**Data Science Job Salaries — Final Report**

**Project Title:** Data Science Job Salaries Analysis

**Author:** Manasa Shetty

**Introduction**

The rapid growth of Data Science as a profession has resulted in diverse career paths, ranging from entry-level analysts to senior executives and specialized AI/ML engineers. Salaries in this field vary significantly depending on multiple factors such as experience level, employment type, company size, remote work opportunities, and geographic location.

This project uses the **Data Science Job Salaries dataset** to analyze these variations. The goal is to identify trends and derive meaningful insights about compensation in the industry. The analysis is performed using Python in a Jupyter Notebook environment and includes both exploratory data analysis (EDA) and a predictive financial modeling approach.

**Dataset Description**

The dataset contains **607 records** and **12 columns** with details about job roles and salaries.

**Key Features:**

* **work\_year** – Year of the job posting
* **experience\_level** – Career level (Entry, Mid, Senior, Executive)
* **employment\_type** – Full-time, part-time, contract, freelance
* **job\_title** – Job designation
* **salary** – Reported salary
* **salary\_in\_usd** – Standardized salary in USD
* **employee\_residence** – Country of the employee
* **remote\_ratio** – Degree of remote work (0 = on-site, 50 = hybrid, 100 = fully remote)
* **company\_location** – Country of the company
* **company\_size** – Small (S), Medium (M), Large (L)

The dataset required minimal cleaning, mainly ensuring consistency in categorical values and handling missing entries.

**Exploratory Data Analysis (EDA)**

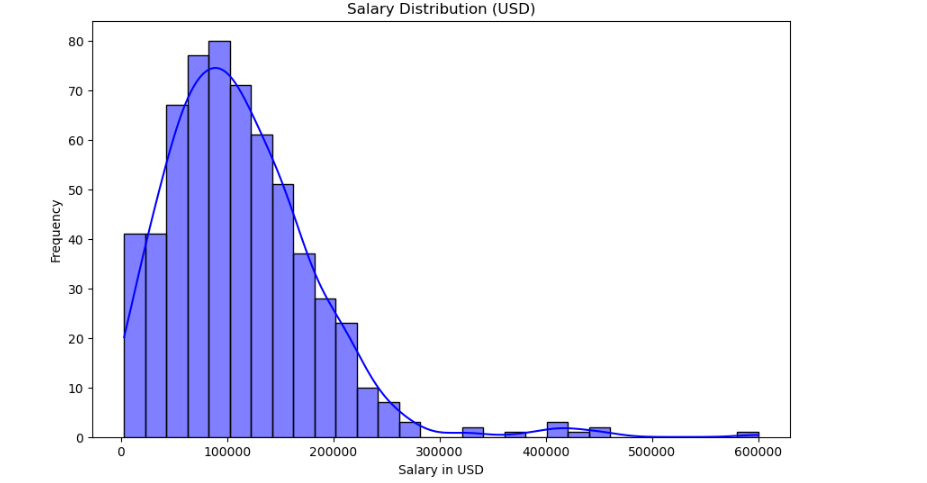
**Salary Distribution**

Mean salary $112,298

Median salary $101,570

Minimum salary $2,859

Maximum salary $600,000

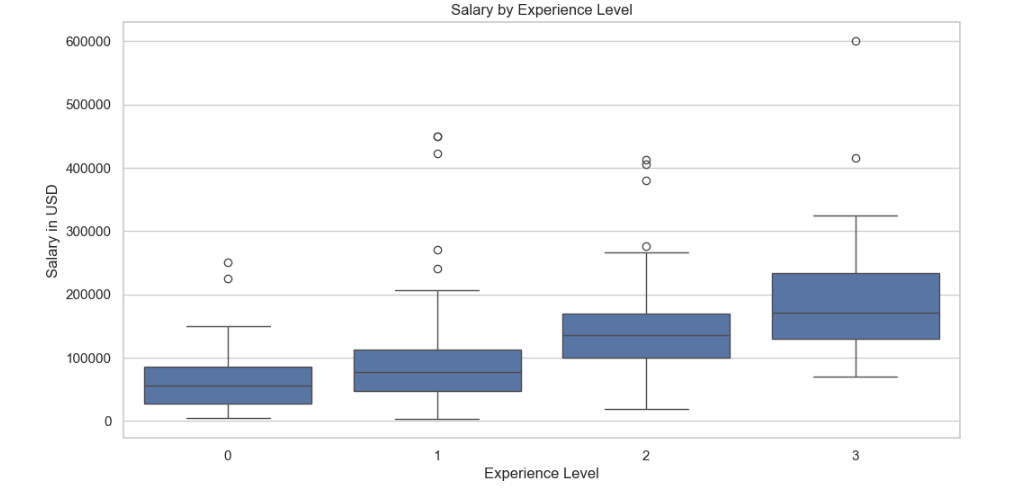


The salary distribution is **right-skewed**, with most salaries concentrated between **$100K–$150K**. A few extremely high salaries create a long tail, making the **mean** much higher than the **median.**

