

Superstore Sales Project Documentation

Project Overview :

Start Data : Oct 6th 2025

Title : Superstore sales project

Dataset: "Superstore Sales" (by [Ishan Shrivastava](#)).

Dataset link : <https://www.kaggle.com/datasets/ishanshrivastava28/superstore-sales>

Problem Statement : Identify the key factors (product, region, customer segment, discount, etc.) that most strongly influence sales and profit performance across the Superstore dataset.

Description : The goal is to explore operational and product factors that drive successful sales periods and product lines. Insights will help optimize marketing strategies, product mix, and pricing decisions.

Deliverables: Cleaned dataset, summary report, and visual dashboard showing sales trends, top-performing products, and profit drivers.

Tools : Google Sheets, SQL / SQL-Server, Power Bi

Initial Research Questions.

What factors drive sales for a retail store?

What factors are associated with lower sales or profitability?

What are the top winning products ?

What are the periods of peaks and falls of sales ?

Which factors contribute to higher profits versus just higher sales volume?

Emerging Questions from Exploration.

Which Segment represents the biggest Sales Share and Which represents the biggest Profit Share?

What is the recommended Discount range?

Is there a relationship between the number of sold items and profitability?

Which customer segments and regions deliver the best profitability?

How do shipping times affect profitability?

How can we improve our data collection system to get better understanding and performance?

Data Description

The “Superstore Sales” dataset contains detailed transactional data from a U.S. retail store, including information on customers, products, regions, and profits.

It enables exploration of trends, patterns, and correlations in sales and customer behavior. Key variables include order and shipping dates, customer segments, product categories, sales, quantity, discounts, and profits.

Analyzing these dimensions helps identify high-demand products, evaluate the effect of discounts, and understand regional and temporal sales performance — providing actionable insights for improving profitability and business strategy.

Analysis Plan

1. Success Metrics (KPIs)

The analysis will focus on identifying factors that drive sales and profit. Key metrics include:

- Total Sales (USD)
- Total Profit (USD)
- Profit Margin (%)
- Average Discount (%)
- Sales Growth Over Time

- Top-performing Categories and Subcategories
- Regional Sales Share

2. Analytical Plan (Workflow Steps)

1. Inspect the dataset and understand column meanings (Google Sheets)
2. Clean the data — handle missing values, verify dates, remove duplicates
3. Import into SQL Server for efficient querying
4. Explore data through SQL queries (sales by region, category, discount level, etc.)
5. Export results to Google Sheets for visualization
6. Build charts or dashboards (Google Sheets or Tableau)
7. Document findings and recommendations (Google Docs report)

3. Success Criteria

- Clean and structured dataset ready for analysis
- Clear answers to each objective question
- Meaningful visuals showing trends and patterns
- Actionable insights explaining what drives or limits sales and profit

Workflow Log

Exploration Phase 1

Tool: Google Sheets

Goal : Cleaning, Organizing and Getting Familiar with the Dataset

Notes:

- The dataset contains **9,994 rows** and **21 columns**.
- **Key variables:** *Order Date*, *Ship Date*, *Sales*, *Profit*, *Quantity*, *Discount*, *Region*, and *Segment*.
- **Date range:** First order on **2011-01-02**, last order on **2014-12-31**.
- In *Order Date* and *Ship Date* fields, some dates were stored in **mm/dd/yy** format and others in **dd-mm-yy**.
- Both *Order Date* and *Ship Date* were standardized to the **mm-dd-yyyy** format.
- Identified **1,781 rows** where *Ship Date* < *Order Date* (invalid delivery timeline). These are likely data entry or extraction errors.
- **Correction Note:** After finishing the analysis, I realized that these 1,781 rows were not actually errors. The raw data used the "Day-First" format (DD-MM-YYYY), but my tool read them as "Month-First" (MM-DD-YYYY). For example, it read February 10th as October 2nd, which messed up the timeline calculation.
- **Impact:** The financial numbers (Sales and Profit) are still 100% correct because the date format does not change the total values.
- Five new columns were added:
 - **Shipping Period:** Calculated as *Ship Date* – *Order Date*. Measures delivery efficiency and highlights potential logistic delays.
 - **Date Validity:** Logical flag indicating *invalid* date sequences where *Ship Date* < *Order Date*.
 - **Price per Unit:** Computed as *Sales* / *Quantity*. Used to analyze pricing consistency and compare product-level sales values.
 - **Profit per Unit:** Computed as *Profit* / *Quantity*. Helps identify the most profitable products and directly supports the main project question, "*What are the winning products?*"
 - **Profit Margin:** Computed as *Profit* / *Sales*. Indicates the profit margin per transaction, offering insight into pricing and discount effectiveness.
 - **Location:** Created by *concatenating Region and State*. Provides a concise geographic identifier that supports regional sales and profit analysis while keeping visualizations clear and interpretable.
 - **Product Category:** Created by *concatenating the main Category and Sub-Category fields*. This combined column enables higher-level product analysis, allowing trends and

performance to be observed across product groups rather than individual items, which improves clarity and efficiency in analysis.

