The Data-to-Decision Pipeline: Transforming Global Superstore's Transactions

By : Dhusyanth R S

Email: dhusyanthrs3@gmail.com | LinkedIn: www.linkedin.com/in/dhusyanth-r-s-

GitHub Repository: Data to Decisions pipeline Project

Executive Summary

This project delivers actionable business intelligence for Global Superstore by analyzing over 50,000 transactional records to drive profitability and operational excellence. The analysis, conducted using an end-to-end data pipeline (Pandas, SQL, Power BI) , identified five critical areas for strategic intervention. Key recommendations focus on cost control, margin optimization, and targeted regional investment.

Introduction

Global Superstore faces the challenge of managing vast transactional data across global markets. The mission of this analysis was to transform this raw data into actionable intelligence. The focus areas included identifying sales masking profit issues, analyzing global trends, optimizing product margins, and evaluating logistical efficiency.

Data Engineering & Preparation (Pandas)

The process began with Python (Pandas), where I performed the initial inspection of the dataset .

Date columns(Order Date, Ship Date) were converted from irregular format(string) to proper Pandas DateTime format.

The postal code column contained **approximately 80% null values (NaNs)**. As this granularity was not relevant to the analysis and would impair data quality, the column was dropped.

Gross Sales columns was created by using mathematical formula (sales /(1-discount)). This served as a critical step for the further validation of the existing metrics .

Once all these are preprocessing steps are done to get the clean data, then I performed the modelling part which is the most critical step of this phase 1 .This allows to get rid of redundant data and get the data stored in a manner in which further analysis will be efficient.

The flat wide table is then decomposed into Fact (central) and Dimension tables(products, customers, locations, logistics).

Dimension tables are created from the existing flat table and dropped the duplicate rows from the dimension tables so that the data will be very concise .

The unique keys(Primary keys) are created by resetting the index .

Then performed the newly created dimension tables with the flat table on the columns which are common .This process is to integrate the keys into the fact table .

Once that is done then the non required columns from the fact table was dropped keeping the respective keys only(foreign key) for their specific dimension.

Few preliminary visualizations are done using matplotlib to ensure the successfull completion of the modelling process since it involves merging across tables to get the job done.

These tables are then migrated to the sql database by connecting python with MySQL in which all the tables are created with integrity rules .

Note : The table is not completely normalized to 3NF . This step is cruicial and intentional as this model (Star Schema) is the most perfect robust way for faster queries (Online Analytical Processing) . This model helps in both way having significantly less redundant data and quick query time for analysis.

The Python(Pandas) code can be found here

Data Integrity and Advanced business querying (SQL)

I used MySQL for this project. This is the place where the data got certified as it is stored in the most efficient way.

Created the fact and dimension tables with integrity rules (primary and foreign key constraints) before the migration of the data from python.

Then I performed 13 business level reporting queries using advanced windows (Rank ,Lag,Ntile etc) mixing with CTEs that gave the valuable insights of the company's metrics across all categories.

The 13 Business problems are listed below followed by the link of the SQL file .

1. Product Profitability Ranking:

    Top 10 Products: Sales vs. Profit Margin Rank

2. Year-over-Year (YoY) Profit Growth:

Customer Segment YoY Profit Growth

3. Customer Lifetime Value (CLV) & Metrics:

Top 10 Customer Lifetime Value (CLV)

4. Cost Efficiency Ratio:

Shipping Cost as % of Sales by Mode & Segment

5. Geographic Profit Anomaly:

Top 5 Cities: High Discount Sales Percentage

6. Sustained Sales Decline:

Sub-Categories with 3+ Months of MoM Quantity Decline

7. Quarterly Trend Analysis & Ranking:

Quarterly Category Sales Rank (Last 8 Quarters)

8. Discount Tier Profitability:

Discount Tier Analysis: Avg Profit & Quantity

9. Regional Performance Quartiles:

High Sales / Low Profit Margin States Anomaly

10. Sales Volatility Detection:

Sales Volatility: MoM Change Exceeding 20%

11. Cumulative Market Growth:

Market-Specific Cumulative Sales Trend

12. New Product Performance:

New Product Performance: First 3-Month Avg Sales & Profit

13. Above/Below Average Profitability:

Order Profit Margin vs. Category Average

The SQL file for setting up the tables can be found here

The SQL file which contains all the queries can be found here

Visualization and Storytelling (PowerBI & DAX)

I connected SQL DataBase to PowerBI to transfer all the tables (Fact and Dimensions). Performed Initial column standardization in Power Query Editor.

Calendar Table was created using DAX and then connected with the fact table to create a Star Schema Model .

I created a main page upfront with insightful visuals which tells what is happening from the top view.

The following subsequent pages tells about more granular details when drilled through .

Some of the Key areas are :

Sales masking underlying profit issues

Global and regional trend analysis

Product performance and margin optimisation

Strategic regional investment opportunities

Logistical efficiency and cost control

The full Executive dashboard can be found here

Key Takeaways of the Project

Control logistics costs

Restrict Same Day Delivery to high-density locations, protecting 13% of profit margins

Focus on margins, not just sales

Furniture's high sales mask poor profitability—prioritise margin-rich categories

Implement granular SKU monitoring

Continuous margin analysis prevents profit-killing products from eroding success

Invest in second-tier cities

Markets like Hamburg demonstrate superior efficiency—double down on these opportunities

Protect top customers

Launch a formal loyalty programme to retain the vital customer segment driving disproportionate profits