Multivariate Statistical Analysis of Salary Data Scientist

Exploratory Analysis of a dataset comprising

Prepared by Rui Candongo

In partnership with Technology training school Alura

September 5, 2025

Abstract

This study presents an exploratory analysis of a dataset comprising over 133,000 observations of global job market information in data-related roles from 2020 to 2025. Variables include work year, experience level, employment type, job title, salary (local and USD), geographic location, remote work ratio, and company size. The report highlights salary distributions, geographical disparities, remote work adoption, and organizational structures. Findings reveal significant variation in compensation based on experience, geography, and job role, with increasing adoption of remote work practices.

Exploratory Analysis

1. Dataset Overview: 133,339 records, 12 columns covering salaries, job roles, experience levels, locations, company size, and remote modality.

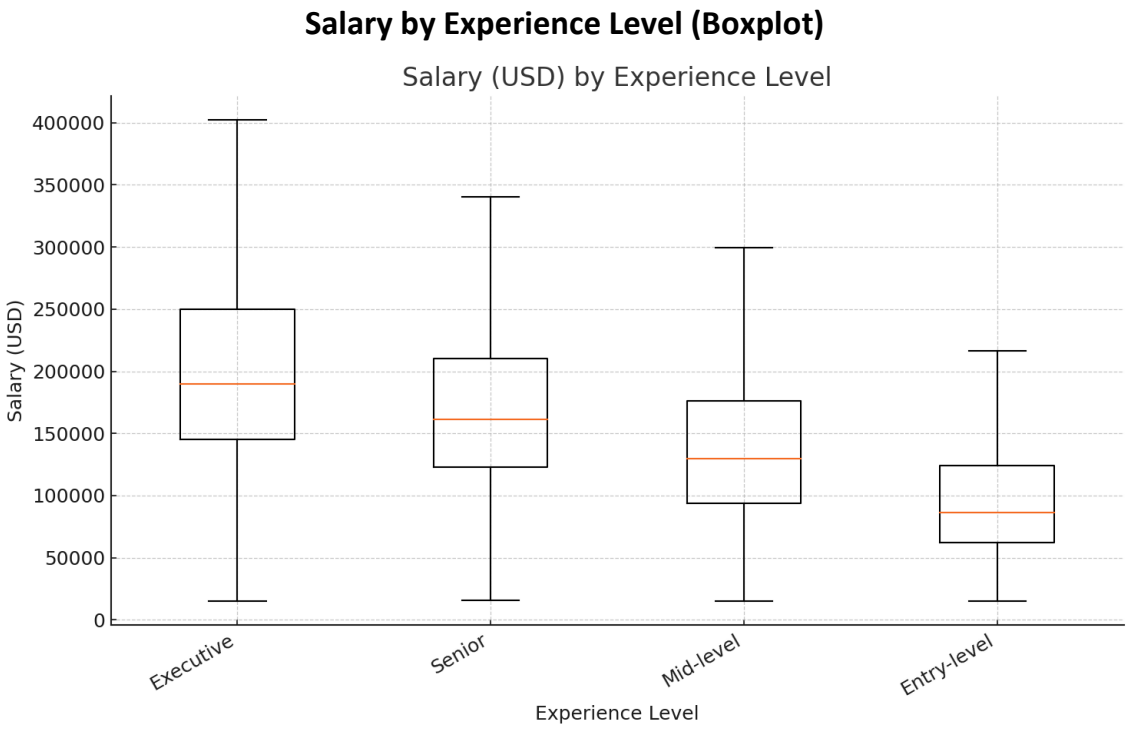
2. Salary Trends: Significant variation by role and geography. Median salaries differ sharply between entry-level and senior professionals.

