DATS 6103: Final Write-up
Team 6
Fyrooz Anika Khan, Rasika Shrikant Nilatkar, Aniruddh Rajagopal, Bairu Likhitha Reddy
Professor Ning Rui
Fall 2024
12/11/2024

Data Science Job Salary Analysis

Introduction

In the rapidly evolving landscape of data science, understanding salary trends and job market dynamics has become increasingly critical for professionals, employers, and academic researchers. The field of data science continues to expand, with emerging technologies, changing work models, and global market shifts significantly influencing compensation structures. This research aims to provide a comprehensive analysis of data science job salaries, examining the intricate factors that contribute to salary variations and offering insights into the current state of the data science job market.

The study leverages a comprehensive dataset from Kaggle containing 6,599 observations across 11 variables, capturing salary information for data science professionals from 2020 to 2023. By employing rigorous statistical methods and machine learning techniques, our research seeks to answer critical questions about salary determinants, including the impact of experience levels, job titles, company sizes, and work models on compensation.

Our analysis goes beyond simple salary comparisons. Through exploratory data analysis, predictive modelling, and detailed visualizations, we unpack the complex ecosystem of data science employment. We investigate how factors such as geographic location, company size, and professional experience intersect to shape salary ranges. Moreover, our logistic regression models provide a predictive framework for understanding what characteristics contribute to high-paying data science roles.

The significance of this research lies not only in its empirical findings but also in its potential to guide career decisions, inform hiring strategies, and provide a nuanced understanding of the data science job market. By revealing trends in salary distribution, work models, and compensation across different professional levels, this study offers valuable insights for data science professionals, recruiters, and organizational leaders navigating the dynamic landscape of technological employment.

SMART Questions

1. What is the average salary of data scientists based on their level of experience in the United States?
2. Has the median salary for entry-level data science positions changed significantly from 2020 to 2024, and by what percentage?
3. What are the highest-paying job titles in the U.S. compared to other countries?

Description of Data

The "Data Science Salaries" dataset, gathered from Kaggle, offers an in-depth look at salary trends for data science professionals around the world from 2020 to 2023. It captures a range of information about salaries in various data science roles, making it easier to understand how different factors influence compensation. There is a total of 6599 observations and 11 columns. We don't have any duplicate values in the dataset.

Here are some key elements of the dataset -

- **Job Title:** This details the specific role, such as Data Scientist, Machine Learning Engineer, or Data Analyst.
- **Experience Level:** Salaries are categorized based on experience, whether you are just starting out (entry-level), have some experience (mid-level), are a veteran in the field (senior-level), or hold executive positions.
- **Employment Type:** The dataset indicates if a position is full-time, part-time, contract-based, or freelance.
- **Salary in USD:** To maintain consistency, all compensation figures are normalized to USD.
- **Company Location:** This feature reveals the geographic location of the hiring company, helping to identify regional salary differences.
- **Employee Residence:** It shows where the employees live, which is particularly relevant in the context of remote work and its impact on salaries.
- **Company Size:** This categorizes companies as small, medium, or large based on their employee count, highlighting how company scale can affect pay.

This dataset is a fantastic resource for anyone looking to understand how salaries in the data science field are changing. It sheds light on how various factors like job location, company size, and experience influence salary levels. Whether you're a job seeker, a hiring manager, or just curious about the industry, this data provides valuable insights into the evolving landscape of data science compensation.

