Global Data Science Salary Market Analysis (2020–2025)

Executive Summary

This report presents a comprehensive analysis of global data science salary trends between 2020 and 2025.
The objective of this project was to identify the major drivers of salary variation in the data science labor market and develop a predictive model capable of estimating salaries based on job-related attributes.

Using Python-based analytics tools including Pandas, NumPy, Matplotlib, Seaborn, SciPy, and Scikit-Learn, the dataset was cleaned, transformed, and analyzed through multiple phases including exploratory analysis, feature engineering, correlation analysis, and predictive modeling.

The findings indicate that experience level, job specialization, company size, geographic location, and industry demand are the strongest predictors of compensation. Salaries in the data science industry have increased steadily during the period analyzed due to increasing adoption of artificial intelligence and advanced analytics across industries.

Problem Statement

Organizations increasingly rely on data-driven decision making. As demand for skilled data professionals grows, understanding salary trends becomes essential for talent acquisition strategies, compensation benchmarking, and workforce planning.

However, salary structures vary significantly based on experience, job roles, geographic location, and company size. The goal of this project is to analyze a global dataset of data science salaries and uncover patterns that explain these differences.

Key questions addressed in this analysis include:

- What factors influence salary levels in the data science profession?
- Which roles command the highest compensation?
- How do salaries vary across regions and company sizes?
- Can salaries be predicted using job-related features?

Dataset Overview

The dataset used for this analysis contains information about data science jobs between 2020 and 2025.

Key attributes include:

- Work Year
- Experience Level
- Employment Type
- Job Title
- Salary
- Salary in USD
- Employee Residence
- Remote Work Ratio
- Company Location
- Company Size

The dataset provides a global view of compensation patterns within the data science industry.

Data Architecture Diagram

The analytical workflow followed a structured data pipeline:

Raw Dataset
↓
Data Cleaning and Preparation
↓
Feature Engineering
↓
Exploratory Data Analysis
↓
Correlation Analysis
↓
Machine Learning Model (Linear Regression)
↓
Business Insights and Reporting
↓
Dashboard and Visualization

This architecture ensures a systematic approach to data analytics and reproducible results.

Data Cleaning and Preparation

The following steps were performed to prepare the dataset for analysis:

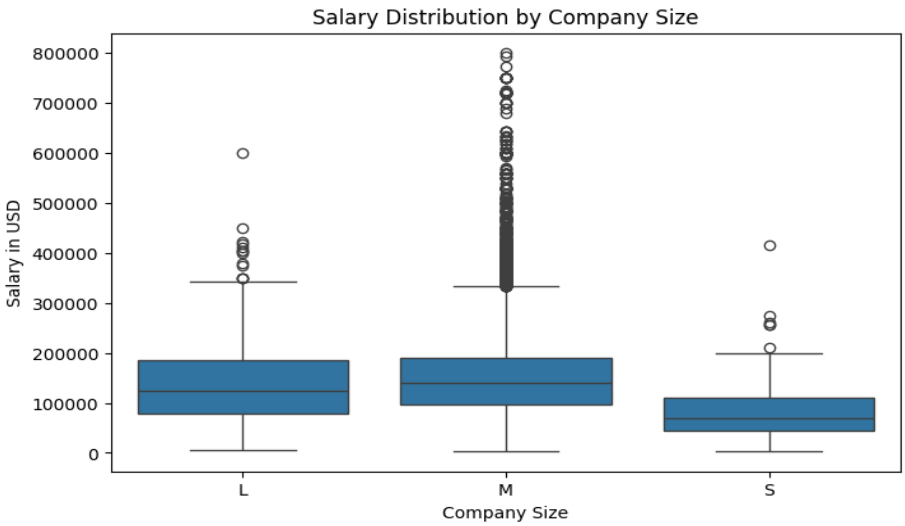
1. Loaded the dataset using Pandas.
2. Inspected the dataset structure using head(), info(), and describe().
3. Checked for missing values and data inconsistencies.
4. Removed duplicate records.
5. Standardized column names and data types.
6. Converted categorical variables into appropriate formats.
7. Verified salary values using the standardized salary_in_usd column.
8. Prepared datasets for visualization and modeling.

