

Does the Level of Data Engineer Impact Salary?

Akshaj Kammari

Data 101 HW 7 Data Exploration

https://www.kaggle.com/datasets/milanvaddoriya/data-science-job-salary

Overview

Hey! I’m Akshaj. I’m currently a junior here at Rutgers and I’m double majoring in Computer Science and Data Science. The reason I chose to analyze this dataset is because I believe that Analyzing data from previous data science job positions is crucial for data science students as it provides invaluable insights into industry trends and expectations. Questions such as the most common job titles, salary distribution based on experience, and emerging roles help students glean actionable information for career planning and development. Ultimately, this data-driven approach empowers students to make informed decisions, ensuring they are well-prepared for the dynamic landscape of the data science job market, as they are nearing joining the workforce whether it be internships or full time jobs.

The Dataset

This dataset has been uploaded to Kaggle, sourced from ai-jobs.net, which is a job board for AI, ML, data science and big data jobs. By examining columns such as job type, job title, experience level, location, and salary in this dataset, I will be able to discern patterns in demand for specific skills, salary benchmarks for different roles and locations, and the evolution of job titles over time. This analysis aids in aligning skill development with industry needs, setting realistic salary expectations, and understanding the geographical distribution of opportunities.

Hypothesis

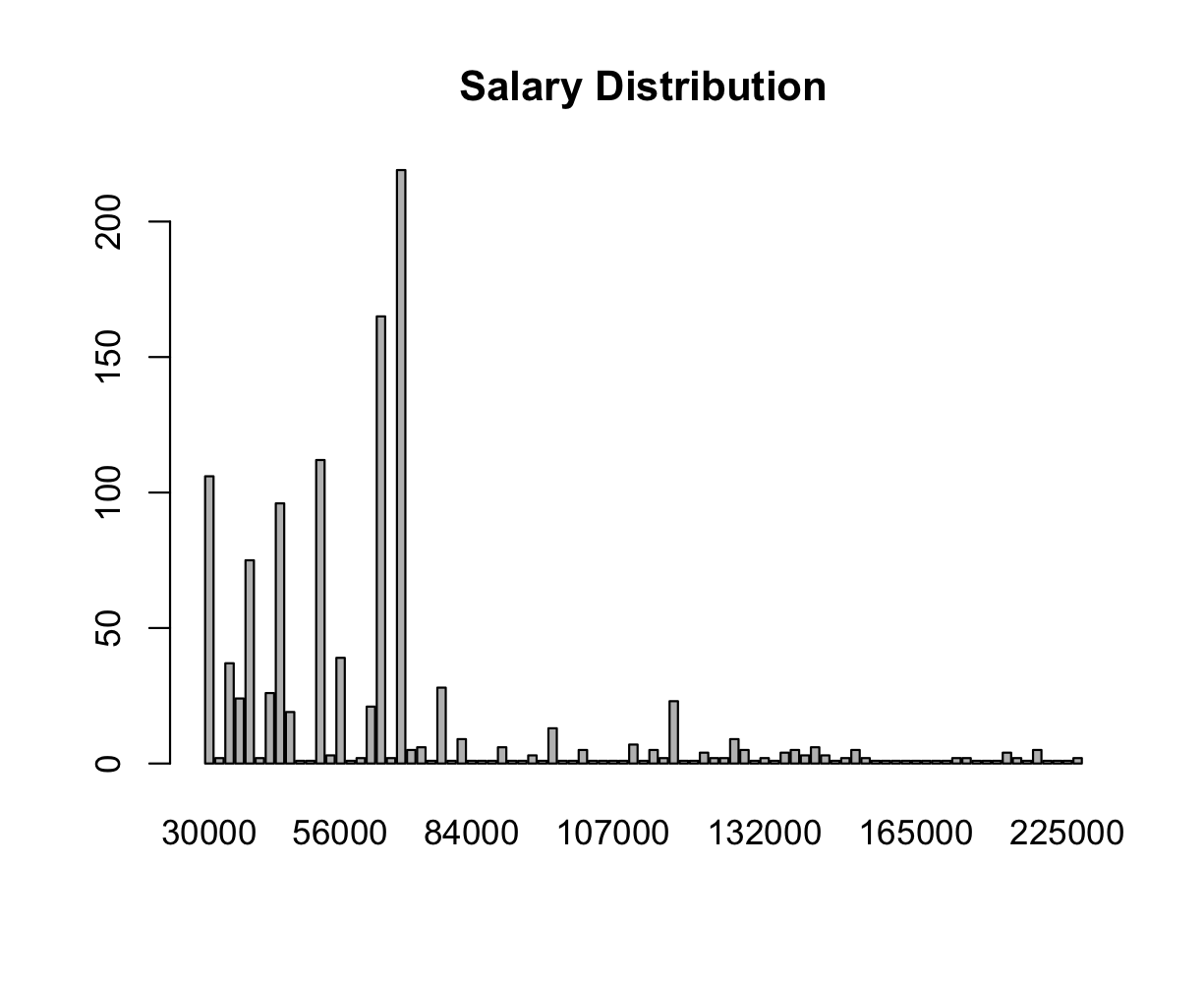
Salary is a critical factor for data science professionals, influencing career choices and financial well-being. In this blog post, we will explore the relationship between experience levels (entry-level and mid-level) and salaries within the data science field. By analyzing a dataset with columns specifying job type, experience level, and salary, we aim to uncover insights into how salaries vary across different experience levels, providing valuable information for both aspiring and seasoned data scientists as they navigate their career trajectories.

