Robust Multi-Source Sales Data Pipeline

Built by **Vangara Yaswanth Sai** for Flipkart Task-1

Last Updated: **17 July 2025**

Overview

This project implements a **custom sales data pipeline** that ingests, cleans, analyzes, and visualizes sales-related data from multiple sources: **CSV, JSON, and Excel**.
It is designed to be:

- Modular
- Insight-rich
- Visualization-capable
- Extensible for large-scale or real-time pipelines

File Structure

```
Sales_Data_Pipeline_Main.py # Main Python script with the complete pipeline
custom_merged_data.csv # Auto-generated merged dataset (output)
auto_insights_dashboard.png # Auto-generated visualizations (output)
custom_insights_report.txt # Auto-generated insights (output)
```

Features

- Multi-source data loading (CSV, JSON, Excel)
- Auto-cleaning and type conversion
- Smart merging using auto-detected keys
- Insight extraction: stats, revenue, category analysis, time series
- Visualizations: histograms, bar charts, time plots, correlation matrix
- Exports: Merged CSV, PNG dashboard, Text report
- Custom handling for quirky real-world datasets

Requirements

Install dependencies via pip:

```
pip install pandas numpy matplotlib seaborn tabulate openpyxl
```

Usage

1. Configure Your File Paths

Edit the following in the __main__ block:

```
your_csv_file = r"path\to\your\sales_data.csv"
your_json_file = r"path\to\your\product_metadata.json"
your_excel_file = r"path\to\your\region_info.xlsx"
```

2. Run the Pipeline

```
python Sales_Data_Pipeline_Main.py
```

You will be prompted whether to generate visualizations.

Output Files

After successful execution, you'll get:

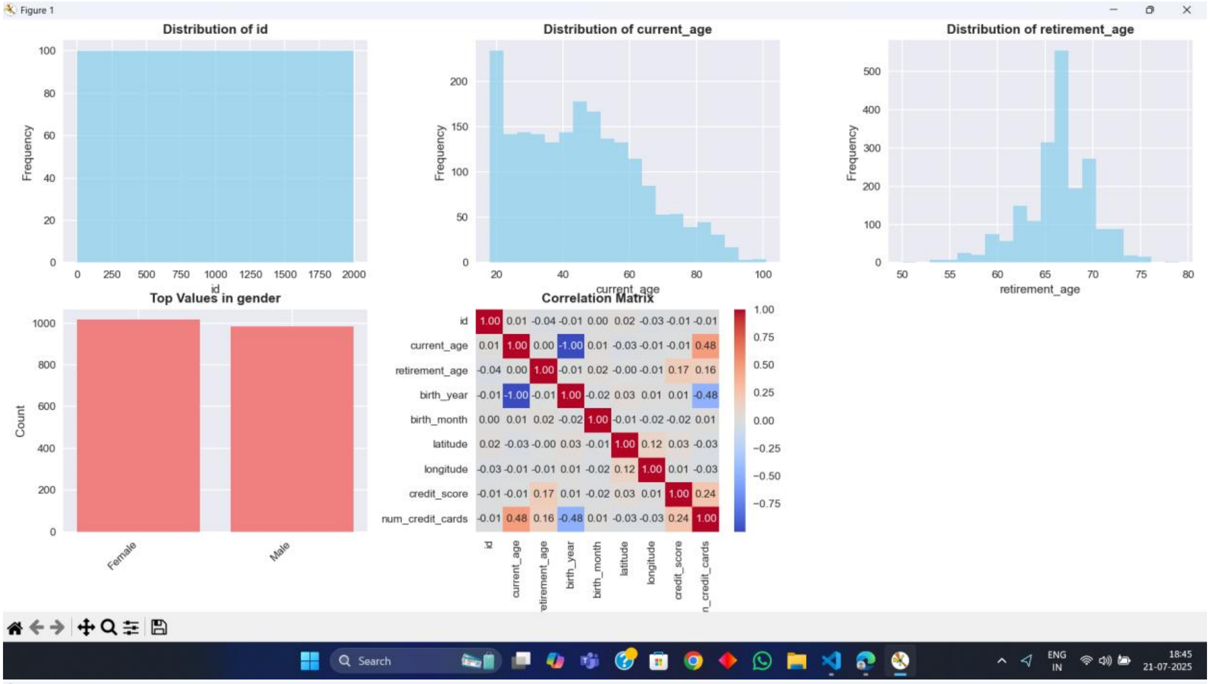
- custom_merged_data.csv: Merged, cleaned dataset
- custom_insights_report.txt: Readable report with stats and findings
- auto_insights_dashboard.png: Visual dashboard

Example Insights Extracted

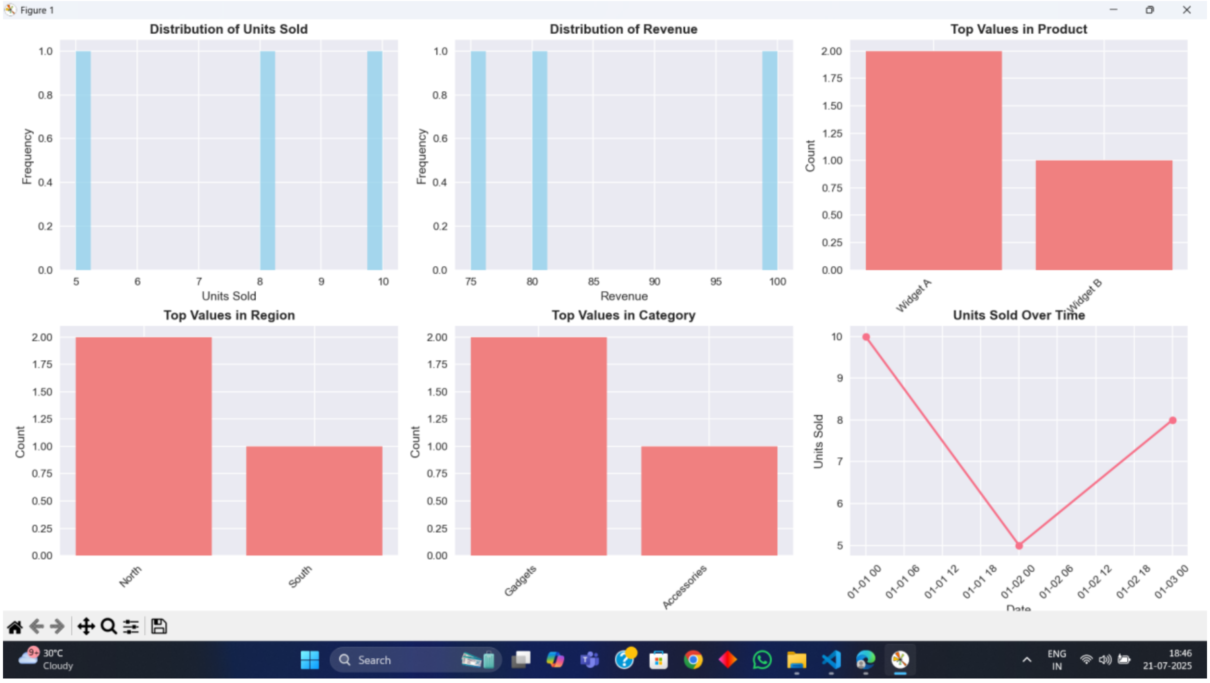
```
DATA OVERVIEW
Total Records: 2000
Total Columns: 14
Numeric Columns: 9
Text Columns: 5
Date Columns: 0


CATEGORICAL ANALYSIS
GENDER:
Unique values: 2
Top categories:
Female: 1016
Male: 984
```

- Total records, column types
- Financial summary: total revenue, average cost, etc.
- Top categories by count
- Daily performance over time
- Strong correlations between metrics

Sample Visuals



	id	current_age	etirement_age	birth_year	birth_month	latitude	longitude	credit_score	n_credit_cards
id	1.00	0.01	-0.04	-0.01	0.00	0.02	-0.03	-0.01	-0.01
current_age	0.01	1.00	0.00	-1.00	0.01	-0.03	-0.01	-0.01	0.48
retirement_age	-0.04	0.00	1.00	-0.01	0.02	-0.00	-0.01	0.17	0.16
birth_year	-0.01	-1.00	-0.01	1.00	-0.02	0.03	0.01	0.01	-0.48
birth_month	0.00	0.01	0.02	-0.02	1.00	-0.01	-0.02	-0.02	0.01
latitude	0.02	-0.03	-0.00	0.03	-0.01	1.00	0.12	0.03	-0.03
longitude	-0.03	-0.01	-0.01	0.01	-0.02	0.12	1.00	0.01	-0.03
credit_score	-0.01	-0.01	0.17	0.01	-0.02	0.03	0.01	1.00	0.24
num_credit_cards	-0.01	0.48	0.16	-0.48	0.01	-0.03	-0.03	0.24	1.00



- Sales over time plot
- Category distribution

- Revenue distribution histogram
- Correlation heatmap

Error Handling

The pipeline uses specific exceptions:

- Handles missing or corrupt files gracefully
- Warns about unsupported formats or merge conflicts
- Logs failed conversions per column

Version Control

This project is under Git version control:

- All major updates and commits are tracked
- Code is production-ready and AI-clean

Author

Vangara Yaswanth Sai
Flipkart Data Pipeline Task-1 Project