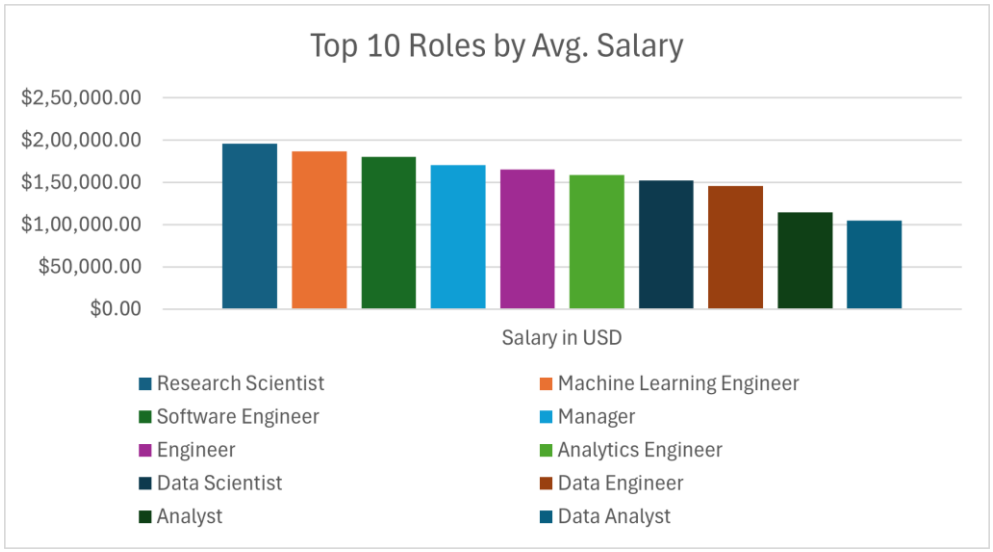
Exploratory Data Analysis (EDA) – Visual Storytelling

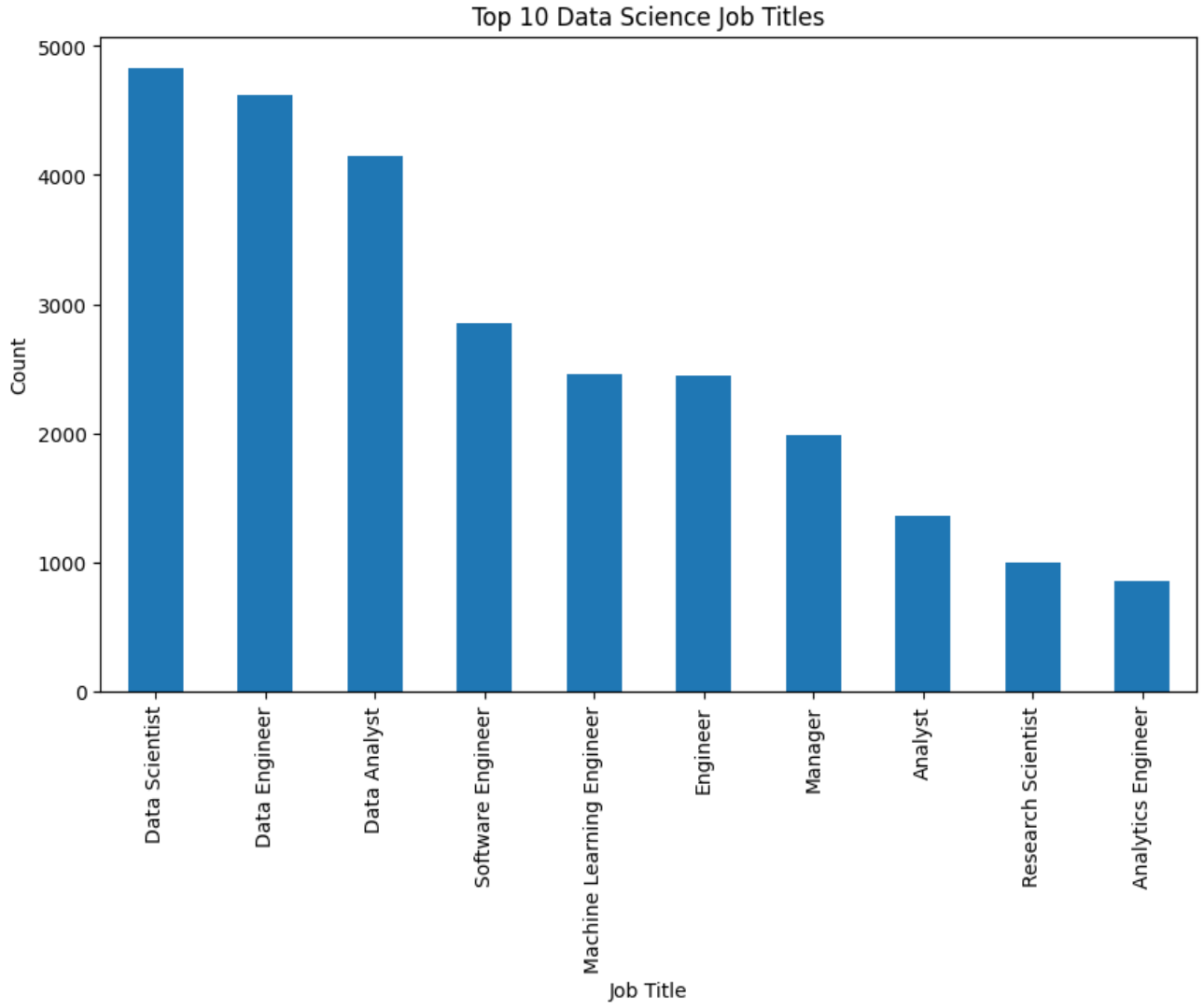
Exploratory Data Analysis was conducted to understand patterns and relationships within the dataset.

