Evolution of Data science salaries across countries from 2020-2023 through various companies

Arjun Sahasrabuddhe

Abstract

The analyst wants to examine and better understand which countries pay their employees more, and also gain information on whether the size of companies affects the pay. They also wanted to confirm their notion that salaries in this field have increased as a whole from 2020 to 2023, as well as note on relationships between categorical variables. After data examination, it was found that salaries did increase from 2020 to 2023 besides a drop in 2021, and that the US had the highest mean salary. The size of the company didn't appear to have an effect on employees' salaries and there wasn't a clear relationship between categorical variables.

Introduction

Many factors can play a role in influencing data science salaries, which include experience levels, educational backgrounds, and more. By 2019, postings for data scientists on Indeed had risen by 256%, and the U.S. Bureau of Labor Statistics, predicts data science will see more growth than almost any other field between now and 2029 (Davenport & Patil, 2022). The goal of this analysis was to discover what specific factors, whether it be location or a company characteristic, made a clear and visible impact on one's salary. In order to do this, descriptive statistics, charts, and tables were constructed based on multiple variables that are explained in further detail below.

Data Description

The data that was analyzed is from the researcher, Saurav Banerjee and is titled: Latest Data Science Salaries: Analyzing the evolution of Data Science Salaries from 2020 to 2023. His data was found on kaggle.com and was created from ai-jobs.net, which is a fast and lean job search site in the field of data science. The dataset includes 83% medium sized companies, 13% large companies, 4% small sized companies, and contains 11 variables which are listed in Table 1 below. It includes 3,300 observations, and there are no missing values recorded.

Table 1: Variable Descriptions

Variable Name	Description
Job Title	Title of Occupation
Employment Type	Full-time, Part-time, etc.
Experience Level	Executive, Senior, Mid, Entry Level
Expertise Level	Junior, Intermediate, Expert
Salary	Salary based upon type of currency used by company
Salary Currency	Form of currency used by the company
Company Location	Country of company
Salary in USD	Salary converted to USD
Employee Residence	Employee country of residence
Company size	Small, medium, large
Year	Year of salary given (2020-2023)

Methods and Analysis

Using the WHERE statement, non full-time workers were removed, to better normalize the data by removing potential outliers who don't work normal hours and 39 observations were removed here. The Salary column was dropped as there already exists a Salary in USD column which standardizes the Salary across various countries. The Expertise level column was dropped as this goes hand in hand with Experience Level. When the data was separated creating classes by year, it was noted that most of the observations were from 2023, as 2020 only had 17 observations whereas 2023 had 1,834 observations (see table 2). The salary in USD from 2020 to 2023 increased over the years aside from a decrease from 2020 to 2021 (see figure 1).

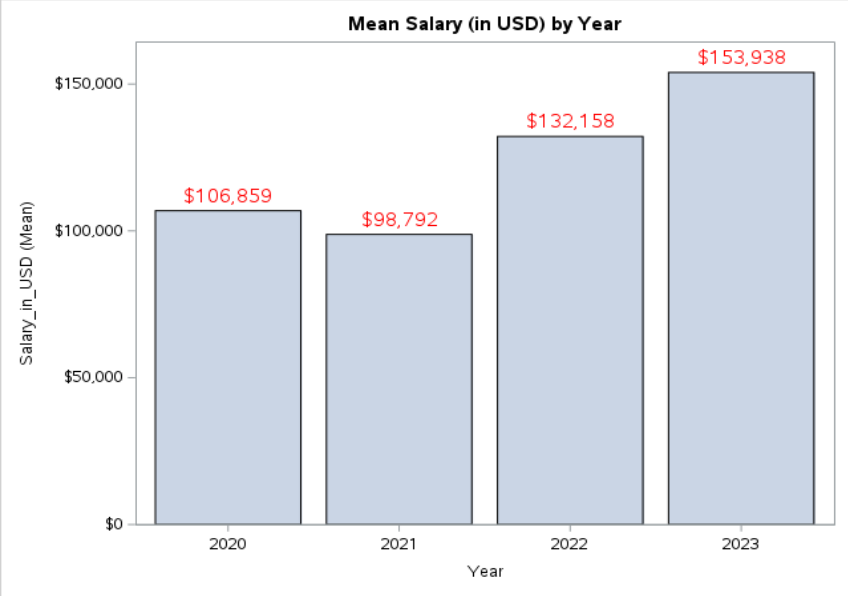
Table 2

Salaries in USD by Year

The MEANS Procedure

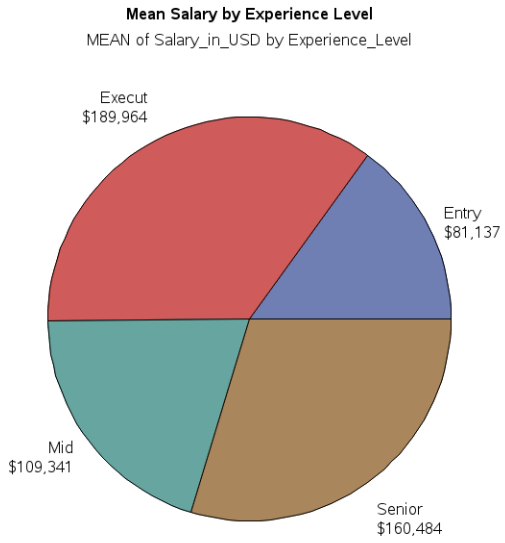
Analysis Variable : Salary_in_USD						
Year	N Obs	N	Mean	Std Dev	Minimum	Maximum
2020	68	68	106859.22	85003.21	15000.00	450000.00
2021	205	205	98791.71	62976.89	15000.00	423000.00
2022	1003	1003	132158.44	62489.60	15000.00	430967.00
2023	1985	1985	153937.90	68335.35	15680.00	430640.00

Figure 1



The data was also sorted by Experience Level and observed as predicted that Entry had the lowest mean salary, followed by Mid, Senior, and Executive (see Figure 2). However, there was a wide range of salaries as the standard deviation was in the $50K-$60K range. The minimum value fell around $15K across all four experience levels, which was much different from the means and stood out as a clear outlier.

