Technical Report: Amazon Sales Analytics Pipeline

Case Explanation

The project involved building a complete data pipeline to standardize and analyze Amazon sales data for an e-commerce company. The primary challenge was transforming raw, inconsistent sales data into a clean, reliable dataset.

Assumptions Made

Data Structure & Content

- The dataset contained essential e-commerce columns: product details, pricing, ratings, categories, and customer reviews
- Indian currency symbols (₹ and â‚¹) were present in price columns and needed special handling
- Rating counts contained comma separators for thousands
- Discount percentages were stored as strings with '%' symbols

Data Quality Handling

- Products without `product_id` or `product_name` were considered invalid and removed
- Missing ratings were filled with the median value to avoid skewing averages
- Zero values for rating counts were accepted as valid (products with no ratings)
- Duplicate entries based on product-user-review combinations were removed

Business Logic

- Profit margin calculated as `actual_price - discounted_price`
- Discount ratio computed as `discounted_price / actual_price`
- Rating count used as a proxy for product popularity/sales volume

Solution Approach & Methodology

Chosen Approach

I selected a **Python-based ETL pipeline** with **Streamlit for visualization** because:

1. **Python Ecosystem**: Rich libraries for data processing (pandas) and visualization (plotly)
2. **Rapid Prototyping**: Streamlit allows quick dashboard development without frontend complexity
3. **End-to-End Pipeline**: Single technology stack from data cleaning to visualization
4. **Portfolio Demonstration**: Shows full-stack data engineering capabilities

Implementation Strategy

- **Incremental Development**: Built and tested each component separately (ETL → DB → Dashboard)
- **Defensive Programming**: Added comprehensive error handling and data validation
- **User-Centric Design**: Created interactive filters and intuitive visualizations
- **Production-Ready Code**: Included logging, configuration management, and documentation

> Although some fixings were approached using AI to find solutions for uncommon circumstances.

Key Results & Visualizations

Data Processing Outcomes

- Successfully cleaned **1,465 product records** with complex currency formatting
- Reduced data inconsistencies by **standardizing 15+ columns**
- Generated **2 new business metrics** (profit margin, discount ratio)

Dashboard Insights

- **Top Products**: Identified highest-rated and most-reviewed products
- **Category Distribution**: Revealed dominant product categories and their revenue contribution
- **Pricing Strategy**: Showed correlation between discounts and customer ratings
- **Profitability Analysis**: Visualized margin distribution across product portfolio

Technical Achievements

- Built a **fully functional ETL pipeline** handling real-world data challenges
- Created an **interactive dashboard** with 6 distinct analytical views
- Implemented **robust data cleaning** for international currency formats
- Delivered **production-quality code** with proper error handling and documentation

Future Improvements & Adjustments

Immediate Enhancements

- Add automated data validation rules and quality checks
- Implement unit tests for critical data transformation functions
- Create scheduled pipeline execution (e.g., daily data refreshes)

Scalability Considerations

- Database migration from SQLite to PostgreSQL for production use
- Add data partitioning for handling larger datasets
- Implement caching mechanisms for dashboard performance

Personal Reflections

This project demonstrated the importance of **practical data engineering** - beyond just theoretical knowledge. The most challenging aspect was handling the real-world data inconsistencies, particularly the Indian currency symbols that required multiple encoding approaches.

I particularly enjoyed the **end-to-end nature** of this project - from raw data to business insights. It reinforced how data cleaning decisions directly impact analytical outcomes and business decisions.

The experience highlighted that **user-friendly visualization** is as crucial as robust data processing. Building the dashboard helped me appreciate how technical work translates into business value through accessible insights.