Analysis into Global Data Science Jobs

EMMA D. AKUE
KODY LAM
JONATHAN RODRIGUEZ-PEREZ

Description of Research Problem

As we know, AI technologies will most likely create a different set of employability-related challenges for data scientists in certain parts of the world while generating multiple opportunities in other parts of the world. Indeed, experts are predicting that for some regions of the world in South Asia, generative AI will create relief and opportunities for many entry-level workers. At the root of this, we believe that data science roles and their responsibilities differ across regions, impacting salary disparities within countries. Our analysis examines text to uncover trends and relationships between job roles, salaries, and geographic variations, providing insights into the evolving data science industry.

Our project questions are as follows:

- **What are the skills/salary trends for data-related roles when considering country and GDP?**

- **Do variables such as job category, company size, and education level affect salary-related findings and interact with job location?**

Text Mining

Text mining is a powerful tool for extracting valuable insights from textual data, a process that's particularly relevant when analyzing datasets consisting of job descriptions in the data industry. By deploying text mining techniques, an individual can effectively decipher the large volumes of unstructured data to identify key terms, discern patterns, and uncover underlying themes that might not be immediately apparent. This method not only facilitates the identification of frequently occurring words but also aids in the aggregation of similar terms to reveal prevalent themes and trends.

The primary advantage of using text mining in the context of this project is that it allows those new to the data sector, to gain a deep understanding of the landscape of job opportunities. Analyzing the most commonly mentioned words in job descriptions can highlight the roles that are in highest demand, guiding job seekers towards the most lucrative and demanded positions. Similarly, focusing on the skills section of these jobs can inform aspiring data professionals about the specific competencies that are currently valued in the industry. This knowledge is instrumental in directing their personal development efforts through self-study or formal education.

Text mining can also shed light on the educational qualifications that frequently appear in these job listings. This is particularly useful for individuals contemplating further education; understanding which degrees are most often required allows them to strategically plan their education path to align with career goals. For instance, if a significant number of job descriptions prioritize candidates with a master's degree or a Phd, prospective students can consider these programs over others.

Moving into the actual analysis through text mining, we eliminated all stop words to ensure the extraction of only meaningful and impactful terms, leading to more precise conclusions about the nature of these data-related jobs.

From the text mining into job descriptions, common terms such as "data," "network," "security," and "analysis" suggest a strong demand for candidates who are not only proficient in technical skills but also capable of ensuring security and providing insightful research. *(Refer to Figure 1)* Candidates can leverage this information by focusing their career development efforts on these areas, enhancing their desirability to potential employers. In the skills section, the prevalence of words like "database," "management," "troubleshooting," and "visualization" indicates the need for versatile abilities ranging from technical database management to complex problem-solving and data presentation skills. *(Refer to Figure 2)* Job applicants can consider acquiring or sharpening these skills through targeted training or practical experiences to meet job market expectations.

We compared the top 10 most popular programming languages—Python, JavaScript, Java, C++, Swift, Kotlin, Rust, TypeScript, PHP, Ruby, and SQL—against job descriptions and skills. SQL emerged as the standout with 1,386 mentions, highlighting its critical role in job descriptions, while SQL, Python, and Java were the top three skills demanded. *(Khuvaish, 2024)* This analysis underscores the importance for candidates to prioritize learning SQL, Python, and Java, reflecting their high demand in the data industry and enhancing job market competitiveness. Understanding these key trends also helps aspiring professionals tailor their development efforts towards technologies that shape the industry's future, ensuring they remain competitive and well-prepared.

Finally, the qualifications analysis revealed that degrees such as BCA, M.Tech, and MBA are highly valued in data-related job markets. *(Refer to Figure 3)* Knowing which degrees are most sought after can guide individuals in making informed decisions about their education, particularly in selecting specializations that are in high demand as these degrees are extremely costly. Additionally, possessing these recognized

degrees often provides a competitive edge in job applications and can lead to higher starting salaries and more rapid career progression within the data industry.