- **Dates in Months:** Extracted the month and year from the *Order Date* field to enable time-series analysis of sales and profit trends across the dataset. Used for identifying seasonal patterns, growth trends, and monthly performance variations.
- **No missing values** were found.
- **The country** column contains only “United States” → provides no analytical value and can be removed.
- **No duplicates** were found → all 9,994 rows are unique.
- **No negative values** were found in **Sales**, **Quantity**, or **Discount**.
- **Negative values** detected in the **Profit** field (indicating losses on some transactions).

Descriptive Statistics:

- **Total Sales:** \$2,297,200.86
- **Total Profit:** \$286,397.02
- **Total Profit Margin:** 12.47%
- **Sold Items:** 37873
- **Sales:** Min = 0.444, Avg = 229.86, Max = 22,638.48
- **Quantity:** Min = 1, Avg = 3.79, Max = 14
- **Discount:** Min = 0%, Avg = 15.62%, Max = 80.00%
- **Profit:** Min = -6,599.98, Avg = 28.66, Max = 8,399.98
- **Profit Margin :** Min = -275.00%, Avg = 12.03%, Max = 50.00%.

Exploration Phase 2

Tool : Google Sheets

Goal : Discover relationships and patterns — especially factors affecting *sales*, *profit*, and *profit margin*.

Notes :

Sales and Profit Trends Over Time (Phase 2 – Google Sheets)

Objective:

To analyze monthly sales and profit performance from 2011 to 2014 in order to identify overall trends, seasonal patterns, and potential periods of inefficiency or high performance.

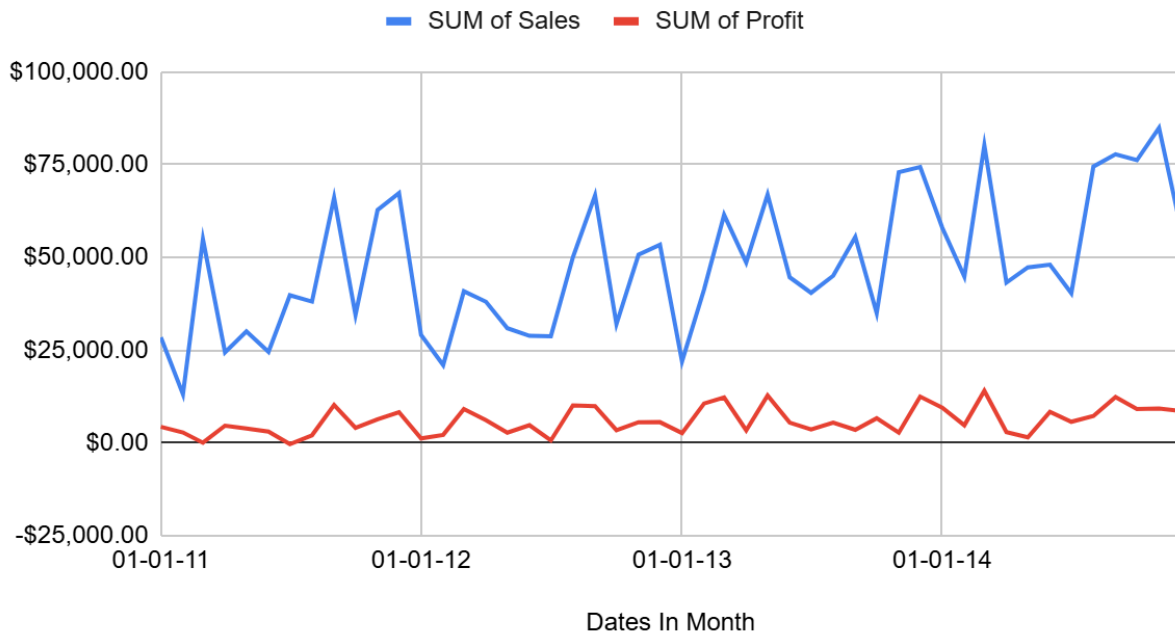
Methodology:

A new column, “**Dates in Months**,” was created by extracting the month and year from the **Order Date** field. A **pivot table** was then used to aggregate total **Sales** and **Profit** by month, and a **line chart** was generated to visualize their progression over time.

Findings:

- Both **sales and profits show a clear upward trend** from 2011 through 2014, reflecting growth in business volume and financial performance.
- **Seasonal peaks** are consistently observed in **October through December**, suggesting strong end-of-year consumer demand.
- Early 2011 exhibits **profit instability**, with some months showing high sales but negative or minimal profit — indicating possible **inefficiencies or over-discounting** during that period.
- From **2013 onward**, profit margins stabilize, and both sales and profits increase together, pointing to **improved pricing or cost control** practices.
- The **best-performing months** in terms of total sales and profit were **November and December 2014**, while **March and July 2011** showed the weakest profit performance despite solid sales volumes.

SUM of Sales and SUM of Profit



Customer Segment Analysis (Phase 2 – Google Sheets)

- Pivot Table was created to summarize the (Sales, Quantities, Discount, Profit, Profit Margin) for each segment

Objective:

To compare the three customer segments (*Consumer*, *Corporate*, *Home Office*) across multiple performance metrics (Sales, Profit, Sold Quantities, and their relative contribution percentage) to identify which segment drives revenue and profitability.

Chart 1 – Sales & Profit per Segment (Column Chart)

Purpose: Visualizes total Sales and Profit side-by-side for each segment.

Insight: Consumer generates the highest sales volume but only moderate profit margin; Home Office, though smallest in sales, shows the highest profit ratio.

Comment: Indicates margin optimization opportunity within the Consumer segment.

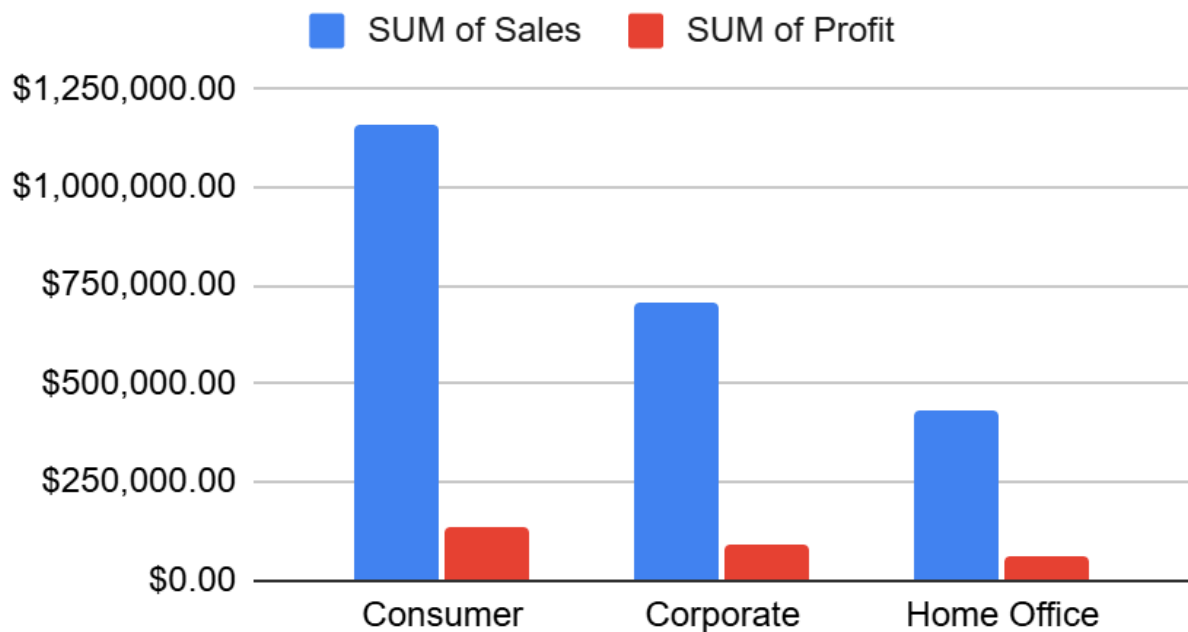


Chart 2 – Sold Quantities per Segment (Column Chart)

Purpose: Displays total number of sold units by segment.

Insight: Quantity trend mirrors sales trend—Consumer dominates in units sold, confirming volume-driven revenue pattern.

