🗂️ DataSense – Final Sprint Challenge Summary

🧑‍💻 Overview

The goal of this project was to analyze the U.S. Adult Income dataset to identify the key socio-demographic factors influencing whether a person earns more than $50K annually. This was part of the final sprint challenge to apply data cleaning, EDA, and storytelling skills.

🧹 What I Did

1. Cleaned the Data:
   - Replaced all placeholder values (?) with NaN.
   - Dropped rows with missing values in key columns like workclass, occupation, and native-country.
   - Final dataset: 45,222 clean records and 15 columns.
2. Performed EDA Using:
   - Python libraries: pandas, seaborn, matplotlib.
   - Visuals created:
     - Salary Distribution
     - Occupation vs Salary (Countplot)
     - Education Level vs Salary (Countplot)
     - Gender vs Salary (Countplot)
     - Hours-per-week vs Salary (Boxplot)

📊 Key Insights

- Most people in the dataset earn ≤50K.
- People earning >50K tend to be older (typically 40s–50s).
- Higher education levels (e.g., Bachelors, Masters, Doctorate) strongly correlate with higher income.

- Males are more likely to earn >50K compared to females, indicating a gender income gap.

- Individuals earning >50K generally work more than 40 hours per week.

🛠️ Tools Used

- Python: pandas, numpy, seaborn, matplotlib

- Jupyter Notebook / Google Colab

📌 Conclusion

This project shows how basic demographic and occupational attributes can be leveraged to understand income trends. These insights are not only useful for policy and workforce strategy but can also be fed into predictive models for classification problems.