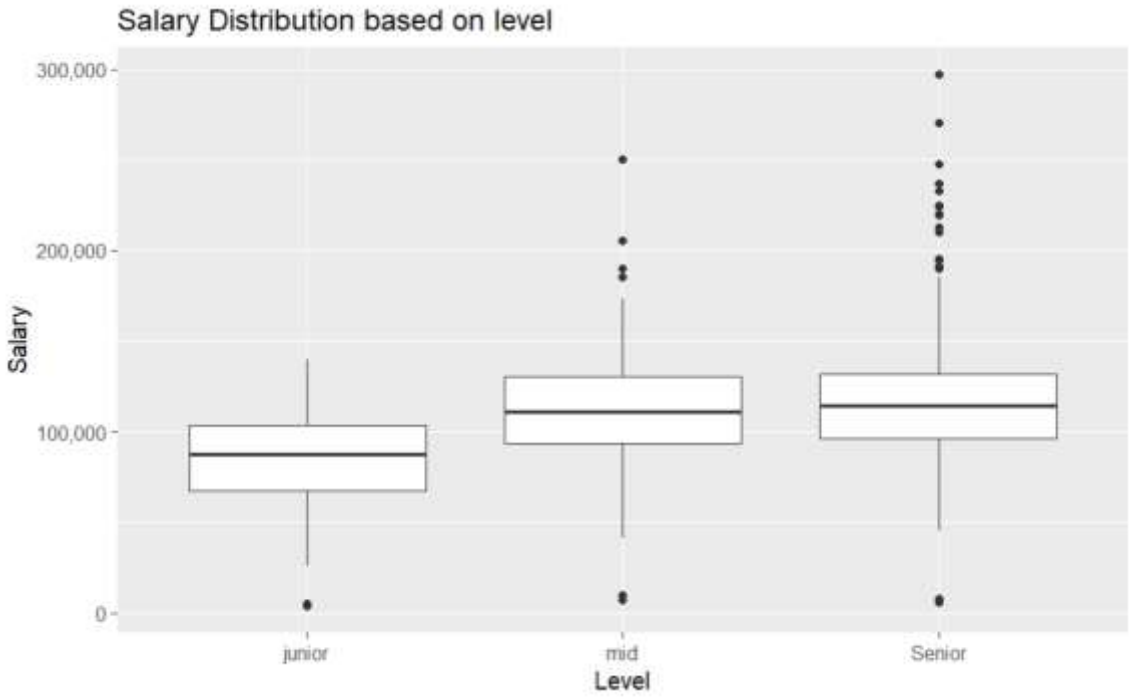


Figure 2.3. Relationship between salary and job simplification

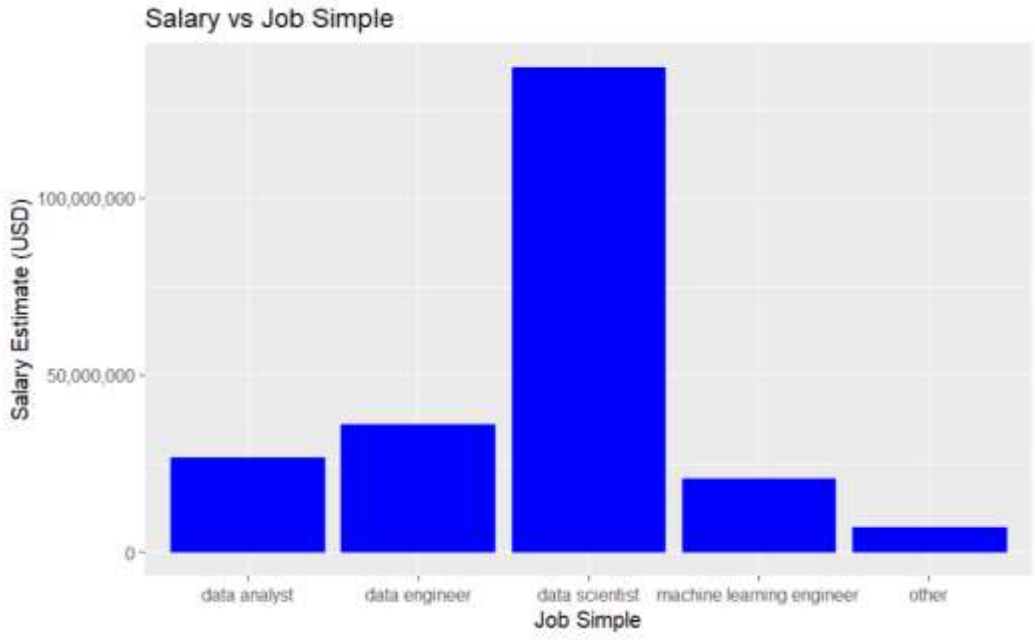

Figure 2.4. Skill distribution by job simplification

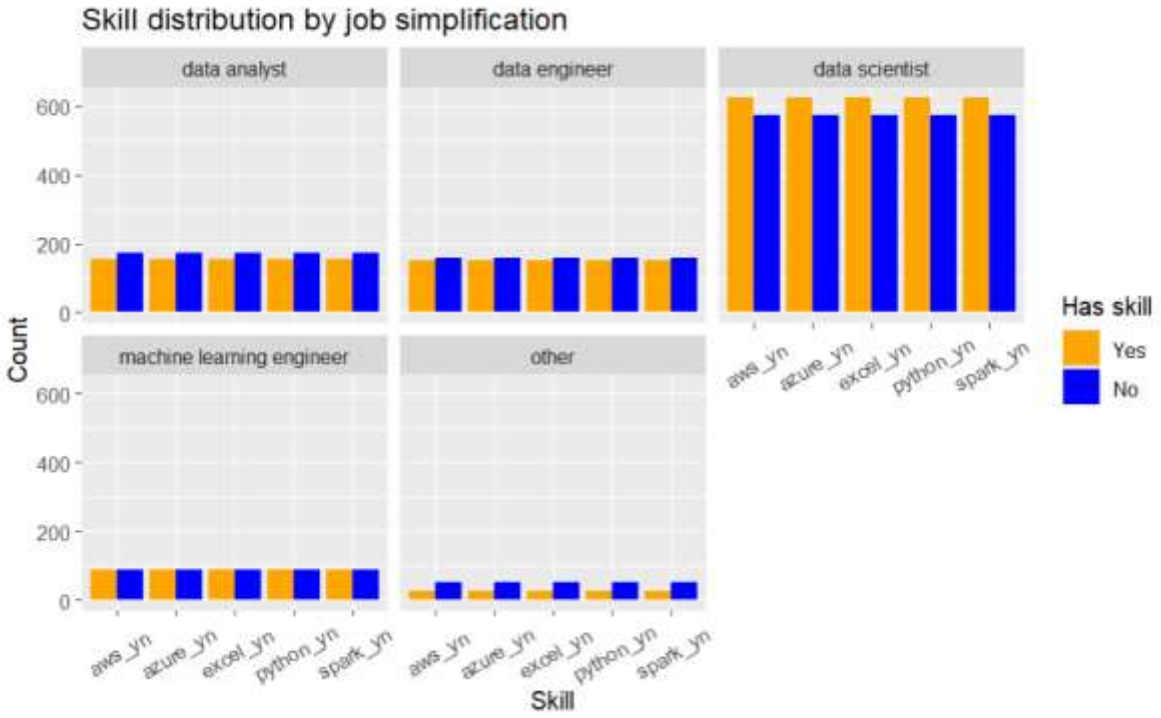

Figure 2.5. Relationship between rating company and salary

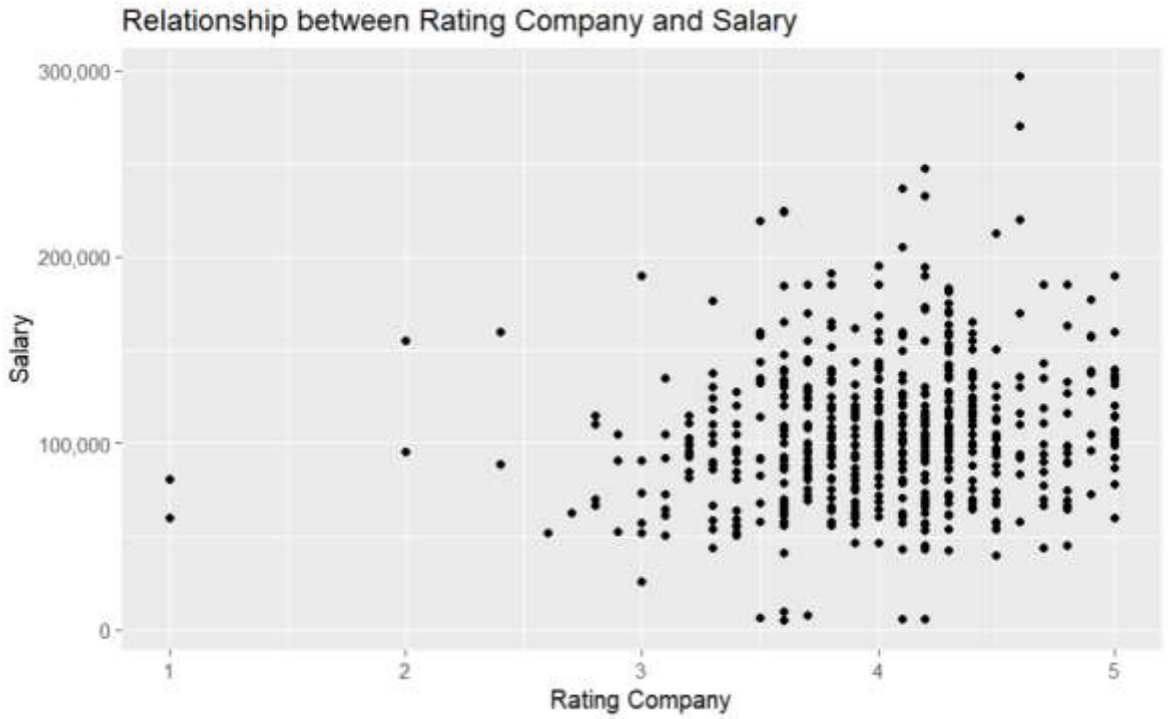


Figure 2.6. Relationship between salary and company sector

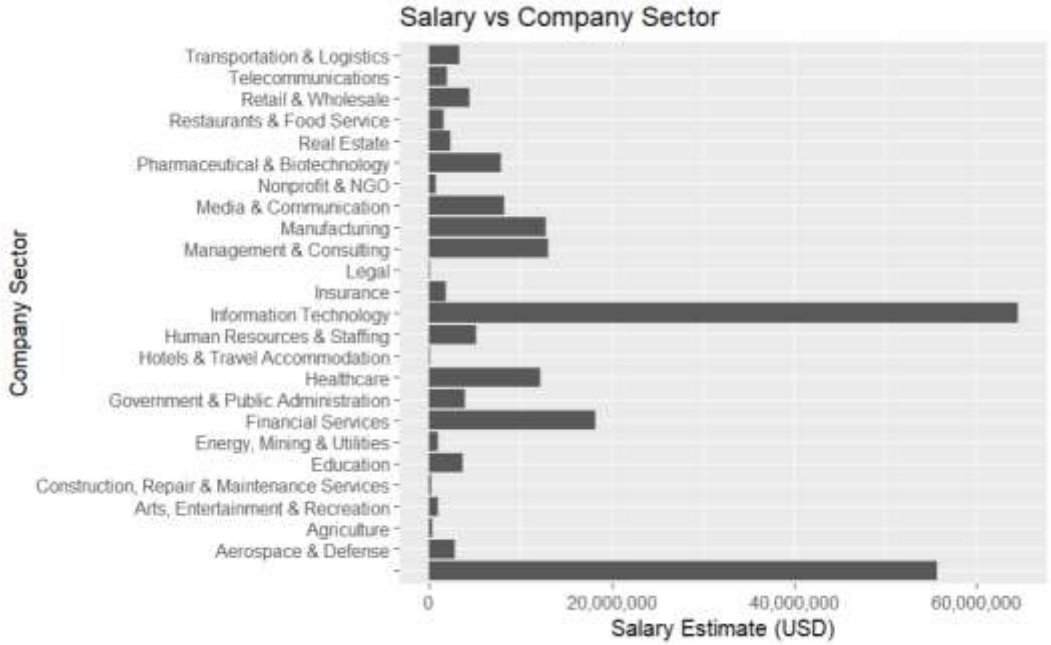

Table 5.1. The coefficients of model 1

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 19892.190 | 10755.319 | 1.850 | 0.06460 |
| rating | 10700.295 | 2098.184 | 5.100 | 3.88e-07 *** |
| description_len | 17.362 | 4.639 | 3.743 | 0.00019 *** |
