Analysis of Data Science Jobs Between 2020 and 2023

DATA 601 Final Presentation

Akinyemi Apampa & David Fakolujo

February 13, 2025

Table of Content

- Introduction
- Datasets
- Guiding Questions
- Data Cleaning and Preprocessing
- Exploratory Data Analysis
- Conclusion
- References

Introduction

- Data Science is a broad field that uses scientific methods such as programming and statistics to extract meaningful insights from different forms of data.
- The field can be applied across diverse sectors to support data-driven decision-making.
- This project explores the attributes surrounding Data Science jobs between 2020 and 2023.
- The insights generated from this dataset's exploration will be valuable to professionals seeking to transition into the data science field, as well as those already in the field, looking to widen their expertise or transform within the field.

Datasets

- The primary analysis will be conducted using a dataset from Kaggle, which contains data on data science jobs between 2020 and 2023.
- Singular, structured dataset contains 3,755 rows and 11 columns consisting of key variables related to jobs in the data science field.
- A complementary analysis will focus on the skills variable from a separate dataset retrieved from Kaggle.
- The dataset contains 12,217 rows that each contain skills required for data science related jobs posted on LinkedIn in January 2024.

Datasets

- **work_year** : the year the salary was paid
- **experience_level**: The experience level in the job during the year
- **employment_type**: The type of employment for the role
- **job_title**: The role worked in during the year
- **salary**: The total gross salary amount paid
- **salary_currency**: The currency of the salary paid as an ISO 4217 currency code
- **salaryinusd**: The salary in USD
- **employee_residence**: Employee's primary country of residence in during the work year as an ISO 3166 country code
- **remote_ratio**: The overall amount of work done remotely
- **company_location**: The country of the employer's main office or contracting branch
- **company_size**: The median number of people that worked for the company during the year

Guiding Questions

1. What has been the growth rate in Data Science jobs in the past four years?

- This would inform people on whether Data Science has been a growing or declining field over the past four years. It would also provide insights to the availability and importance of Data Science jobs globally.

2. What is the overall distribution of salaries in the dataset, as well as the salary distribution across categorical variables, such as experience level, company size, and remote status?

- This analysis would give Data Scientists an idea of the salary expectations as they grow in experience and interact with different companies. It could also help with salary discussions before accepting job offers from companies.

3. What has been the trend over time of remote jobs in the Data Science field?

- Many companies now offer the option to work remotely. This analysis would give a job seeker an idea of the likelihood of securing a remote job.

Guiding Questions

4. What has the distribution of Data Science job titles been over the years, as well as the distribution per experience level and remote status?

- This analysis can inform job seekers which roles are more popular and which roles are scarce. It can also guide people on which jobs to apply for, based on the demand for certain job titles.

5. What are the in-demand skills in the Data Science field?

- This would help job seekers have an idea of which skills to focus on to increase employability and to stay relevant in the Data Science field.

6. What are the proportions of employment types in the Data Science field?

- This insight can help job seekers understand the distribution of Data Science roles in organizations and identify which employment types are most and least preferred. This can also improve job search by targeting specific job platforms that post jobs with specific employment types.

Data Cleaning and Preprocessing

We performed some data cleaning and transformation processes on our dataset before performing the exploratory data analysis. Some of these included:

- Dropping Irrelevant Columns
- Grouping the job titles into major job categories
- Replacing abbreviations in the columns with their full meaning
- Converting remote ratio column to categorical column
- Filtering to include just the jobs in North America

Exploratory Data Analysis (EDA)

Our Exploratory Data Analysis on the Dataset included:

- Generating summary statistics for numerical features
- Using the Inter Quartile Range (IQR) method to identify outliers
- Data Visualizations that answer the guiding questions
- Statistical Analysis (T-Test, Chi-Squared Test)

Exploratory Data Analysis - Summary Statistics

The following observations were gotten from the summary:

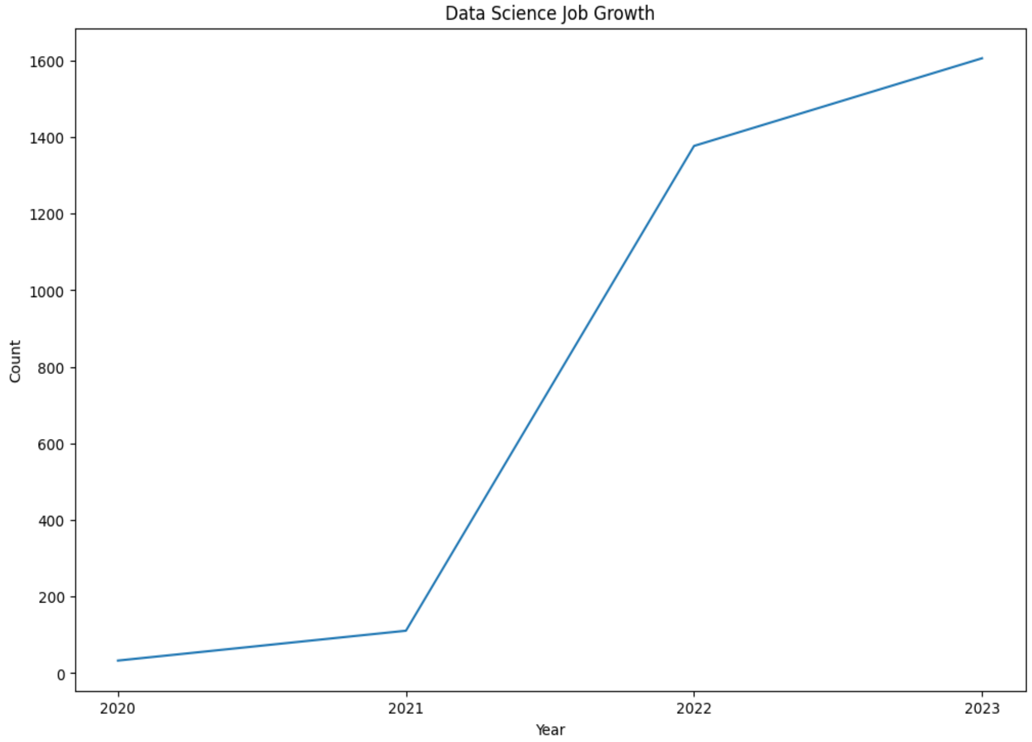
- Average salary in cad is 216,313.57 while the standard deviation is 80,442.33. The high standard deviation shows that the salaries have a high variation.
- The minimum salary is 8,120.97, which could indicate a part-time/contract/freelance job, while the maximum salary is 643,500.00, which could indicate the salary paid by a large company to employees with the highest experience level.
- 25% of the data earn less than 161,447.00, while 25% earn more than 264,550.00. The median salary is 207,350.00.
- The mean being greater than the median indicates that the salary distribution is right -kewed. This is reasonable, as fewer employees are expected to earn the highest salaries.

Exploratory Data Analysis - Identifying Outliers

We used the IQR method to determine and identify outliers in the salary_in_cad column, the salaries of each position in Canadian dollars. Based on our analysis, we determined that there were about 60 outliers in that column. The minimum outlier was $423,159 and the maximum outlier was $643,500. Based on the minimum and maximum outliers above, these values likely represent expected salaries that large companies could pay employees at the highest experience levels. As a result, we removed these values from the dataset, as they could also provide meaningful and important insights.

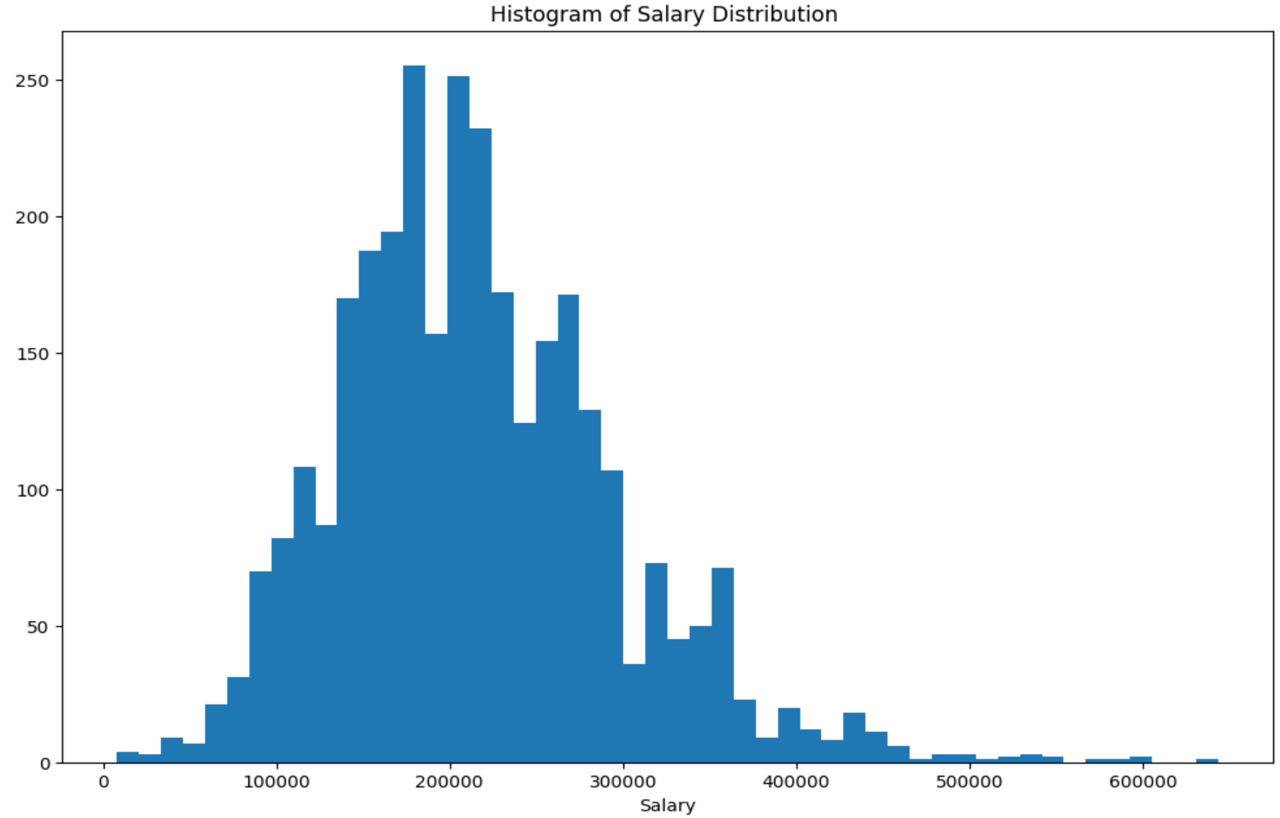
Exploratory Data Analysis - Data Visualization

Question 1: What has been the growth rate in data science jobs over the past four years?

