Data Science Job Salaries — Final Report

1) Basic Details
- Name: Vaibhav Parmar
- Email: vaibhavparmar.180708@gmail.com
- Institution: Atmiya University
- Project Title: Data Science Job Salaries — Cleaning & EDA (No ML)
- Tech Stack: Python, Pandas, Matplotlib

2) Problem Statement
Clean and explore a dataset of data science job salaries to understand roles, experience levels, geographic trends, remote work, and salary ranges using simple, clear analysis — no machine learning.

3) Dataset
- Source: Data Science Job Salaries dataset (CSV provided)
- Size: 607 rows after cleaning (removed duplicates: 0)
- Time coverage: 2020–2022

4) What I Did
- Loaded the CSV and checked its basic structure (columns and types).
- Standardized column names, removed exact duplicate rows, and filled key text fields with 'Unknown'.
- Converted numeric columns (salary, salary_in_usd, remote_ratio, work_year) safely.
- Summarized key fields (job titles, locations, experience levels).
- Created simple charts: top job titles, average salary by experience, median salary trend by year, top company locations, and remote ratio distribution.

5) Insights
- Most frequent roles include: Data Scientist, Data Engineer, Data Analyst, Machine Learning Engineer, Research Scientist (top among 10 shown).
- Average salary by experience (USD) shows higher pay for senior roles — EX: $199,392, SE: $138,617, MI: $87,996, EN: $61,643.
- Common company locations: US, GB, CA, DE, IN.
- Remote work mix based on 'remote_ratio': 0% remote: 127, 50% remote: 99, 100% remote: 381.
- Median salaries by year indicate a trend across 2020–2022.

6) Challenges
Some rows had missing or inconsistent values in text columns (job titles, locations). There were also differences in numeric fields (currency conversions are already reflected in 'salary_in_usd'). Another minor challenge was ensuring that remote work ratios and experience levels were interpreted correctly across the dataset. Overall, light cleaning was enough for a reliable exploratory analysis.

7) Conclusion
This project cleaned and explored a real dataset of data science job salaries without machine learning. The analysis highlights common roles, geographic distribution, remote work patterns, and how salaries vary with experience and over time. These findings give a straightforward picture of the data science job market and can help students understand how to approach EDA on salary datasets.

8) How to Run
- Open and run the Jupyter notebook: DS_Salaries_EDA.ipynb
- Charts and the cleaned CSV are saved to outputs_salaries/ folders.

9) Project Links
- GitHub repo link: https://github.com/vaibhavparmar96/Internship-projects