ASSIGNMENT REPORT – Week 3

Use of LLM API for analysis: requirements of job postings

Hong Anh Chu

Introduction:

The objective of this assignment is to look at the relationship between the estimate median salary and the years of experience amongst candidate with specific competency on programming languages of R and Python.

The data analysis is done using LLM APIs with the open AI tool of OpenRouter. The data set for this analysis was scraped from Glassdoor in June 2020 by the Graph course team. The fulle data set can be found here https://github.com/picklesueat/data_jobs_data

Analysis:

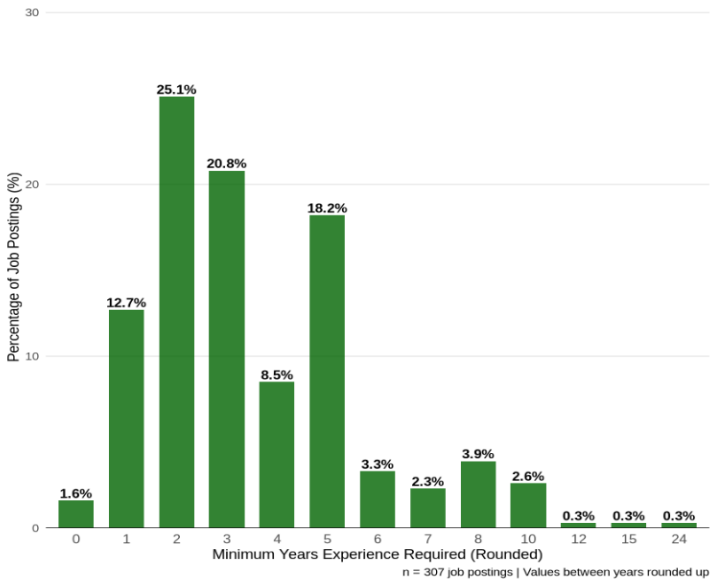
The information on the following requirements in the job description was extracted from the dataset, using LLM API directly on the excel sheet:

- Minimum year of experience
- Data languages, only select R, Python, both or neither

The data set then uploaded to Julius.ai for analysis.

1. Years of Experience requirements among the job postings:

The chart below show the distribution of years of experience requirements in the job posting:



1. [1] "- Most common experience requirement: 2 years (76 postings)"
2. [1] "- Second most common: 3 years (63 postings)"
3. [1] "- Average experience required: 3.6 years"
4. [1] "- Range: 0 to 24 years"

2. Programming languages requirements

The first visualization shows the relationship between minimum years of experience required and median salary estimates. The analysis shows a scatter plot with 307 data points after removing missing values. Interestingly, the correlation between experience and salary is very weak (r = 0.01), suggesting that years of experience alone doesn't strongly predict salary in these data analyst positions. In fact, majority of job posting are within the median salary and below (ranging between 0-5 years of experience and $30,000 to $90,000 median salary estimates.

With the median year of experience at 2.6 years, and about half of the requirements are between 1-4 years of experience which complement the above suggestion that experience alone doesn't predict salary.

Box Plot: Salary by Programming Language Skills

The second visualization displays salary distributions grouped by required data languages. This reveals some interesting patterns in how programming language requirements relate to compensation.

Key Findings

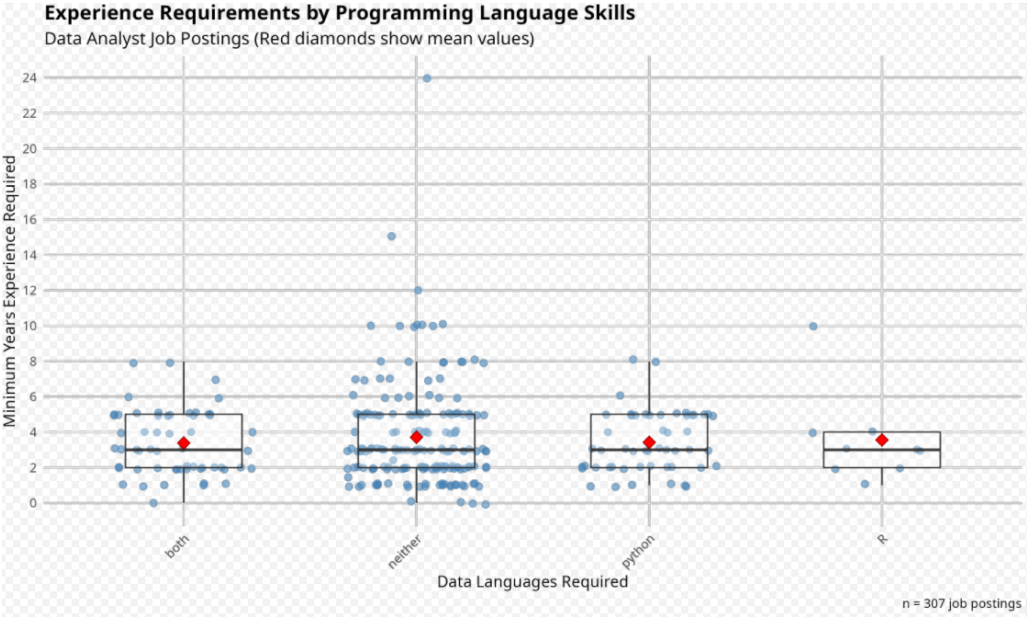
	Data.Languages	Count	Mean_Salary	Median_Salary	Min_Salary	Max_Salary
1	R	9	67222	68000	52500	81000
2	both	57	76974	68500	40000	150000
3	neither	189	70598	68000	33500	150000
4	python	52	71644	67250	36500	150000

- **"Both" (R + Python)**: Highest mean salary at $76,974 with 57 positions
- **"Python only"**: Mean salary of $71,644 with 52 positions
  "Neither": Mean salary of $70,598 with 189 positions (largest group)
- **"R only"**: Lowest mean salary at $67,222 with only 9 positions

The data suggests that knowing both R and Python commands a salary premium, while R-only positions are relatively rare and lower-paying in this sample.

With the two programming languages in the requirements took up less than 50% of the job requirements for the post of data analysis.

I also studied the relationship between the requirements for data languages and years of experience.



The chart confirms the following conclusions:

- Experience does not determine the salary level.
- R is the least required programming language and brings the lowest salary.
- About half of the job posting require both R and python for data language, with higher (second and third) salary .

Reflections

It requires the use of several tools for using LLM API, which is interesting for those whose work doesn't involve data analysis often. The prompts and steps provided by the instructors are very helpful. However, this will be challenge for real life use of the tools, in term of knowing where to get the tool and what to look for.

It's important that before using analytic tool like Julius.ai, a clear plan and requirements of the process and outputs should be thought through. I think it would help to write clear, specific and good prompts.