Cluster Analysis

Since we recognize that cluster analysis is mainly an exploratory technique, the core of our project focused on text mining and spatial analysis. However, we have chosen to utilize cluster analysis as an initial point into deciphering the data due to two main advantages that this method provides. One, by clustering data, we can summarize the large dataset into meaningful clusters, making it easier to understand and interpret the underlying patterns and relationships. Two, clustering provides insights into the categories present in the data, which may not be apparent through simple observation. Particularly to this project, we are interested in exploring the patterns related to a few variables of interest, mainly: job title, mean salary and mean years of experience. In our case, we had to use means because the data provided ranges for these variables, and we felt the averages would be easier to handle in the analysis.

We also recognize that cluster analysis results have no statistical basis. For this reason, we supplemented the results with outside research. The cluster analysis results show us that the data points are indeed assigned to a cluster based on its proximity to the cluster center as seen in Appendix Figure 4. When using hierarchical clustering on mean_salary and avg_experience using job titles as the color, two groups emerged split on low experience vs. high experience, however the interesting thing was that the salary was not split, both groups had high and low salary. When reviewing the research in this field, we found that there are indeed entry level roles that provide a very competitive salary for “skills in demand” such as proficiency in programming languages, data querying, and familiarity with visualization tools (O’Reilly Media, 2021). On the other hand, we also found examples of jobs that require a high amount of experience but may not offer high salaries. These roles were mainly found in nonprofit organizations, education sector, and smaller organizations with budget constraints. This insight helped us understand why we were getting low strength of an association when computing the correlation between certain variables in our dataset - such as no correlation between company size or salary.

Spatial Analysis

Finally, for the purpose of our problem, we also believe that spatial analysis could offer us the possibility to visually uncover certain specificities about data-science-related jobs in different geographic regions.

Because data science is a field in which different professionals use essentially universal tools and coding languages to do their jobs, we know that location isn't as much of an obstacle when it comes to applying for jobs as much as it is in different fields such as Education, Health Care or Law. Especially since many of those jobs can also be done remotely.

Our data contains information on location, countries, latitude, and longitude which lets us know that our dataset is suitable for partial analysis. Additionally, the data contains different variables available to us such as mean salary, average experience required, and gender preferences therefore it is a great way for us to make recommendations to seasoned professionals as well as recent grads who study in the data science field. It is important to note that generic data frames aren't ideal when it comes to doing spatial analysis, for that reason we turned our data in the sf (simple feature) format to use ggplot 2 to create our maps with layers.

For the purpose of our spatial analysis, it is important to point out the different economic situations the countries of our dataset are associated with since job demands and professional opportunities are dependent on a region's economic prosperity and specificities.

Therefore we included a built-in map as part of the analysis *(Refer to Figure 5)*. This map will be important when discussing our results.

Our next map, *(Refer to Figure 6)* shows the global distribution for the average year of experience required. The map shows that the average experience required for data science jobs tends to be higher (around 9 years) in G7 developed countries such as the UK and the USA but we also note that non-G7 developed countries such as the UAE will on average also require over 7 years of work experience. This can indicate that more companies located in developed and highly competitive countries either outsource entry-level positions in other countries or replace those jobs with AI. We can also point out that India is the only country that is part of the BRICs for which many years of experience are required (over 9 on average). Which not surprising since India is expected to become the world's third-largest economy by 2030 (CNBC, 2024). For countries that require less work experience (under 7 years), we see that many of them are located in South and Central America and include Chile, Paraguay, Grenada, Trinidad and Tobago.

What's interesting to note is that even though the average experience required is different across countries, we see that the most popular job titles are very similar from one country to another and often time include **Data Analyst**, **Database Administrator,** and **Network engineer** in the top 5 for most in-demand jobs in data science.

Thanks to the universality of our dataset, we were also able to have access to gender preferences for the job postings*(Refer to Figure 7.*.When looking at the map, we see that Western countries such as the United States, Canada, and Australia do not discriminate based on gender which is to be expected since national laws and policies are put in place to avoid this phenomenon, especially in G7 Developed Countries. However, we see that countries located in South America, North Africa as well as the Middle East will more often prefer males for different job postings.

