Amazon Dataset Analysis

📌 Overview

This project focuses on data cleaning and preprocessing for a structured dataset containing financial and business-related information. It demonstrates how to transform raw, unstructured values (e.g., currency formats) into clean, analysis-ready data. The goal is to ensure the dataset is consistent, accurate, and ready for further analysis or modeling.

📁 Project Structure

bash

CopyEdit

```
├── Part1.ipynb   # Main Jupyter Notebook with cleaning and preprocessing steps
└── README.md     # Project documentation
```

📊 Dataset

The dataset contains:

- **Financial data:** values with currency symbols (₹) and formatting
- **Categorical fields:** business classifications, labels, or categories
- **Date/time values:** transaction or reporting periods
- **Numerical metrics:** revenue, quantity, or other measurable data points

(Dataset source not specified — can be added if available)

🛠️ Tools & Libraries

The analysis uses:

- **Python 3**
- **pandas** – data manipulation and cleaning
- **numpy** – numerical operations
- **re** – regular expressions for text and pattern processing
- **Jupyter Notebook** – interactive development

🔍 Analysis Highlights

The notebook covers:

- **Currency Formatting Cleanup** – removing ₹ and commas from numerical fields

- **Datatype Conversion** – converting string-based numbers into numeric types
- **Missing Value Handling** – identifying and filling or removing nulls
- **Text Normalization** – standardizing categorical values
- **Initial EDA** – checking dataset statistics after cleaning

---

📈 **Key Insights**

- Many financial values were stored as text due to currency symbols and commas.
- Regex-based cleaning significantly improved data consistency.
- After cleaning, the dataset was ready for statistical analysis and visualization.
- Proper preprocessing reduced the risk of errors in later analysis stages.