Figure 2



When the data was sorted by country and proc means was performed, the US didn't have the highest mean, falling behind Puerto Rico as this was quite a noteworthy observation. However Puerto Rico only consisted of 4 recorded observations in the dataset as opposed to 2,475 from the US. In addition, the minimum for Puerto Rico was significantly higher than the US which will be addressed as well (see Table 3).

Table 3

Salaries By Company Location

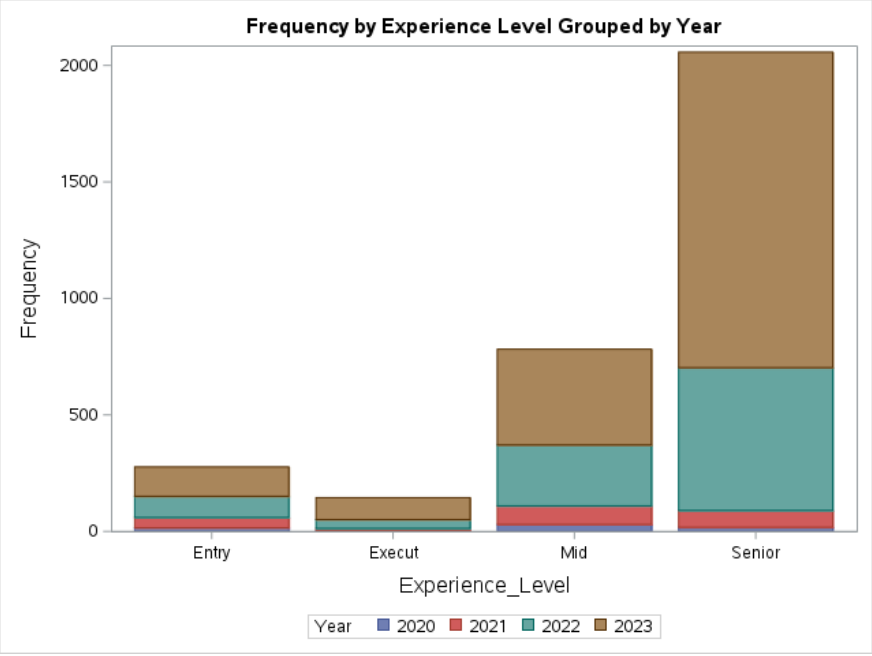
The MEANS Procedure

Analysis Variable : Salary_in_USD

Company_Location	N Obs	N	Mean	Std Dev	Minimum	Maximum
Puerto Rico	4	4	167500.00	37527.77	135000.00	200000.00
United States	2475	2475	158829.36	62514.69	20000.00	450000.00

When studying various frequency tables, it was observed that over 75% of the data came from the US, with the second most frequent data coming from the UK (7.70%). In addition, 67.59% of the jobs in this dataset are held by senior level employees, and the most common subgroup in this dataset were Senior Level Employees in 2023, followed by Senior Employees in 2020 (see Figure 3). The most frequent job titles across countries were found to be Data Engineer (627), Data Scientist (513), Data Analyst (402), Machine Learning Engineer (245), and Analytics Engineer (129).

Figure 3



Discussion

Although there was a decrease in mean salary from 2020 to 2021 (likely due to COVID), and then an increase until 2023, it wouldn't be fair to conclude anything from this due to the shortage of data in 2020 as well as 2021. When studying the data when sorted by experience level, it was expected that the means would increase more exponentially as employees gained experience but senior and executive titles were paid a similar amount, whereas entry and mid level were also paid similar. It's possible that these titles can be used interchangeably depending on the company (especially when comparing various sizes of companies). A likely reason that the US didn't have the highest mean salary when compared to Puerto Rico was due to insufficient data from those countries as depicted above. Also, it is probable that the sample data for Puerto Rico was extracted from a wealthy small pocket, as the minimum was at $135,000. A reason for a high salary in the US can be due to major tech hubs in the country, as the sought-after job is generally paid quite well; the median salary for an experienced data scientist in California is approaching $200,000 (Davenport & Patil, 2022).

Conclusion

It wasn't completely fair to compare certain countries with another due to some countries only having a couple observations, and the data not being evenly distributed from 2020 to 2023. However, it was clear that the US had the highest mean salary aside from a couple countries with mean salaries that weren't representative of the population. The data gives evidence that mean salaries across all countries have increased over the 21st century. In future studies, the salaries of specific job titles can be studied in more detail but as for the scope of this report, there were too many job titles to consider, and it wasn't the goal of this analysis.

Citations

1. Banerjee, Sourav. “Latest Data Science Salaries.” Kaggle, 22 July 2023, www.kaggle.com/datasets/iamsouravbanerjee/data-science-salaries-2023.

2. Davenport, Thomas, and DJ Patil. “Is Data Scientist Still the Sexiest Job of the 21st Century?” *Harvard Business Review*, 21 July 2022, hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century.