It is interesting to note that among Western countries the more progressive ones such as France, Japan, Iceland, and Poland may sometimes favor females for certain job postings which reflect HR strategies commonly used in Europe with companies such as L OREAL (L'Oreal.com, 2022). We also seem the same phenomenon in India with the US-India Alliance for Women's Economic Empowerment (The Economic Times, 2024) designed to promote Women in STEM

Recommendations

On improving data analysis:

To better understand the problem and enhance our analysis, additional steps could have been taken:

1. Data Appending: We could have incorporated more variables or attributes into the dataset, such as industry type, or job posting dates. This enriched data can provide deeper insights into the factors influencing job distribution and salaries.
2. Machine Learning: Utilized advanced machine learning models, such as neural networks combined with text analysis. These models can capture complex relationships within the data.
3. External Data Sources: Additionally, we could have integrated external datasets, such as labor market reports, economic indicators, or social media trends related to job postings. This external context could have enriched our analysis and provided benchmarking against broader industry trends.
4. Industry Research: Seek feedback from domain experts or stakeholders to validate our findings and interpretations. Incorporating their insights can enhance the relevance and applicability of our analysis to real-world decision-making.

By critically evaluating the results and incorporating additional data and methodologies, we could have developed a better understanding of the complexities of the job market particularly related to our research question of skill trends and variable influence on job parameters.

On insights from analysis/to job seekers:

1. Focus on Keywords: Many companies use applicant tracking systems (ATS) to filter resumes. Make sure to include keywords from the job description to

increase your chances of getting noticed. Consider including the most popular keywords from the text mining section into resumes.

2. Research Market Rates: Use online resources like Glassdoor, LinkedIn Salary, or industry reports to research salary ranges for data science roles in your location. Companies in the private sector have higher salary posts than the public sector.
3. Remote Opportunities: With the rise of remote work, do not not be limited to jobs in specific geographic locations. Our data maps show that Latin American countries are more open to hiring female applicants.
4. Consider Experience and Education: Your level of experience, education, and specific skills can influence your salary expectations.Our analysis showed that candidates should prioritize learning SQL, Python, and Java.
5. For job seekers looking to get employed by companies in sought after G7 and non-G7 countries, a higher level of experience might be required which is why supplementing studies with bootcamps and hackathons will become more essential in the years ahead. Simultaneously, there is also the possibility of remotely applying to jobs located in South America in order to gain more experience.
6. Despite job disparities in STEM for women, many G7 could start seeking out women employees in data science to combat inequity, additionally it is important to seek out companies that promote women in stem regardless of location in order to increase chances.

Conclusions

The methods of cluster analysis, text mining, and spatial analysis applied to our jobs dataset have provided valuable insights into the relationships and patterns within the job market.Through cluster analysis, we were able to group similar jobs based on their attributes, revealing distinct categories within the data. Text mining techniques allowed us to extract key information from job descriptions, such as important skills and qualifications, providing deeper understanding into the job market's demands. Additionally, spatial analysis enabled us to explore geographical patterns, uncovering regional variations in job distribution. By combining these methods, we gained a comprehensive view of the job landscape, identifying trends and opportunities. We hope these insights are invaluable to job seekers, such as our fellow APAN cohort, that will be entering a job market with similar parameters described in this report.

