**Internship Assignment: “Decode the Data – Real-World Chaos to Clarity”**

**🗂️ Assignment Title:**

**"Reverse Engineering a Mystery Dataset"**

**🎯 Objective:**

You are given a **raw, unlabeled dataset** (mystery\_dataset\_complex.csv) that mimics real-world data chaos. It includes:

* Inconsistent formats
* Missing or misleading values
* Embedded JSON
* Mixed data types
* No headers

Your goal is to **make sense of the data**, **clean it**, **understand the underlying structure**, and **generate meaningful insights** using **Python**.

**📦 What’s Inside:**

* 1000 rows
* 20 columns, unlabeled
* No documentation or column descriptions
* Contains names, locations, dates, income, categories, booleans, skills, and job-related data — all mixed

**🛠️ Your Tasks:**

**1. Decode the Data Structure**

* Load the dataset and analyze each column
* Assign **appropriate names** to all columns (e.g., name, age, income, etc.)
* Justify your interpretation of what each column means

**2. Clean and Normalize**

* Fix inconsistencies in data formats (e.g., "yes", "Y", 1)
* Handle missing/invalid values
* Extract embedded JSON fields
* Convert all values into consistent, analysis-friendly formats

**3. Perform Exploratory Analysis**

* Use pandas, numpy, and visual libraries (matplotlib, seaborn, or plotly)
* Identify **at least 5 meaningful patterns or insights**, such as:
  + What cities or categories dominate?
  + Any correlation between income and experience?
  + Who tends to have more skills?
  + Differences between employment types?
* Bonus: Try clustering profiles using logic or unsupervised learning (e.g., KMeans)

**4. Document Your Process**

* Use Jupyter Notebook or clean .py scripts
* Add comments and markdown explanations
* Clearly list all **assumptions**, **issues found**, and **fixes applied**

**Create a Dashboard**

**Dashboard Guidelines:**

* Use Plotly Dash, Streamlit, or Flask + Plotly
* Provide:
  + Filters (e.g., by city, job title, category)
  + Visuals for insights (bar, pie, scatter, etc.)
  + Clean summary statistics
* Bonus if you deploy it (locally is fine)

**✅ Deliverables:**

1. **Cleaned & labeled dataset** (cleaned\_dataset.csv)
2. **Python/Jupyter analysis** (notebook.ipynb or .py file)
3. **Dashboard app** (optional but encouraged)
4. **README or report** summarizing:
   * Column descriptions
   * Cleaning steps
   * Key insights
   * Screenshots (if dashboard included)

**⏱️ Duration:**

**2–3 days**

**🧠 Tip for Interns:**

This is not a right-or-wrong problem — it’s about how you **think**, **reason**, and **present**. Do not copy code blindly or use ChatGPT for direct answers — we want **your approach**, even if imperfect.