Final Project

IST652 – Scripting for Data Analysis

Toby Anderson

12/09/2022

Table of Contents

[Introduction 2](#_Toc121570340)

[Data Description and Pre-Processing 3](#_Toc121570341)

[Method of Analysis 5](#_Toc121570342)

[Analysis Results 6](#_Toc121570343)

[Company Size 6](#_Toc121570344)

[Employee Location 7](#_Toc121570345)

[Remote Work 8](#_Toc121570346)

[Conclusions 9](#_Toc121570347)

# Introduction

The demand for Data Science jobs is growing worldwide. As Data Scientists continue to provide value through new Machine Learning techniques, companies have been struggling to keep up and leverage this value by finding the best positions for the best candidates. Businesses have been navigating this competitive job market by increasing compensation, defining new roles within their company, and experimenting with remote work.

However, as companies are still determining the most effective roles and responsibilities for these new positions, compensation is highly variable in the industry. A lack of standardization has led to a wide range of salaries, based on level of experience and geographic location. In this study, we will examine the impacts of different factors on the distribution of salaries for Data Science related positions.



Figure 1 – Growth of Data Science job postings from Indeed.com

# Data Description and Pre-Processing

The data was collected from Kaggle (<https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries>). The data was originally pulled from ai-jobs.net and collected together into the data source linked above. A description of the relevant columns is given below in Table 1. Only the columns which were used in this analysis are described.

|  |  |
| --- | --- |
| Column | Description |
| experience\_level | Category representing level of experience for the employee. Ranked from lowest to highest: EN, MI, SE, EX |
| Salary\_in\_usd | Gross yearly amount paid in US dollars. |
| employee\_residence | Country code of employee’s primary country of residence. |
| company\_residence | Country code of companies location. |
| remote\_ratio | The amount of overall work done remotely. (0 to 1) |
| company\_size | Number of employees that work for the company.  S (<50) M (50-250) L (>250) |

Table 1 – Description of data fields

An employee with a degree in Data Science may perform the same duties under similar job titles, such as “Analyst” or “Engineer”. For this reason, the dataset contains a variety of job titles which are traditionally similar to Data Scientist. However, mixed within these jobs are management positions and consultants, which represent outliers. In order to improve the quality of the analysis, only the top 4 positions will be used as representative samples for this study.

Job titles included:

* Data Scientist
* Data Engineer
* Data Analyst
* Machine Learning Engineer

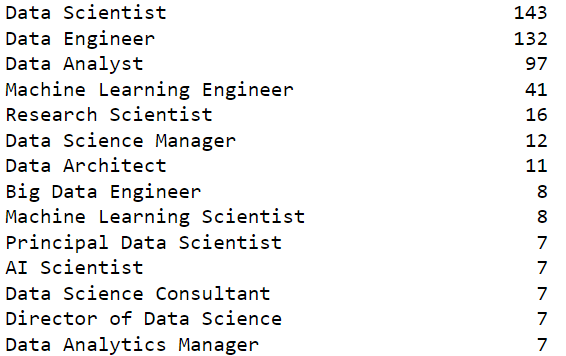


Figure 2 – Sample counts of positions included in the raw data

After filtering for job title, there still appear to be outliers in the salary distribution, see figure 3 below. These outliers were removed by filtering out any salary above $350,000 per year in USD.