APPENDIX

A1.Figures:

Figure 1: Texting Mining for Job Descriptions

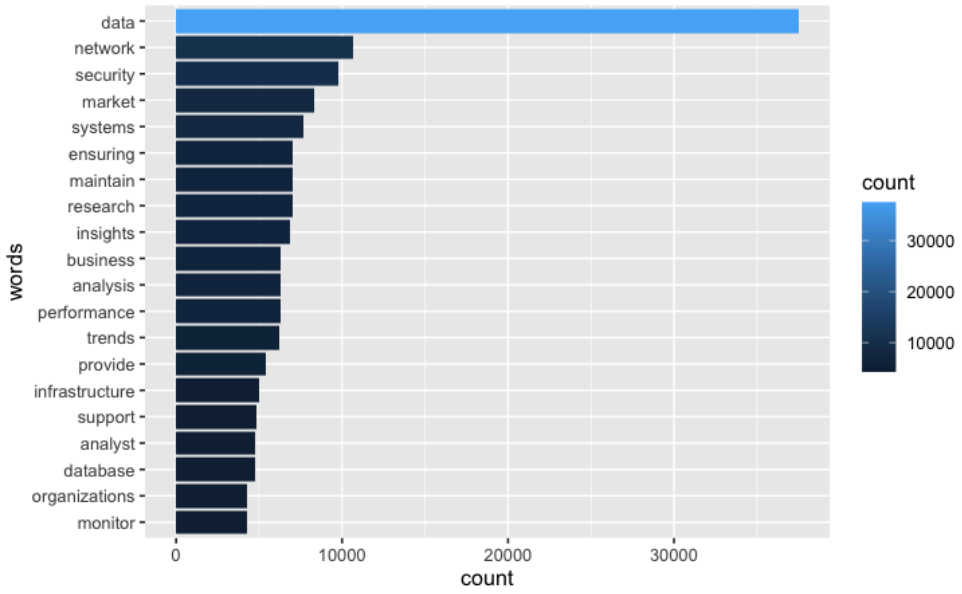


Figure 2: Texting Mining for Skills

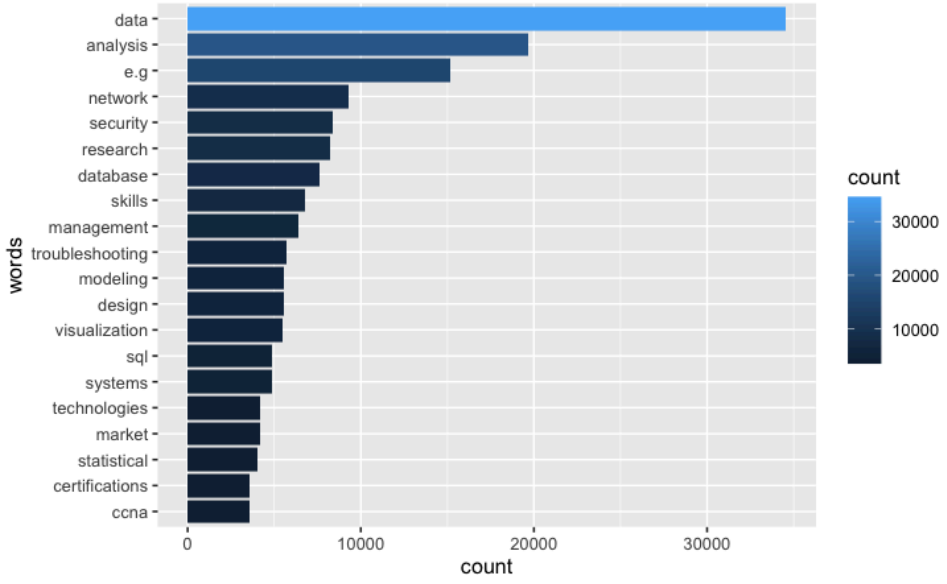


Figure 3: Text Mining for Qualifications

word <chr>	count <int>
bca	2728
m.com	2702
m.tech	2677
mca	2676
mba	2669
b.tech	2664
ba	2660
bba	2558
phd	2557
b.com	2549