**Salaries by Experience Level**

| **Experience Level** | **Median Salary (USD)** |
| --- | --- |
| Entry (EN) | 68,500 |
| Mid (MI) | 85,000 |
| Senior (SE) | 140,000 |
| Executive (EX) | 187,500 |

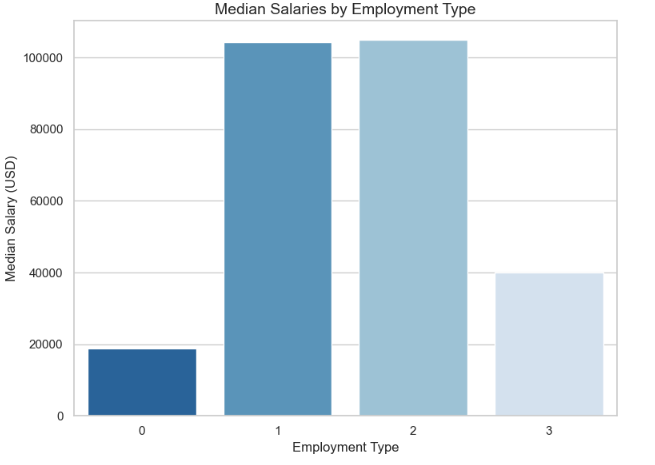
Salaries increase steadily with experience, showing a clear upward trend from entry-level to executive roles.Executives earn nearly 3 times more than entry-level professionals. Outliers are present at each level, but higher experience consistently leads to higher median pay.



Salaries increase steadily with experience, showing a clear upward trend from entry-level to executive roles.Executives earn nearly 3 times more than entry-level professionals. Outliers are present at each level, but higher experience consistently leads to higher median pay.

**Salaries by Employment Type**

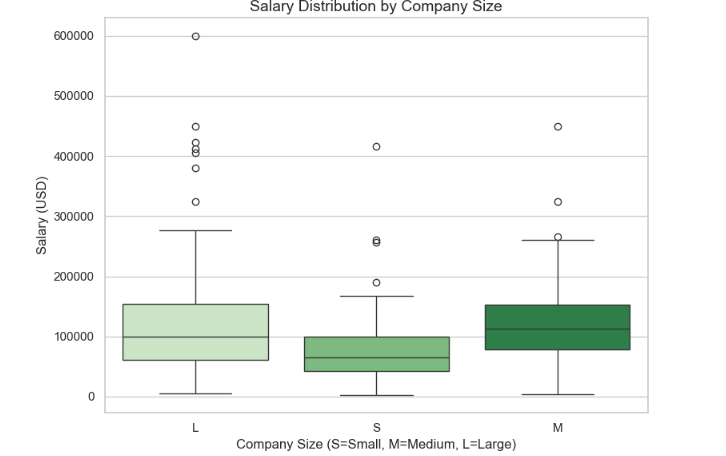
| **Employment Type** | **Median Salary (USD)** |
| --- | --- |
| Full-time (FT) | 115,250 |
| Contract (CT) | 105,000 |
| Freelance (FL) | 40,000 |
| Part-time (PT) | 34,500 |



Full-time roles dominate and provide the highest consistent salaries. Freelance and part-time roles offer significantly lower pay.

**Salaries by Company Size**

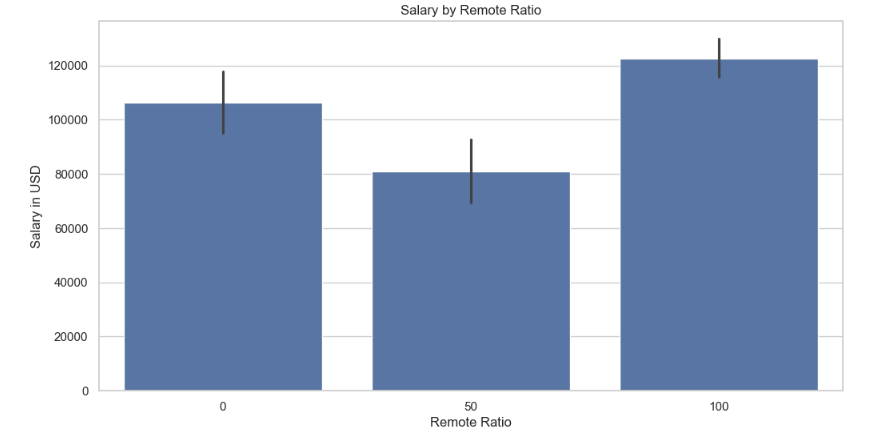
| **Company Size** | **Median Salary (USD)** |
| --- | --- |
| Small (S) | 80,000 |
| Medium (M) | 116,075 |
| Large (L) | 120,250 |



Larger companies generally offer higher compensation packages than smaller firms.