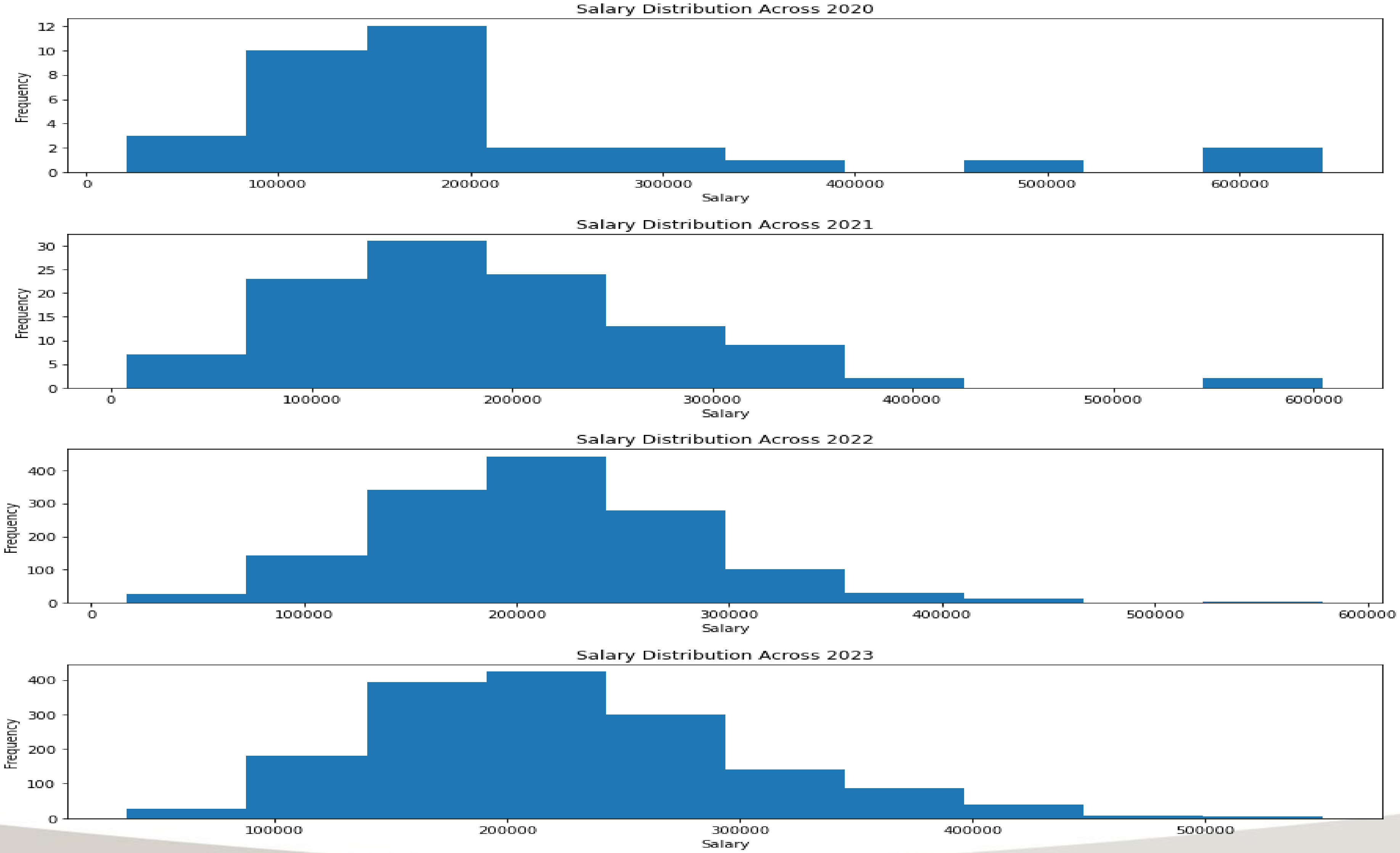


Exploratory Data Analysis - Data Visualization

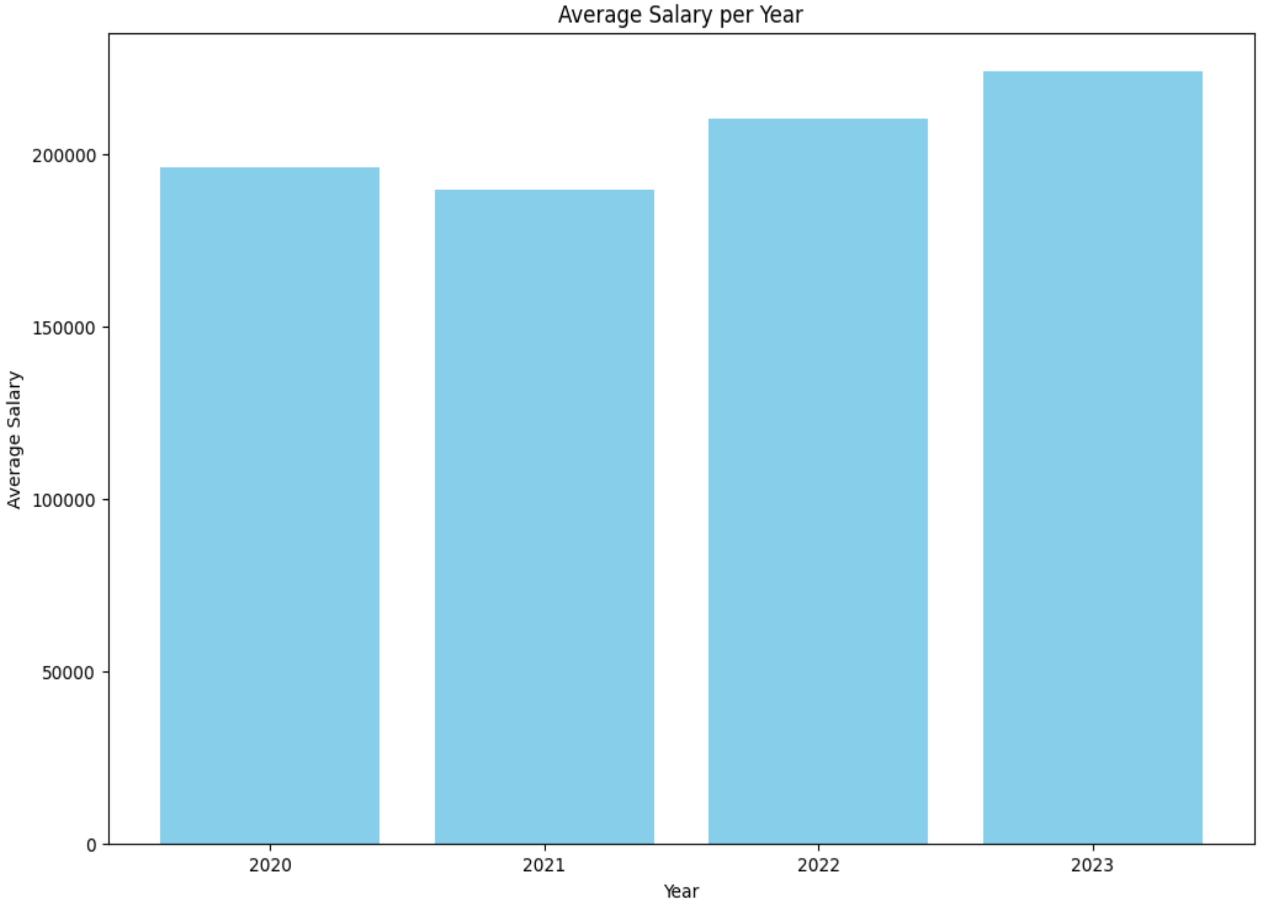
- Question 2: What is the overall distribution of salaries in the dataset, as well as the salary distribution across categorical variables, such as experience level, company size, remote status?



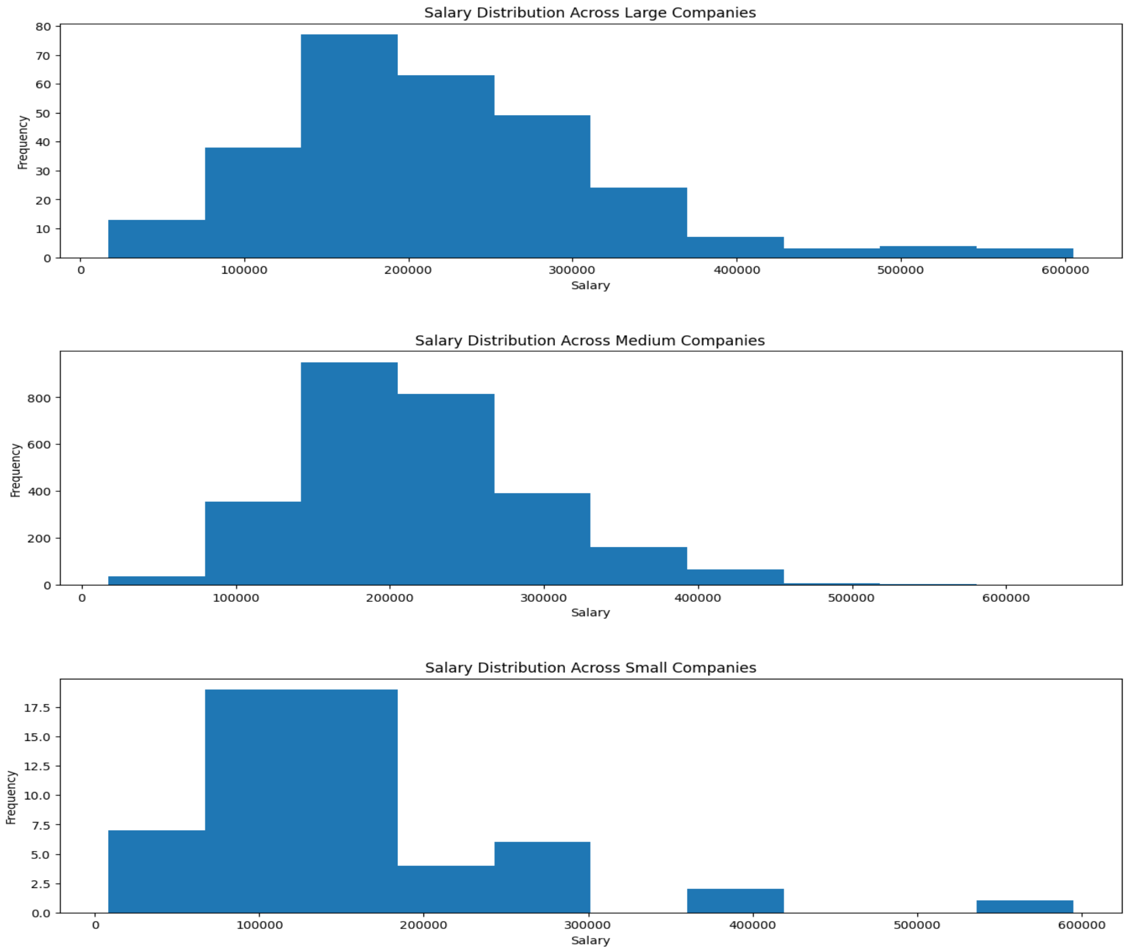
Exploratory Data Analysis - Data Visualization



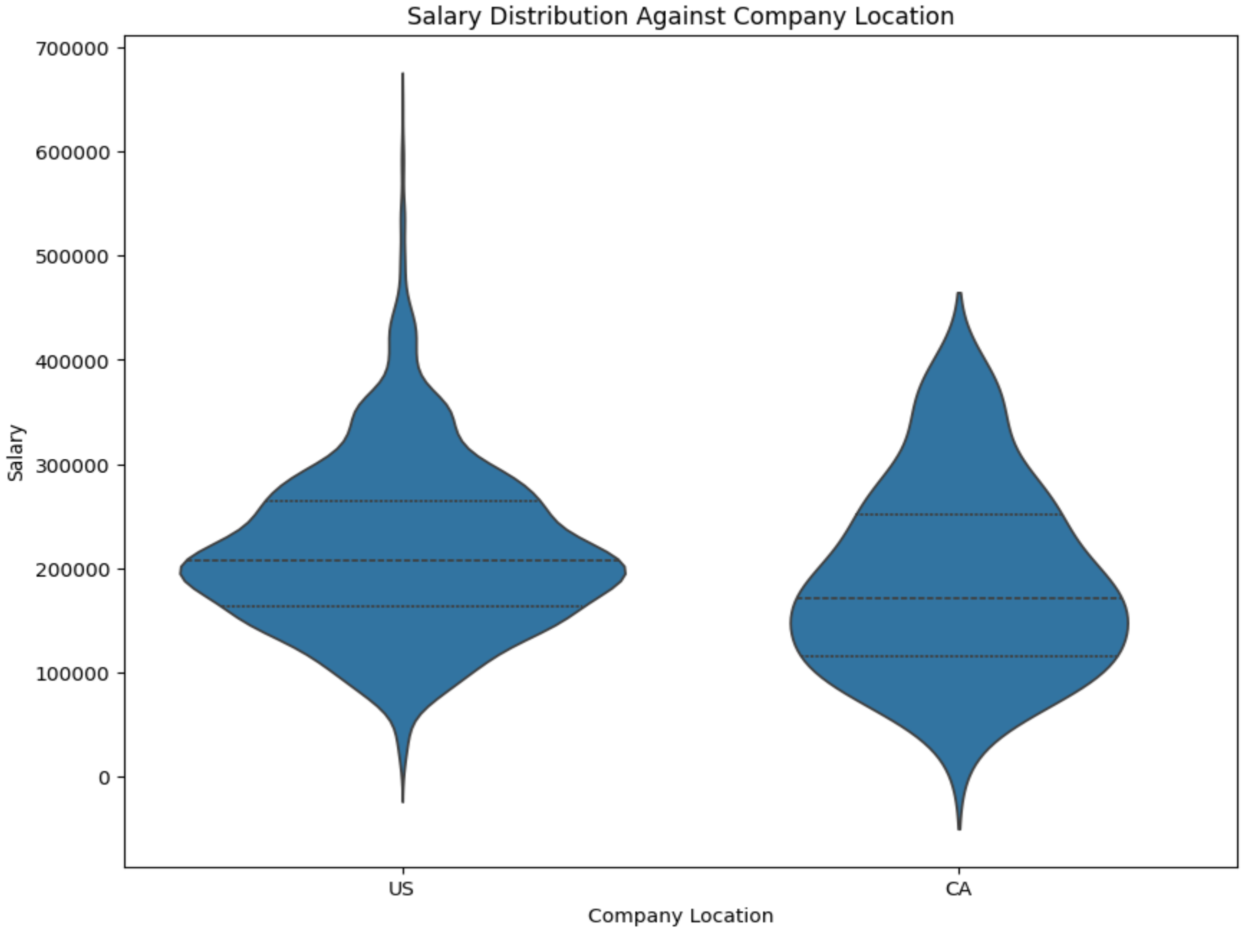
Exploratory Data Analysis - Data Visualization



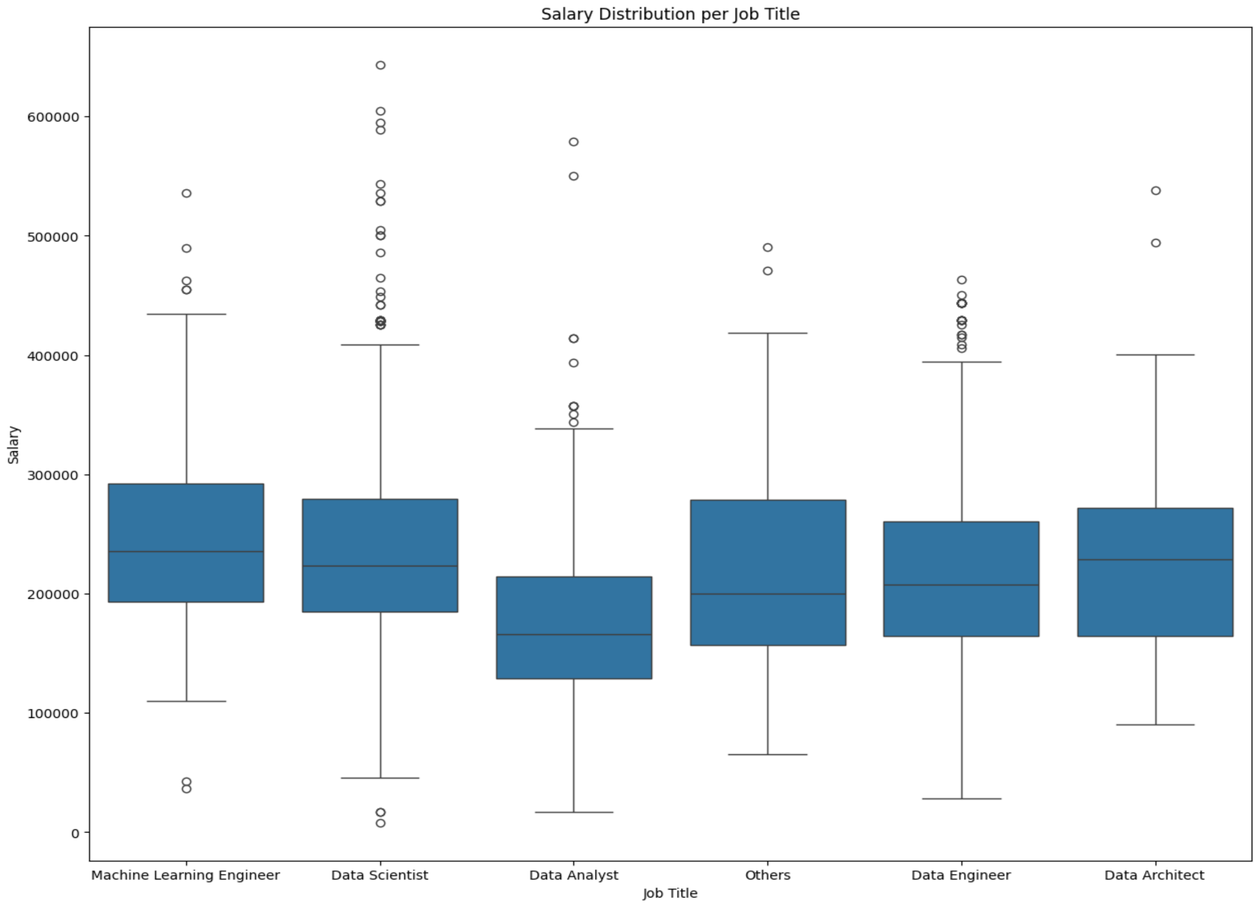
Exploratory Data Analysis - Data Visualization