Box Plot



This box plot reveals a discernible trend in which mean salaries exhibit a notable linear increase corresponding to higher experience levels. Additionally, the presence of numerous outliers in the dataset underscores the variability and dispersion within the salary distribution across different experience levels.

Bar Plot



This bar plot reveals prominent peaks in the salary distribution, concentrated within the $30,000 to $80,000 range, aligning with the central tendency of mean salaries. Notably, the plot also illustrates outliers surpassing the $100,000 threshold, indicative of instances where salaries deviate significantly from the mean.

P-Test

To further analyze the test prep and math score, I conducted a p-test in regards to experience\_level and salary, resulting in an assertive value of 0. Based on this, my **null hypothesis** is that there exists no correlation between experience level and salary. My **alternative hypothesis** would be that those with higher experience levels (mid-level) earn more than those with lower ones (entry-level). With this information, I am confidently able to reject the null hypothesis. I also manually checked the mean salary of entry-level and mid-level jobs, giving me the information that the salary for entry level is $36111.11 and mid-level is $51786.89, reinforcing my alternative hypothesis.

Bayesian Odds

Step by step bayesian analysis:

* The dataset has 1171 observations.
* Among them, 126 are labeled as having "Entry" level experience.
* The prior probability of a data scientist having "Entry" level experience is calculated as 0.1076 or 10.76%.
* The prior odds of a data scientist having "Entry" level experience is calculated as 0.12.
* The true positive rate is calculated as 0.9, indicating that 90% of data scientists with "Entry" level experience earn less than the mean mid-level experienced data scientist.
* The false positive rate is calculated as 0.37, suggesting that 37% of data scientists without "Entry" level experience also earn less than the mean mid-level experienced data scientist.
* The likelihood ratio is calculated as 2.43, representing the ratio of the probability of observing the given evidence (earning less than mid-level) under the hypothesis of "Entry" level experience to the probability under the alternative hypothesis.
* The posterior odds are calculated as 0.2916.
* The posterior probability of a data scientist having "Entry" level experience, given the observed salary information, is calculated as 0.2258 or 22.58%.

In summary, based on Bayesian analysis, the posterior probability suggests a moderate confidence that a data scientist earning less than the mean mid-level salary is likely to have "Entry" level experience. The analysis incorporates both the prior belief and observed evidence to update the probability estimate.

Conclusion

In conclusion, the correlation between experience level and salary, as examined through various analyses and visualizations, sheds light on nuanced patterns within the data science job market. By scrutinizing a Bayesian framework and exploring the salary distribution, it becomes evident that experience level plays a pivotal role in determining earnings. The box and bar plots illustrate a clear progression in mean salaries corresponding to ascending experience levels, affirming the traditional expectation of salary advancement with increased expertise. Additionally, the Bayesian analysis delves into the probability of a data scientist earning an entry-level income, offering a nuanced understanding of salary dynamics within the context of observed evidence. As the bar plot underscores concentrated salary spikes within certain ranges, it becomes apparent that while there is a general alignment between experience level and earnings, outliers persist, challenging the conventional norms and emphasizing the diverse nature of compensation in the data science field. This comprehensive exploration provides valuable insights for data science professionals and aspiring practitioners, informing strategic career decisions and fostering a nuanced understanding of the relationship between experience level and salary in the dynamic landscape of data science employment.