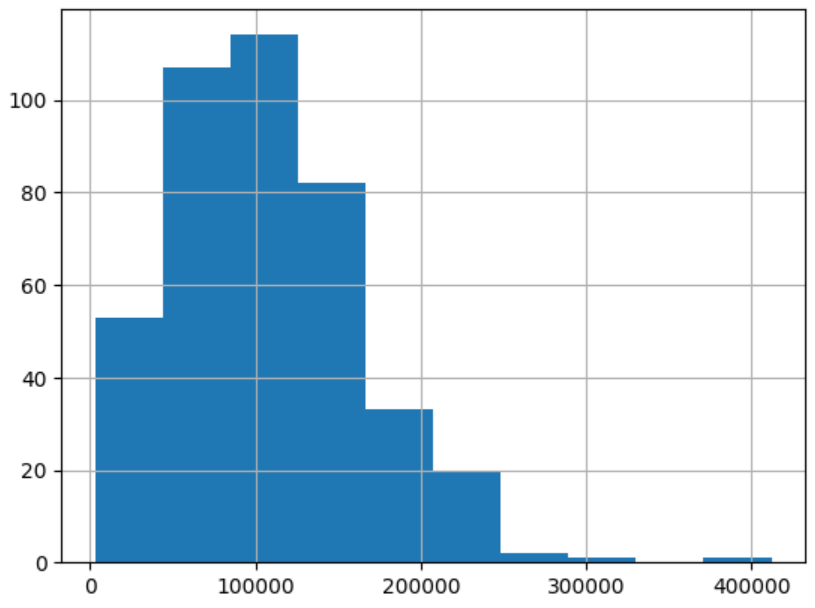


Figure 3 – Salary distribution

After cleaning the data, a single feature was created to enable the full analysis in this study. A flag was created to determine whether a worker was within the same country as the employer, called ‘in\_country’. The majority of workers in the sample are in country, as shown by figure 4 below.

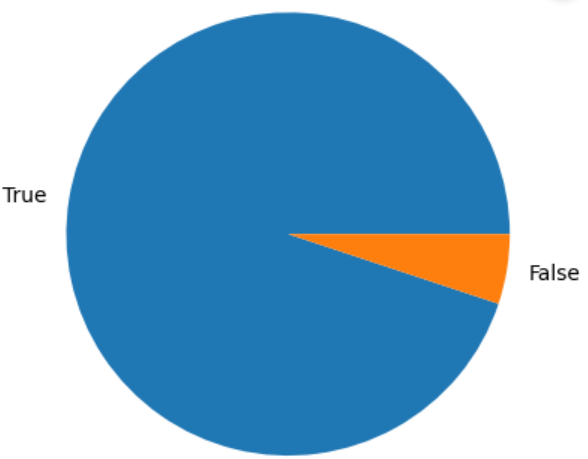


Figure 4 – Portion of employees that work in the same country as the employer

Although this data is pulled from multiple countries, the majority of the sample is from the United States. Average income varies widely by country, so additional consideration will be given to employee residence during the analysis. The top four countries within the sample are the United States, Great Britain, Germany, and India, as shown below.

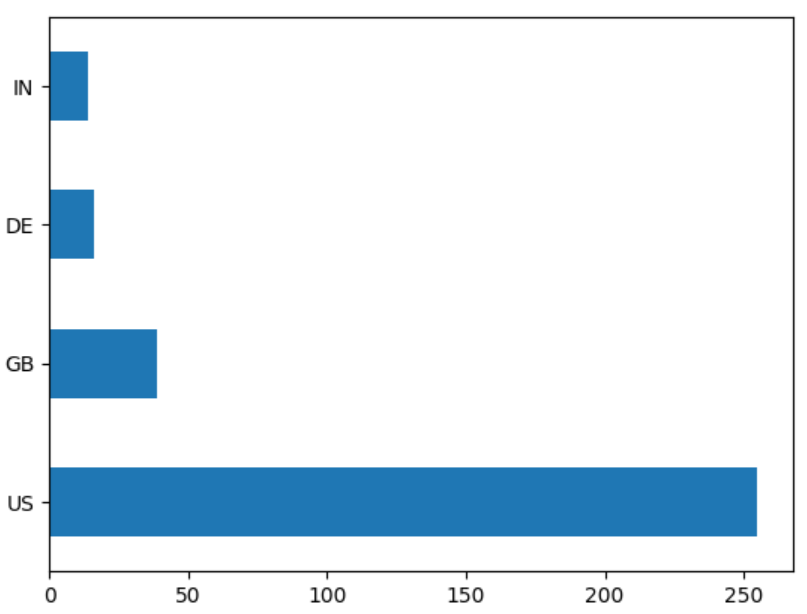


Figure 5 – Distribution of employee country of residence

# Method of Analysis

The data in this report will be analyzed to answer three questions. The questions and analysis approach are detailed in the table below.

|  |  |
| --- | --- |
| Question | Approach |
| Does company size impact salary offered? | Since the location of the company is likely to make a large impact on salary, the data will be stratified by country, and the salary distributions of the top four countries will be compared by company size. |
| Do companies pay more for employees working outside the country? | A feature was created, labeled “in\_country”, to compare the salaries of employees who work outside of the employers country. The intended result will be a list of countries ranked by the increase or decrease in salary paid to external employees compared to in-country employees. |
| Are remote workers more successful in certain countries? | To answer this question, the data will be aggregated by country, and the ratio of the average salaries of remote vs in-person workers will be ranked to show which countries pay remote workers a higher boost. |

Table 2 – Analysis questions and approaches

# Analysis Results

## Company Size

How does company size impact salary distribution? To answer this question, we must stratify the data by company size. The data curators flagged the company size into three bins, small, medium, and large. The data is fairly well distributed, but having the majority of companies in the medium category.