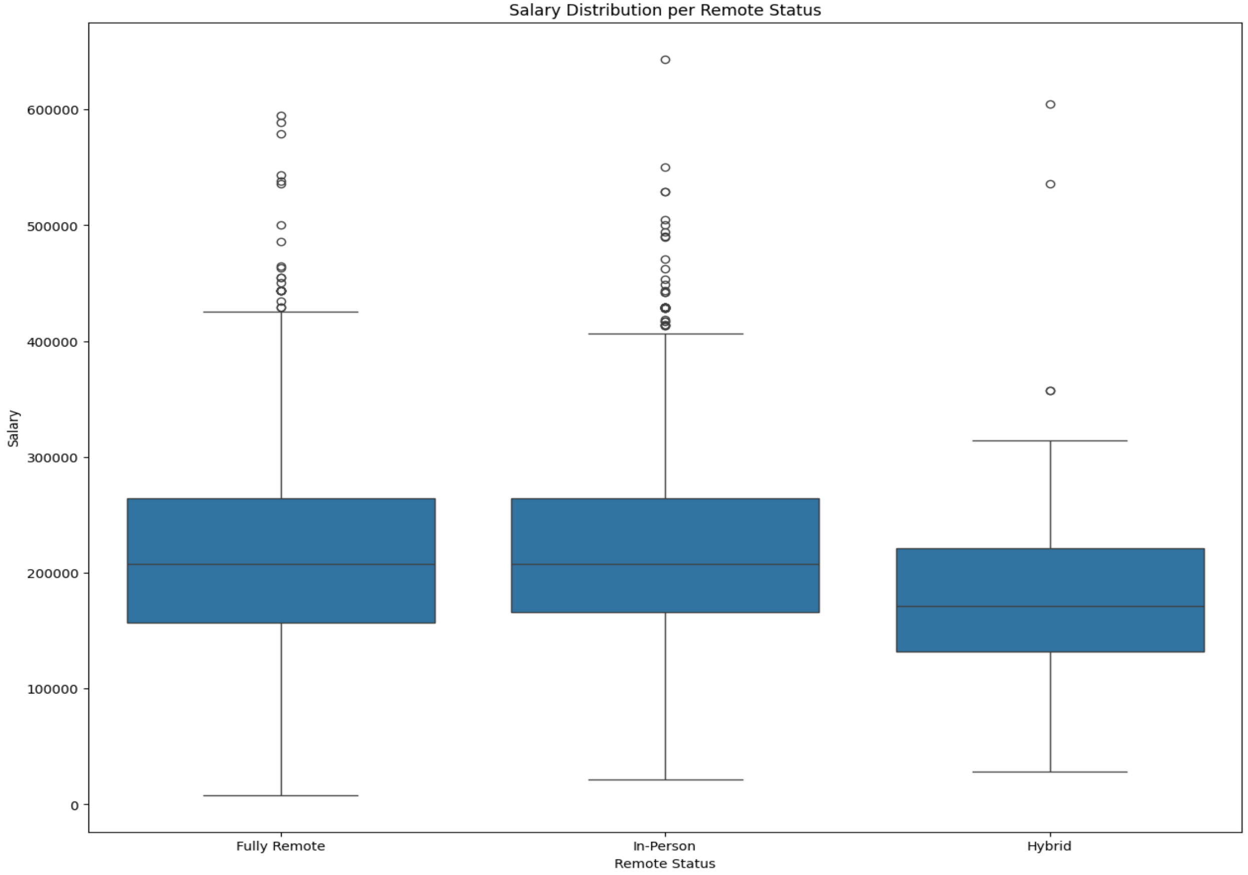
Exploratory Data Analysis - Data Visualization



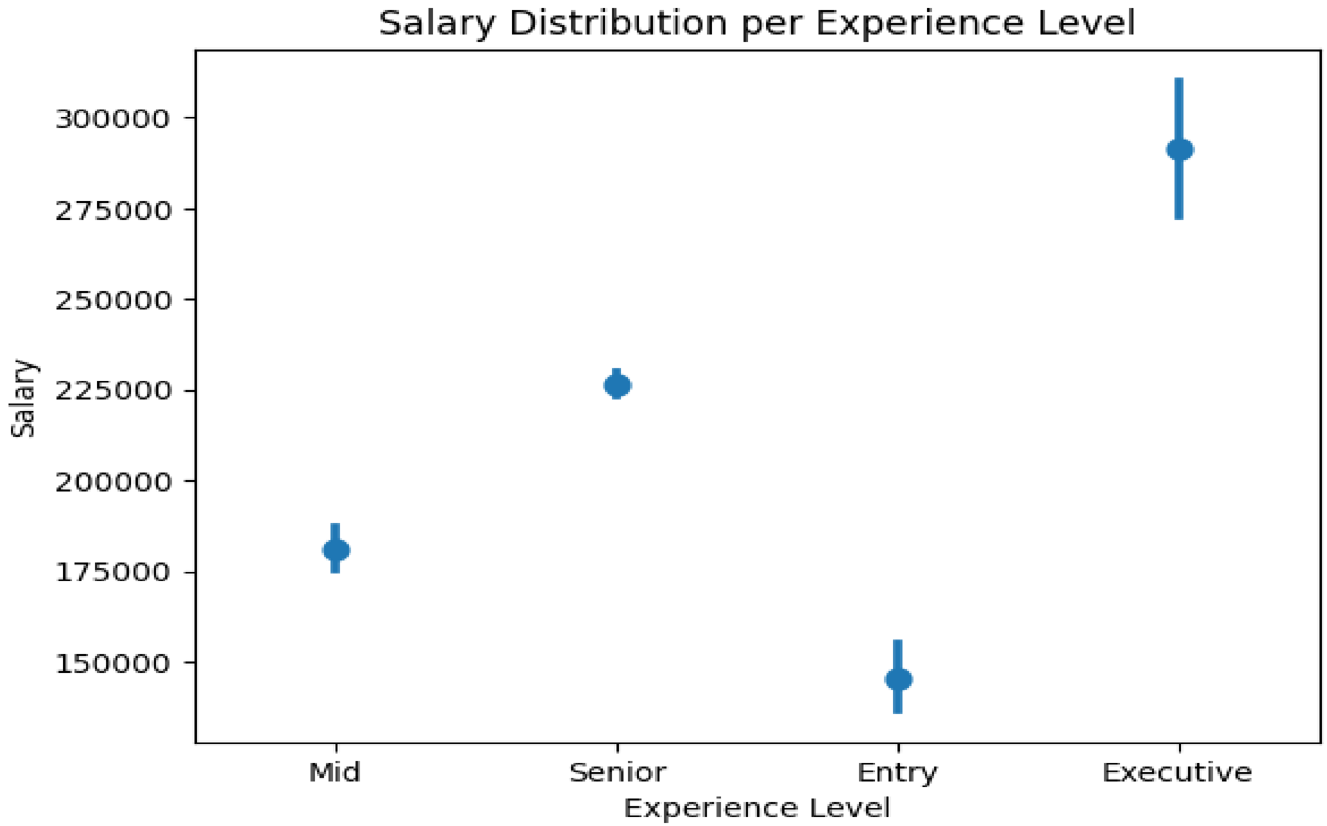
Exploratory Data Analysis - Data Visualization



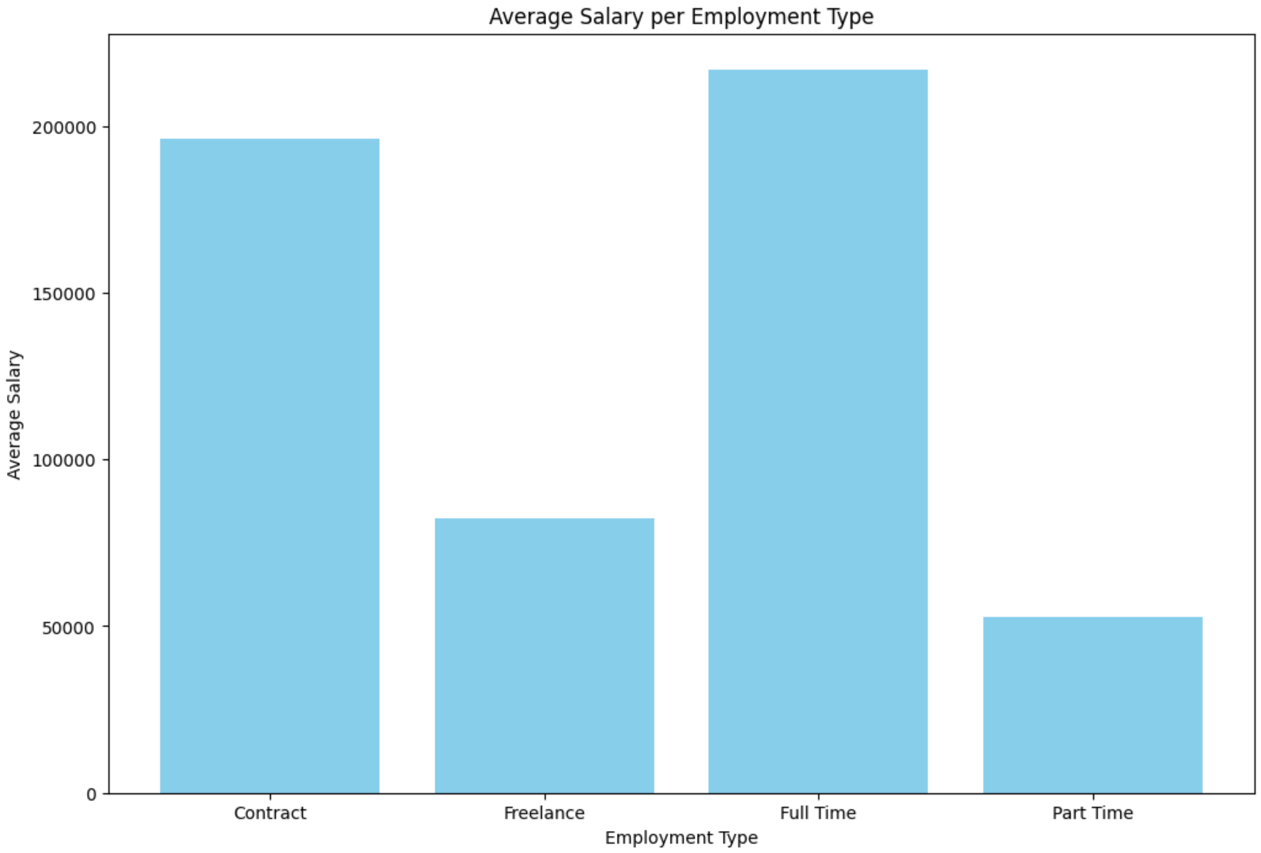
Exploratory Data Analysis - Data Visualization



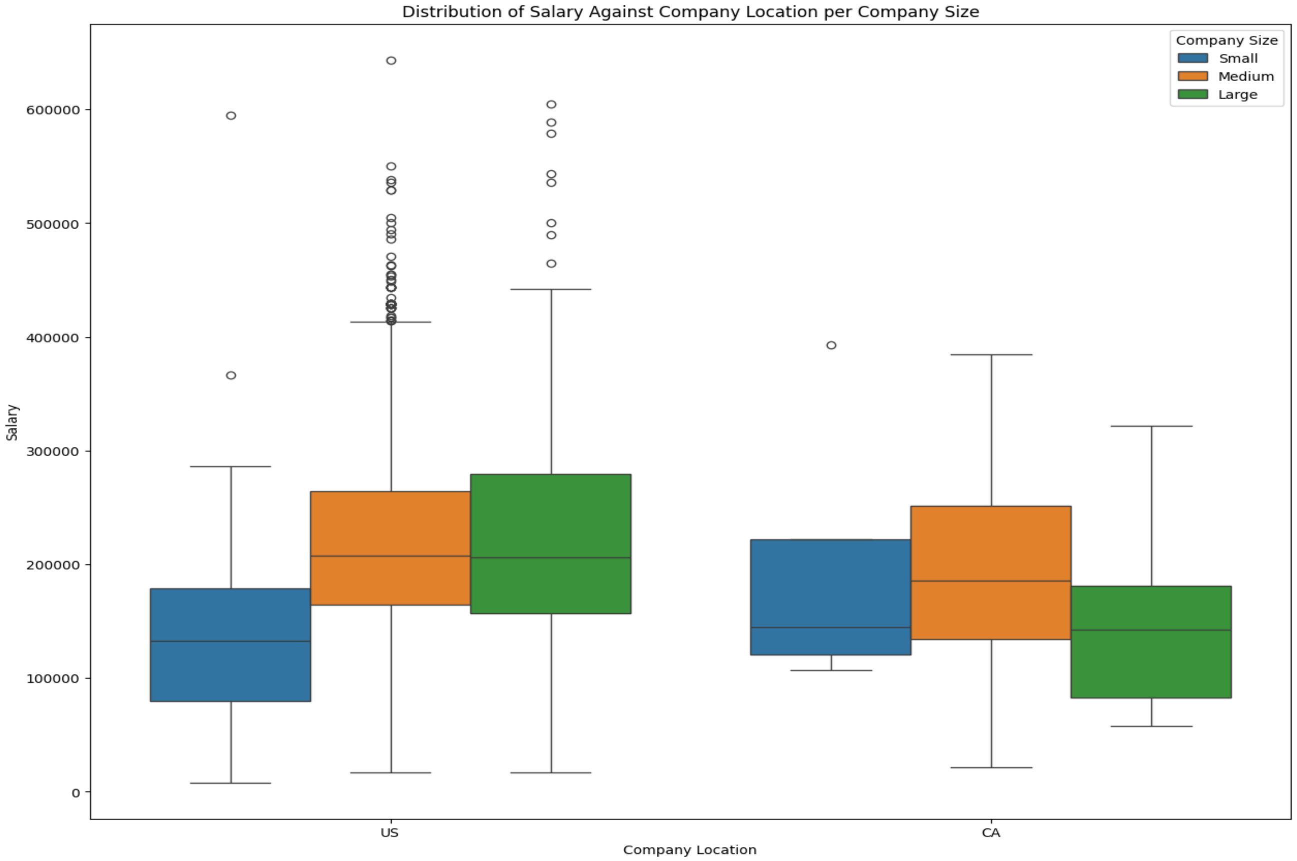
Exploratory Data Analysis - Data Visualization



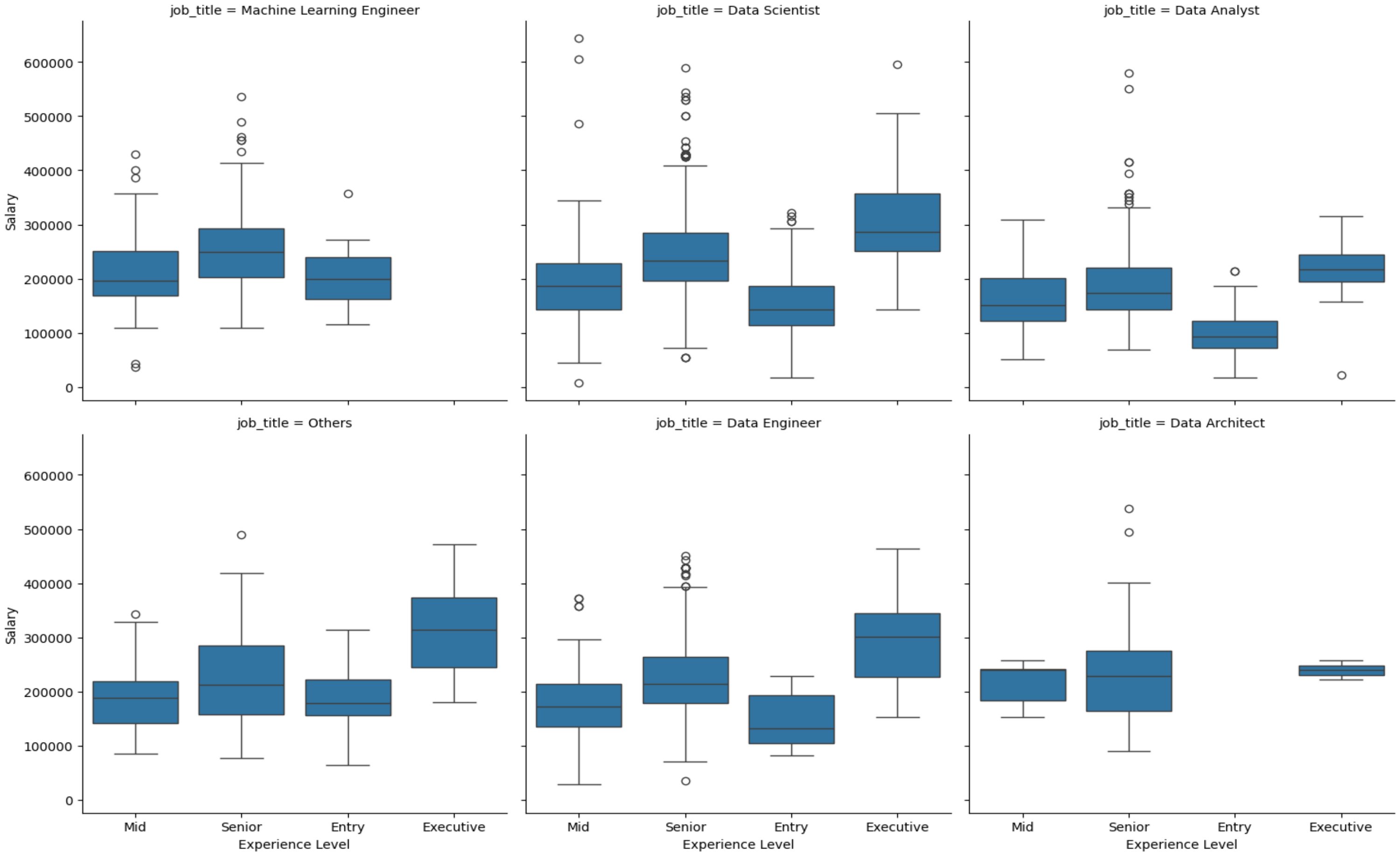
Exploratory Data Analysis - Data Visualization



