EDA: Data Science Salaries

Patrick D. Redington

Bellevue University

DSC530-T302: Data Exploration and Analysis

June 1, 2023

EDA: Data Science Salaries

Data science salaries were analyzed using Exploratory Data Analysis (EDA) to gain insights into the factors influencing compensation for the field. This summary highlights the EDA outcome, identifies missed analysis opportunities, suggests variables for enhanced analysis, discusses incorrect assumptions, and addresses challenges faced during the process.

The EDA revealed a positive correlation between years of experience and the year of the reported salary. There was also a correlation discovered between the origin country of the company and the salary. With United States based companies paying the most on average. The EDA also revealed that the amount of remote work and the company's size play almost no role in determining a position's salary. Overall, around 56% of the variances in salaries were able to be determined using a random forest model.

Although the EDA yielded valuable insights, one area that should have been explored farther was how job titles affected the salary. By examining the salary differences across various job titles within the data science field, a more comprehensive understanding of the compensation could have been obtained. The data set included job titles, but due to the size of the dataset compared to the amount of different job title categories, it did not seem productive.

The EDA could have further explored the impact of specific skills, programming languages, and certifications. These variables could have shed light on the importance of specialization within the field and identified which skills are most highly rewarded in terms of compensation. By examining the affects skills such as Python, R, SQL, or various certification have on salaries, a more comprehensive understanding of the salary range could have been obtained.

An assumption that was proven incorrect during the analysis was that the salary of a position would be influenced by variables such as company size and the remote work ratio. However, after conducting analyses using multiple methods, it was determined that there is little to no correlation between these variables and the salary of a position. Despite initial expectations, the data revealed that variations in company size or the ratio of remote work did not significantly impact the compensation of data science positions.

The biggest challenge for the EDA was learning how to perform all the tests and creating the plots using mainstream packages. Due to using the 'thinkstat' and 'thinkplot' packages during the semester, I had to spend additional time to learn various packages to perform the same functions. However, this challenge was self-imposed as I wanted to explore alternative packages for the final project. Overall, though there were no issues with understanding the concepts of the EDA. Some of the usefulness of the plots for this particular data eluded me though. Such as the Cumulative Distribution Function (CDF), did not seem useful by itself with this data.