Sold Quantities Per Segment

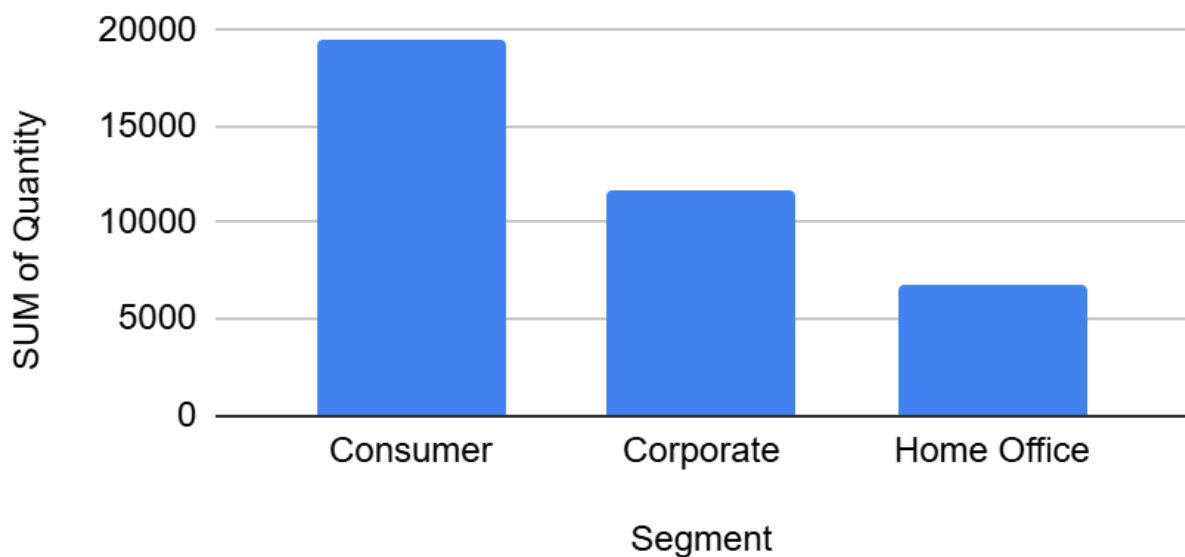


Chart 3 – Sales Share by Segment (Pie Chart)

Purpose: Shows relative contribution of each segment to total Sales.

Insight: Consumer $\approx 50\%$, Corporate $\approx 31\%$, Home Office $\approx 19\%$ (approximate).

Comment: Consumer is the dominant market share.