Key visualizations included:

- Salary distribution histogram
- Experience level distribution
- Employment type distribution
- Remote work distribution
- Salary distribution by company size
- Salary distribution by employment type
- Salary trend from 2020 to 2025
- Top job roles by frequency
- Top job roles by average salary




Top 10 Roles by Avg. Salary
$2,50,000.00
$2,00,000.00
$1,50,000.00
$1,00,000.00
$50,000.00
$0.00
Salary in USD
Research Scientist
Machine Learning Engineer
Software Engineer
Manager
Engineer
Analytics Engineer
Data Scientist
Data Engineer
Analyst
Data Analyst


Top 10 Data Science Job Titles
Count
5000
4000
3000
2000
1000
0
Data Scientist
Data Engineer
Data Analyst
Software Engineer
Machine Learning Engineer
Engineer
Manager
Analyst
Research Scientist
Analytics Engineer
Job Title

These visualizations revealed significant patterns in compensation structures across roles and industries.

Feature Engineering

Feature engineering was applied to improve analytical insights and predictive modeling.

New features created include:

• experience_level_num – numerical encoding of experience levels
• company_size_num – numerical encoding of company size
• salary_band – salary segmentation categories
• job_category – grouping of job titles into broader categories
• years_since_2020 – trend variable for time-based analysis

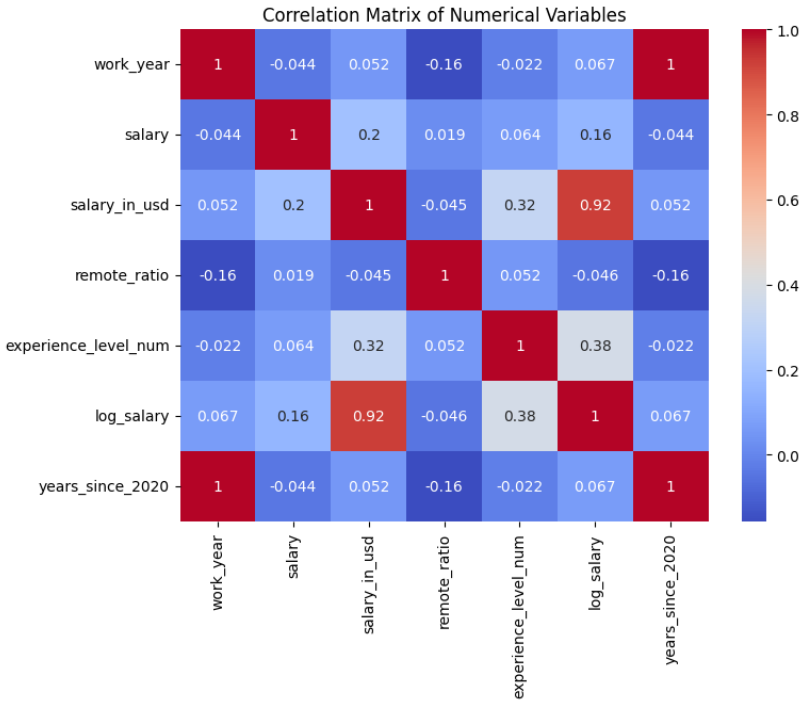
These engineered features enhanced the model's ability to capture salary patterns.

Correlation Analysis

Correlation analysis was conducted to identify relationships between numerical variables.

Key findings include:

• Strong positive relationship between experience level and salary
• Moderate correlation between company size and compensation
• Increasing salary trends over time
• Limited correlation between remote work ratio and salary