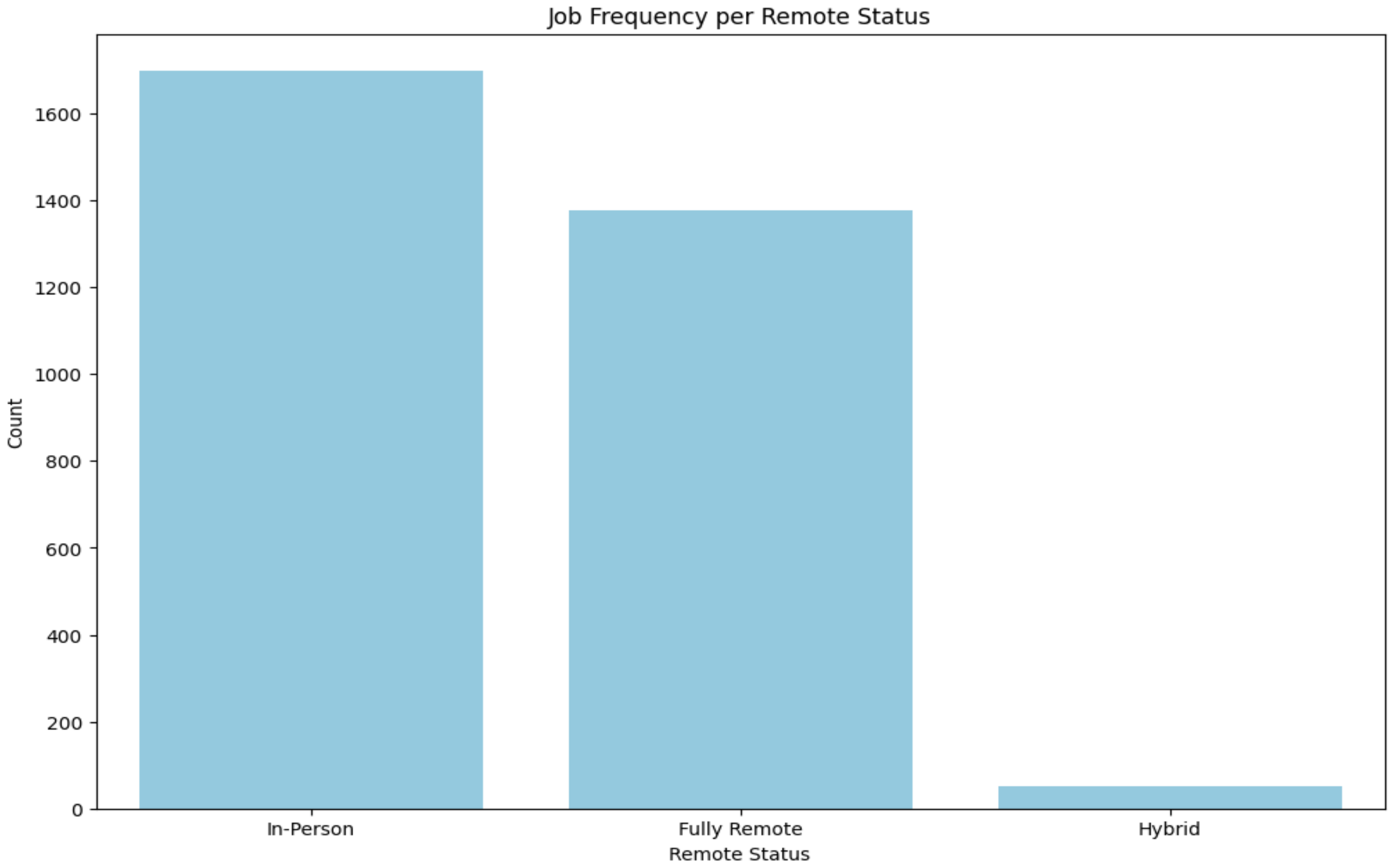
Exploratory Data Analysis - Data Visualization



Exploratory Data Analysis - Data Visualization

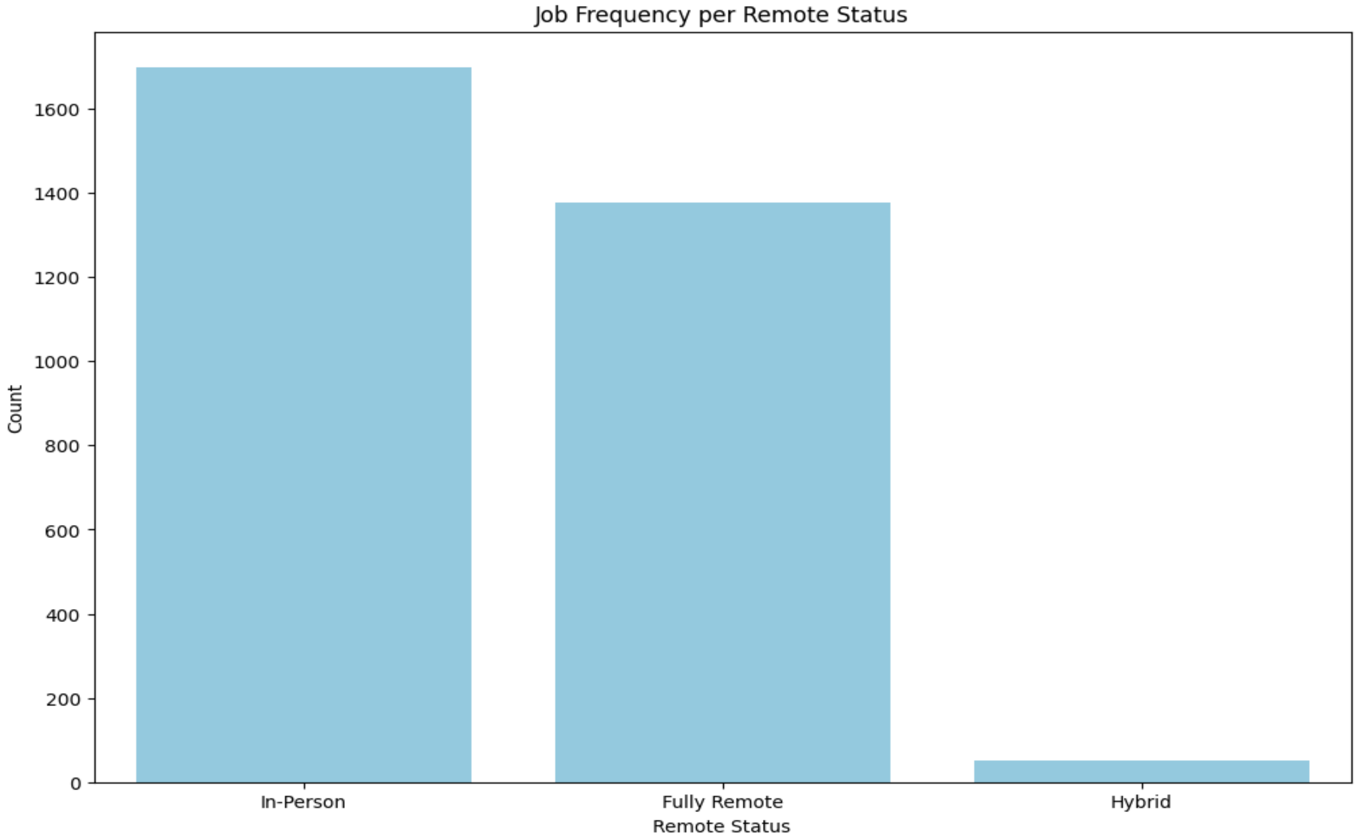


Exploratory Data Analysis - Data Visualization

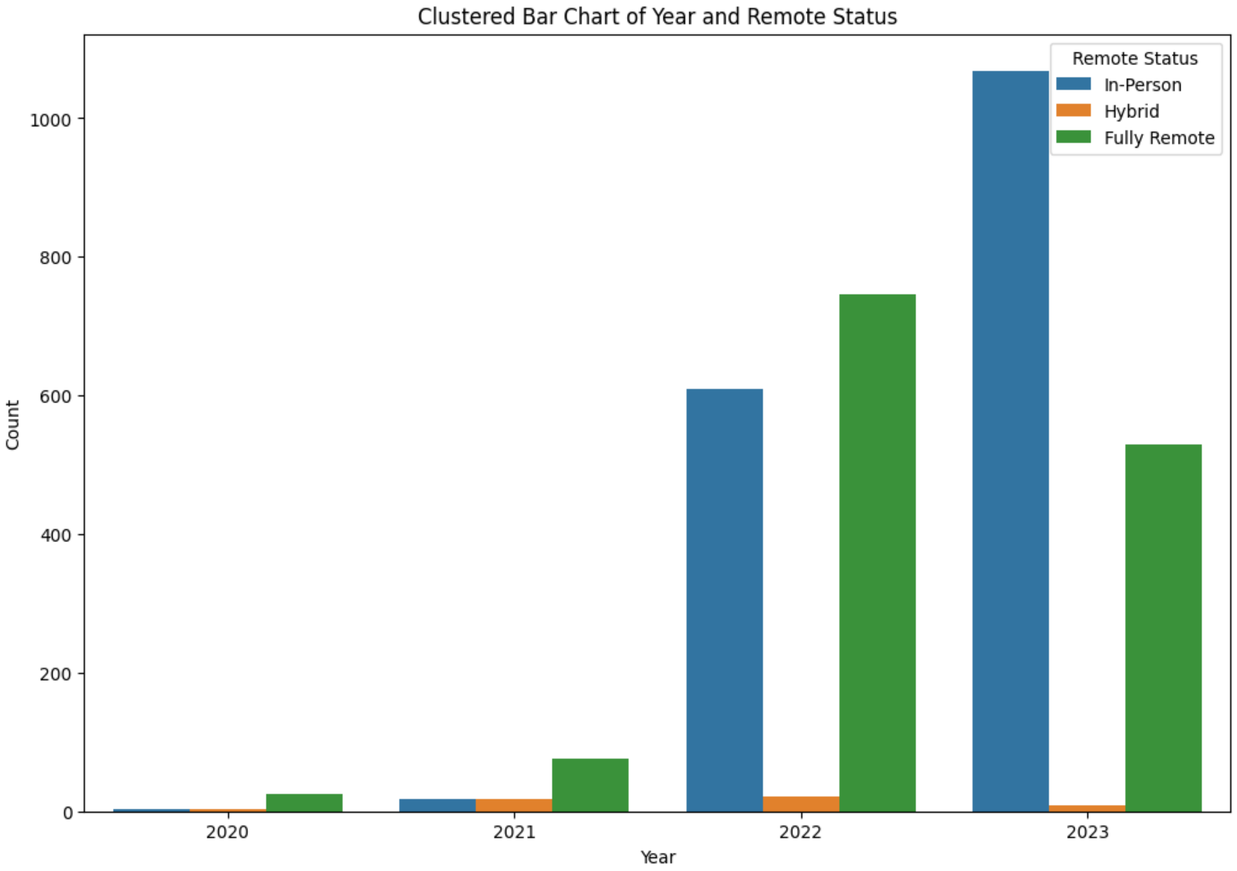


Exploratory Data Analysis - Data Visualization

- Question 3: What has been the trend over time of remote jobs in the Data Science field?