3. Geographical Insights: Disparities across regions. US-based roles generally command higher salaries.

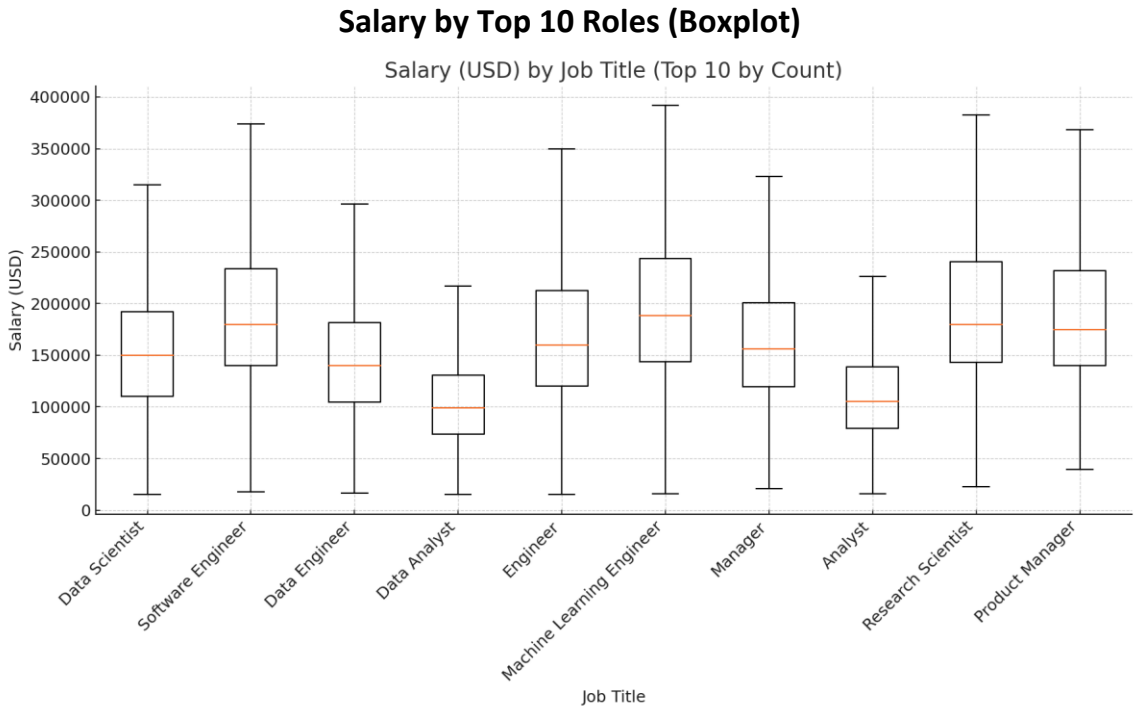
4. Remote Work Adoption: Sharp rise in remote roles post-2020.

5. Job Titles: Top roles include Data Scientist, Data Engineer, Machine Learning Engineer.

6. Company Size: Larger firms generally offer higher compensation.

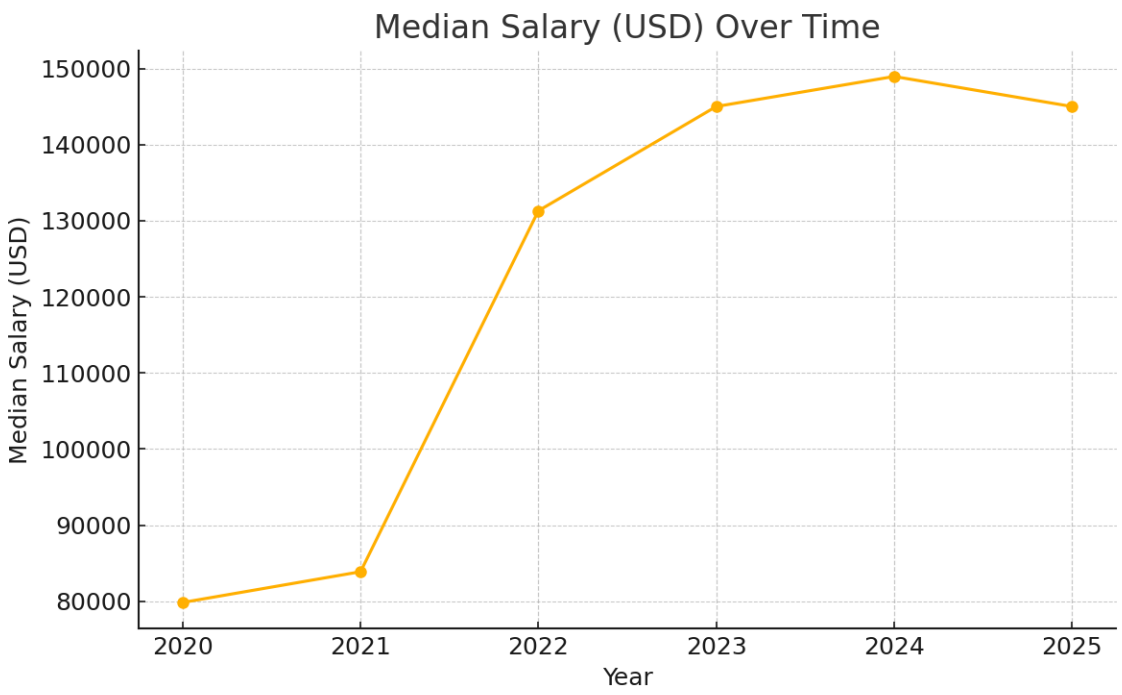


Salary by Experience Level (Boxplot)