Figure 4: Hierarchical Cluster Analysis

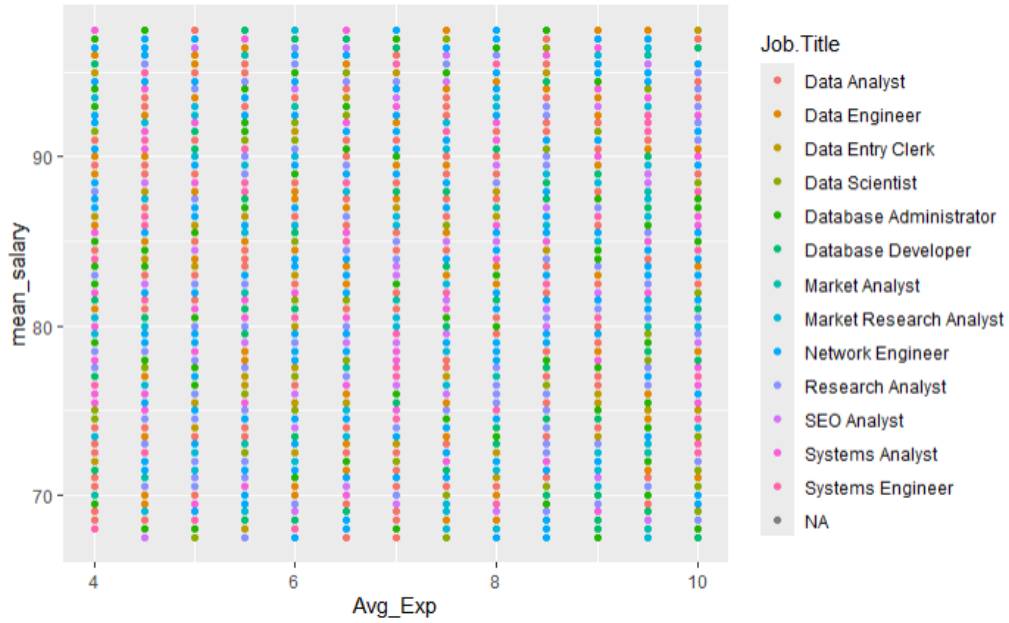


Figure 4.2: Hierarchical Cluster Analysis

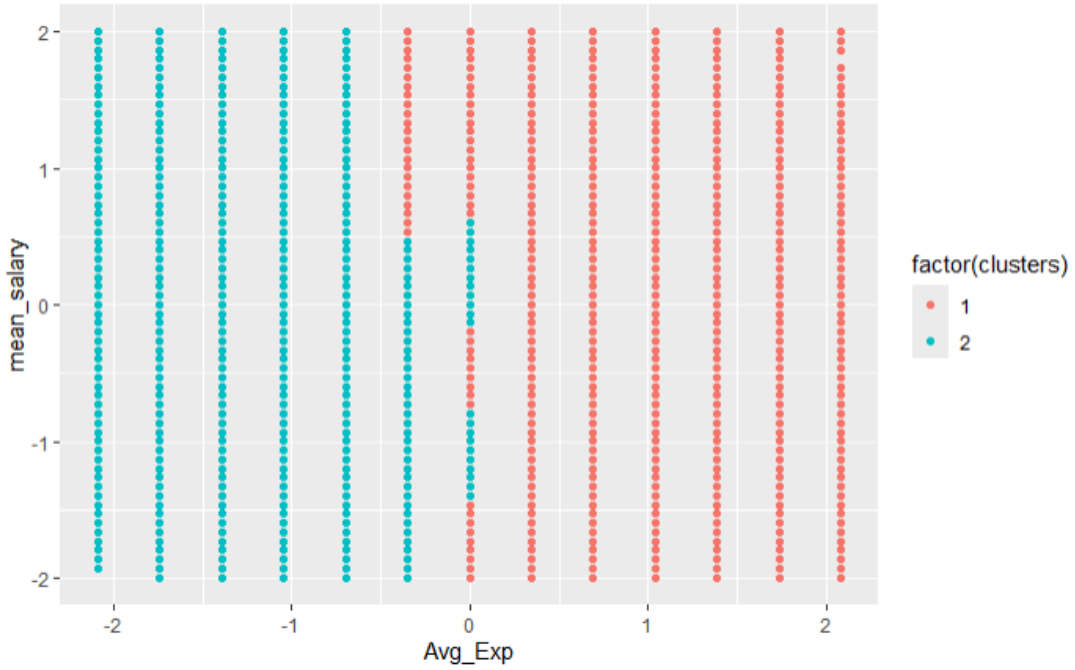


Figure 5: Spatial Analysis : World Economy

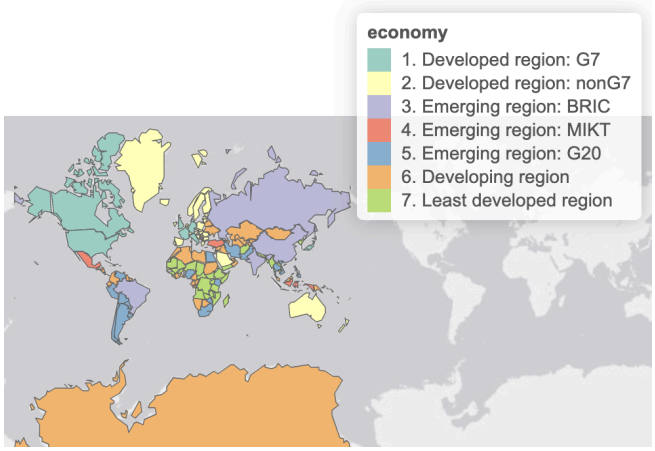
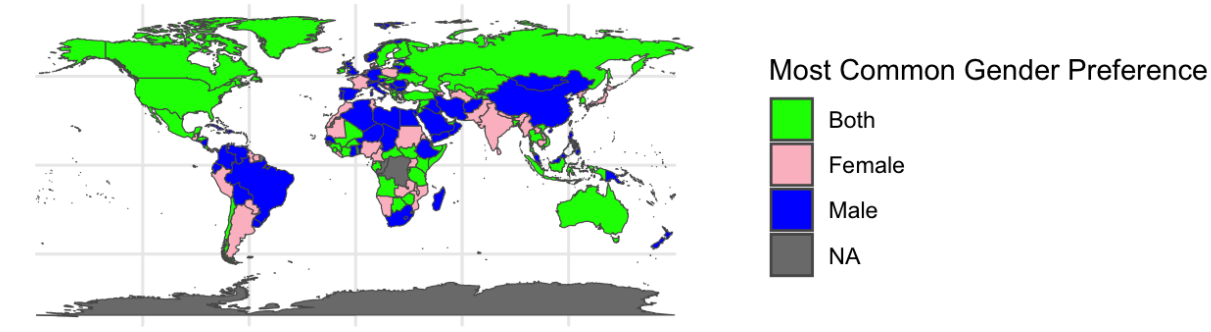


Figure 6

Figure 7



A2.Works Cited:

1 - Jhuvaish. (2024). 10 Best Programming Languages to Learn in 2024, Linkedin https://www.linkedin.com/pulse/10-best-programming-languages-learn-2024-khuvaish-undefined-lwa4c?trk=public_post_main-feed-card_feed-article-content

2 - O'Reilly Media. (2021). 2021 Data Science Salary Survey. Retrieved from https://www.oreilly.com/radar/2021-data-ai-salary-survey/

3 - L'Oréal celebrates its women in engineering at Operations https://www.loreal.com/en/articles/group/operations-celebrates-its-women-in-engineering/

4 - Cover Photo: Fresenius SE & Co. KGaA. (n.d.). Focus on Data Science. Fresenius Karriere. Retrieved from https://karriere.fresenius.de/en-US/digital-careers/focus-on-data-science

5- US, India join forces to promote women's inclusion in STEM (2024) https://economictimes.indiatimes.com/tech/technology/us-india-join-forces-to-promote-womens-inclusion-in-stem/articleshow/107857436.cms?from=mdr

6-How India is challenging China as Asia's tech powerhouse (2024) https://www.cnbc.com/2024/04/05/how-india-is-challenging-china-as-asias-tech-powerhouse.html

7-Data Source:Kaggle. (n.d.). Job Description Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset