📄 Project Documentation: Data Cleaning & Automation Pipeline

📌 Project Title: Retail Sales Data Cleaning & Automation Pipeline

---

🎯 Objective:

This project demonstrates how to **automate the cleaning and transformation** of messy, multi-source retail sales data using Python. It converts raw, unstructured sales data into a clean and structured format for use in dashboards, reporting, and analytics.

📁 Dataset Overview:

- **Raw Data File**: retail_store_sales.csv *(simulated sample dataset)*
- **Output File**: retail_store_sales_cleaned.csv
- **Data Size**: ~1,000–10,000 rows (customizable)
- **Industry Use Case**: Retail / E-commerce

📜 Columns in the Dataset:

Column	Description
Transaction_Date	Date of purchase
Quantity	Number of items bought
Price_per_Unit	Unit price of each item
Category	Product category
Location	Store or delivery location
Payment_Method	Cash / Card / UPI / Other

🛠 **Tools & Technologies Used:**

- **Python**
- **Pandas, NumPy**
- **Jupyter Notebook** (for exploration and testing)
- **Standalone Python Script** (cleaning_pipeline.py) for automation

---

⚙️ **Automation Workflow:**

The cleaning logic is implemented in a Python function:

def clean_retail_data(input_file, output_file):

...

✅ **Cleaning Steps Performed:**

Task	**Description**
✅ Column Standardization	Strip spaces, convert to lowercase, replace spaces with underscores
✅ Date Conversion	Convert transaction_date to datetime format
✅ Numeric Conversion	Handle invalid quantity & price_per_unit with median imputation
✅ Feature Engineering	Created new column: total_spent = quantity * price_per_unit
✅ Text Cleaning	Standardize category, location, and payment_method text
✅ Missing Value Handling	Replaced nulls with 'Unknown' or median
✅ Duplicate Removal	Dropped exact duplicate rows
✅ Output Generation	Saved cleaned dataset to new .csv file

📦 Folder Structure:

RetailDataPipeline/

├── cleaning_pipeline.py # Automation script

├── cleaning_pipeline.ipynb # Notebook for development/testing

├── retail_store_sales.csv # Raw dataset

└── retail_store_sales_cleaned.csv # Cleaned output

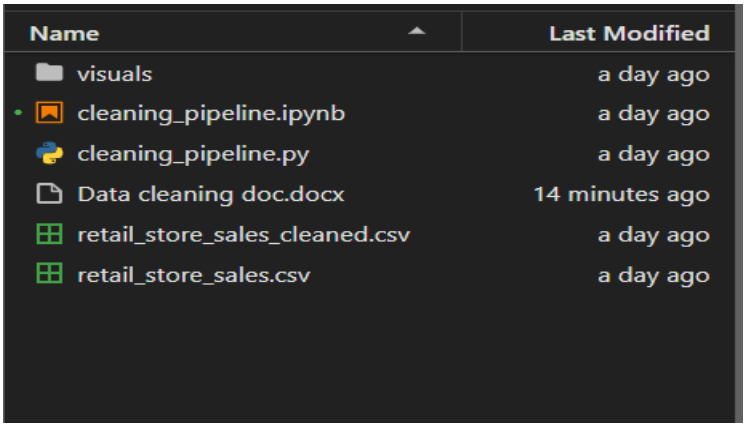
🔄 Execution Instructions:

To run the automated script:

python cleaning_pipeline.py

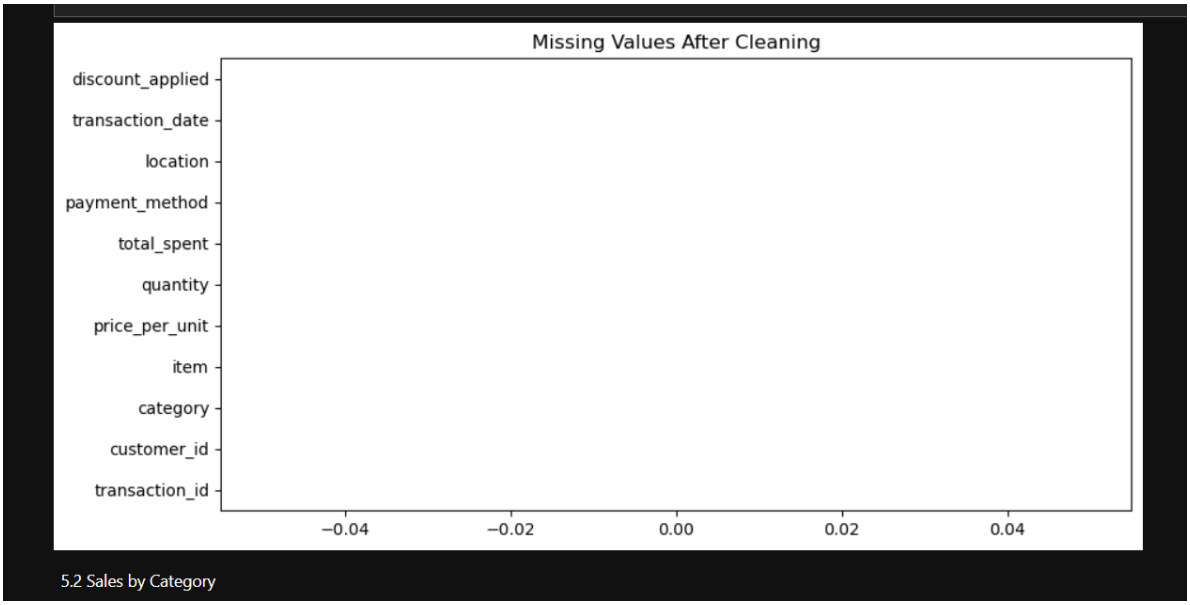
This will:

- ***Read***: retail_store_sales.csv
- ***Clean and transform*** the data
- ***Save result as:*** retail_store_sales_cleaned.csv



📈 Use Cases & Applications:

- ✅ Retail / eCommerce companies preparing raw transactional data
- ✅ Analysts automating pre-processing before building dashboards
- ✅ Data pipelines where clean CSVs are needed for Tableau/Power BI
- ✅ Preprocessing step for machine learning models



🧠 Key Skills Demonstrated:

- Programmatic Data Cleaning
- Automation of Manual Reporting Tasks
- Handling Missing/Corrupt Values
- Data Type Standardization
- Efficient Pipeline Structure (Function-based + CLI executable)
- Reusability (Can be applied to other retail datasets)

📬 Contact & Portfolio:

If you're looking for a results-driven **Freelance Data & Business Analyst** who can turn raw data into clear strategy — I'd love to help you grow!

📧 Email: vaishnaviganeshkar15@gmail.com

🔗 LinkedIn: https://www.linkedin.com/in/vaishnavi-ganeshkar/