EDA and Graphs

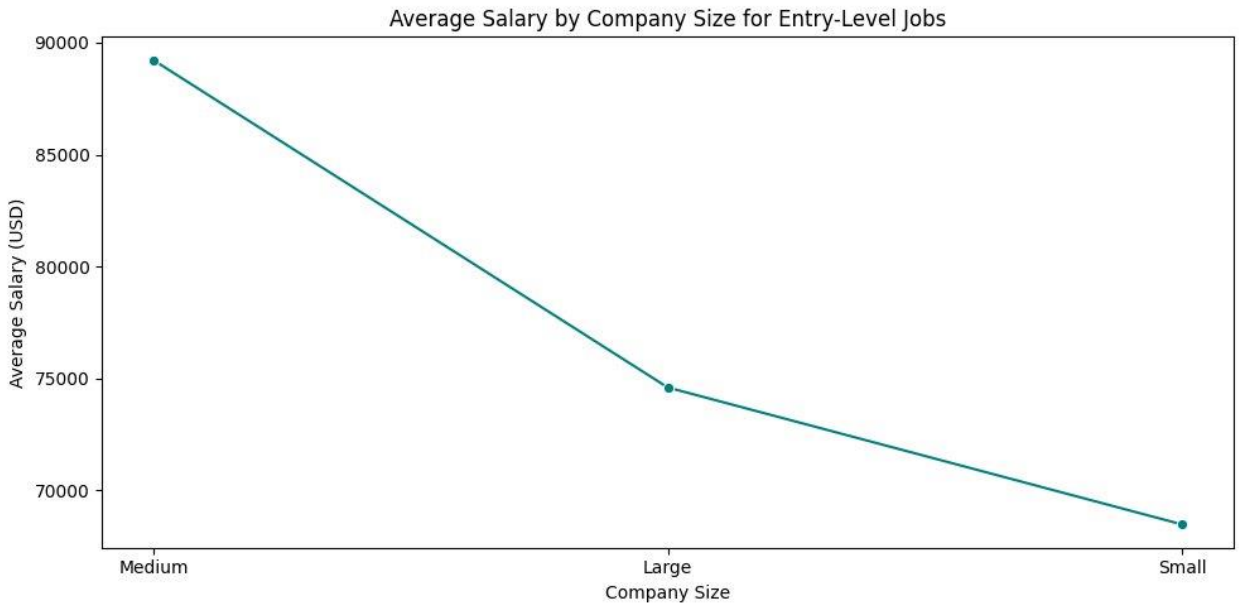

Figure 1: Average Salary by Company size

This analysis compares company size to the average salaries for entry-level positions. It follows that medium-sized companies have the highest average wage, large companies come in second, while small companies offer the least compensation. This trend is shown in a line plot, which shows a clear decline in salaries as company size decreases. This is where medium-sized companies might consider competitive pay to balance talent with scaling. Smaller companies may have limited resources, which restricts them from paying more. It indicates that the size of the company will be one of the main factors when choosing a job, at least for those who have just started their careers and seek the best possible compensation.

Company size is an important factor in determining salary structures, as highlighted in this analysis. Medium-sized firms often invest more in competitive salaries to attract better skill profiles as they grow and become more established in competitive markets. In larger companies, while they are similarly well-resourced, the average salary may be slightly lower due to more structured pay scales, a wider range of non-monetary benefits, or a broader distribution of salaries across different pay grades. In contrast, smaller companies tend to prioritize operational efficiency and resource allocation over offering high salaries, often compensating by providing opportunities for rapid advancement, diverse roles, or equity incentives.

The line plot effectively illustrates these findings, showing a significant salary decline from medium to small company sizes. This information can greatly assist job seekers by allowing them to evaluate their options and enable organizations to benchmark their pay strategies, helping them remain competitive in the market.

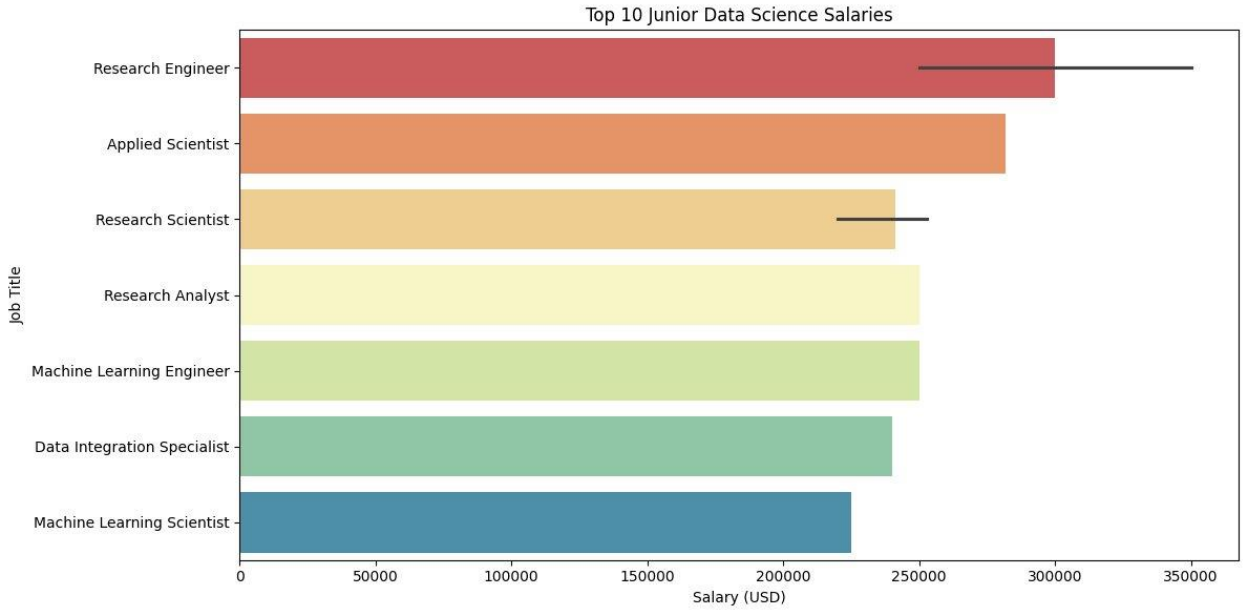

Figure 2: Top 10 Highest-Paying Entry-Level Data Science Roles

It analyses the top 10 best-paying junior data science jobs and focuses on entry-level professionals. The dataset is filtered for the roles labelled "Entry-level" under the experience level column, and the results are sorted by salary_in_usd in descending order. A bar chart to visualize these roles, with the salary level plotted against job titles, uses a vibrant color palette to enhance clarity.

This shows that, on average, "Research Engineer" among the junior data science positions fetches the highest salary; it is closely followed by "Applied Scientist" and then "Research Scientist." Other very lucrative positions that make their way to this list are "Machine Learning Engineer" and "Data Integration Specialist," indicating an increasing need for specialized skills in the domain of data science. From this bar chart, which is in the form of horizontal bars, the differentiations in salaries can be depicted very clearly and thus enable clear comparability across the various roles.

This analysis will be key to aspiring data scientists trying to find out which careers pay most, and for organizations that want to benchmark compensation for junior-level hires. It also shows which technical roles in research and applied machine learning are crucial in driving innovation and business value.

We now explore the relationship between company sizes and the salaries of the data scientists.

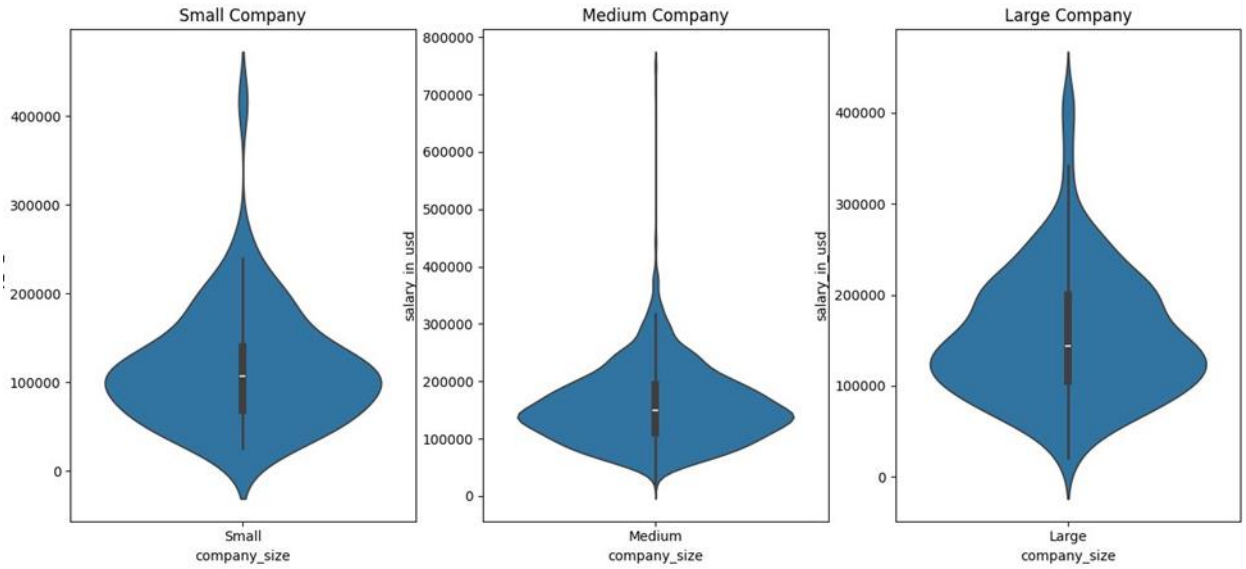

Figure 3: Violin plots showing salary distributions across different company sizes

The plots in Figure 3 reveal that there is not much difference in the range of salary distribution among different companies. The median distribution of salaries among the different sized companies is more or less, although the large companies have more people getting paid over 200,000 USD.

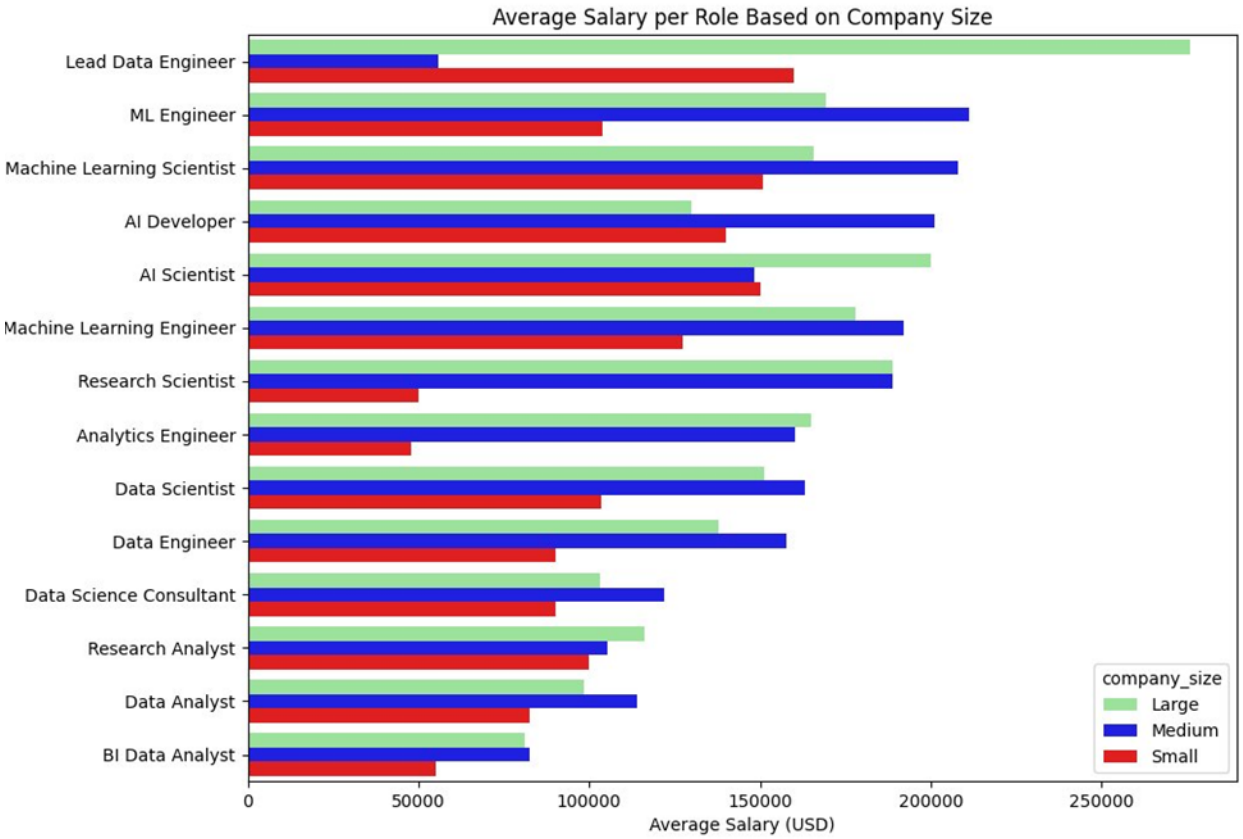

Figure 4: Violin plots showing salary distributions across different company sizes

This figure depicts that the salary range of data related roles is between 50,000 to 275,000 USD. It also shows that the Lead Data Engineers in large companies and ML Engineers in both medium and small sized companies are the highest paid. In contrast, Analytics Engineer in small companies, Lead Data Engineers in medium companies (surprisingly) and BI Data Analysts in large companies are found to be the least paid.

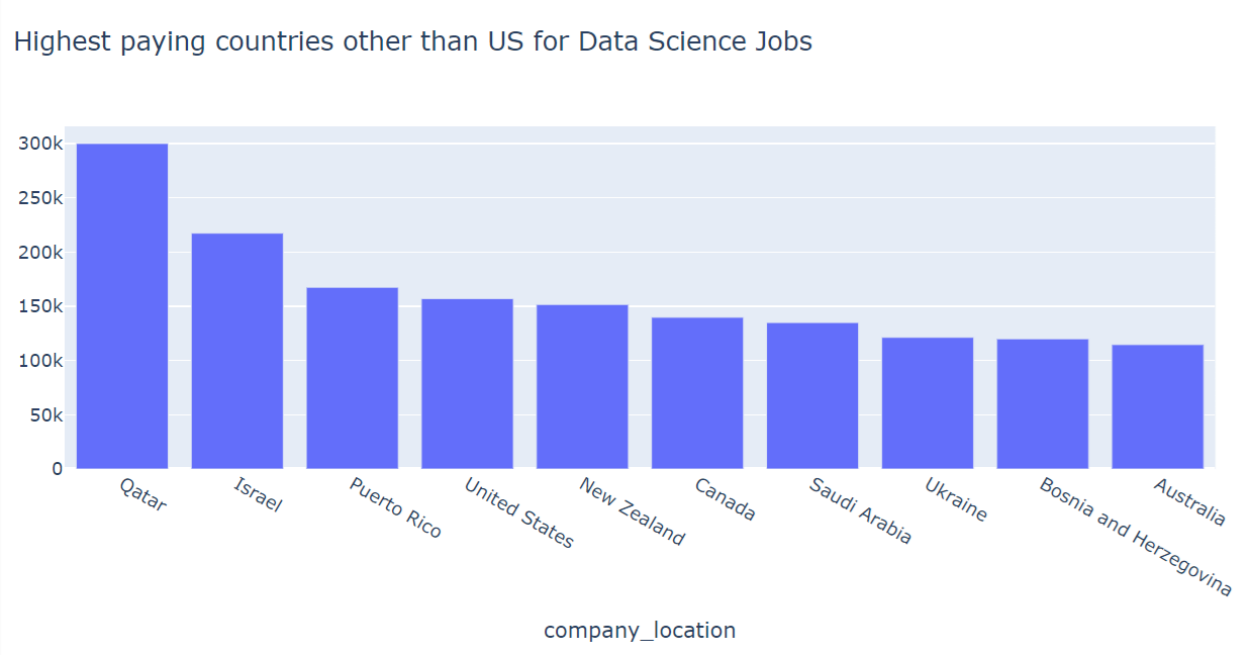

Figure 5: Bar plot showing the highest paying countries in the US for Data Science jobs

This figure illustrates that Qatar offers the highest-paying data science jobs, with the United States ranking fourth. However, it is important to note that, in practice, the U.S. is typically among the top-paying countries. This discrepancy might be attributed to the prevalence of contract-based jobs for workers from the top three countries, which could have skewed the results. Additionally, the presence of outliers in the real-time database used for recording data science salaries may have influenced the observed rankings.

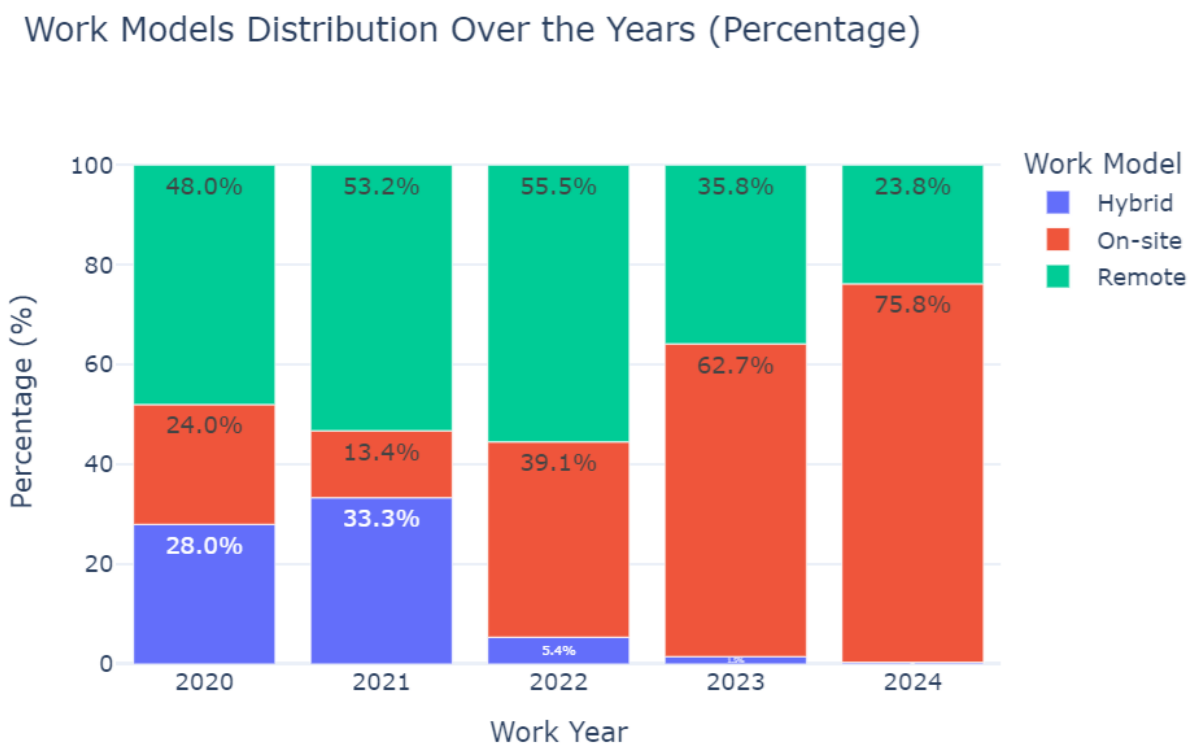

Figure 6: Stacked Bar plot showing the difference in work model (in percentage) from 2020

The figure depicts the distribution of work models for data scientists over the past four years, presented as percentages for on-site, remote, and hybrid arrangements. The data reveals a noticeable increase in remote jobs in 2022 compared to other years, while on-site work dominates in 2024. Additionally, the graph highlights the presence of hybrid work models from 2020 to 2022, demonstrating the evolving nature of workplace flexibility in the field.

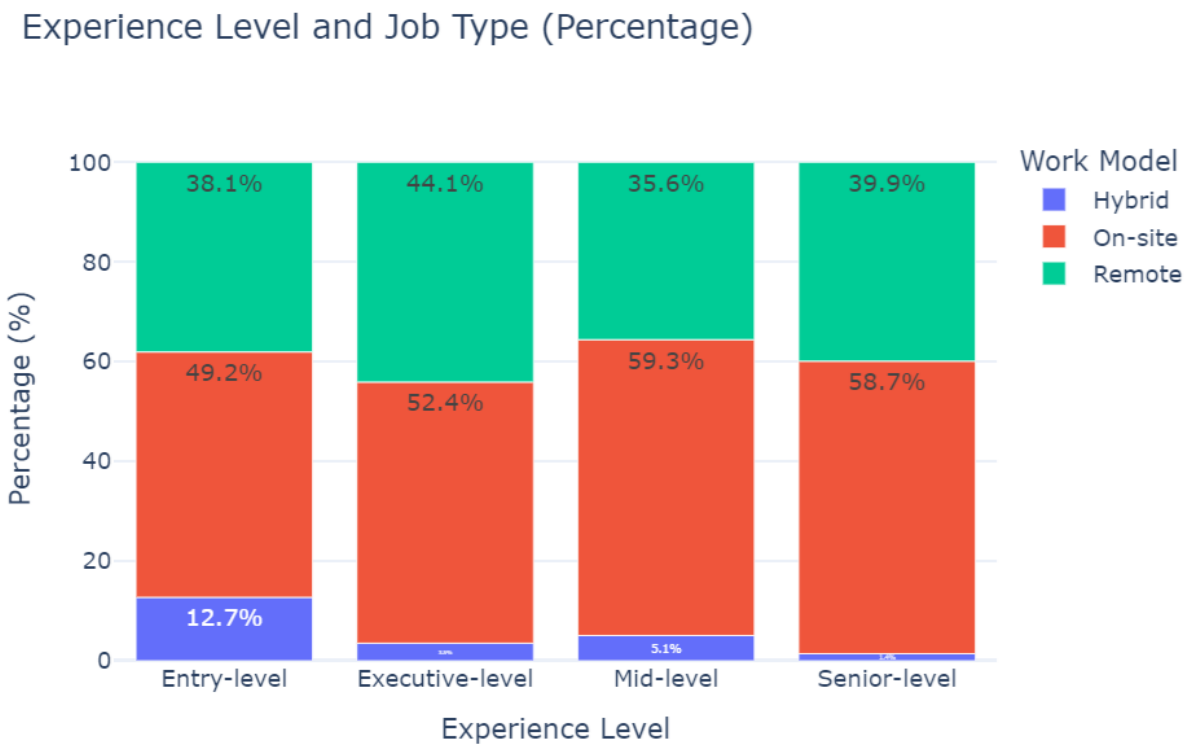

Figure 7: Stacked bar graph showing the work model based on experience level

The figure illustrates the distribution of work models based on employment experience levels. The data reveals an increase in on-site work for individuals at the experienced level, accompanied by a lower percentage of hybrid work. Notably, the executive level demonstrates a higher percentage of remote work compared to other experience levels, reflecting distinct preferences or opportunities at this professional tier.

