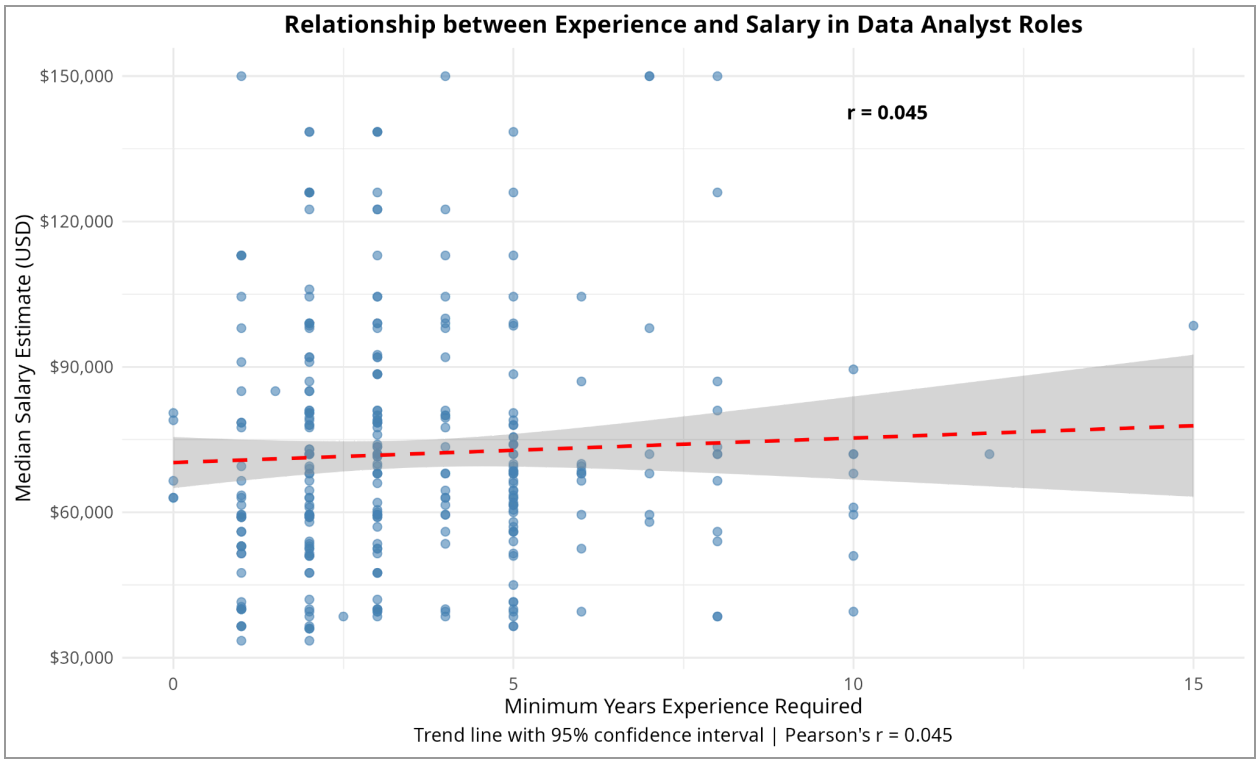
Data Analysis Report Using Google Sheets, LLM API and Julius

Phyo Wai Thaung
May 28, 2025

I. Introduction

The analysis seeks to explore to what extent ‘Years of experience’ and ‘Programming language requirements’ are associated with ‘Median Salary Estimates’ in the data analyst job market. The data was originally scraped from Glassdoor in June 2020. The analysis was started in Google Sheets, which then called an LLM API through OpenRouter to extract specific information from job descriptions. Then, the data file was imported into Julius, AI-powered code editor, to analyse data by R language and produce the report.