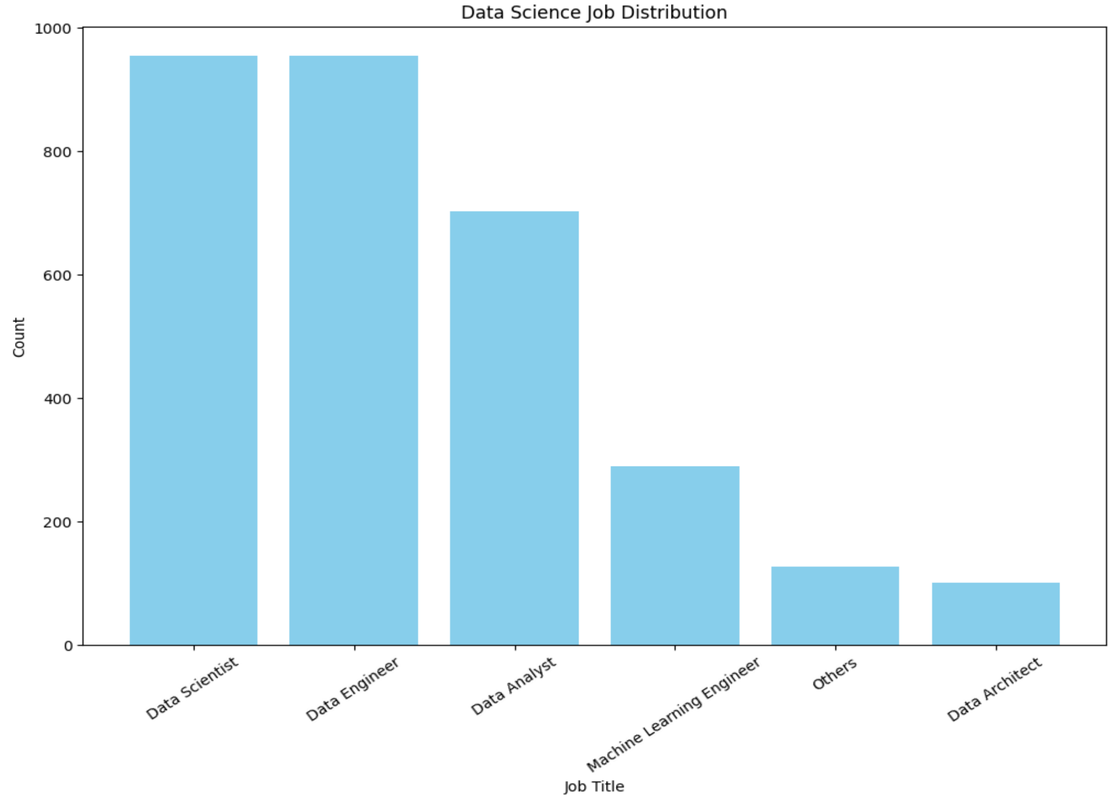


Exploratory Data Analysis - Data Visualization

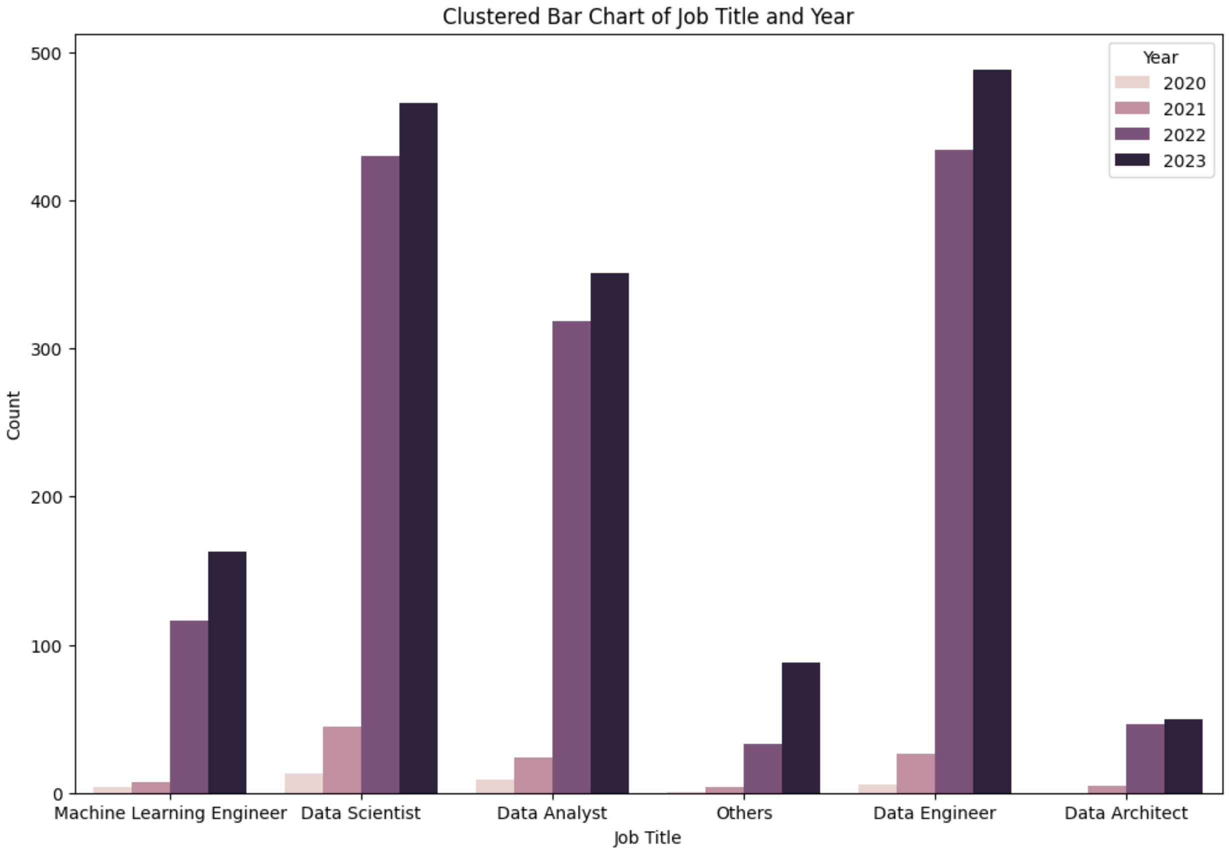


Exploratory Data Analysis - Data Visualization

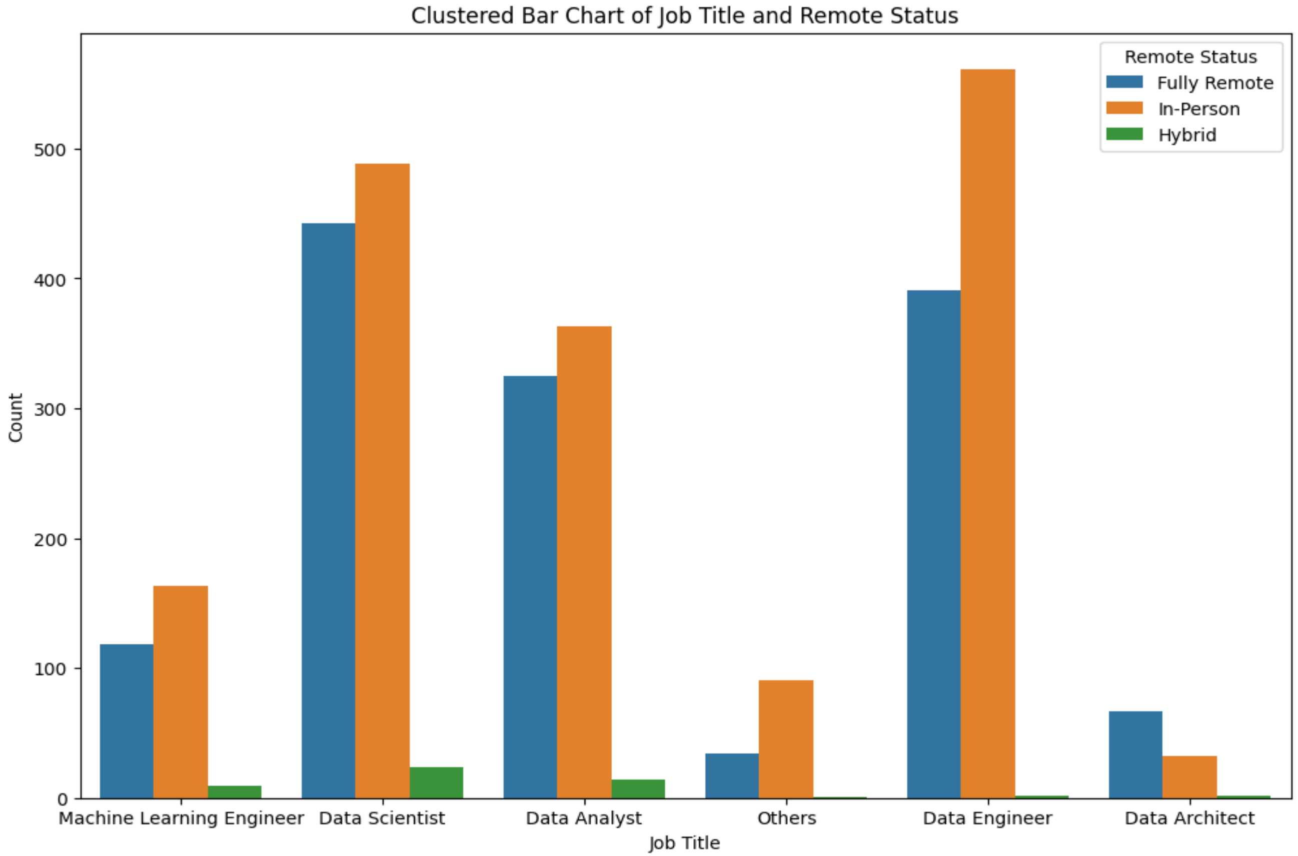
Question 4: What has the distribution of Data Science job titles been over the years, as well as the distribution per experience level and remote status?



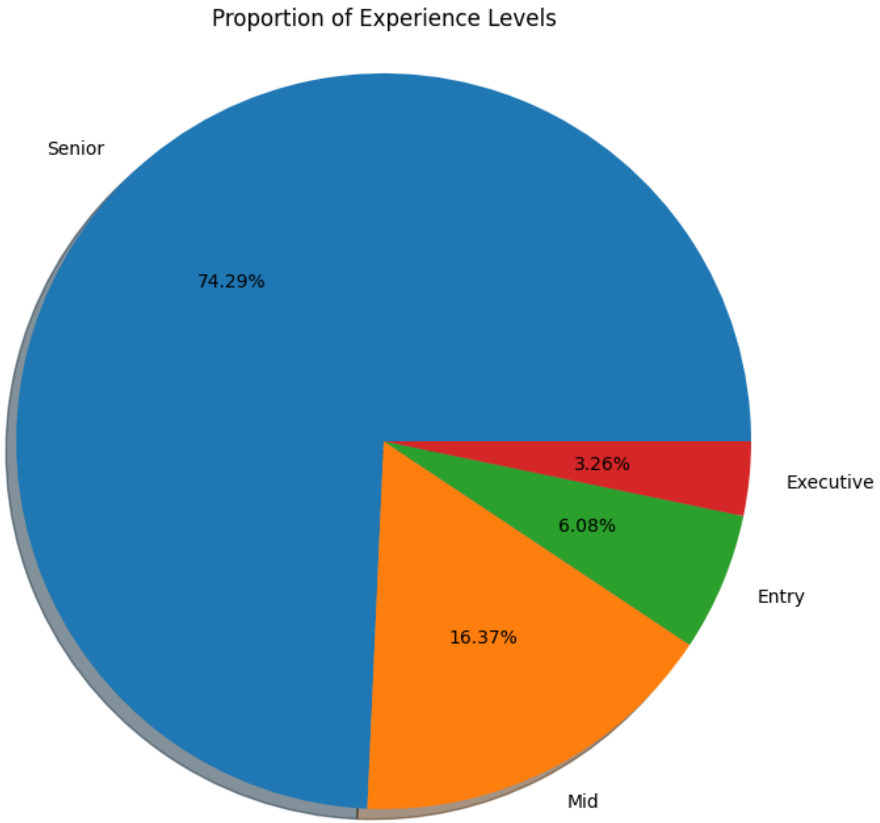
Exploratory Data Analysis - Data Visualization



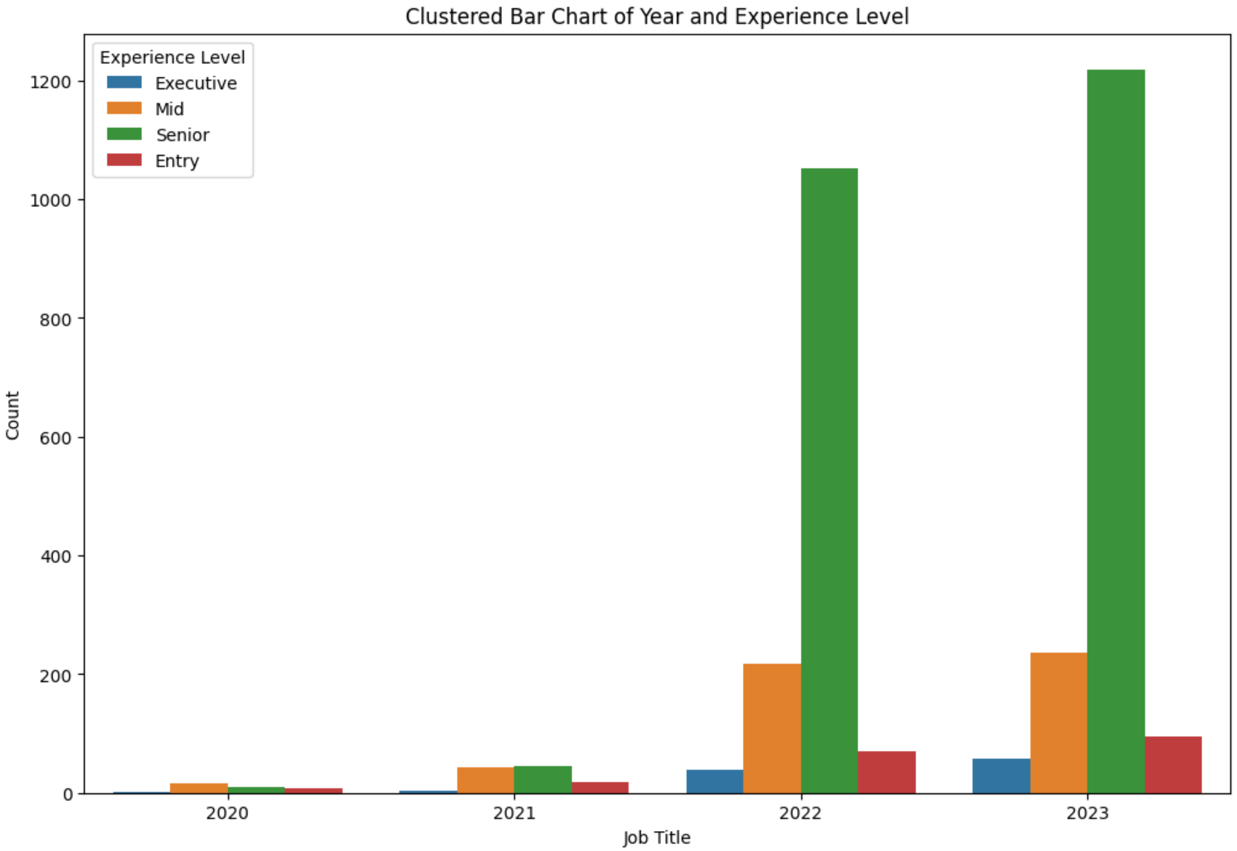
Exploratory Data Analysis - Data Visualization



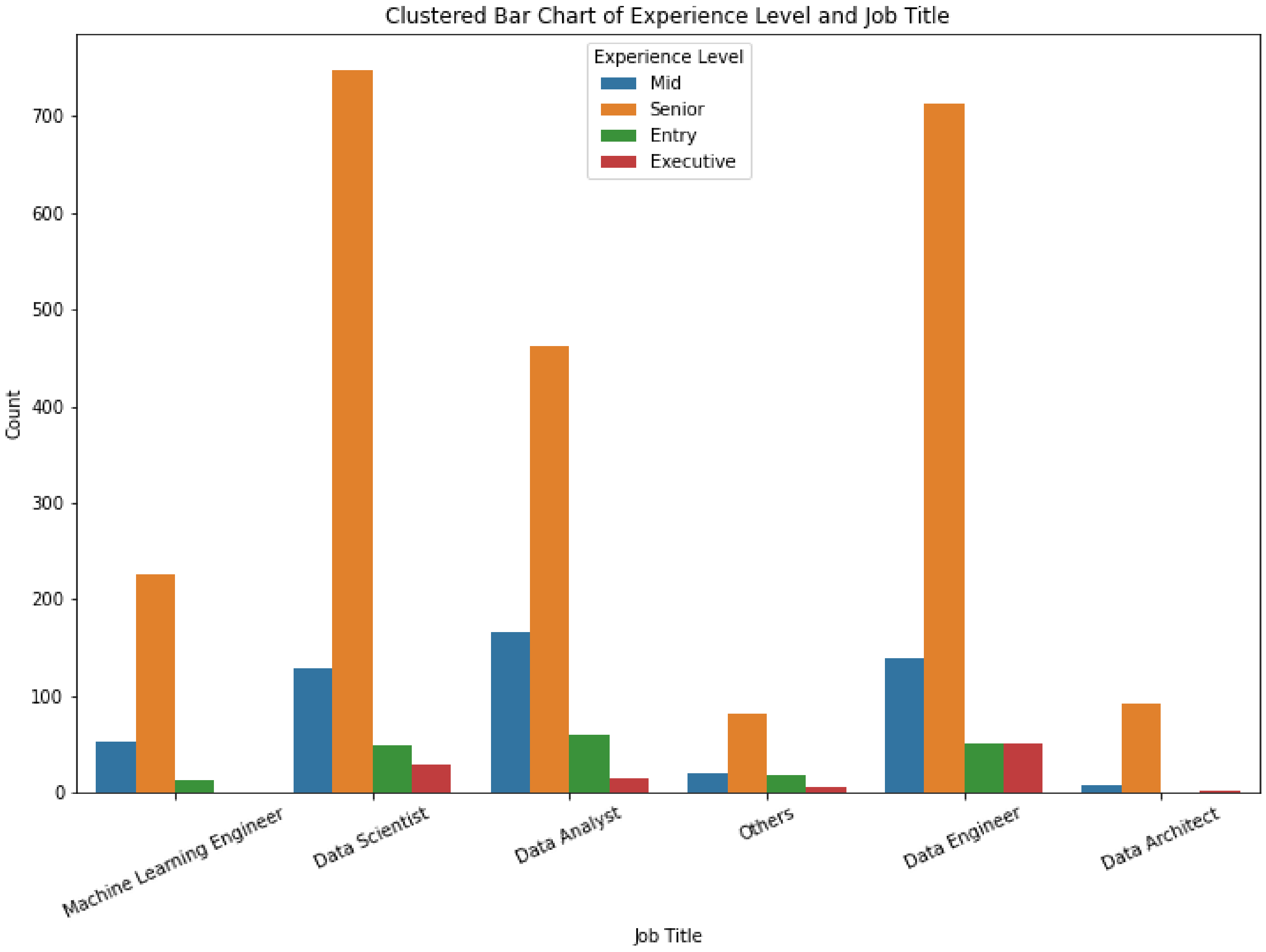
Exploratory Data Analysis - Data Visualization