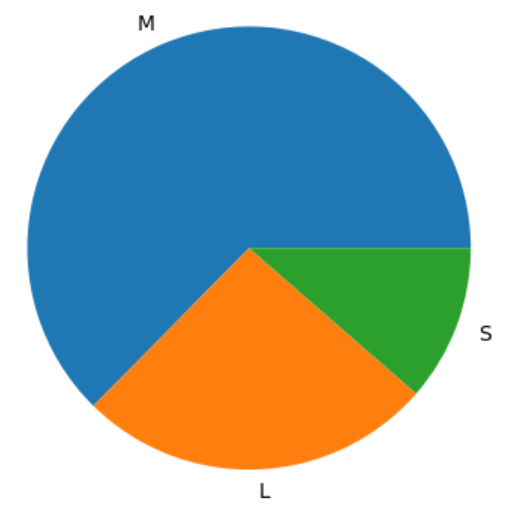


Figure 6 – Distribution of Company Sizes

|  |  |  |  |
| --- | --- | --- | --- |
| Country | Small Company | Medium Company | Large Company |
| United States | $81,496 | $137,813 | $132,677 |
| Great Britain | $73,271 | $80,162 | $78,235 |
| Germany | $55,631 | $58,950 | $77,256 |
| India | $13,110 | $25,571 | $37,147 |

Table 3 – Average salaries by company size and country

As shown in table 3, in every country, medium and large companies pay more than smaller companies. The difference is strongest in the United States and India. Additionally, it appears that there is little difference in pay between medium and large companies.

This might imply that smaller companies have more difficulty leveraging Data Science talent, and that a company needs to reach a certain critical size before it has enough technical infrastructure and data to properly utilize Machine Learning techniques. However, once companies reach a certain size, they are able to utilize Data Scientists just as well as large companies.

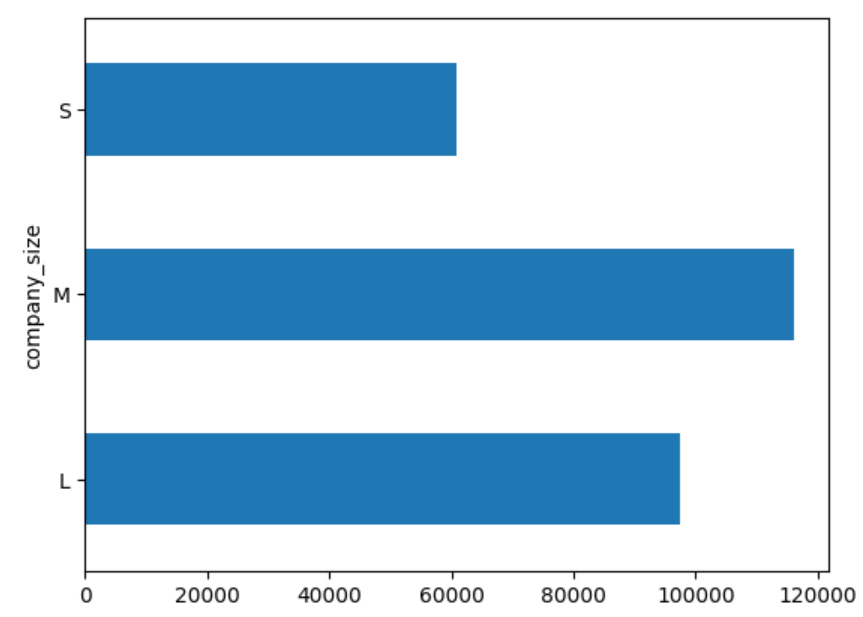


Figure 7 – Overall average salaries by company size.

## Employee Location

Do companies pay more to employees who live outside the country?

This question turned out to be more complex than anticipated. It appears that the employees’ country of residence has a larger impact on salary than the company’s location. Additionally, there are so few remote workers outside of the country, that the sample size is prohibitively small in this dataset.

To isolate the impact of external employment, each example was compared against the average income for an in-country employment with the same level of experience.