Modeling

The goal was to develop logistic regression models to predict high salary jobs based on various job-related factors. The dependent variable was high salary, defined as salaries above the median value of $138,666. The independent variables included experience level, job title, company size, and work models, all of which were encoded for the analysis. We aimed to check the assumptions of logistic regression and assess the models' performance using different metrics.

The assumptions for logistic regression were checked, including linearity of the logit and multicollinearity. Linearity of the logit revealed that job title had a linear relationship with the logit, while other variables did not. Multicollinearity was assessed using the Variance Inflation Factor (VIF), and all VIF values were below 6.5, indicating that multicollinearity was not a

significant issue. A leverage vs. Cook's Distance plot showed that most data points had low leverage and influence, suggesting no undue impact from specific data points.

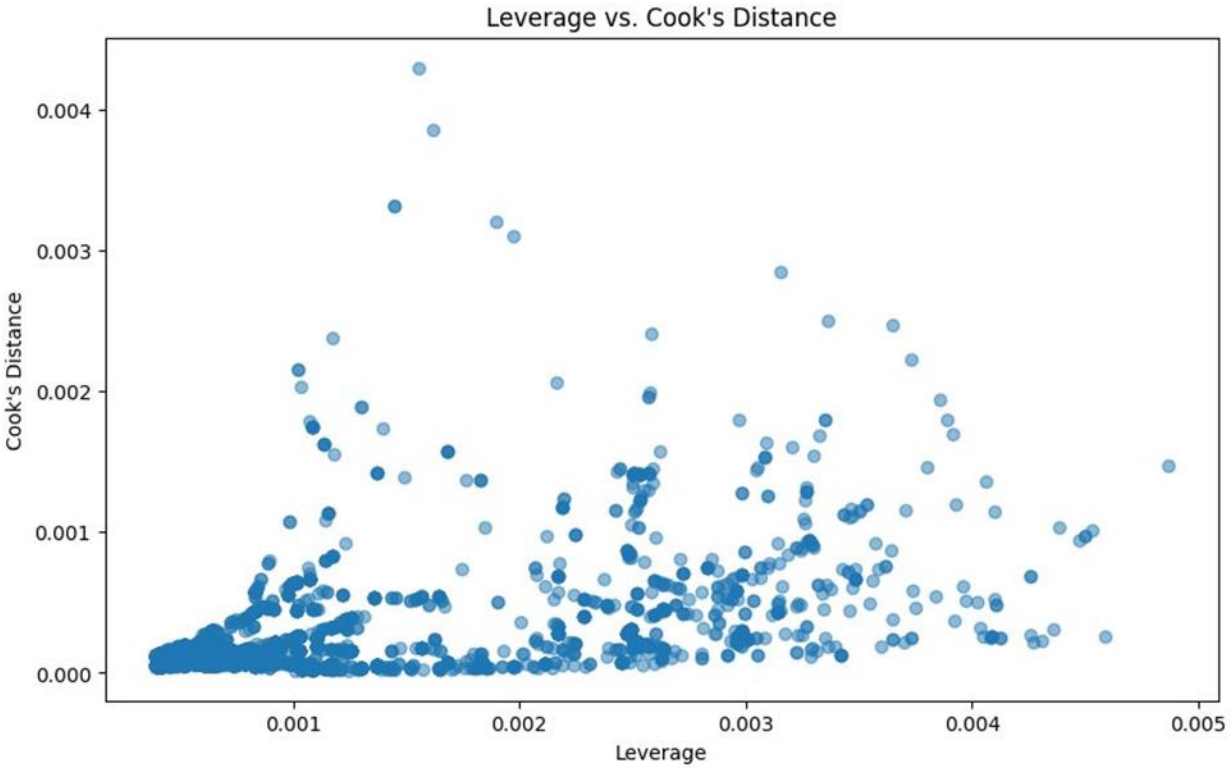

Figure 8: Leverage vs. Cook's distance

We conducted feature selection to identify the most significant predictors. In Model 1, all independent variables were included, with experience level and job title being significant. Model 2 included only experience level and job title, maintaining similar explanatory power (McFadden's R2: 0.086) and statistical significance (LLR p-value < 0.0001). Table 1 provides a clear comparison between the two models, highlighting the significance of the predictors and the model performance.

Table 1: Feature comparison between Model 1 and Model 2

Feature	***Model 1***	***Model 2***
Experience Level	Positive and significant ($p < 0.001$)	0.7305 ($p < 0.0001$)
Job Title	Positive and significant ($p < 0.001$)	0.0118 ($p < 0.0001$)
Company Size	Not significant ($p = 0.396$)	Not included
Work Models	Borderline significant ($p = 0.057$)	Not included

Model Performance	Explains about 8.6% of variance in high salary	Explains approximately 8.6% of variance in high salary
Overall Model Fit	Not specified	Statistically significant (LLR p-value < 0.0001)

Both models indicate that experience level and job title are significant predictors of high salary. Removing non-significant predictors does not affect the explained variance significantly.