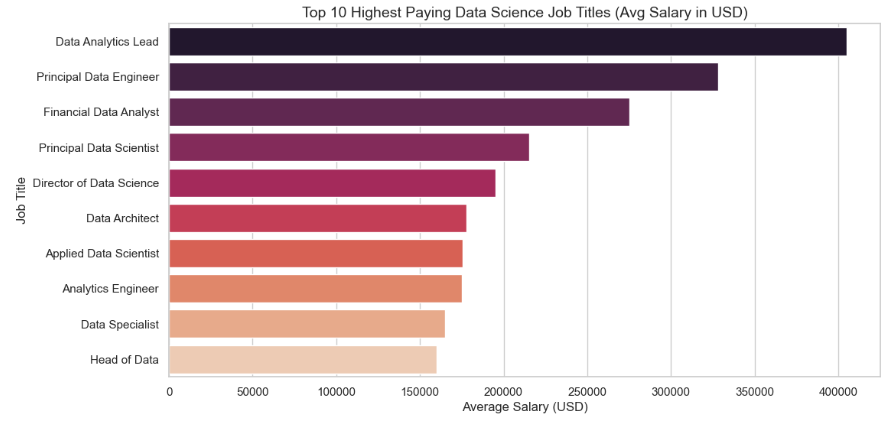
**Salaries by Remote Ratio**

| **Remote Ratio** | **Median Salary (USD)** |
| --- | --- |
| 0% (On-site) | 102,100 |
| 50% (Hybrid) | 76,760 |
| 100% (Remote) | 123,000 |



Fully remote jobs provide higher median salaries compared to on-site or hybrid roles.

**Top 10 Paying Job Roles**



The role of Data Analytics Lead commands the highest average salary among data science job titles, exceeding $400,000 USD. Entry-level and mid-tier roles like Data Specialist and Analytics Engineer have noticeably lower average salaries, highlighting a significant compensation gap based on seniority and specialization.

**Insights**

* Experience level is the strongest determinant of salary growth.
* Full-time employment dominates with stable and high pay.
* Large companies generally offer better compensation packages.
* Remote opportunities pay better than on-site or hybrid roles.
* Specialized roles Data Analytics Lead, Principal Data Engineer, Financial Data Analyst  consistently rank among the highest paid.

**Financial Modeling (Salary Prediction)**

To extend the analysis, a Linear Regression model was built to predict salaries based on job-related features.

**Methodology**

* **Features:** experience\_level, employment\_type, remote\_ratio, company\_size
* **Target:** salary\_in\_usd
* **Preprocessing:** Categorical encoding via OneHotEncoder, numerical features passed through
* **Pipeline:** ColumnTransformer + LinearRegression
* **Split:** 80% training, 20% testing

**Results**

* Mean Squared Error (MSE): 5,356,012,602

**Insights**

* The high MSE indicates Linear Regression struggles with prediction accuracy, mainly due to wide salary variability and extreme outliers.
* Features like experience level and company size still showed meaningful predictive power.
* More advanced models (e.g., Random Forest, Gradient Boosting) could better capture non-linear salary patterns.

**Key Findings**

* Salaries increase sharply with experience and specialization.
* Full-time roles dominate and are better compensated than freelance/part-time.
* Large companies pay more than smaller firms.
* Remote jobs command higher pay, reflecting global flexibility.
* Top-paying roles are leadership and ML/AI-focused.
* Geography matters: significant disparities between regions.
* Predictive modeling with Linear Regression is limited; advanced ML is required for reliable salary forecasting.

**Conclusion**

The analysis confirms that Data Science is a lucrative field with wide-ranging opportunities. Salaries, however, vary significantly by experience, employment type, company size, remote ratio, and geography.

While exploratory analysis revealed key patterns, predictive modeling showed that simple regression is insufficient for accurate salary forecasting due to extreme outliers and non-linear trends. Future models must leverage advanced machine learning techniques.

For professionals: gaining experience, specializing in advanced roles, and exploring remote opportunities can maximize earning potential.  
For employers: offering competitive salaries and remote flexibility can help attract top talent globally.

**Future Work**

1. Apply Random Forest and Gradient Boosting for better predictive modeling.
2. Expand dataset with more years to analyze salary trends over time.
3. Integrate skills and education levels as additional features.
4. Improve outlier handling for more stable salary predictions.

**References**

1. Dataset: *Data Science Job Salaries (Kaggle)*
2. <https://github.com/Manasashetty01/Data_Science>

**CODE & OUTPUT**

