**Data Scientists Salary Project**

**Introduction**

I chose a dataset that focused on the salaries of Data Science employees for this final project. With this research, I want to see if there's a link between greater salaries and criteria like skills, job titles, company names, and employee ratings.

This dataset contains 41 features and 742 observations. Some of the features were average salary, different skills, job titles etc.

**Initial Predictions**

Simply by looking at the dataset, I arrived at the conclusion that people with specific skill sets will be paid more. Another forecast was that higher-rated enterprises will pay a higher average compensation. Furthermore, persons with a professional degree were expected to earn more money.

**Research and Analysis**

One of the visualizations I created showed us that the top three job titles with the highest average income were director, data scientist, and machine learning engineer. Job titles like data analytics and data analyst, on the other hand, were in the bottom two. In comparison to all of the companies, I determined the average wage of the higher rated companies. In addition, I created a visual representation of the top five organizations that were employing data engineers. The average income for employees with various degrees was compared, as well as the top five regions with greater salaries. Lastly, I wanted to compare three different aspects of the dataset, which was job title, degree compared to 2 top skills set i.e. Python and Sql calculating their average salary and occurrence.

**Feature Selection Strategy**

I expected incomes to be influenced by diverse skill sets, job titles, and other factors when deciding on the aspects that I believed to be the most influential. I utilized Recursive Feature Elimination to locate possible candidates to test my assumptions. Python, AWS, and SQL were the top three features that influenced salaries, according to our analysis. Python is the best of these three features. In addition, I ran a correlation between these characteristics to check how closely the average pay is linked.

**Conclusion**

Our goal, based on our study of this dataset, appears to have been met. I was able to uncover some fascinating facts about our employee salary data. To summarize, employees with a Python skill set are more likely to earn a higher salary. Furthermore, the job trend in Data Scientist over the years suggests that there are more career chances currently than in the past. People who work for higher-rated companies are also more likely to get paid more. The relationship between average salary as compared to skill set and degree are volatile which refers to ability to get higher average salary despite missing professional degree or master’s degree.

**Shortcomings of the analysis**

In the dataset I found that it was missing couple of vital data. For example: years of experience and years in the company was missing. I think it will impact the average salary of the employee. Absence of these factors can be vital as the whole prediction of average salary depends on years of experience.

**Directions for future study**

If we were able to add valuables data regarding years of experience in this dataset, prediction regarding average salary and many more other predictions can be more accurate. Also, an exact time stamp dataset is more accurate and helpful for data analysis and data wrangling purposes. This data set was missing timestamp which led to me not being able to make any predictions about how the average salary is going to move forward in future.

**Link to Kaggle Dataset and csv file**

<https://www.kaggle.com/datasets/nikhilbhathi/data-scientist-salary-us-glassdoor?select=data_cleaned_2021.csv>