Model Assessment

We now consider Model 2 as base model and consider another model with polynomial features to add higher-degree terms that help model non-linear patterns more effectively. Table 2 summarizes the model assessments.

The base model included all the independent variables. Our first model was decent at predicting which jobs offer high salaries. It got about 63% of its predictions right, and it was good at finding most of the high salary jobs, but it also had a fair number of false alarms. The model could still be improved, especially in distinguishing jobs that don't offer high salaries. Our second model, which included more complex features, didn't show much improvement over the first one. It had similar accuracy and was just as good at finding high salary jobs, but it didn't significantly reduce the number of false alarms. The added complexity didn't make a big difference.

Table 2: LR Model Assessment using Confusion matrix and ROC-AUC values

Metric	***Base Model (Model 2)***	***Model with Polynomial Features***
Accuracy	62.97%	63.02%
Precision	61.06%	61.15%
Recall	73.46%	73.31%
F1 Score	66.69%	66.68%
Specificity	52.28%	52.54%
ROC-AUC	0.696	0.696
McFadden's R2	0.085	0.085
AIC	6699.33	6701.24
BIC	6719.04	6727.53

These results indicate that the base model had a fair balance between precision and recall, capturing most high salary jobs, though the specificity and ROC-AUC suggested moderate discriminative ability. Adding polynomial features did not significantly enhance model performance, as indicated by similar metrics and slightly higher AIC and BIC values.

To improve model performance, we explored adjusting the threshold value for classifying high salary jobs. By identifying the optimal threshold (0.552), we achieved the following performance metrics:

- **Accuracy:** 65.94%
- **Precision:** 66.93%
- **Recall:** 64.26%
- **F1 Score:** 65.57%
- **ROC-AUC:** 0.696

The results suggest that the adjusted threshold improved the balance between true positives and false positives, leading to better overall performance.

We generated Confusion Matrix to compare the two models (Figure 11), where:

True Positives (TP)	**Top-left cell - correctly predicted high salary jobs.**
False Positives (FP)	Top-right cell - jobs predicted as high salary but are not.
False Negatives (FN)	Bottom-left cell - high salary jobs predicted as low salary.
True Negatives (TN)	Bottom-right cell - correctly predicted low salary jobs.

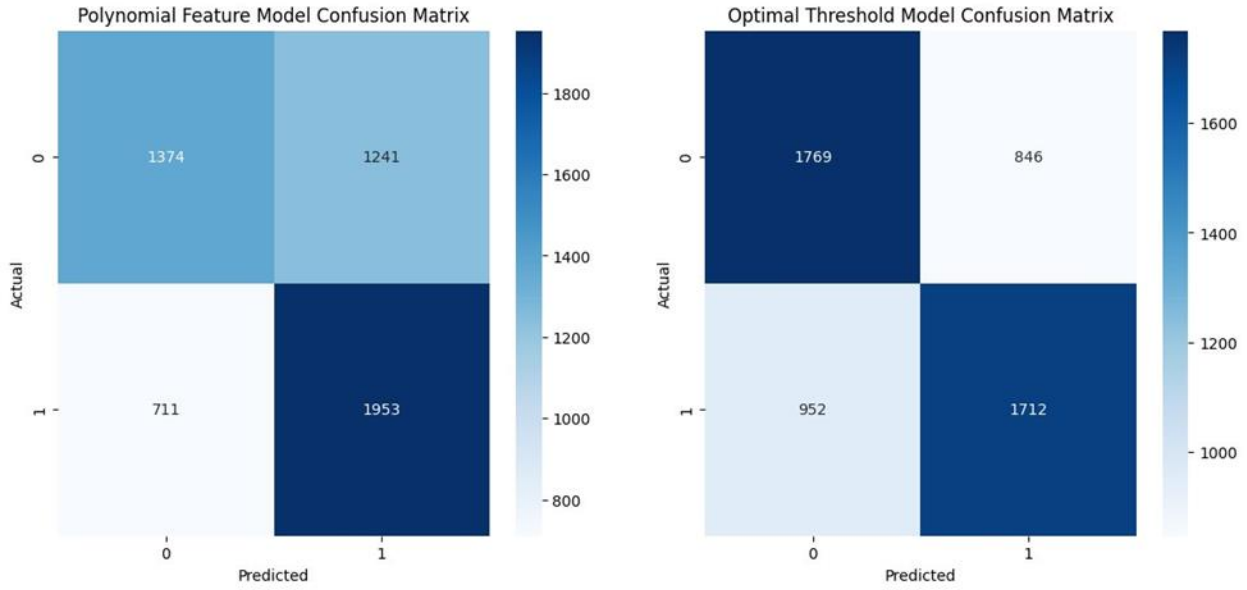

Figure 9: Confusion Matrix comparison of the two models

The confusion matrix indicates that adjusting the threshold improved the model's ability to predict high salary jobs correctly (more TP), but it also resulted in missing some high salary jobs (more FN). However, fewer false positives (FP) were made, indicating that fewer low salary jobs were misclassified as high salary.

The ROC Curve was drawn to evaluate the overall model's performance (Figure: 12). ROC curve area of 0.7 indicates that the model has moderate discriminatory ability, meaning it correctly distinguishes between positive and negative classes about 70% of the time.

