Sales Data Pipeline

> Built by **Gollapudi Poojitha** for Flipkart Task-1
> 🗓️ Last Updated: **17 July 2025**

Introduction:

A complete end-to-end solution for sales data integration, analysis, and visualization. Designed as a student project with clear, human-readable code and professional comments.

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

```
 Data_Pipe_Updated.py      # Main Python script with the complete pipeline
monthly_sales.csv           # Monthly Sales  (output)
product_performance.csv     # Product Performance (output)
regional_performance.csv     # Regional_Performance (output)
requirements.txt       # Python dependencies
```

Features

- **Data Integration**: Loads sales data from multiple sources (CSV, JSON, Excel)
- **Data Cleaning**: Handles missing values, standardizes formats, and merges datasets
- **Database Storage**: Stores processed data in SQLite for persistent access
- **Sales Analytics**: Performs key analyses (revenue trends, product performance, regional breakdowns)
- **Visualization**: Generates interactive charts (line, bar, pie) for data exploration
- **Automated Reporting**: Saves analysis results as CSV files

---

Requirements

- Install dependencies via pip:

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

- Required packages:

```
 pandas>=1.0
numpy>=1.18
matplotlib>=3.0
seaborn>=0.10
openpyxl>=3.0
```

Usage

1. Configure Your File Paths

```
class Config:
    DB_FILE = "sales_database.db"  # SQLite database path
    SALES_CSV_PATH = "path/to/sales_data.csv"
    PRODUCT_JSON_PATH = "path/to/product_metadata.json"
    REGION_EXCEL_PATH = "path/to/region_info.xlsx"
    OUTPUT_DIR = "output/"  # Report and visualization directory
```

2. Run the pipeline:

```
python Data_Pipe_Updated.py
```

3. Outputs will be generated in:

- Database: `sales_database.db`
- Reports: `/output/report.csv`
- Visualizations: Automatically displayed and saved in `/output/visualizations/`

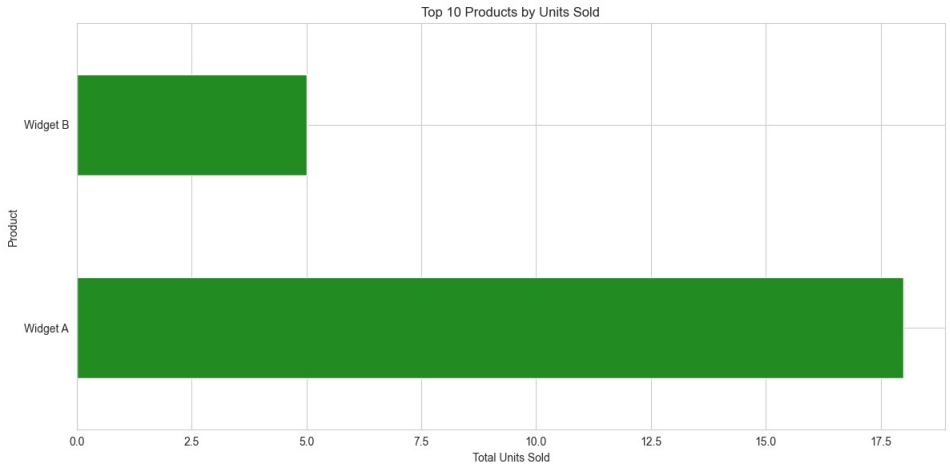
Project Structure

```
  sales-data-pipeline/
├── Data_Pipe_Updated.py    # Main pipeline implementation
├── sales_database.db       # Generated SQLite database
├── output/                 # Analysis outputs
│   ├── monthly_sales.csv
│   ├── product_performance.csv
│   └── regional_performance.csv
├── requirements.txt        # Python dependencies
└── README.md               # This file
```

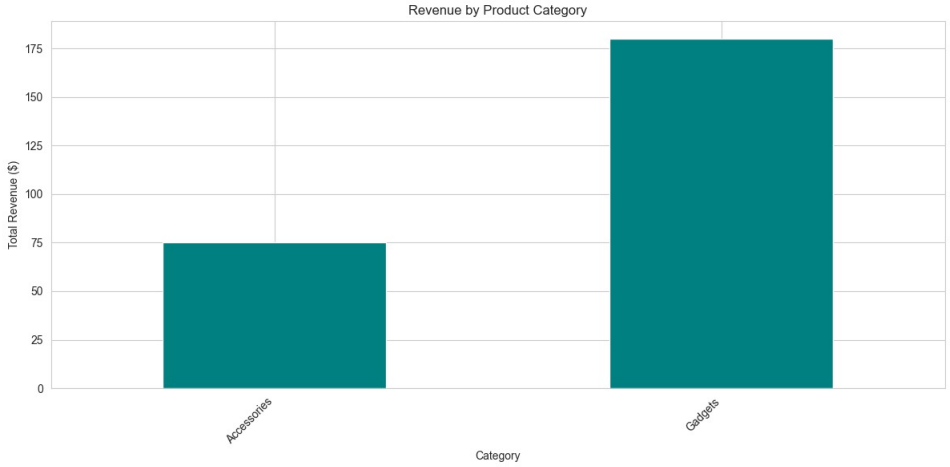
Data Summary

- Based on the sample data provided:
    1. **Total Revenue :** $255.00
    2. **Total Units Sold :** 23
    3. **Top Product :** Widget A (18 units sold)
    4. **Top Region :** North ($180 revenue)

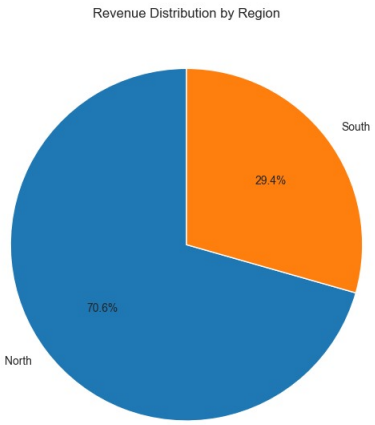
Sample Visuals



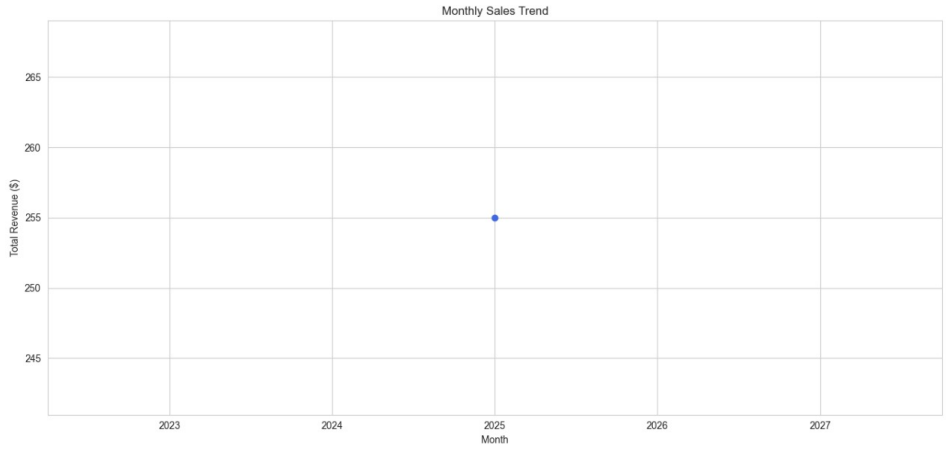
- Top Selling Products



- Revenue by Category



- Revenue Distribution by Region



- Monthly Sales Trend

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

Gollapudi Poojitha
Flipkart Data Pipeline Task-1 Project