Sales Per Segment

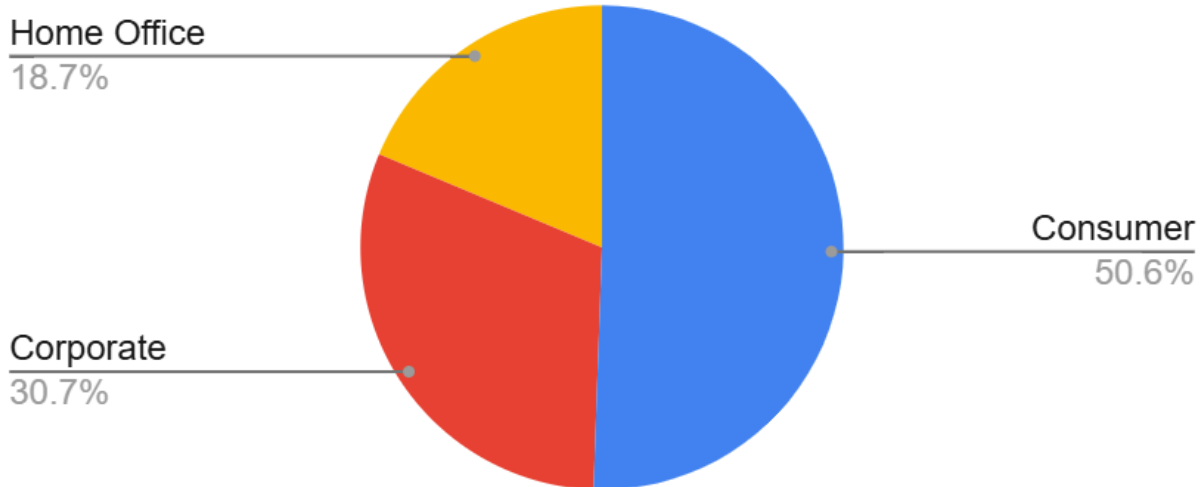


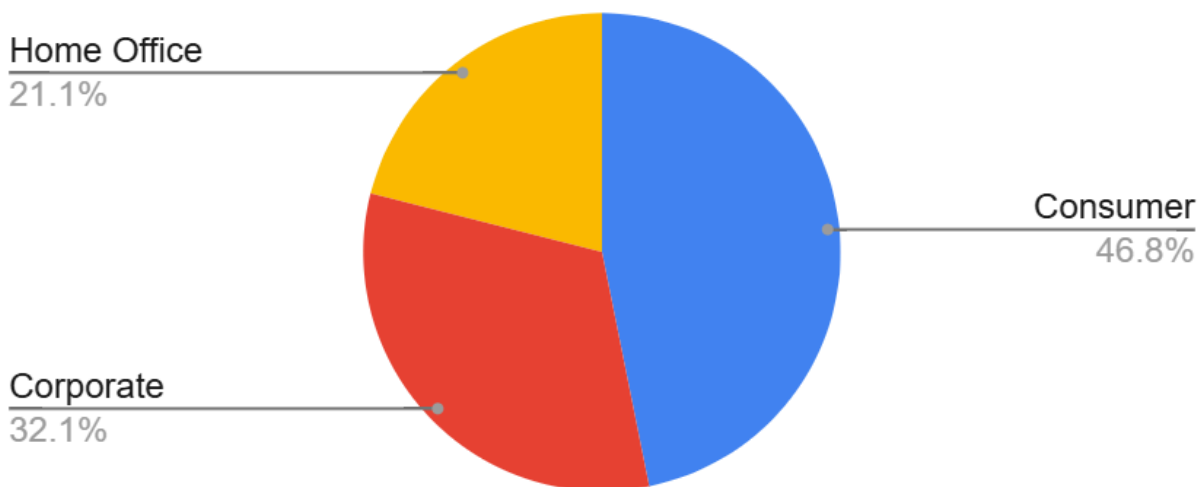
Chart 4 – Profit Share by Segment (Pie Chart)

Purpose: Shows relative contribution of each segment to total Profit.

Insight: Home Office contributes a higher profit share than its sales share, confirming efficiency.

Comment: Reinforces finding from Chart 1 that Home Office yields better profitability per sale.

Profit Per Segment



Summary interpretation:

- Consumer segment drives most revenue but has lower profitability.
- Corporate performs steadily in both.
- Home Office, while smallest in size, achieves the best margin efficiency.

Geographical Analysis (Phase 2 - Google Sheets)

- Pivot Table was created to summarize the (Sales, Quantities, Discount, Profit, Profit Margin) for each Region

Objective:

To compare Regions across multiple performance metrics (Sales, Profit, Sold Quantities, and their relative contribution percentages) to identify which Regions have better performance and which do not, and try to understand the causes that made these differences in performance

Chart 1 – Sales & Profit per Region (Bar Chart)

Purpose: Visualizes total Sales and Profit side-by-side for each Region.

Note:

Top Performing Regions in terms of Sales are (West California, East New York, **Central Texas**, West Washington, **East Pennsylvania**)

Top Performing Regions in terms of Profit are (West California, East New York, West Washington, Central Michigan, South Virginia)

Lowest Performing Regions in terms of Sales are (Central North Dakota, East West Virginia, East Maine, Central South Dakota, West Wyoming)

Lowest Performing Regions in terms of Profit are (**Central Texas**, East Ohio, **East Pennsylvania**, Central Illinois, South North Carolina)

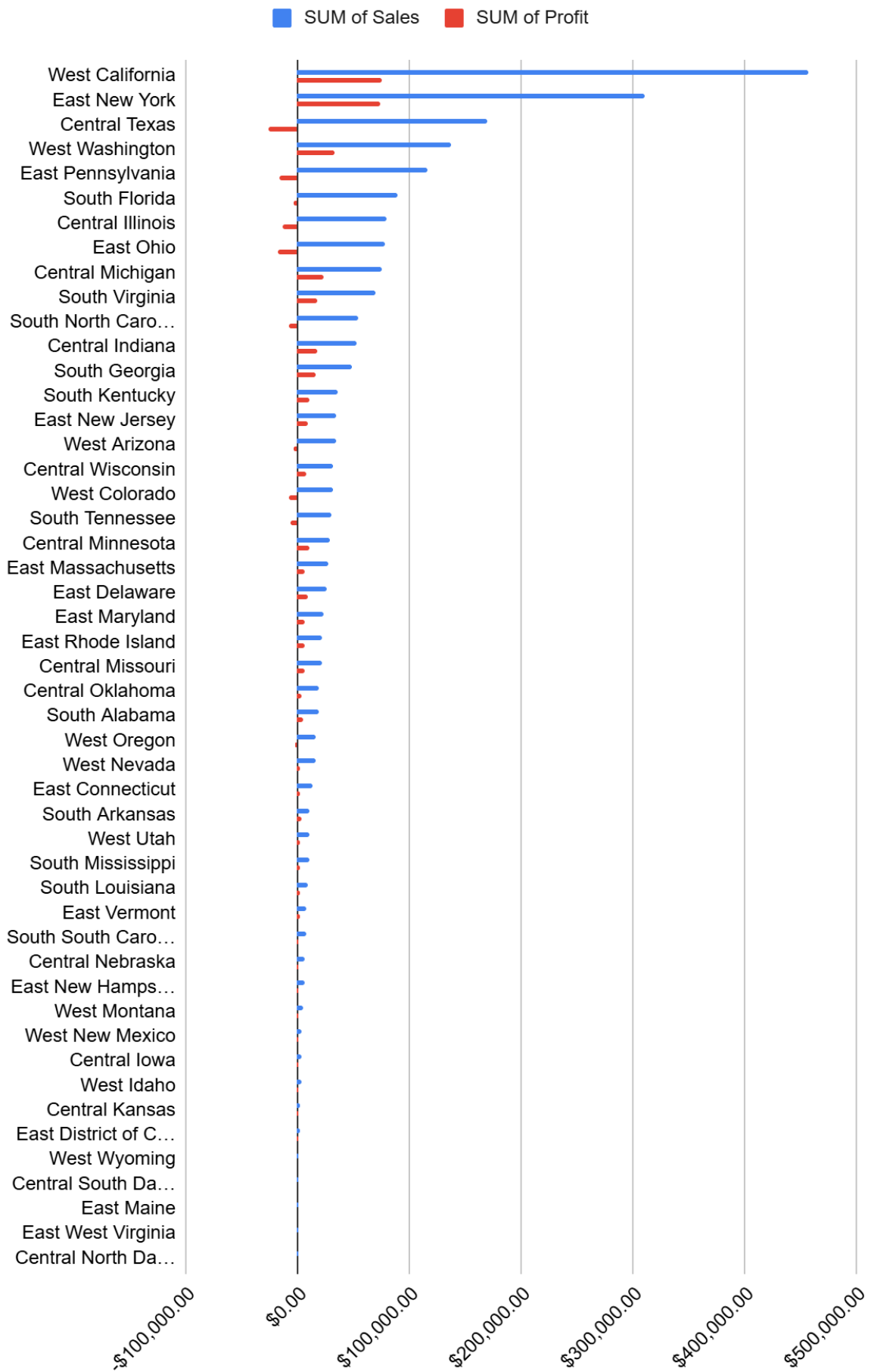
Insight:

West California and East New York are top performers both in sales and profit, indicating strong pricing and product mix strategies.

Central Texas and East Pennsylvania rank **high** in sales but show **negative profits**, suggesting high discounting, high costs, or unprofitable product categories in these regions.

West Washington, Central Michigan, South Virginia achieve a good balance between sales and profitability.

Comment: Top and Bottom performing regions will be analyzed more to understand factors that influence the performance .



Product Performance Analysis (Phase 2 - Google Sheets)

A Pivot Table and three Bar Charts were created to analyze **Sales**, **Profit**, and **Sold Quantities** across product subcategories.

Top Sales: Phones, Chairs, Storage, Tables, Binders

Top Profit: Copiers, Phones, Accessories, Paper, Binders

Top Sold Quantities: Binders, Paper, Furnishings, Phones, Storage

Lowest Sales/Profit/Quantities:

Sales – Supplies, Art, Envelopes, Labels, Fasteners

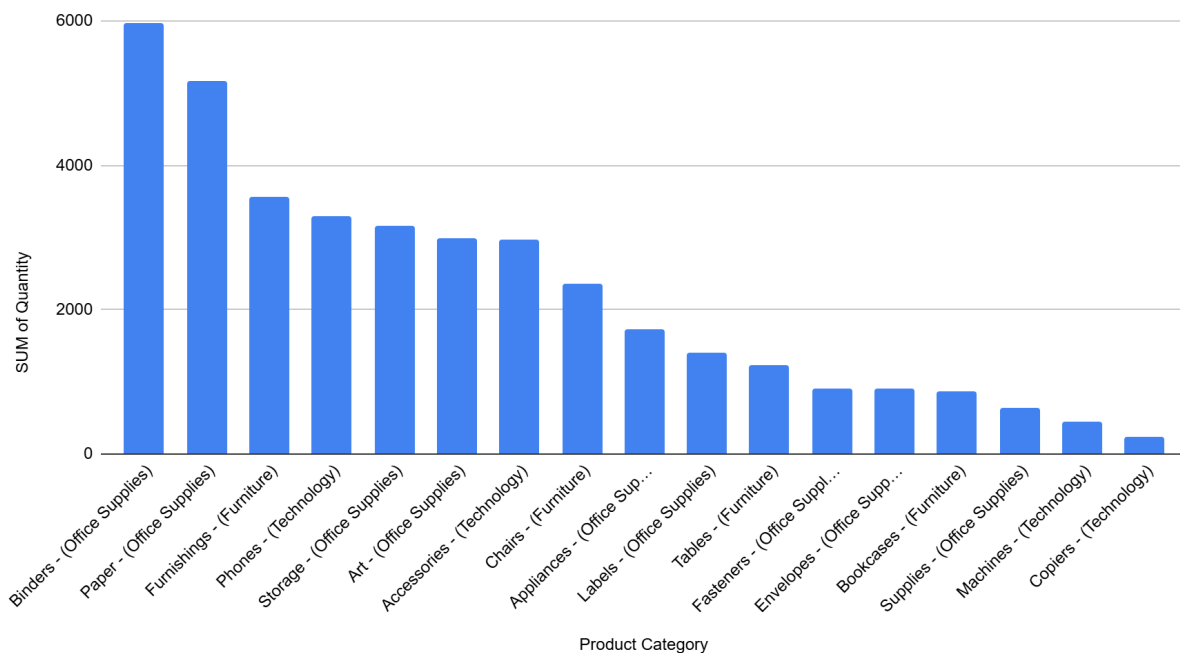
Profit – Machines, Fasteners, Supplies, Bookcases, Tables

