Regression Analysis of Glassdoor Data Science Salaries

Matthew Chen, Jasper Tsai, Mark Faynboym

# Abstract

Write a bunch of stuff here. Wow cool results.

# Introduction

According to the U.S. Bureau of Labor Statistics, the data science industry is projected to grow 36% from 2021 to 2031, which is much higher than the national average [1]. Many organizations are increasingly benefiting from data science and statistical knowledge as the availability of data grows exponentially. As such, we would like to explore salaries relating to this in-demand industry to better understand what factors may influence salary the most. This may give insight in understanding the current state of the industry.

Some of the research questions that we hope to answer are:

* What factors affect salaries of data related industries the most?
* What skills or education level do the highest paid data scientists have?
* Are there significant differences in the location or size of the company?
* How much do data science salaries vary naturally?

To answer this question, we use data sourced from Kaggle.com [2] (last updated in 2021) which originally scraped data-related job postings from Glassdoor.com. This data includes information, in no particular order, about the average salary, the company size, employee ratings (from a scale of 0-5), age of the company, the seniority of the role, the degree requirements, and the location of the role. We perform a thorough regression analysis on this data to answer our research questions.

# Methods and Results

## Data processing

In the original raw data, there were columns that were either difficult to interpret, difficult to process, or contained high amounts of missing data, and therefore were dropped from the overall data. There were also duplicate observations that were deleted from the data. Further, there were data entries that were nonsensical, in particular, containing negative ages, negative ratings, and “unknown” locations. These were also deleted from the data to make it more amenable for analysis.

While job location (state) is important, there are 50 potential categories which may be too numerous to use as dummy variables in multiple regression. However, we see that the only state with a significant difference in salary compared to the others is California (Figure 1). Thus, we create a binary dummy variable based on whether or not a job is in California. Similarly, we combine machine learning skills (keras, pytorch, scikit, tensorflow), data visualization skills (Tableau, Power BI), whether or not the position is senior standing or not into single dummy variables to reduce the total number of required classifiers. Table 1 contains a full description of the final variables selected.

## Exploratory Data Analysis

We conduct an exploratory data analysis to help us better understand the data collected and inform future decisions when we fit our models.

For the quantitative variables, we see in the scatterplot matrix (Figure 2) that rating and age are not very correlated with average salary. However, relationships between rating and age are also weak so there is little worry about multicollinearity. Further, we see that the distribution of average salary is slightly right skewed (this may suggest that a transformation is needed later), the distribution of age is approximately symmetric, and that age is heavily right skewed. For average salary, we see that the distribution of the square-root transform is much more symmetric (Figure 3). Also, we see that the median average salary is approximately 100,000 (Figure 4).

For the qualitative variables, that PhD holders earn the highest median salary, followed by MS holders (Figure 5) and that senior positions have higher median pay (Figure 6). Further, the size of the company does not noticeably affect average salaries (Figure 7). In Figure 8, we plot side-by-side boxplots of the different skills and find that Python, machine learning, Spark, AWS, and Hadoop skills have noticeably higher median salaries.

## First Order Multiple Regression

We initially fit a first order model based on all of the available predictor variables.

## Second Order Multiple Regression with Pairwise Interactions

## Logistic Regression Model

In sections 2.3 and 2.4, we see from our multiple regression models that the data cannot be fully explained from our available X-variables which prevents us from making good predictions. This could potentially be because of high error variance, noisy data, or lack of important unknown X-variables. Thus, to fill in the lack of predictability, we train a logistic model to classify whether a salary is above $100k (approximately the median salary) or not to reduce the prediction difficulty.

Logistic regression utilizes a log-linear model that models prediction probabilities given a binary Y. To do this, it fits a sigmoidal function on the observed y, in this case an indicator of whether or not a salary is greater than $100k (Eq. 1). Note that it is possible to transform P(y) in Eq.1 to be linear in coefficients β, which indicates that it is a generalized linear model.

We fit the model based on our available X predictors (except for size, which was not found to be important in our EDA or the stepwise regression analyses). Then, the model is validated using k-fold cross-validation (k=10), and find that the mean training accuracy (0.73) and mean testing accuracy (0.70) is reasonable close which indicates that there is no severe overfitting. Note that our data is approximately balanced since we created our binary y variable using the median as the threshold, so using overall accuracy as a criterion is reasonable.

Training the model on all available data, we obtain an overall accuracy of 74%, which is very decent given the low predictability of our linear models. In Figure X, we see through the trained coefficients that Python and machine learning skills, being in California, and having a senior position is highly rewarded in terms of prediction probability. This is consistent with what we found in our multiple regression analysis.

# Discussion and Conclusion

Here we talk about our cool conclusions

# Appendix A: Figures and Tables

Chart, box and whisker chart

Description automatically generated

Figure . Salary of the top 10 states based on occurrence frequency in the data

Table . Description of selected variables after data cleaning



Chart, scatter chart

Description automatically generated

Figure . Scatterplot matrix of quantitative variables

Chart, histogram

Description automatically generated

Figure . Histogram of square-root transformed average salary.

Chart, box and whisker chart

Description automatically generated

Figure . Boxplot of average salary

Chart, box and whisker chart

Description automatically generated

Figure . Side-by-side boxplots of average salary vs.

Chart, box and whisker chart

Description automatically generated

Figure . Side-by-side boxplots of average salary and senior status

Chart, box and whisker chart

Description automatically generated

Figure . Size of the company vs. average salary

Chart, box and whisker chart

Description automatically generated

Figure . Side-by-side boxplots of different skills vs. average salary

# Appendix B: Code and Notebooks

# References

1. <https://www.bls.gov/ooh/math/data-scientists.htm>
2. Kaggle data