Exploratory Data Analysis - Data Visualization

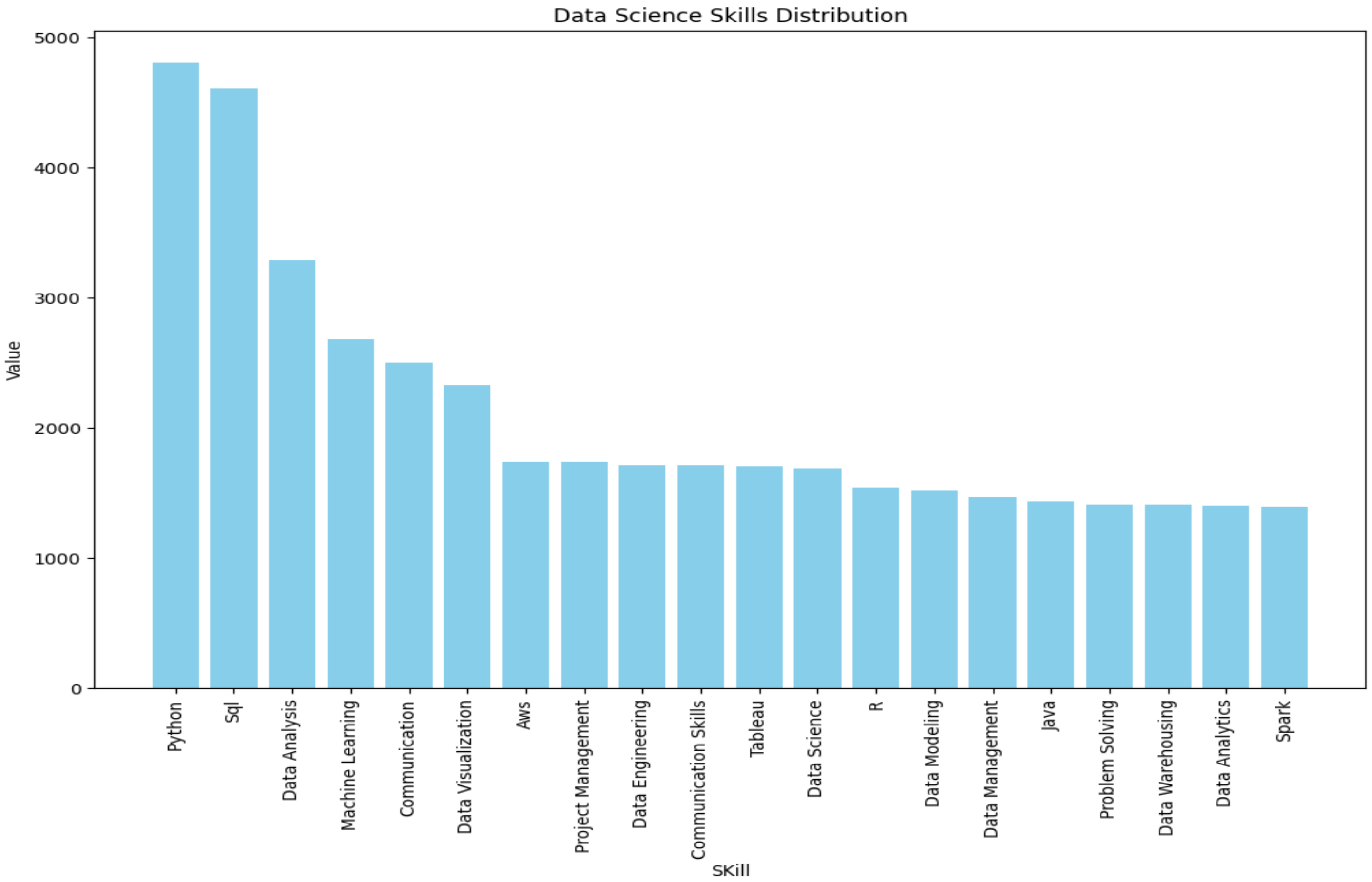


Exploratory Data Analysis - Data Visualization

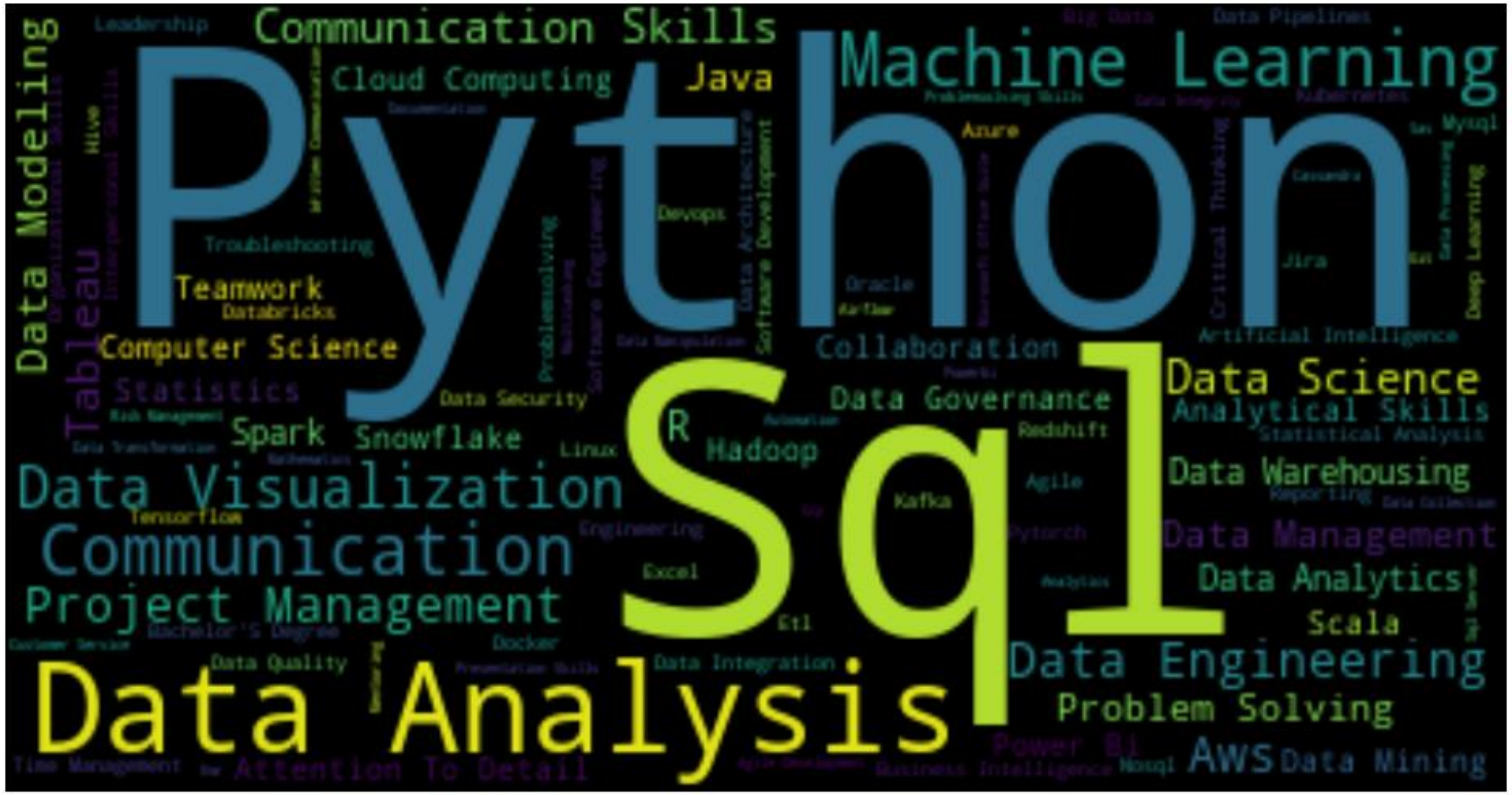


Exploratory Data Analysis - Data Visualization

Question 5: What are the in-demand skills in the Data Science field?

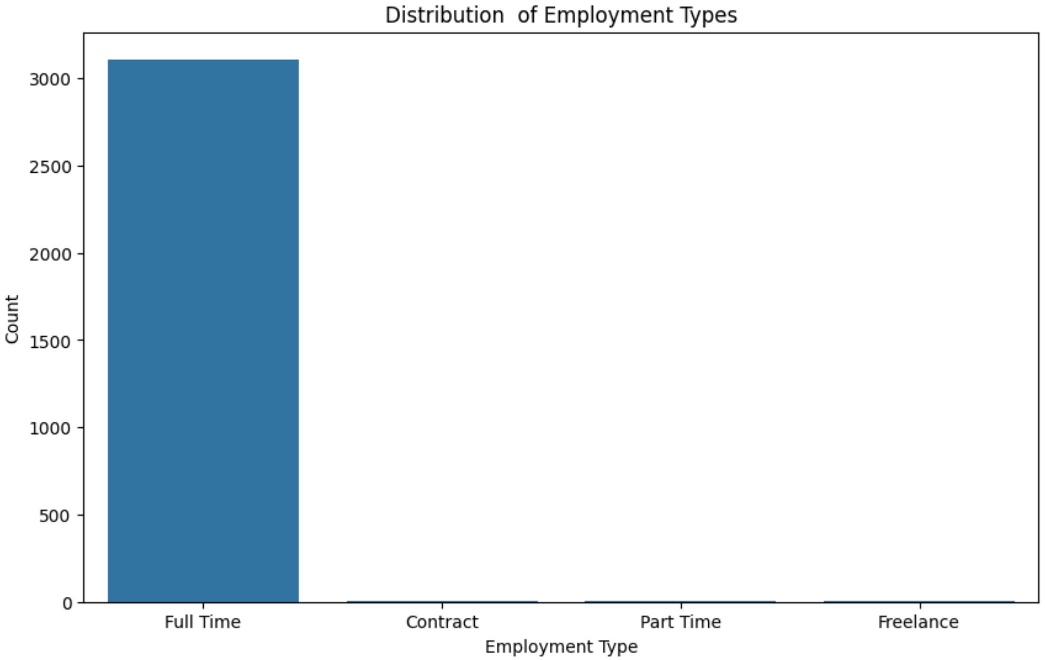


Exploratory Data Analysis - Data Visualization



Exploratory Data Analysis - Data Visualization

Question 6: What are the proportions of employment types in the Data Science field?



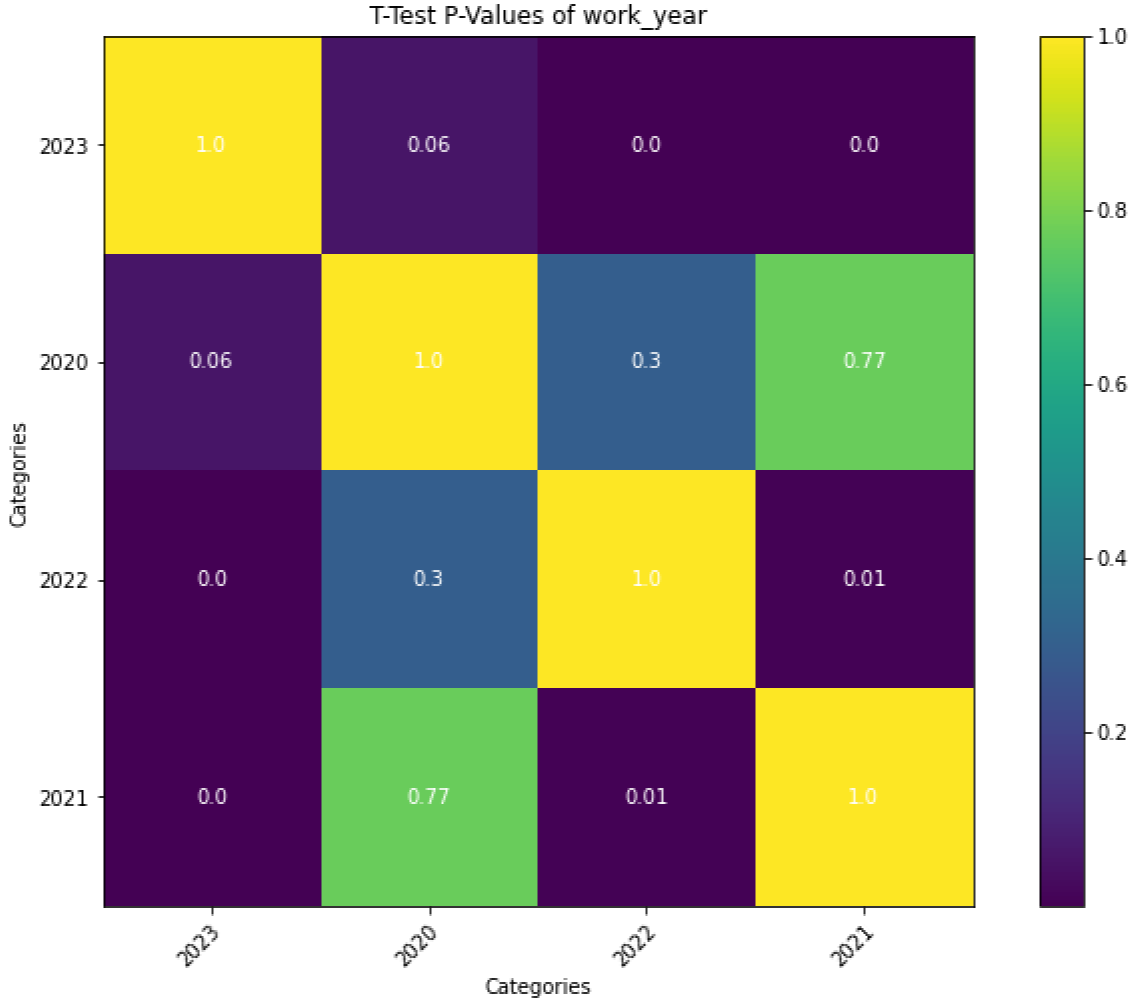
Exploratory Data Analysis (T-Test)

