Project Report: Data Science Job Salaries Analysis

---

1. Introduction

The demand for data professionals has grown rapidly in the digital era. As businesses become more data-driven, roles like Data Scientist, Data Analyst, and Machine Learning Engineer are gaining prominence. However, salaries for these roles vary significantly based on multiple factors such as location, experience, employment type, and company size.

This project analyzes a dataset comprising salary information for various data roles collected globally. The dataset provides insights into compensation patterns and helps understand how different features impact salary. The aim is to analyze these patterns, clean and engineer the dataset, and explore insights that can assist job seekers and employers in making data-driven decisions.

---

2. Objective of the Project

The key objectives of this project include:

- ✧ Cleaning and preparing the dataset for analysis.
- ✧ Mapping and transforming complex data points into understandable categories.
- ✧ Performing Exploratory Data Analysis (EDA) to uncover salary trends.
- ✧ Segmenting data by job title, region, company size, and work mode.
- ✧ Extracting insights that reveal the current salary structure in the data industry.
- ✧ Preparing the dataset for potential modeling applications like salary prediction.

---

3. Dataset Description

The dataset used in this project is titled **"Data Science Job Salaries"** and includes the following features:

- ➢ **Work Year**
- ➢ **Experience Level**
- ➢ **Employment Type**
- ➢ **Job Title**
- ➢ **Salary (Local Currency)**

- **Salary in USD**
- **Employee Residence**
- **Remote Ratio**
- **Company Location**
- **Company Size**

The data spans different regions and job types, making it suitable for comprehensive analysis.

---

4. Tools and Technologies Used

This project is implemented using Python in a Google Colab environment. The primary libraries used include:

Pandas – for data manipulation and preprocessing.

NumPy – for numerical operations.

Matplotlib & Seaborn – for data visualization.

Google Colab – for an interactive cloud-based Python environment.

---

5. Data Cleaning and Preprocessing

Data cleaning involved the following steps:

Dropping Irrelevant Columns: The index column was removed as it added no value.

Renaming Columns: To improve readability, columns were renamed to full words (e.g., "experience_level" to "Experience Level").

Mapping Coded Values:

1. "EN" to "Entry-level"
2. "MI" to "Mid-level"
3. "SE" to "Senior-level"
4. "EX" to "Executive-level"

Similar mappings were applied for employment type and company size.

Handling Remote Work Categories:

Remote Ratio of 0 → "On-site"

Remote Ratio of 50 → "Hybrid"

Remote Ratio of 100 → "Fully Remote"

Creating Salary Bands:

Salary (USD) column was divided into Low, Medium, and High bands using quantiles.

Job Domain Classification:

Based on the job title, roles were categorized into:

1. Data Science
2. Analytics
3. Engineering
4. Machine Learning
5. Others

Region Grouping:

Country codes were grouped into continents (Asia, Europe, North America, etc.).

---

6. Exploratory Data Analysis (EDA)

Visualizations and summary statistics were used to extract insights.

A. Salary Distribution

1. Most salaries were concentrated under $150,000.
2. A few high-paying roles skewed the mean.
3. Box plots were used to detect outliers.

B. Salary by Experience Level

1. Entry-level roles had the lowest median salary.
2. Executive-level roles earned significantly more.
3. Senior-level professionals had high consistency in pay.

C. Salary by Job Title

1. Machine Learning Engineers and Data Scientists were among the top earners.

2. Data Analysts had a more moderate salary range.

3. A wide variance was observed within similar job titles due to geography and company factors.

D. Remote Work Impact

1. Fully remote roles earned competitive salaries.

2. Hybrid roles were also well-compensated, showing flexibility doesn't reduce pay.

3. On-site roles were more common in lower-paying jobs in some regions.

E. Salary by Company Size

1. Large companies offered higher average salaries.

2. Medium companies had mixed compensation depending on the role.

3. Some small companies offered higher pay for niche positions.

F. Salary by Employment Type

1. Full-time jobs were dominant and better paying.

2. Freelance and contract roles had variable compensation.

3. Part-time positions had the lowest average pay.

---

7. Geographical Insights

A. Employee Residence vs. Salary

North America and Western Europe had the highest median salaries.

Asian countries (especially India) had lower salary ranges.

The geographic location of the company and employee had a strong influence on pay.

B. Company Location

1. Salaries offered by companies in the USA, UK, and Germany were significantly higher.

2. Companies in India and Eastern Europe offered lower compensation comparatively.

3. Global firms with remote teams helped balance regional pay gaps.

---

8. Feature Engineering Summary

Feature engineering played a vital role in simplifying and categorizing the dataset for better interpretation:

Remote Category: Transformed numeric ratios into meaningful categories.

Salary Band: Helped in segmenting pay levels.

Job Domain: Grouped similar job functions for comparative analysis.

Region: Allowed grouping of countries into continents to assess geographical trends.

---

9. Key Insights and Observations

Experience Level is the most important factor in determining salary.

Remote Work is normalized and does not necessarily result in lower pay.

Company Size has a moderate influence, but **Job Role** and **Location** matter more.

High-paying roles are concentrated in North America and Western Europe.

Machine Learning Engineers and **Data Engineers** have higher average salaries compared to other roles.

Salary data shows clear gaps between developed and developing countries.

---

10. Conclusion

This project provided deep insights into the salary structures for data science-related roles. By cleaning, transforming, and analyzing the dataset, patterns emerged that can guide both job seekers and recruiters. It highlighted how experience, company size, and remote work flexibility influence salaries globally.

The dataset is now ready for further applications such as salary prediction models using regression techniques or clustering similar roles.