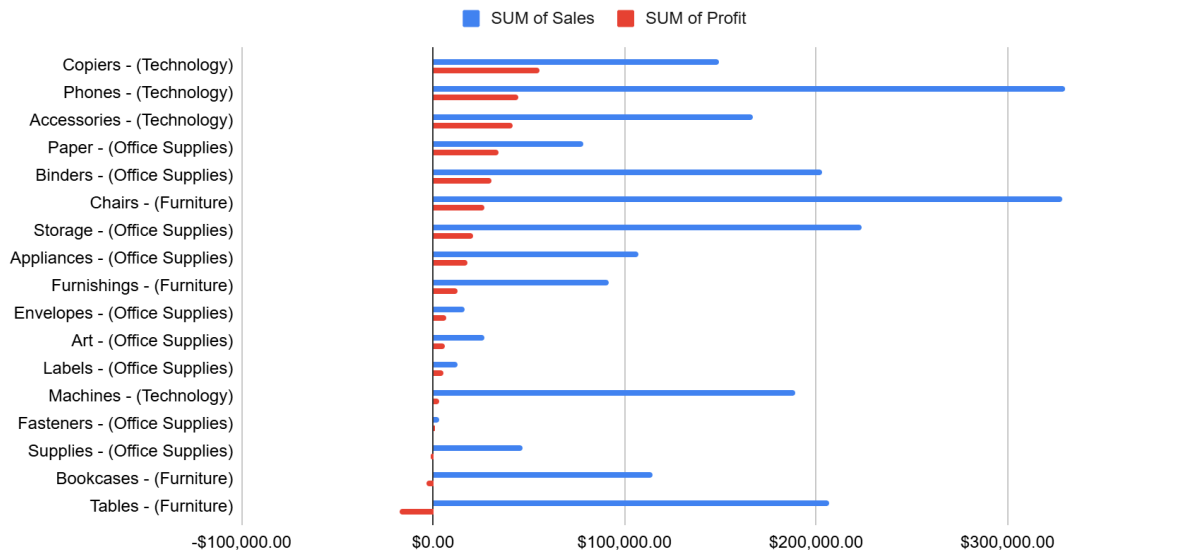
Quantities – Envelopes, Bookcases, Supplies, Machines, Copiers

Finding: Some subcategories such as *Tables* and *Machines* achieve strong sales volumes but generate poor or even negative profits, while others like *Copiers* and *Accessories* achieve high profitability despite lower sales.

Comment: This analysis highlights that **sales volume doesn't necessarily translate to profitability**. Future focus should prioritize **high-margin subcategories** and **evaluate pricing or discount strategies** in low-performing but high-selling categories.

SUM of Quantity vs. Product Category





Discount Impact Analysis (Phase 2 - Google Sheets)

A Pivot Table was created to examine how different discount levels affect Sales and Profit.

Visualization Insight

Scatter plots between Discount vs. Sales and Discount vs. Profit showed no clear linear correlation, suggesting that the relationship is not direct or consistent at the transaction level.

However, when data was aggregated by discount levels and visualized in a column chart (showing total Sales and Profit per Discount), a clear pattern emerged