Define Null and Alternate Hypotheses-

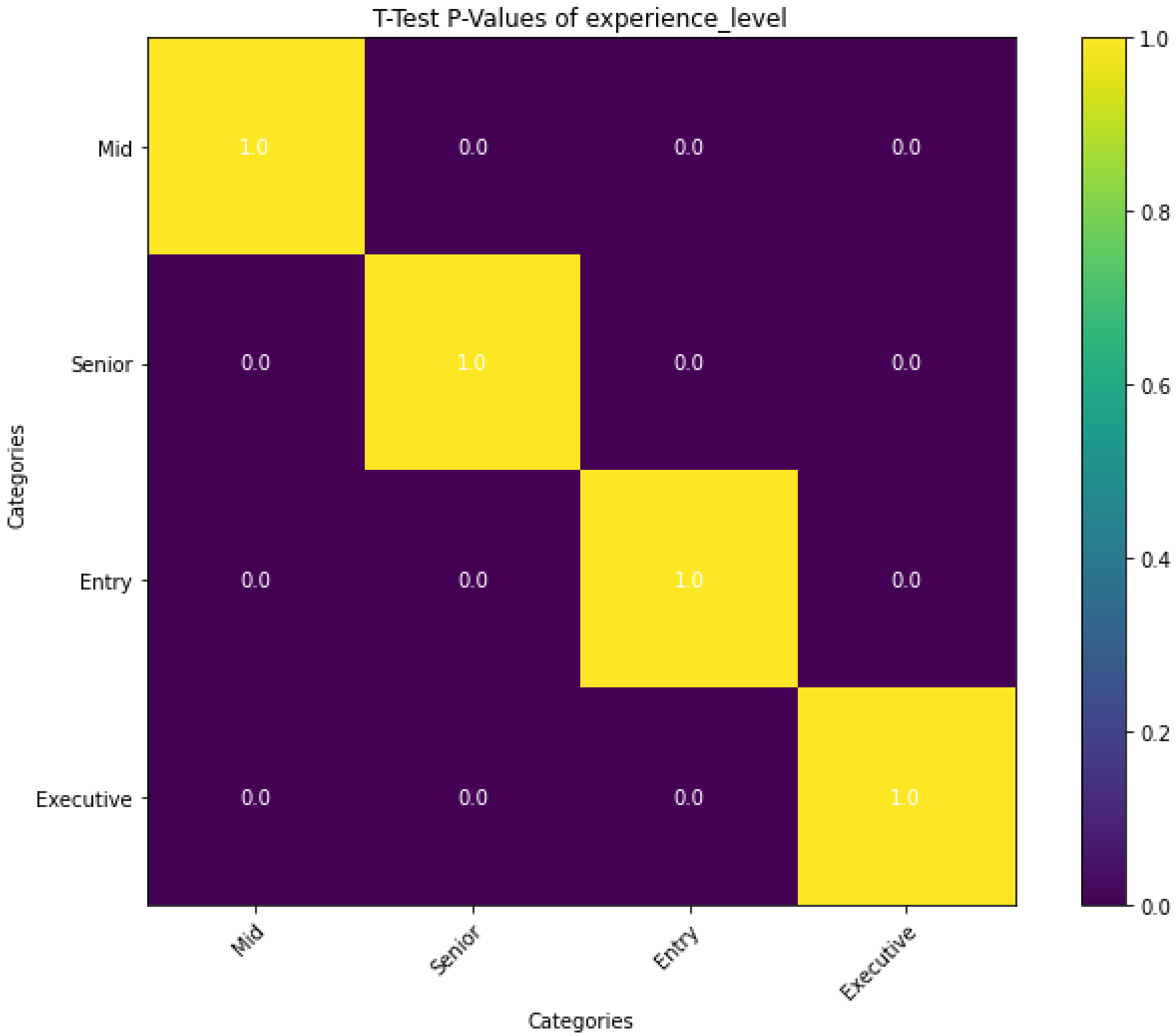
Null Hypothesis: There is no significant difference between the salaries of two categories within a variable

Alternate Hypothesis: There is a significant difference between the salaries of two categories within a variable

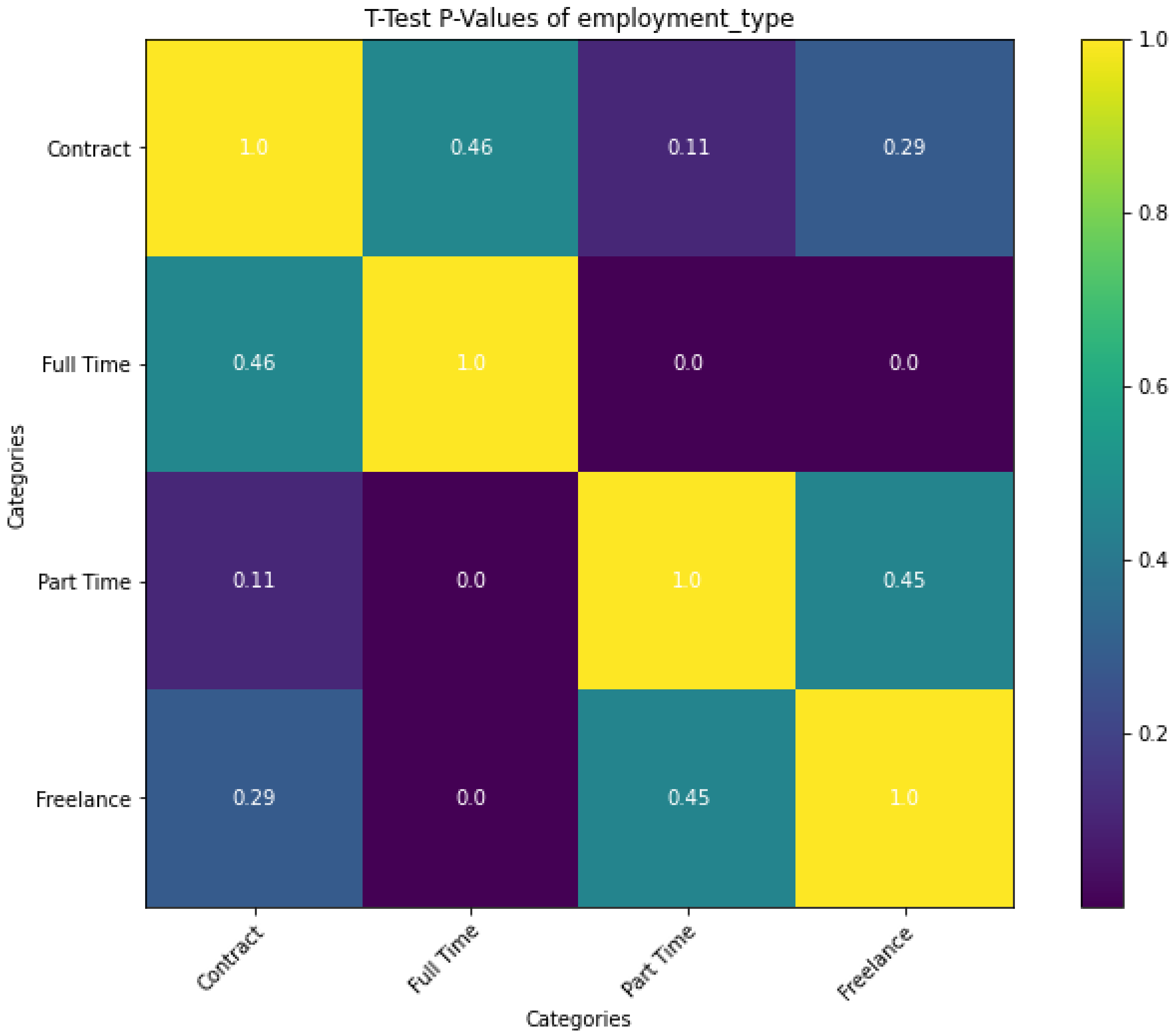
Exploratory Data Analysis (T-Test)



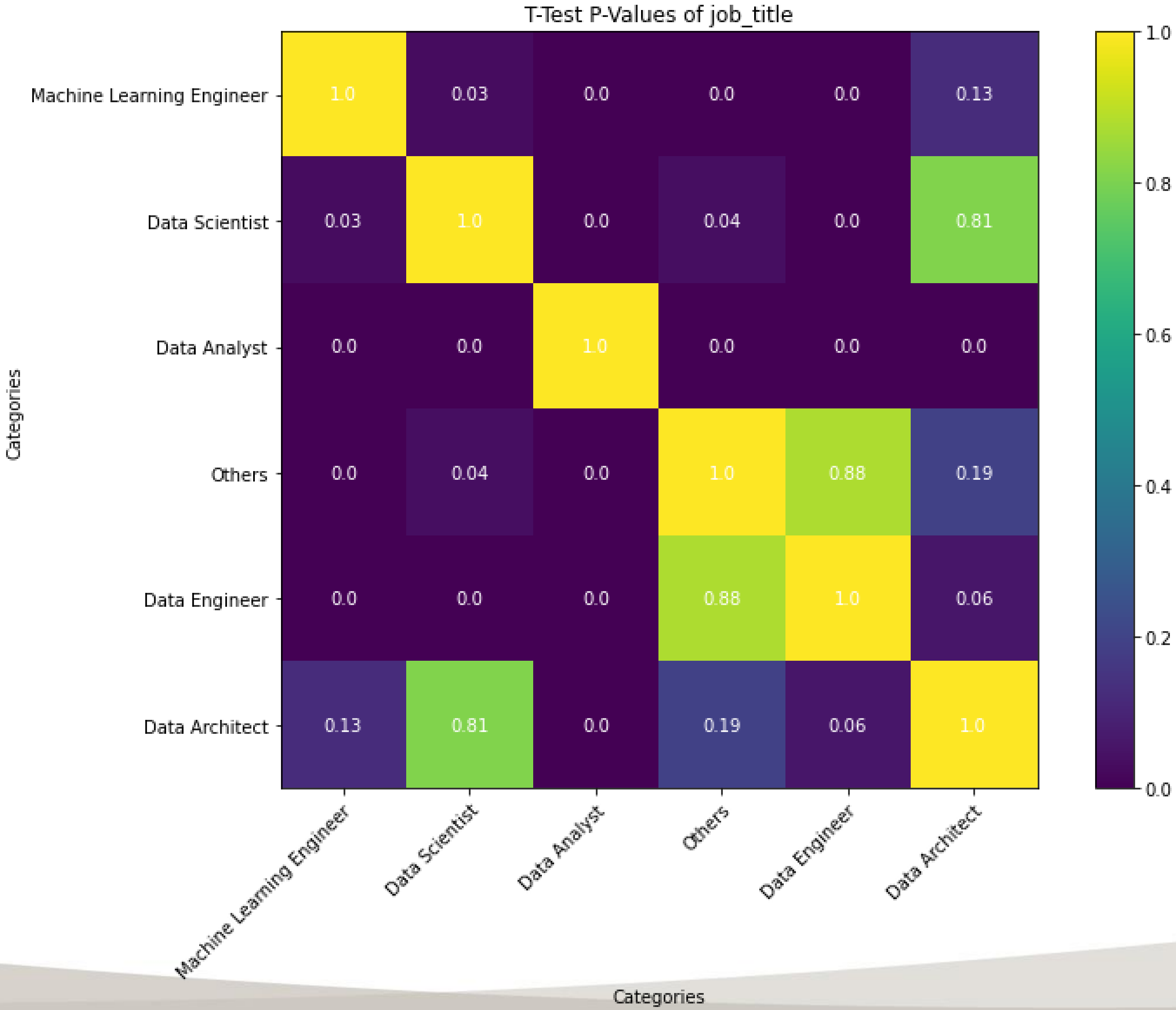
Exploratory Data Analysis (T-Test)



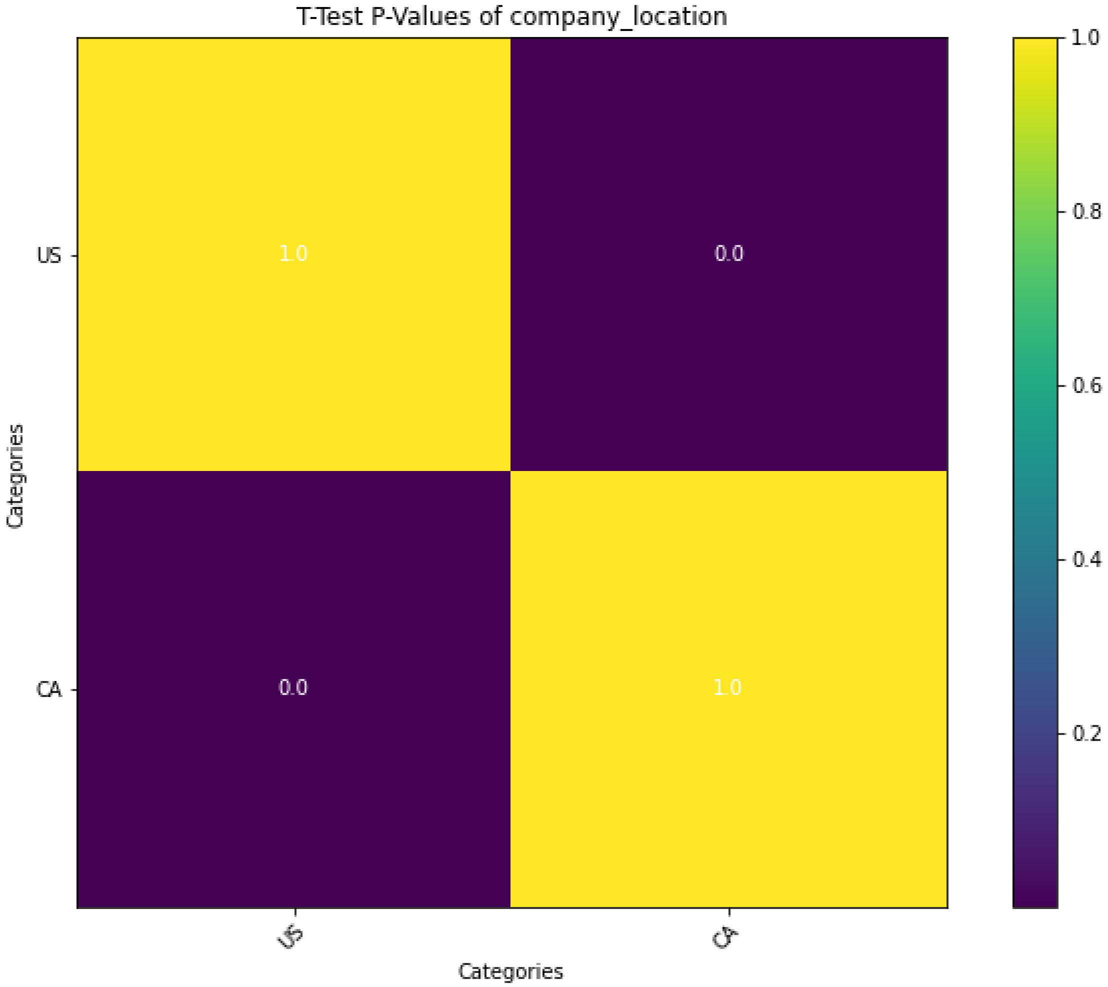
Exploratory Data Analysis (T-Test)