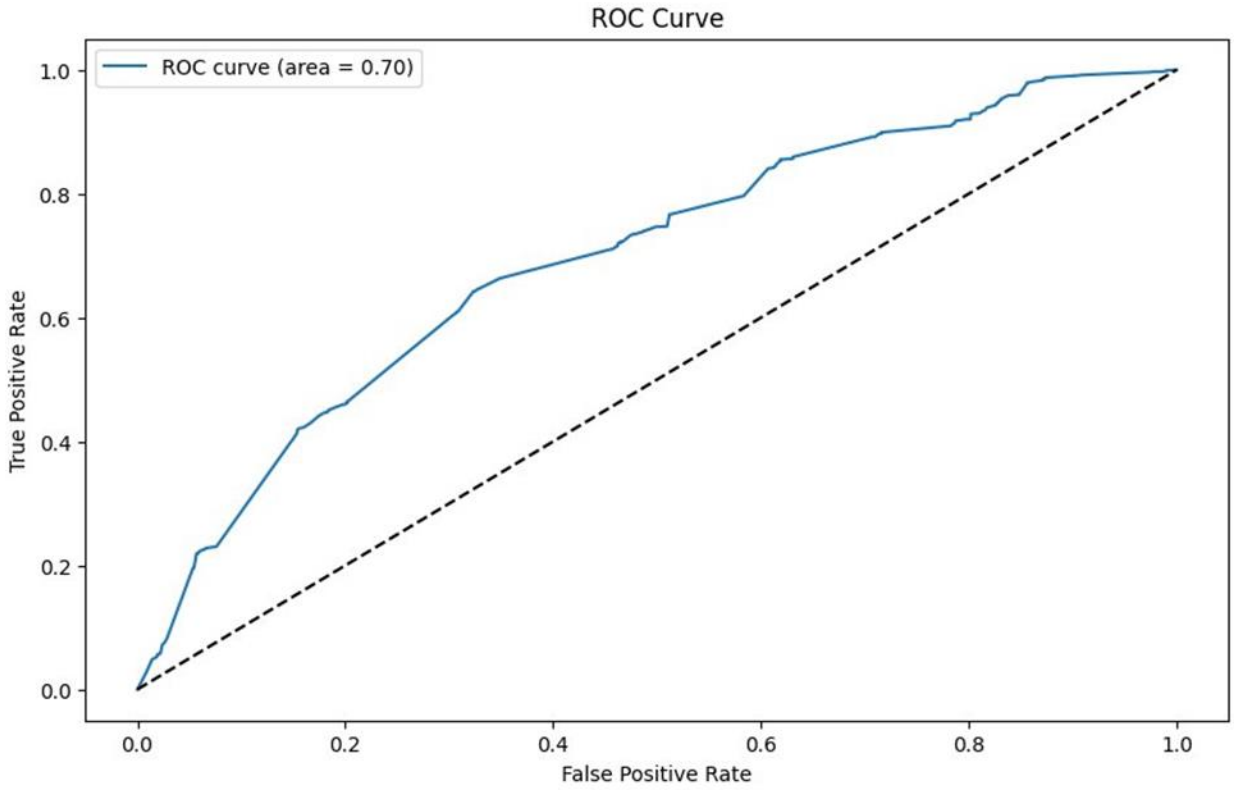

Figure 10: ROC curve of both models

Interpretation

Both models showed moderate effectiveness in predicting high salary jobs. The accuracy, precision, and recall values indicated that the models were fairly accurate in identifying high salary jobs, although they made some mistakes. The ROC-AUC and specificity values suggested room for improvement. Overall, our analysis provides a solid foundation for predicting high salary jobs, but there's potential for further improvements.

Limitations and Scope of Improvement

While they were fairly accurate in identifying high salary jobs, there was room for improvement in terms of discriminative ability. Future steps include exploring alternative machine learning algorithms, further optimizing the threshold, and refining the feature set to enhance predictive power. This analysis lays a solid foundation for predicting high salary jobs but indicates potential for further advancements. We could try more advanced models (e.g., Random Forest, XGBoost) for better performance.

Conclusion

Based on the above results and observations, the classification model and other statistical analysis demonstrates effectiveness in the model's performance to predict new test values. However, there is certainly room for improvement. This paper effectively demonstrates the entire pipeline of handling data and predicting new outcomes based on test variables.

References

1. Kaggle. (2023). Data Science Salaries Dataset. Retrieved from https://www.kaggle.com/datasets/data-science-salaries-dataset

2. Bureau of Labor Statistics (2023) - Occupational Outlook Handbook: Computer and Information Research Scientists. U.S. Department of Labor.

3. GitHub (2023) - Salary Insights Repository. Open-source salary analysis projects.

4. Stack Overflow (2023) - Annual Developer Survey. Developer Salary and Job Market Trends.

5. IEEE (2024) - Computational Intelligence and Data Science Employment Trends Report.

6. O'Reilly Media (2023) - Data Science Salary and Career Development Report.

7. Glassdoor Economic Research (2023) - Tech Salary Reports.

8. Coursera (2024) - Global Skills Report: Data Science and AI Trends.

9. McKinsey Global Institute (2023) - Technology Workforce Insights.

10. World Economic Forum (2024) - Future Jobs Report: Technology and Data Science Sector.