| company_age | -1.022 | 14.922 | -0.068 | 0.94542 |
| company_typeCompany - Private | 39831.808 | 5339.613 | 7.460 | 1.54e-13 *** |
| company_typeCompany - Public | 46098.916 | 5192.905 | 8.877 | < 2e-16 *** |
| company_typeFranchise | -3861.273 | 13014.033 | -0.297 | 0.76674 |
| company_typeGovernment | 20275.706 | 6973.930 | 2.907 | 0.00370 ** |
| company_typeHospital | 33022.140 | 7867.528 | 4.197 | 2.88e-05 *** |
| company_typeNonprofit Organization | 13872.053 | 6727.728 | 2.062 | 0.03940 * |
| company_typePrivate Practice / Firm | 83133.732 | 31941.390 | 2.603 | 0.00935 ** |
| company_typeSelf-employed | 5131.020 | 31910.612 | 0.161 | 0.87228 |
| company_typeSubsidiary or Business Segment | 53156.359 | 6117.173 | 8.690 | < 2e-16 *** |

Table 5.2. The coefficients of model 2

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 94760 | 1519 | 62.385 | < 2e-16 *** |
| python_yn | 18643 | 1742 | 10.705 | < 2e-16 *** |
| spark_yn | 9335 | 2343 | 3.985 | 6.99e-05 *** |
| azure_yn | 1122 | 2220 | 0.505 | 0.6135 |
| aws_yn | 1980 | 1722 | 1.150 | 0.2502 |
| excel_yn | -7940 | 1511 | -5.255 | 1.63e-07 *** |
| machine_learning_yn | 2977 | 1565 | 1.902 | 0.0573 . |

# References


Bruce, P., Bruce, A., & Gedeck, P. (2021). *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. (2nd ed.). O'Reilly Media, Inc, 87 – 194.

Martín, I., Mariello, A., Battiti, R., & Hernández, J. A. (2018). Salary prediction in the IT job market with few high-dimensional samples: A Spanish case study. *International Journal of Computational Intelligence Systems, 5.* Retrieved from https://www.researchgate.net/publication/327080220_Salary_Prediction_in_the_IT_Job_Market_with_Few_High-Dimensional_Samples_A_Spanish_Case_Study.

Papadaki, I., & Tsagris, M. (2019). Estimating NBA players' salary share according to their performance on court: A machine learning approach. Journal of Business Research, 100, 281-292. Retrieved from https://www.researchgate.net/publication/343304438_Are_NBA_players_getting_paid_according_to_their_performance_on_court