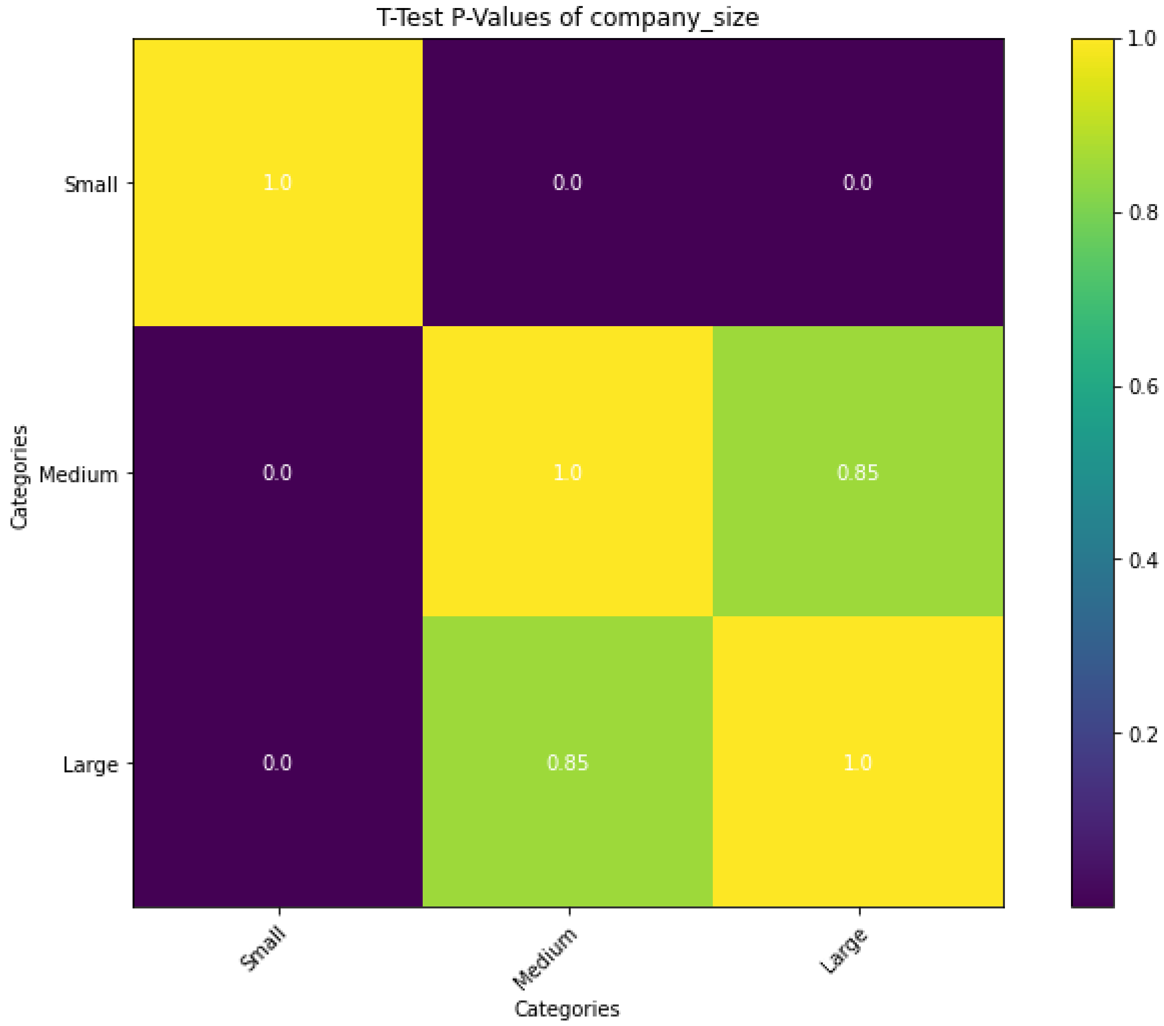
Exploratory Data Analysis (T-Test)



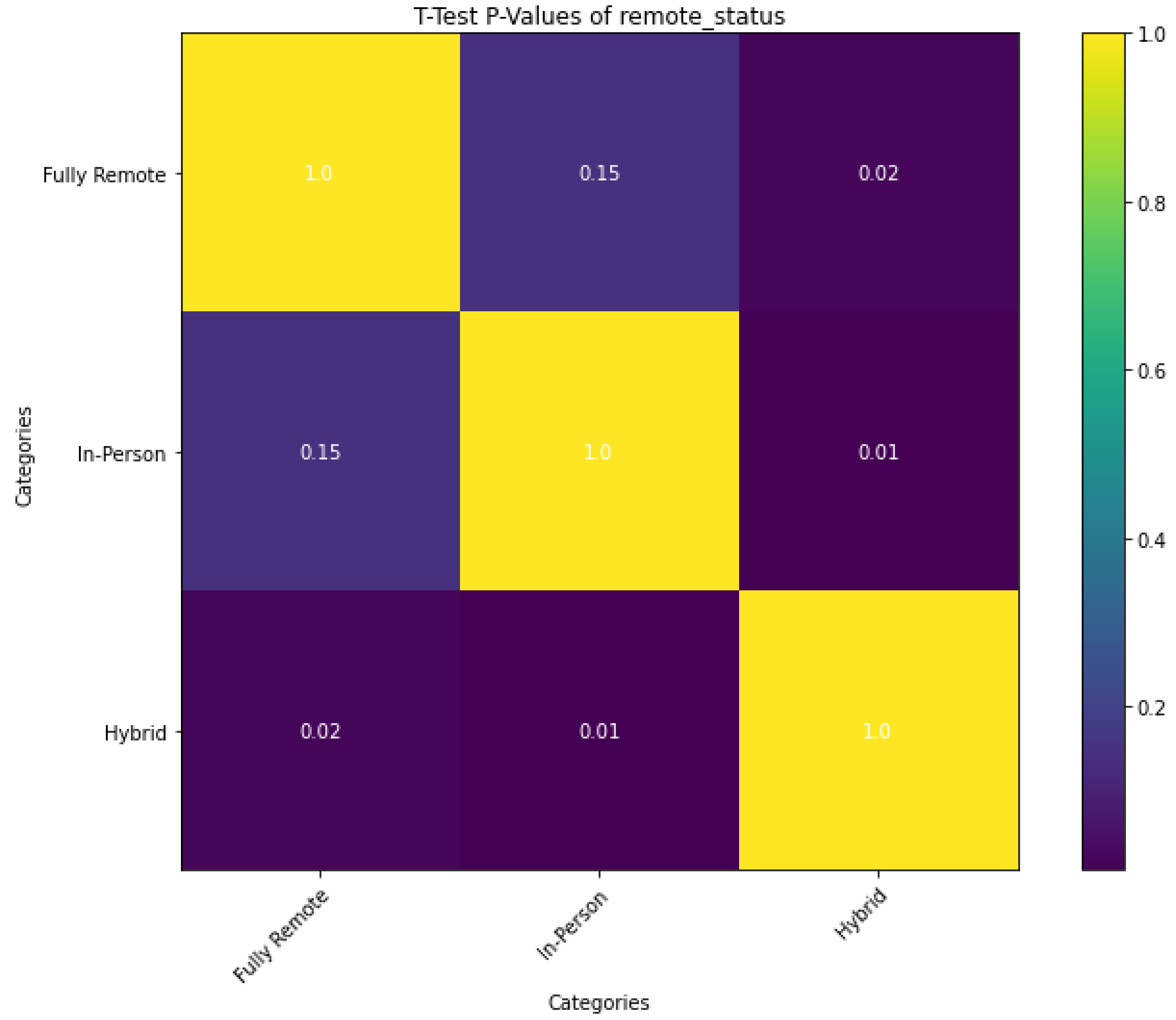
Exploratory Data Analysis (T-Test)



Exploratory Data Analysis (T-Test)



Exploratory Data Analysis (T-Test)



Exploratory Data Analysis (T-Test)

Work Year

Salaries from 2020 and 2021 were similar but changed significantly in later years. The biggest salary shifts happened after 2021, possibly reflecting industry growth or post-pandemic market changes.

Experience Level

The p-values being less than 0.05 across the entire heatmap indicate that there is a significant difference between the salaries of all experience levels, as expected.

Employment Type

Full-time jobs had significantly higher salaries than part-time and freelance roles. Also, contract salaries are not significantly different from full-time or freelance, meaning they can vary widely. In addition, part-time and freelance roles have similar salary structures, which supports the fact that they generally pay less than full-time positions.

Exploratory Data Analysis (T-Test)

Job Title

Most data roles have significantly different salaries, especially Data Analysts, who earn less than other roles. The high p-value between Data Architects & Data Scientists indicate those roles earn similar salaries, suggesting they may be equally valued in organizations. Data Architects & Data Engineers also have some slight similarities between salaries, although not so much. In addition, Data Engineers and Others have similar salaries, suggesting some roles in the "Others" category earn similar salaries to Data Engineers.

Company Location

The p-value of 0 between US and Canadian jobs indicates that there is a statistically significant difference in salaries between the two locations. This confirms that salaries in the US and Canada are not the same, with the US offering higher salaries.

Company Size

The heatmap of company size shows there is a statistically significant difference between the salaries of small companies and both medium & large companies. The heatmap also shows that there is no significant difference between the salaries of medium and large companies.

Remote Status

There is a statistically significant difference between the salaries of fully remote jobs and hybrid jobs, while there is no difference between fully remote and in-person jobs. There is also a statistically significant difference between the salaries of in-person and hybrid jobs.

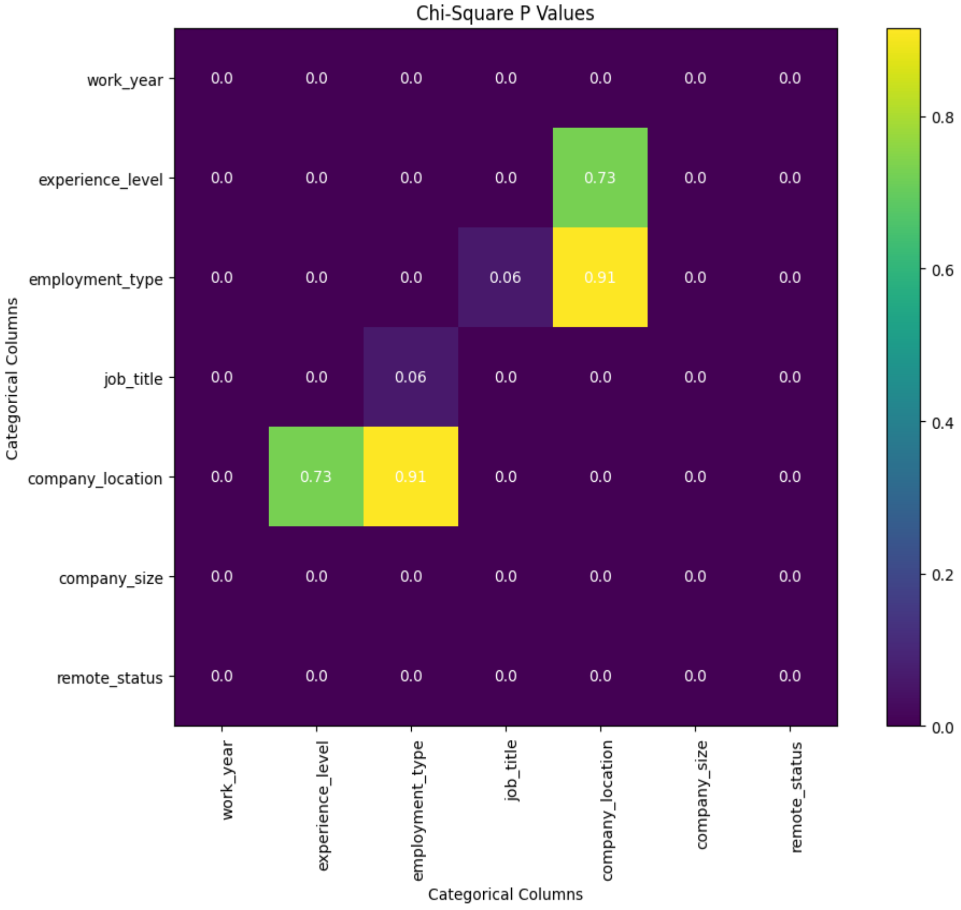
Exploratory Data Analysis (Chi-Squared)

Define Null and Alternate Hypotheses

Null Hypothesis: The two variables are independent and there is no relationship

Alternate Hypothesis: The two variables are dependent and there is a relationship

Exploratory Data Analysis (Chi-Squared)



	work_year	experience_level	employment_type	job_title	company_location	company_size	remote_status
work_year	0.0	0.0	0.0	0.0	0.0	0.0	0.0
experience_level	0.0	0.0	0.0	0.0	0.73	0.0	0.0
employment_type	0.0	0.0	0.0	0.06	0.91	0.0	0.0
job_title	0.0	0.0	0.06	0.0	0.0	0.0	0.0
company_location	0.0	0.73	0.91	0.0	0.0	0.0	0.0
company_size	0.0	0.0	0.0	0.0	0.0	0.0	0.0
remote_status	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Conclusion

This analysis of data science jobs between 2020 and 2023 provides insights into salary trends, job distributions, experience levels, company locations, remote work trends, employment types, job titles, company sizes, and required skills. Below are some key takeaways:

Growth of Data Science Jobs - The number of data science jobs has grown significantly over the past four years, with a notable boom between 2021 and 2022.

- Data Engineering roles have become more prevalent, surpassing Data Scientists as the most common role in 2023.

Salary Insights - Most salaries range between 150,000 and 300,000, with a few roles exceeding 400,000.

- Machine Learning Engineers and Data Architects have the highest median salaries, while Data Analysts earn the least.

- Salaries increase with experience, with Executives earning the most of around 300,000 and above, while entry level professionals earn the least of around 140,000 and below.

- Jobs in the US generally pay higher salaries than jobs in Canada, with wider salary ranges and more high-paying roles. - Larger companies offer higher salaries compared to medium and small companies.

Conclusion

Experience Level Trends

- Senior roles have the highest proportion of jobs in the job market, making up over 70% of the dataset.

- Mid level roles are more common among Data Analysts, while Executive roles are mostly filled by Data Scientists and Data Engineers.

- Entry level positions are relatively scarce, with Machine Learning Engineers and Data Architects rarely having them.

Remote Work Trends

- Fully remote roles were dominant between 2020 and 2022, likely due to the COvid-19 pandemic.

- In-person roles surged in 2023, suggesting companies are shifting back to pre-pandemic work arrangements.

- Hybrid roles have consistently remained low throughout the years.

Conclusion

Relevant Skills

- Python and SQL are the most in-demand skills, followed by Data Analysis, Machine Learning, and Communication.

- Some other technical skills such as AWS, R and visualization tools, such as Tableau, Power BI are also highly valued.

- Soft skills like Communication, Problem-Solving, and Project Management are also emphasized in job descriptions, and are equally important in the data science field.

Statistical Relationships

- There is no statistical relationship between employment type and job title, employment type and company location, or experience level and company location. All other categorical variables are statistically dependent, meaning they influence one another.

Conclusion

Final Thoughts

The Data Science job market is growing, with increasing demand for skilled professionals, rising salaries, a shift back to in-person work, etc. Professionals seeking higher salaries may target US-based jobs and larger companies. In addition, professionals should ensure they continuously develop high-demand skills like Python, SQL, Machine Learning, Communication, etc., to ensure they stay relevant in the field.

The high demand for senior expertise should serve as motivation rather than discouragement. Internships and networking opportunities can be valuable entry points into the field.

This analysis provides valuable insights for data science professionals, job seekers, employers, etc., looking to understand trends in the Data Science industry.

References

Chaki, A. (2023) *Data Science Salaries 2023* 💸 [online]. Available at: https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023?select=ds_salaries.csv (Accessed 17 January 2025)

Asaniczka. (2024) Data Science Job Postings & Skills (2024) [online]. Available at: https://www.kaggle.com/datasets/asaniczka/data-science-job-postings-and-skills (Accessed 31 January 2025)

UNIVERSITY OF CALGARY

THANK YOU

QUESTIONS