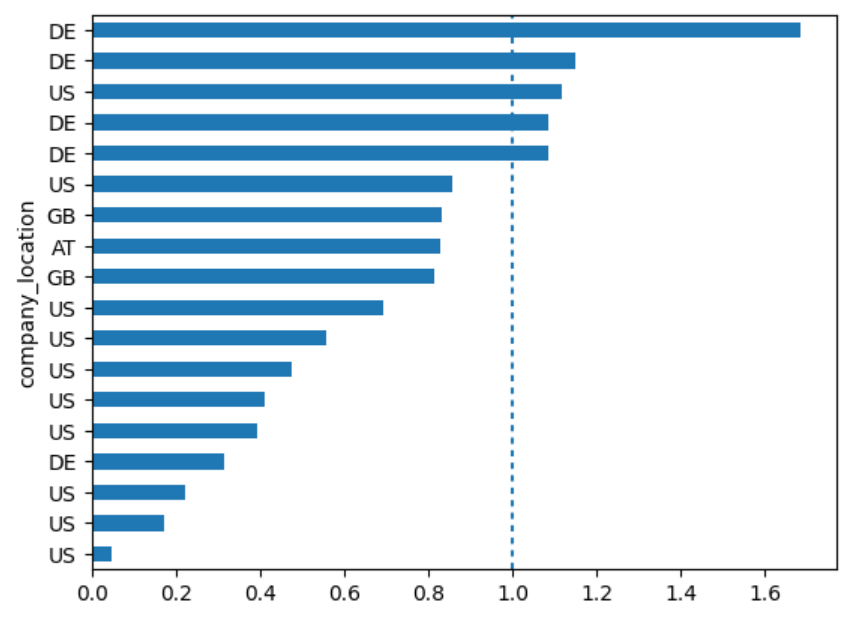


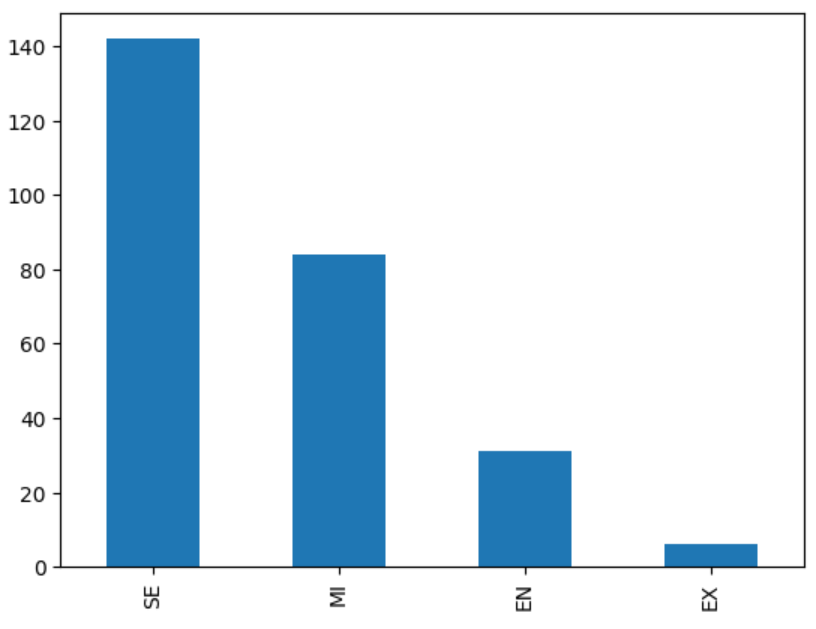
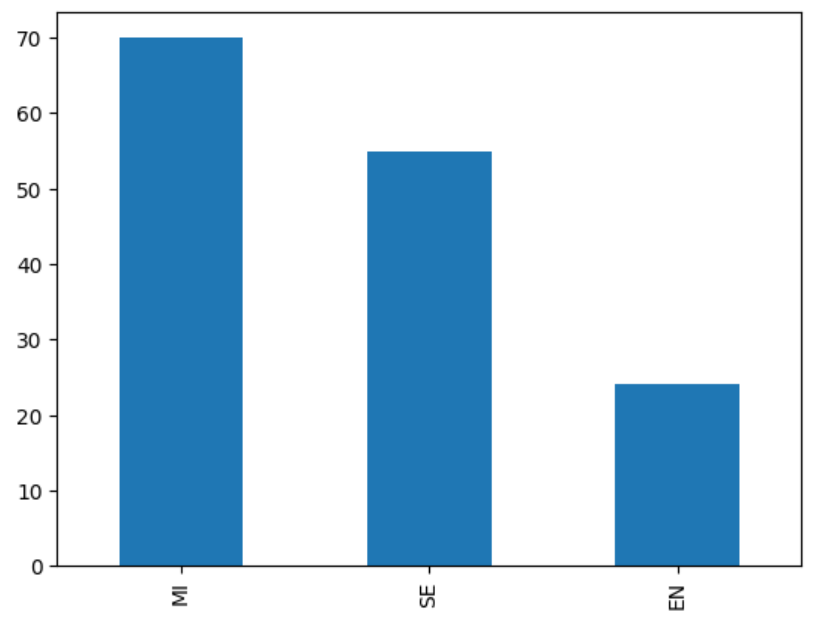
Figure 8 – Pay ratios for external employees