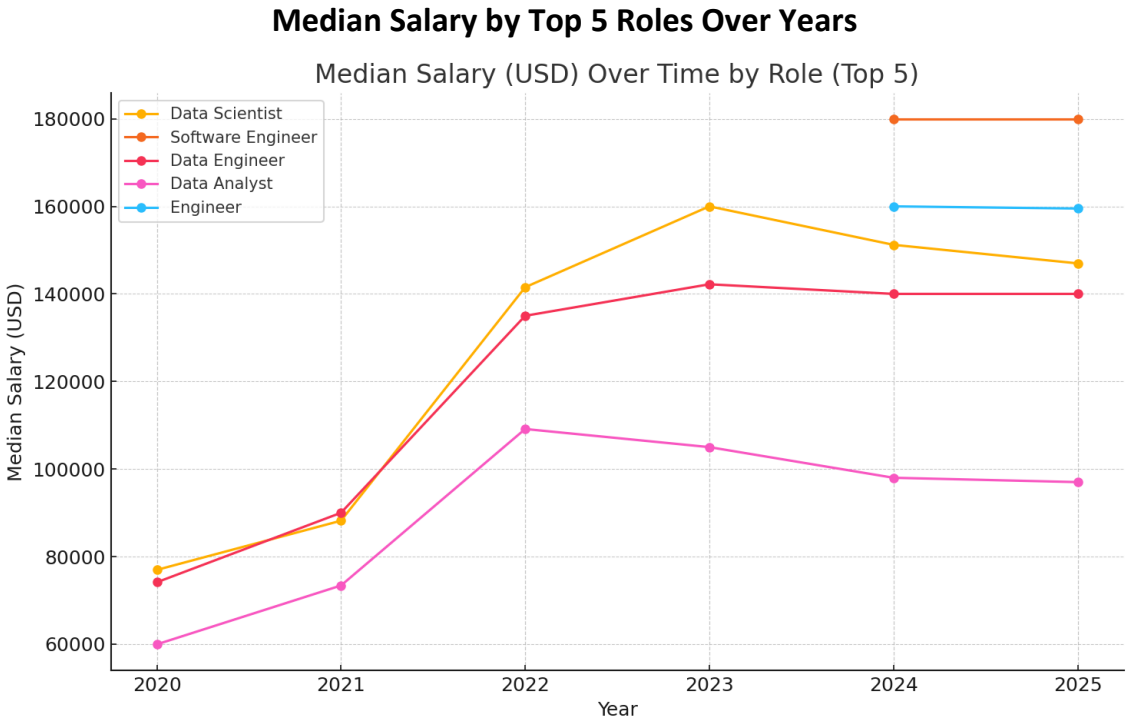


Salary by Top 10 Roles (Boxplot)

