A Regression and Visualization Study of Global Data Science Salary Trends Adjusted for Cost of Living

Team-11

(Course: DS5110 Intro to Data Management and Processing)

Professor: **Kylie Bemis**

Authors:

Abdul Sameer Shaik (002062125)

Deepashree Srinivasa Rao Rannore (002026701)

Jeevith Doddalingegowda Rama (002031889)

Sunidi Vijayakrishna Kumar (002474285)

Summary

This project investigates income trends in the data science profession by analysing salary data in conjunction with cost-of-living metrics across various countries. While many studies focus on gross salaries, we aim to provide a more realistic picture by adjusting salaries for living expenses, offering insights into the *real value* of a data science job across different regions.

We utilized two main datasets: a global dataset on data science jobs and salaries, and a secondary dataset on cost-of-living indices. By merging and cleaning these datasets, we created a unified view that allowed us to analyse regional disparities, temporal trends, and other influencing factors such as company size, experience level, and job title. Our approach involved data wrangling, exploratory data analysis, and linear regression modelling to determine which factors most strongly predict adjusted salary. Key findings were also visualized to uncover patterns in job compensation after adjusting for living costs. This project not only helps aspiring data scientists understand the global job market more deeply but also equips decision-makers with information to compare compensation in a more meaningful way.

Methods

This project involved the integration of two real-world datasets: (1) data science job postings containing salary and role information, and (2) an extended cost of living and income index dataset. The primary objective was to analyse compensation trends by adjusting salaries for cost of living and identifying influential factors using both visual and regression-based techniques.
We used two primary datasets for this project:

1. **jobs_in_data.csv**: Contains global records of data science job roles, salary information, experience levels, and company characteristics.
2. **Cost_of_Living_and_Income_Extended.csv**: Provides yearly cost-of-living indices for different countries, which allows normalization of salaries by living expenses.

Data Preprocessing and Cleaning

The jobs_in_data.csv dataset was first imported and standardized by converting company location to lowercase and casting work year as an integer. Similarly, the Cost_of_Living_and_Income_Extended.csv dataset was cleaned by renaming the Country column to company location and Year to year to enable merging. A left join was performed on company_location and work year to create a merged dataset that included both salary and cost of living data.

To normalize salaries across countries, a new column, adjusted_salary, was computed by dividing salary_in_usd by the Cost_of_Living index. This adjustment allowed for more equitable comparison of salaries across regions. Any rows with missing values were removed using na.omit() to ensure clean analysis and modelling.

Feature Engineering

An additional categorical column job_group was derived using pattern matching on job_title to classify roles into broader groups such as Data, Engineering, Analytics, Management, and Architecture. This grouping facilitated more interpretable visualizations. The Region column was also used to categorize countries into broader geographic areas such as Asia, Europe, North America, Oceania, and South America for regional trend analysis.

Data Visualization

Various ggplot2 visualizations were created to explore patterns:

- **Box plots**: Showed adjusted salary distribution across job roles, regions, company sizes, and work settings.
- **Bar and column charts**: Highlighted the average adjusted salary by experience level.
- **Line plots**: Illustrated salary trends over time (2020–2023) across experience tiers.
- **Scatter plots**: Analyzed relationships between salary and cost of living.

These visuals were used to extract and document key observations on experience-based salary progression, regional pay disparity, and cost-of-living normalization.

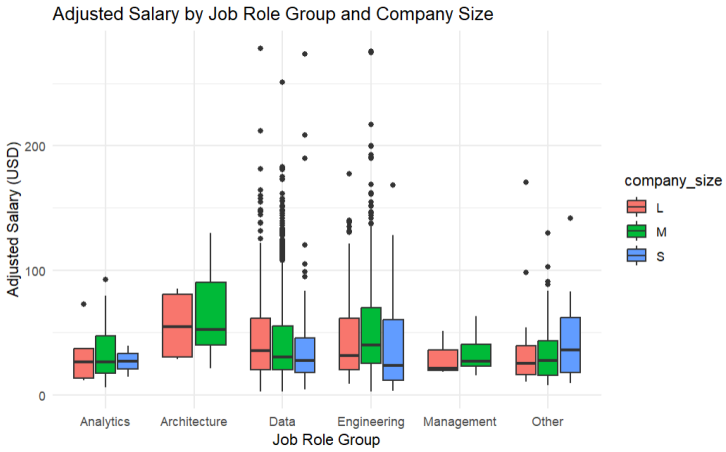
Modelling

For predictive modelling, the cleaned dataset was filtered to include only job titles with at least 10 observations to avoid sparsity. Categorical features including experience_level, job_title, company_size, and work_setting were converted to factors. A **linear regression model** was then built using these features to predict salary_in_usd.

Results

The analysis revealed several key trends in the data science job market by exploring real-world job postings merged with cost-of-living data. Visualizations showed that experience level is a dominant factor influencing salary, with executive roles earning the highest adjusted salaries across regions, followed by senior, mid-level, and entry-level roles. This progression confirms a well-defined career ladder in the industry. Following are the different type of visualizations we have incurred from our dataset.

1. Adjusted Salary by Job Role Group and Company Size

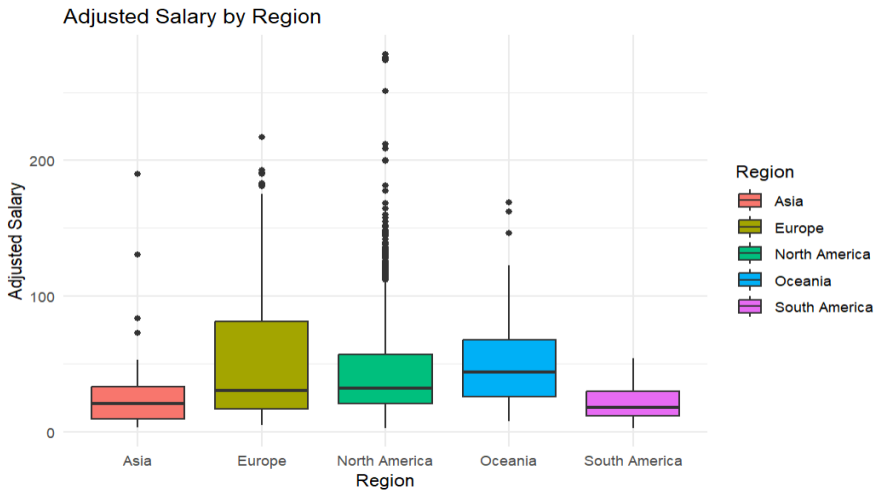


This box plot shows that Architecture and Data roles have the highest median adjusted salaries, especially in medium and large companies, indicating better pay for specialized roles in larger firms. Engineering roles show a wide range of salaries with high outliers, while Analytics and Management roles have lower medians, particularly in small companies. Overall, both job role and company size significantly impact earning potential.

2. Average Adjusted Salary by Experience Level

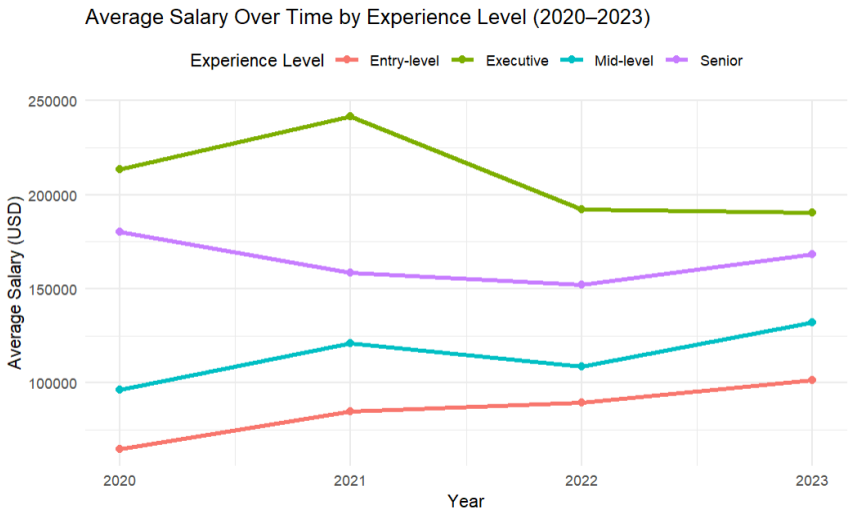
This bar chart shows a clear increase in adjusted salary with experience level. Executive roles earn the highest, followed by Senior, Mid-level, and Entry-level positions. The trend reflects a well-structured, merit-based salary hierarchy, where advancing in experience leads to significant pay growth, regardless of location.

3. Adjusted Salary by Region



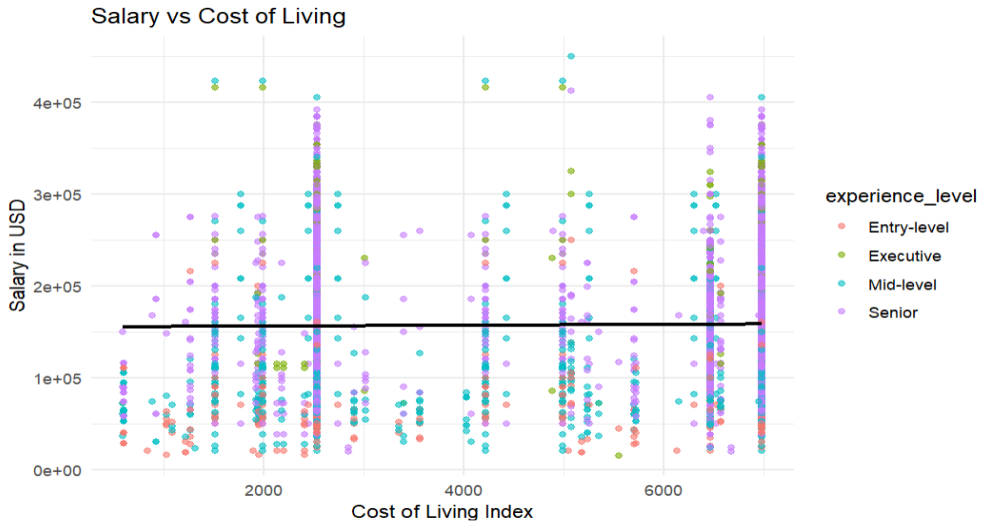
This box plot shows that Europe and Oceania have the highest median adjusted salaries, while Asia and South America have the lowest. North America shows many high outliers, indicating the presence of top-paying roles. Europe's wide IQR reflects high salary variation, and Oceania's consistency suggests uniform pay. Overall, region impacts salary, but individual factors like role and company still play a major role.

4. Average Salary Over Time by Experience Level



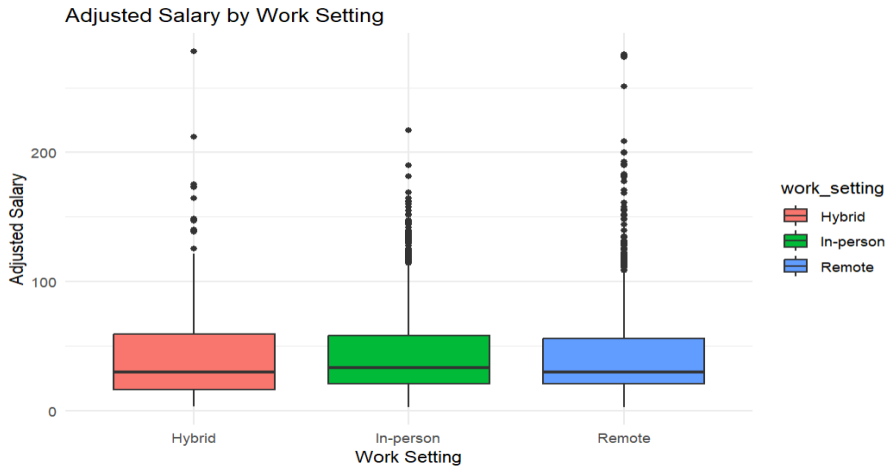
Executive roles consistently earn the highest salaries, peaking in 2021 but dropping sharply in 2022, reflecting sensitivity to market shifts. Senior and mid-level salaries declined slightly before recovering in 2023, while entry-level salaries showed steady growth, eventually crossing 100000 USD. Experience strongly influences salary, with higher roles earning more but experiencing greater volatility, and junior roles demonstrating stable upward trends.

5. Salary vs Cost of Living



There is no strong connection between salary and cost of living, as indicated by the flat regression line. Experience level is a much stronger factor in determining pay. Executive and senior roles consistently earn more across all cost of living ranges, while entry-level positions tend to stay on the lower end regardless of location. The presence of many high-paying outliers suggests that company policies, role specialization, and remote work options may have a bigger impact on salaries than regional expenses. Overall, experience and job role are the main drivers of compensation.

6. Adjusted Salary by Work Setting



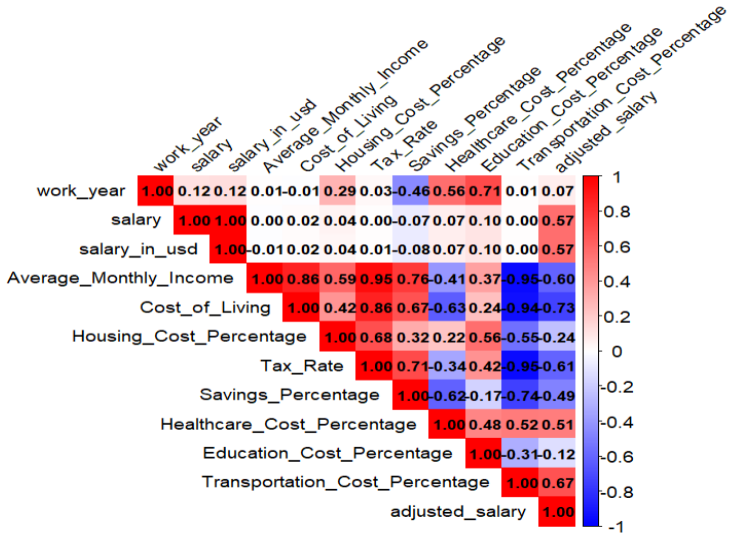
Hybrid roles have the highest median adjusted salaries, followed closely by in-person roles, while remote roles have the lowest median. Hybrid settings show greater variability in salaries, indicating a broader range of compensation. Remote roles have the widest spread of high-paying outliers, suggesting that while average pay may be lower, exceptional remote opportunities do exist. Overall, work setting influences salary to some extent, but other factors likely contribute more significantly to high earnings.

The median adjusted salaries for hybrid, In-person, and remote work settings appear quite similar, indicating that the type of work setting does not significantly impact the central salary value. Both Remote and In-person categories show many high-salary outliers, indicating that some of the highest-paying jobs are available in these two formats. The Hybrid work setting has a narrower distribution of salaries, with fewer extreme values, this may reflect more standardized pay for hybrid roles. Professionals across all three setups have access to a similar range of earnings, though individual cases can vary greatly.

7. Modelbuilding

A linear regression model was built using experience level, job title, company size, and work setting as predictors of salary. The dataset was split 50/50 into training and testing sets using caret::createDataPartition(). The model achieved a **Root Mean Squared Error (RMSE) of 52,056.11 and an R-squared value of 0.279**, suggesting moderate explanatory power with room for improvement. Predictions generated from the model aligned with expected values. For example, for a set of test inputs, predicted salaries included values such as $174,901.60, $167,527.80, and $158,973.70, reflecting realistic salary ranges for senior and executive roles. These results confirm that experience level, job role, and company characteristics can meaningfully influence compensation patterns, although additional factors may be needed to improve model accuracy.

8. Correlation matrix



The correlation matrix shows that adjusted salary has a strong positive correlation with Average Monthly Income (r = 0.86), highlighting income as a key driver of effective compensation. Cost of Living also

correlates with Housing Cost Percentage (r = 0.68), reinforcing housing's major role in living expenses. Negatively, adjusted salary is impacted by Education (r = –0.67) and Healthcare costs (r = –0.52), indicating that rising costs reduce effective earnings.

9. Chatbot:

```
> # Call the function to run the interactive prediction without showing the internal dataframe details
> predictSalary()
Enter your job title
Options:  Data Architect, Data Scientist, Machine Learning Researcher, Data Engineer, Machine Learning Engineer, Data Analyst, Analytics Engineer, Appl
ied Scientist, BI Developer, Business Intelligence Engineer, Research Scientist, Research Analyst, Research Engineer, Data Science Engineer, Data Produ
ct Manager, Machine Learning Scientist, AI Engineer, MLOps Engineer, Data Modeler, Data Science Consultant, Business Intelligence Analyst, ML Engineer,
Head of Data, BI Analyst, Data Specialist, Data Integration Specialist, Data Science Practitioner, Business Intelligence Developer, Data Lead, AI Devel
oper, Data Manager, AI Architect, Data Science Manager, Data Strategist, Decision Scientist, Data Quality Analyst, Computer Vision Engineer, Director o
f Data Science, ETL Developer, Data Analytics Manager, Machine Learning Infrastructure Engineer, Principal Data Scientist, Data Developer, Data Infrast
ructure Engineer, Machine Learning Software Engineer, Data Science Lead, Data Operations Analyst, Business Data Analyst, Data Operations Engineer, BI D
ata Analyst, Deep Learning Engineer, Head of Data Science, AI Scientist, NLP Engineer, Applied Machine Learning Scientist, Lead Data Engineer
> Data Science Practitioner
Enter your experience level
Options:  Senior, Mid-level, Executive, Entry-level
> Entry-level
Enter your company size
Options:  M, L, S
> M
Enter your work setting
Options:  In-person, Remote, Hybrid
> In-person

Predicted Salary (USD): 76522.16
>
```

Using this chatbot the user will be prompted with the question as well with the options available for them to choose which makes it easier for them to interact with Bot. This feature of our project is very useful for any data science graduate or a student who wishes to take a particular role or field of study for their master's or to pursue PhD.

Coming to the technical part:
The code is optimized to hide the coding details to make sure its secure to the end and safe. And restricting the showcase of the underlying code.

Taking advantage of the model that is already build in the previous step the bot makes the predictions based on the input provided by the user.

Discussion

This project helped us better understand how salaries in the data science field vary across different roles, experience levels, and parts of the world. By adjusting salaries based on the cost of living in each country, we were able to compare them more fairly and reveal differences that are not visible when using raw salary numbers alone. One of the main findings was that experience level plays a big role in determining salary executive-level employees earn more than entry-level ones, no matter where they work.

Interestingly, the cost of living in a country doesn't seem to strongly affect salaries. This means that many companies may not be adjusting pay based on how expensive it is to live in certain places. This is important for both job seekers and companies, especially in today's remote and global work environment. Still, there are high-paying jobs in every region, showing that skilled professionals can find good opportunities almost anywhere. These insights can help job seekers make smarter career decisions and guide companies in creating fairer pay

structures.

In the future, this analysis could be improved by adding more details like education, technical skills, and the type of industry. Also, trying more advanced models could help us better understand the complex patterns in salary data. Overall, this study shows that using data analysis is a powerful way to learn about the job market and support better decisions for both employees and employers.

Statement of Contributions

Abdul Sameer Shaik contributed to data merging, salary normalization using cost of living indices, model building using linear regression, and drafting the Discussion and Methods sections of the report.

Deepashree Srinivasa Rao Ranmore led the development of visualizations, performed exploratory data analysis, and contributed to the Observations and Results sections, ensuring clarity in communicating insights.

Jeevith Doddalingegowda Rama assisted in data preprocessing, handled missing value treatment and data filtering, and contributed to the Summary and Dataset Description sections.

Sunidi Vijayakrishna Kumar was responsible for deriving engineered features (e.g., job_group), fine-tuning visual presentation of graphs, and contributed significantly to writing the Introduction and polishing the final report.

All team members collaboratively reviewed and edited the final report to ensure coherence, accuracy, and alignment with project goals.

References

1. Kaggle. Data Science Jobs Dataset (jobs_in_data.csv). Retrieved from: https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries

2. Kaggle. Cost of Living and Income Extended Dataset (Cost_of_Living_and_Income_Extended.csv). Retrieved from: https://www.kaggle.com/datasets/thedevastator/cost-of-living-and-average-income-per-country

3. Predictive Salary Modelling: Leveraging Data Science Skills and Machine Learning for Accurate Forecasting: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10859447

4. Salary Prediction Using Regression Techniques: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3526707

5. Predictive Insights: using Machine Learning to Determine Your Future Salary: https://www.ijsce.org/wp-content/uploads/papers/v13i2/B36050513223.pdf