Global Superstore Analysis

Github: https://github.com/allenadron3-lang/data-insights-portfolio

This investigation focuses on the exploration of retail operational efficiency and managerial accountability using the Global Superstore Dataset. Businesses that are operating in a competitive global environment, it is essential for them to understand the drivers of sales success and profit erosion to achieve sustainable growth. With this in mind, the investigation addresses a key business challenge: the difference between high-volume sales and sustainable high margin profits. The significance of this investigation lies in its ability to quantify profit loss caused by operational factors like shipping methods and returns. Which I then directly correlate these operational issues with regional managerial performance. The findings will provide actionable insights for adjusting pricing strategies, reallocating resources, and implementing targeted manager-specific training programs to maximize the company's net profit.

This report addressed numerous interconnected research questions designed to provide a holistic view of the superstores performance:

1. Manager Profit Performance : Which sales manager generates the highest total net profit for the company, after accounting for product returns?
2. Returns Accountability: What are the top 5 sub-categories that contribute the most to total loss profit due to returns?
3. Loss Leaders in Profit Categories: Which ship mode is associated with the lowest profit margin, and is this low margin caused by higher discount rates or by higher return rates?
4. Operational Efficiency Tradeoffs : Is there a correlation between a manager's performance in a given year and the return rate in their assigned region?

My approach utilized Python with the pandas library for data preparation and analysis, in combination with Matplotlib and seaborn for visualization. The initial critical step involved combining three distinct datasets, Orders/Returns/People, using left joins operations based on common keys like 'Order_ID' and 'Region'. I then had to create a simple 'Is_Returned' flag column to link returns to individual sales and transactions. The merged dataset was then subjected to aggregation, filtering, and correlation analysis to acquire quantitative answers to the outlined research questions.

The data used for this analysis was derived from the Global Superstore Dataset, a quite new retail transaction sample designed for business intelligence and data science training. The data was provided in a single excel workbook, containing three distinct sheets, each representing different data entities:

1. Orders: Contains detailed records for all sales transactions, including product information, prices, discounts, and location.
2. Returns: A list detailing which Order_IDs were returned by the customer.
3. People: A lookup table linking ‘Region’ names to the corresponding ‘Person’(Sales Manager)

I acquired this data via .xlsx download on Kaggle. Regarding data origin, the data is likely fabricated. Meaning that it lacks external validation through real world benchmarks or live operational data. This limitation must be acknowledged when interpreting the findings, as certain patterns may not reflect true market dynamics. To further evaluate the fitness of the Global Super Store dataset for operational analysis, I assessed the data using the 5 V’s framework:

Volume: The Global Super Store contains over 50,000 individual transactions. I believe this volume of data is sufficient enough to provide statistically significant insights when breaking down performance by manager or product sub category without the results being skewed by a few isolated sales.

Velocity: The data covers a four year period (2014-2017). While the data is purely historical and not a real-time stream, the velocity of the data allows sequential analysis to see if a manager's performance improved or declined over time.

Variety: The datasets contain a variety of different types of data:
Categorical Data : Region, Segment, and Ship Mode
Numerical Data: Sales, Quantity Discount
Temporal Data: Order Dates and Ship Dates to calculate operational lead times

Veracity: As noted previously the data is fabricated for data science training, though the accuracy of the data is high the reliability is limited because of the lack of real-world nuances. However for demonstrating a data pipeline and managerial accountability logic, the internal consistency is excellent.

Value: The data is meaningful because it bridges the gap between accounting and operations. By combining the returns table and orders table, I transformed the simple sales data into a “Net Probability” model, which is a vital metric to measure a business's health.
The accuracy of the analysis primarily relied on transforming three Excel pages into a single table. The first challenge that I ran into was that the profit column in the orders table did not account for returns. To address this, I performed a left join between the Orders table and Returns table using the Order_id as the common key and created a new column: Is_Returned. The Is_Returned column marked returned orders as 1 and non-returned orders as 0. I then performed

a second left join between the resulting table and the People table on the Region column, successfully assigning a specific sales manager to every transaction record. There was minimal additional cleaning needed, so I proceeded directly to the exploratory data analysis process. My initial step was to use .info() and .describe() functions to verify that no data was lost during the joins and to identify outliers in the discount and profit columns. This step ensured that the following findings regarding ‘Loss Leaders” were based on accurate, non duplicated records.

Analysis + Findings

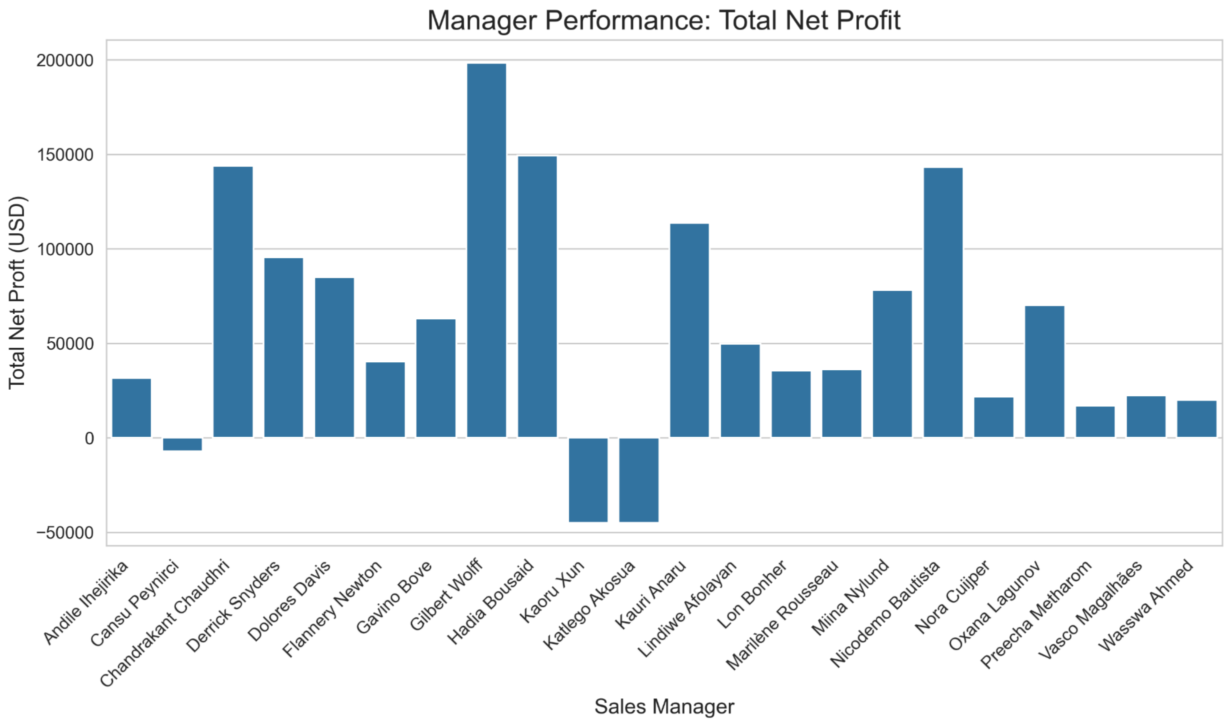
The following visualizations and analysis summary will address the core research questions regarding regional accountability, managerial performance, and operational inefficiencies. The analysis combines aggregated metrics with charts to highlight trends, identify profit leaks, and evaluate efficiency across managers, product categories ,and shipping methods.

Managerial Accountability (Question 1 & 4)

To evaluate managerial accountability, I combined each manager's total net profit with their regional return rates to determine if poor performance was linked to product returns.

Top Performers : Gilbert Wolff generated the highest total net profit at $198,283, while Oxana Lagunov proved to be the most efficient with a 25% profit margin.

Return Correlation: Managers with negative margins, for example Cansu Peynirci, showed a direct correlation with high return rates in their regions. This indicates that losses are not solely based on sales, but may also reflect quality control or post sale issues.

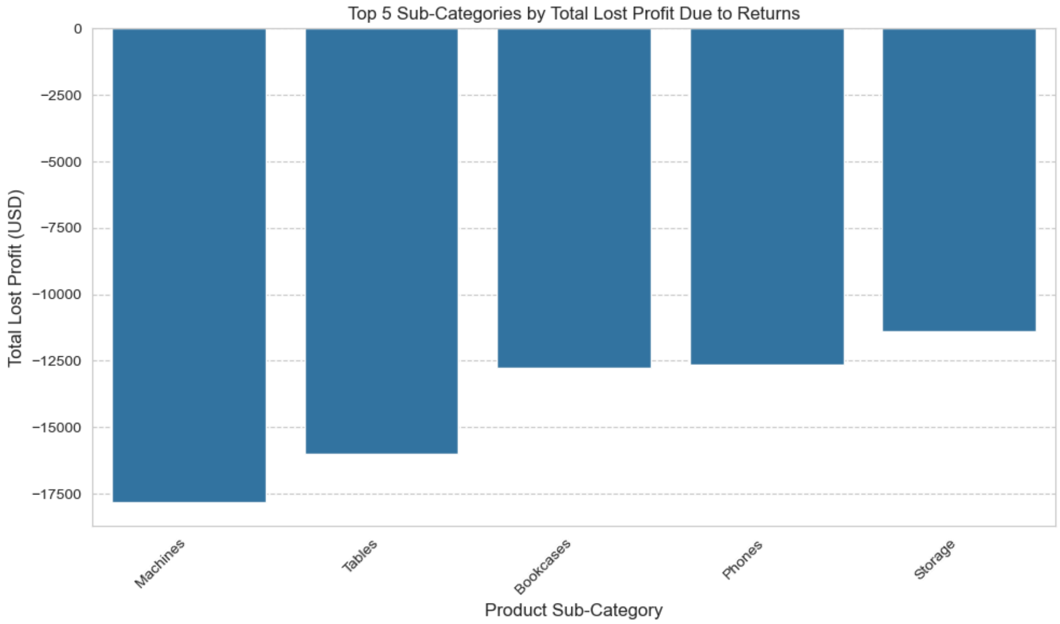


Identifying Profit Erosion (Question 2)

By isolating returned items, I identified the product categories that contribute most to profit loss.

Findings: Machines and Tables account for the highest total losses, at $17,829 and $16,018 , respectively

Insight: These items likely suffer heavy freight and handling costs during returns, which the company absorbs, leading to significant profit loss

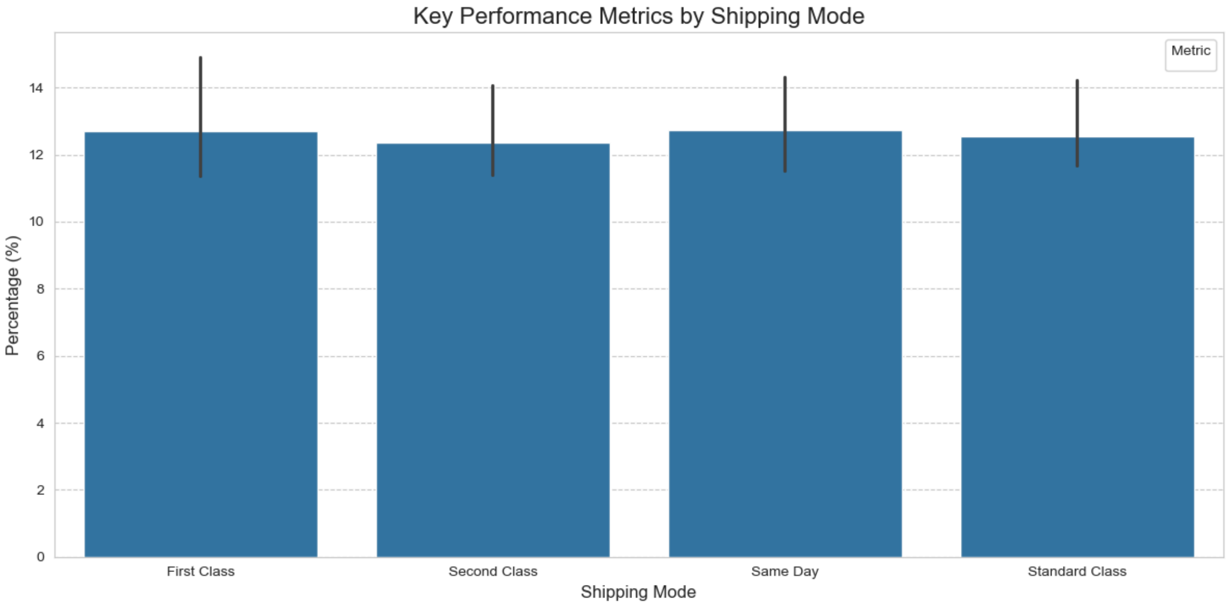


Operational Efficiency Tradeoffs (Question 3)

I further examined whether shipping methods influenced probability and loss patterns.

Findings: First class shipping has the lowest profit margin at 11.35%

Insight: This low margin is not primarily due to returns but to high discount rates at 14.9%. It appears the sales team is sacrificing margin to close First class deals.


Key Performance Metrics by Shipping Mode
Metric
Percentage (%)
First Class
Second Class
Same Day
Standard Class
Shipping Mode