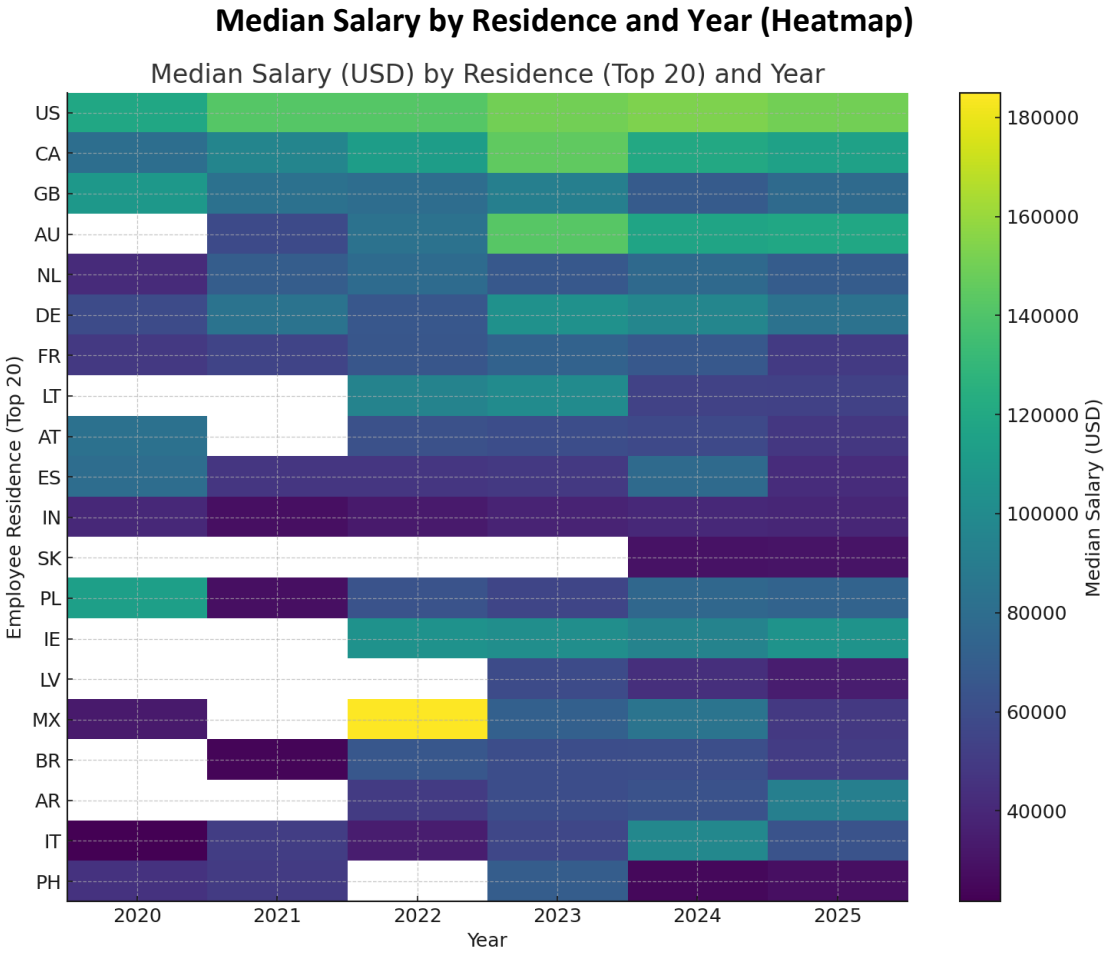
Median Salary Over Years



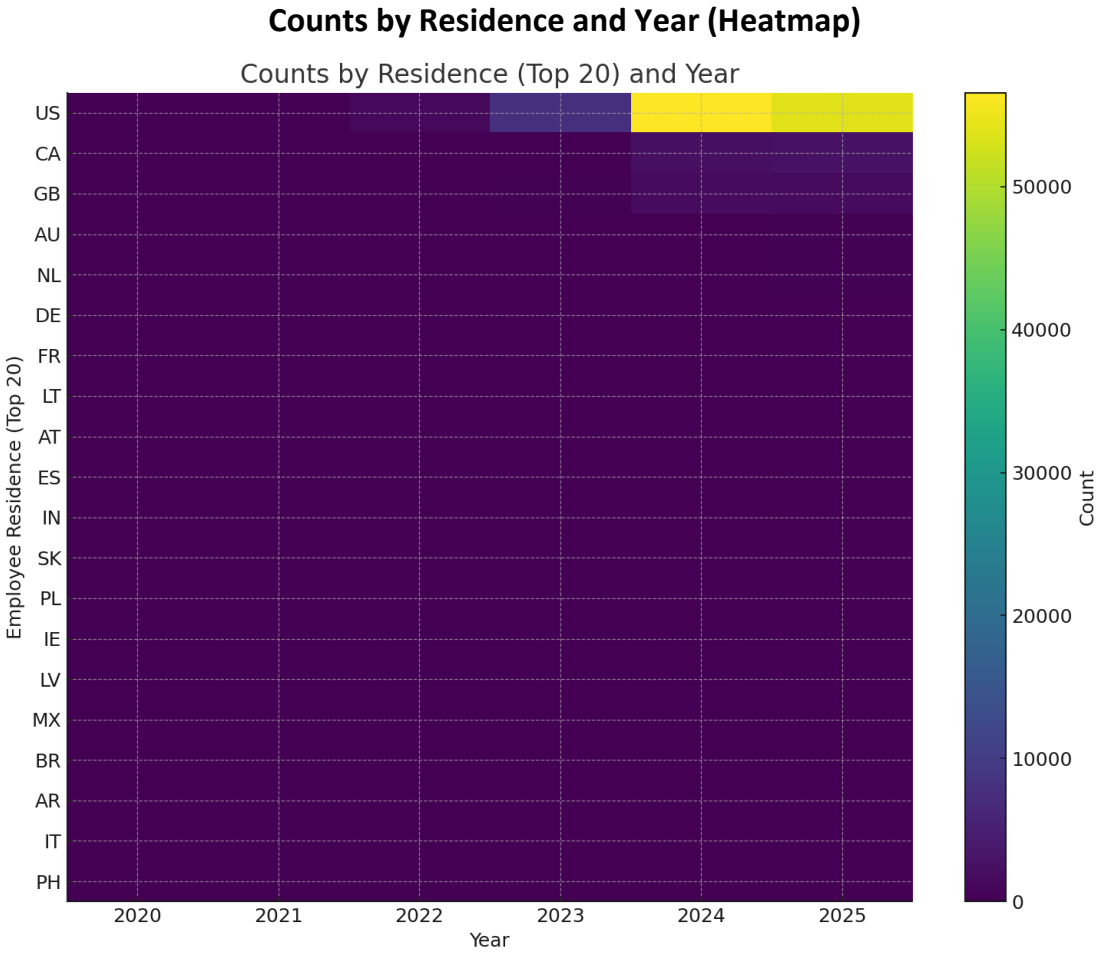
Median Salary Over Years



Median Salary by Top 5 Roles Over Years

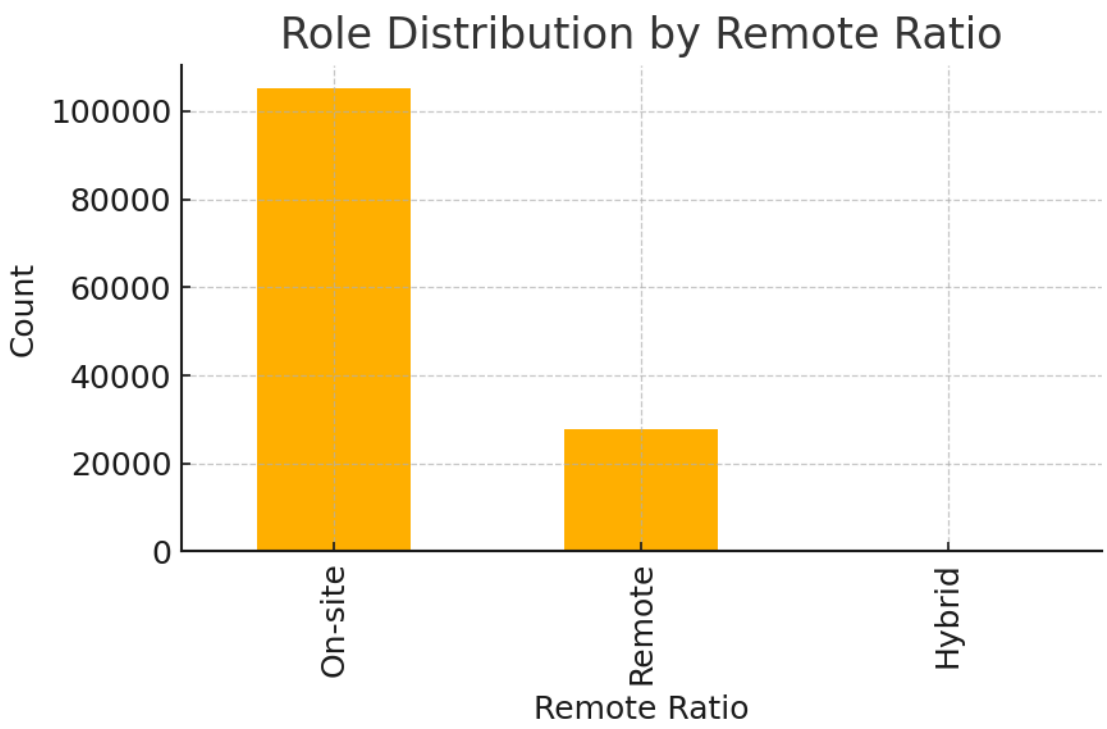


Median Salary by Residence and Year (Heatmap)



Counts by Residence and Year (Heatmap)

Remote Ratio Distribution



Remote Ratio Distribution

Conclusions

1. Experience and salary are strongly correlated, with senior roles earning 3-4x more than entry-level.

2. Remote work adoption is reshaping global hiring and salary distribution.

3. Geographic location remains a key determinant of compensation.

4. Company size impacts salary offers significantly.

5. Data reflects a globalized and increasingly internationalized data workforce.

References

Dataset: Data-analysis-final.csv (user-provided)

Methodology: Exploratory Data Analysis (EDA) using Python (pandas, matplotlib)

Context: International labor statistics and reports on remote work adoption.