From figure 8, we can see that German companies pay external employees higher than the average income they could achieve working in a local position. However, in the United States, quite the opposite is true. Nearly all external positions paid lower than local positions. This result is surprising, since the US pays local employees generously compared to other companies, and would seem to have more funding to pay higher wages to external employees compared to their local wages.

## Remote Work

Are remote workers more successful in certain countries?

The biggest confounding factor in this analysis is the distribution of experience levels for in-person and remote positions. There is a much higher portion of experienced and senior workers in remote positions, possibly because more skilled employees are capable of performing independently, whereas newer employees require more in-person mentorship to perform effectively.

Figures 9 & 10 – Count of remote (left) and in-person (right) positions by experience level

To control for experience level, remote salaries will be compared across positions with the same experience level. This would potentially provide a list of nearly 50 ratios, for each country and experience level. To provide more accessible summary data, only positions within the US will be compared between in-person and remote salaries.

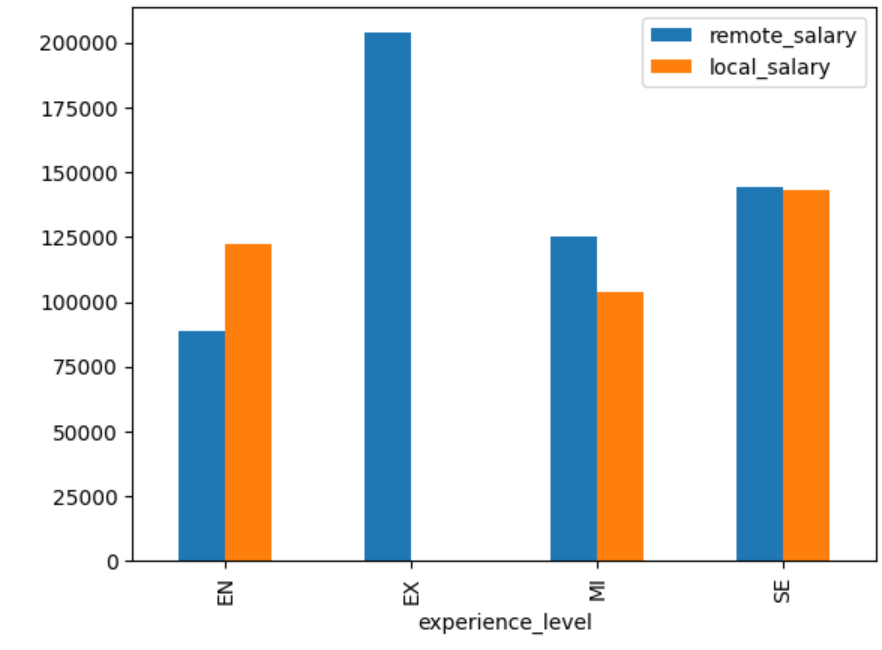


Figure 11 – Remote and Local salaries by experience level

In figure 11, we see the results of the comparison between remote and local salaries. While entry level positions appear to pay significantly higher for local positions, middle and senior level positions have similar pay. The most notable difference between remote and local positions, is that there are no local positions offered for expert level employees in this survey. This implies that companies must look further and wider to find expert level talent that can be difficult to find locally. Additionally, this implies that highly experienced workers need to look beyond local positions to find competitive compensation.

# Conclusions

By analyzing salary data, employees can make informed decisions about their career paths. Employees with information about hiring trends are empowered to find more competitive salaries. In this analysis, we determined whether company size, remote work, or in-country status had impacts on compensation.

New Data Scientists should search for companies with at least 50 employees, as smaller companies are unable to leverage value from advanced data analysis. Data Scientists outside the US should either search for local positions, or move to the US to receive local salary levels; external workers are under-paid in the US. And finally, entry-level employees should find local positions to gain mentorship and provide value, while more experienced workers should seek out remote work, as they are more capable of performing independently and participating in a wider pool of competitive salaries.