Analysis & reporting with LLMs in Google Sheets and Julius AI as reporting agent of a small sample of job postings from Glassdoor, June 2020

Section 1: Introduction

We used Large Language Models (LLMs) to extract the requirements Years of Experience and Programming Languages from the job descriptions of a small sample of 400 job Data Science/Engineer/Analyst postings from the website Glassdoor from June 2020 (see https://github.com/picklesueat for more information regarding the complete dataset).

The analysis was performed in Google Sheets using OpenRouter API, an LLM Application Programming Interface, and two prompts, one given and the other own crafted (`=MYGPT(CONCATENATE("Extract the programming language requirements from this job description. Return ONLY one of these four specific strings: R, Python, Both, or Neither. Here's the job description: ", C2), "google/gemini-2.5-flash-preview", $M$2)`).

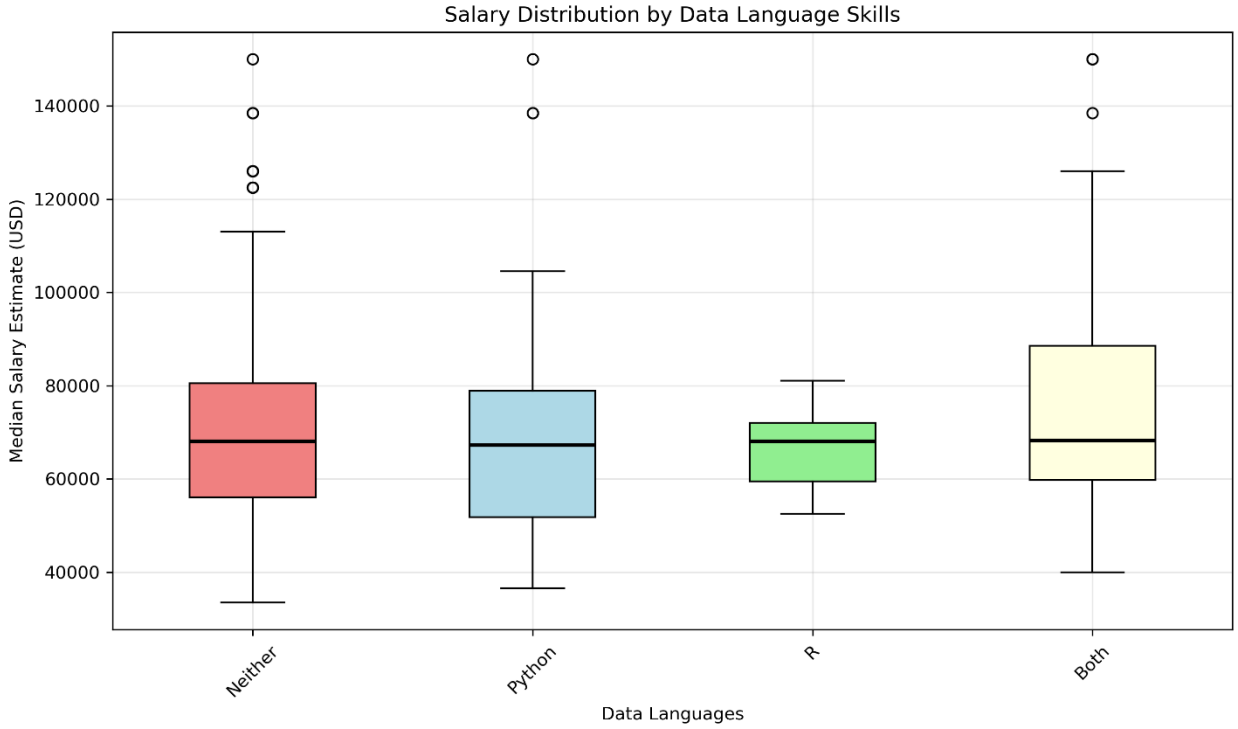
The reporting ensued with the AI agent Julius AI and included a visualization for each the two requirements Years of Experience and Programming Language versus Median Salary Estimate.

Section 2. Experience versus Salary

This scatter plot includes 305 job postings after cleaning, with experience ranging from 0-12 years and salaries from $33,500-$150,000.

The straight red trend line indicates a slight positive correlation. However, the majority of points are clearly grouped between 1 and 5 years of minimum experience and between $35,000 and $80,000 of median salary. Accordingly, it is this mass of lower experience and salaries, which seems responsible for the slight upward slope. As of year 6, the trend would certainly be downward, which is why I asked Julius AI to add a curved trend line that seems to confirm it.

Section 3. Programming Languages versus Salary



The data languages categories included in this box plot contain 188 jobs for “Neither”, 50 jobs for “Python” (including the single “SQL” entry), 58 jobs for “Both”, and 9 jobs for “R”.
All four programming language seem to correlate with a salary range of $55,000-$85,000.
In the “Python” group, but particularly in the “Neither” one, the median in the middle of the box indicates symmetry in the distribution between lower and higher salaries. In the “R” group, data are tightly grouped and the median is skewed towards higher salaries probably due to the small amount of data. This is also the only distribution with no outliers. The “Both” group shows greater variability, a wider spread of data and a median closer to Q1 (skewed towards lower salaries) even if Q3 points towards higher salaries than the rest.
Across the four groups, the median value lies just below the $70,000 salary mark.

Section 4. Reflection

The instructions given are clear and detailed, and therefore of great help. In general, I found the assignment doable, dare I say easy.
However, due to my background and lack of experience working with data, the interpretation of the plots was a challenge. This added to my lack of AI knowledge also means that I am always unsure if my results are correct.
In this assignment, I learned the importance of a specific and concise prompt. Furthermore, I realised that LLMs are quite forgiving in terms of how your request is written. They are able to ignore typos and understand the essence of one's babbling.