II. Analysis of Years of Experience



Interpretation

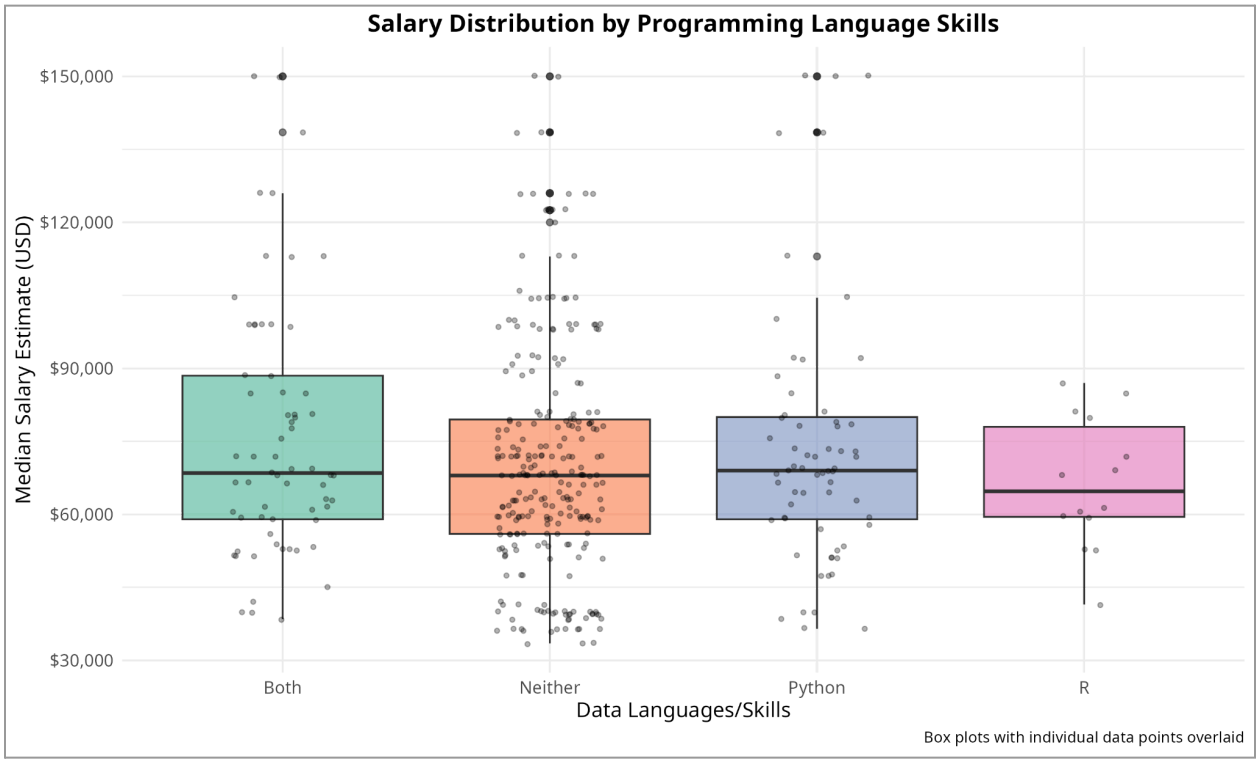
The correlation between years of experience and salary is very weak (r = 0.045), indicating that experience requirements don't strongly predict salary levels in this dataset. It shows that the relationship is not statistically significant (p = 0.427), suggesting other factors beyond minimum experience requirements drive salary differences.

Limitations

- **Missing experience data**: 92 out of 400 job postings (23%) lack experience requirements, potentially biasing the scatter plot analysis toward certain types of positions.

- **Statistical power:** The weak correlation between experience and salary may partly reflect insufficient sample size to detect subtle relationships, especially when controlling for other factors like location, company size, or industry.

III. Analysis of Programming Language Requirements



Interpretation

The data suggests that programming language skills, particularly knowing both R and Python, may be more predictive of salary than minimum experience requirements in data analyst roles.

Data_Languages	Mean_Salary	Median_Salary	Std_Dev
Both	76123	68500	25962
Neither	70307	68000	23878
Python	74477	69000	27444
R	66429	64750	13462

Mean vs. Median Comparison:

"Both" (R + Python): Mean $76,123 vs. Median $68,500 (difference of $7,623)
"Python": Mean $74,477 vs. Median $69,000 (difference of $5,477)
"Neither": Mean $70,307 vs. Median $68,000 (difference of $2,307)

"R": Mean $66,429 vs. Median $64,750 (difference of $1,679)
The fact that means are consistently higher than medians suggests right-skewed salary distributions (some high-paying outliers). The "Both" and "Python" categories show the largest mean-median gaps, indicating more salary variability and potentially more high-paying opportunities in these skill areas.

For a more robust interpretation, the median values might be more representative since they're less affected by outliers. Using medians, the ranking becomes:

- Python: $69,000
- "Both": $68,500
- "Neither": $68,000
- R: $64,750

Limitations

- **R-only positions (n=14):** The small sample size makes conclusions about R-specific roles highly uncertain. The apparent lower salary could be due to sampling bias rather than true market conditions.

- **Generalizability:** Results may not represent the broader job market, particularly for less common skill combinations or specialized roles.

IV. Reflection

As compared to last week, it has become easier to ask Julius to do what I want. The first outputs were quite close to my expectations, with prompting 2 more times producing nice visual outputs and good interpretations. One pitfall I noted was that there was a mismatch between its table output and interpretation in the salary by data language analysis where it used ‘mean’ in the interpretation though these figures are actually ‘median’ in the table. When it comes to using LLM through Google Sheets, it is quite convenient and time saving.

One thing I’m impressed with Julius is that it used both ‘mean’ and ‘median’ in the interpretation of boxplot results and highlighted the risk of relying on ‘mean’ when there are some outliers in the data set. The deeper the instructions go, the better and more detailed outputs it can produce.

What I mainly learnt from this exercise is that AI-generated outputs seem too perfect for one to thoroughly check the errors or mis-interpretations. However, human data analysts are still needed, at least for the time being, to diagnose and fix some quality issues in data analysis,

visualization and interpretation. In the new future, I feel that human data analysts may become more in the roles of AI supervisors or AI quality monitors.