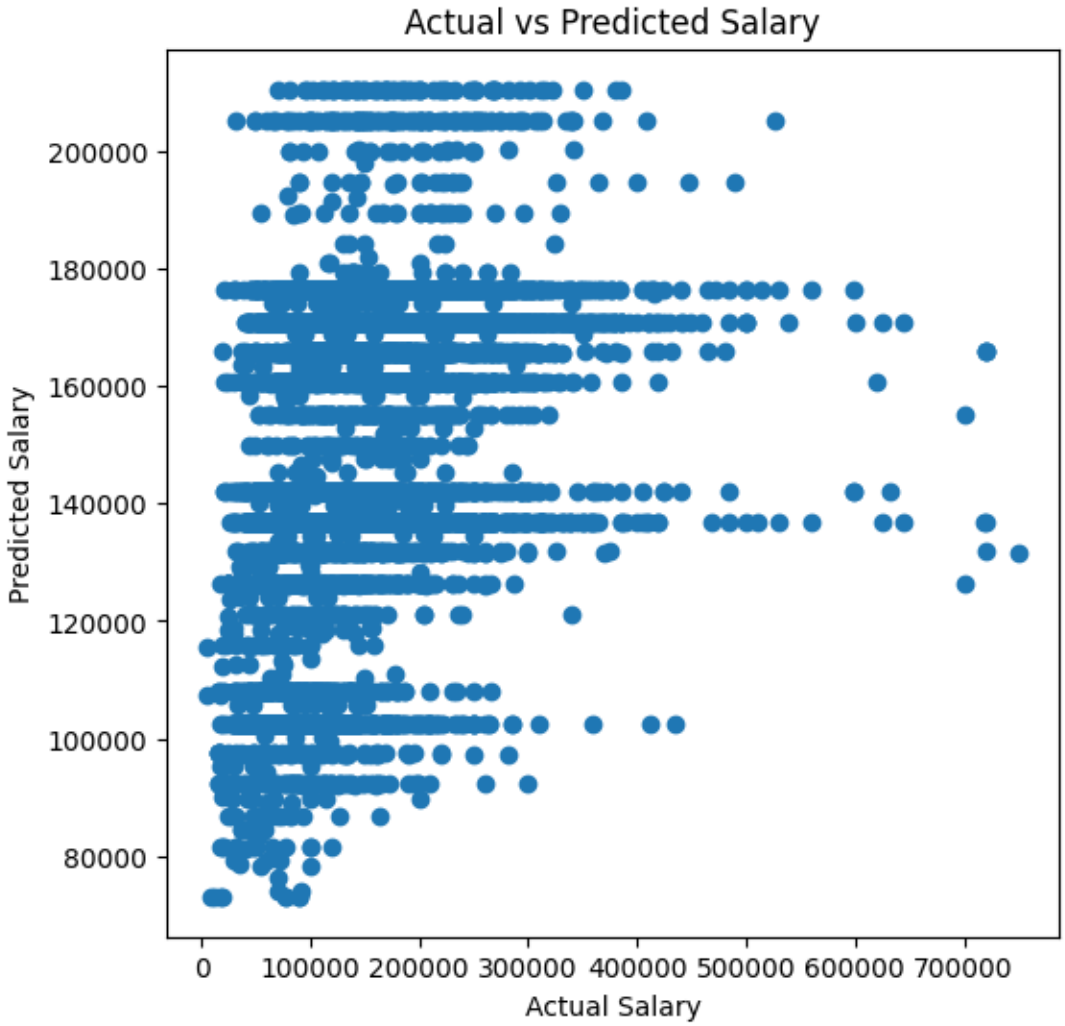


Predictive Modeling

A linear regression model was implemented to predict salaries based on selected features.

Features used in the model:

- Experience level
- Company size
- Remote work ratio
- Work year



The dataset was split into training and testing datasets using an 80/20 ratio.

Model Performance Evaluation

Model performance was evaluated using the following metrics:

• Mean Absolute Error (MAE)
• Root Mean Squared Error (RMSE)
• R^2 Score

These metrics measure the difference between predicted and actual salary values and indicate how well the model captures salary variation in the dataset.

Business Insights

Key insights derived from the analysis include:

1. Experience level is the strongest determinant of salary.
2. Machine learning and AI-related roles command premium salaries.
3. Large companies offer higher salaries compared to small organizations.
4. Remote work does not significantly reduce compensation levels.
5. Data science salaries have increased steadily between 2020 and 2025.

Strategic Recommendations

Organizations should consider the following strategies:

• Invest in experienced data professionals.
• Offer competitive compensation for specialized AI roles.
• Expand remote hiring to access global talent.
• Use salary benchmarking models for compensation planning.

Limitations of the Analysis

While the analysis provides valuable insights, several limitations exist:

• Dataset may not represent all industries equally.
• Salary values may vary depending on benefits and bonuses not included in the dataset.
• Geographic cost-of-living differences are not fully captured.

Future Improvements

Future improvements to this analysis may include:

• Incorporating additional datasets from job portals.
• Using advanced machine learning models such as Random Forest or XGBoost.
• Building interactive dashboards for real-time salary analysis.

Appendix – Python Code Overview

Key Python libraries used:

• Pandas – data manipulation
• NumPy – numerical operations
• Matplotlib & Seaborn – visualization
• Scikit-Learn – machine learning

Example code snippet:

```
df = pd.read_csv("data_science_salaries.csv")

X = df[['experience_level_num','company_size_num','remote_ratio']]
y = df['salary_in_usd']

model = LinearRegression()
model.fit(X_train, y_train)
```