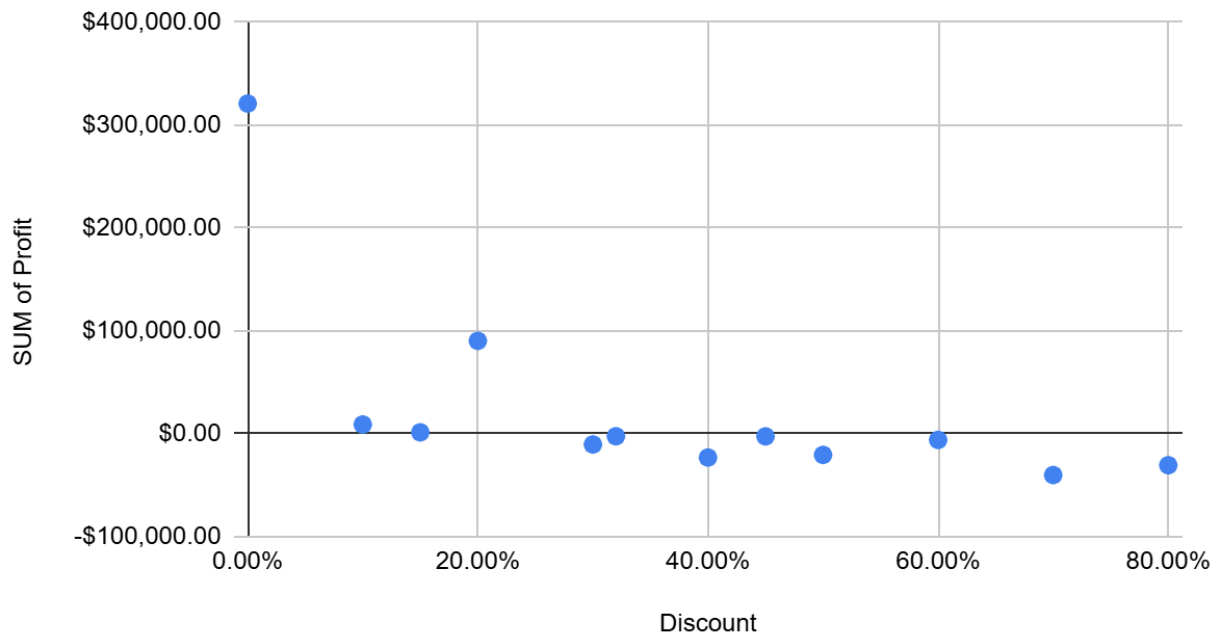
- Moderate discounts (10–20%) are associated with positive profits, with the 20% discount achieving the highest total profit (\$90,337.31) and strong sales volume.
- Discounts above 30% consistently result in negative profits, indicating losses despite some sales activity.
- Transactions with no discount (0%) generated the largest sales (\$1,087,908.47) and the highest overall profit (\$320,987.60).

Comment:

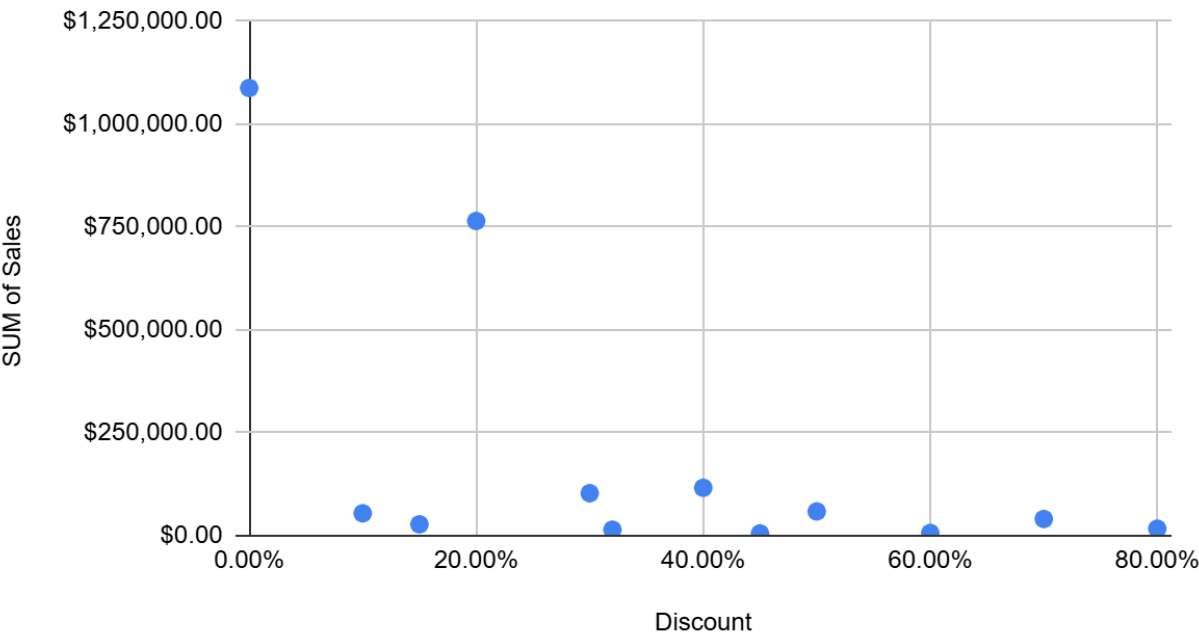
Higher discounts significantly erode profitability, while low to moderate discounts (especially around 10–20%) seem optimal for maintaining both sales momentum and profit margins. This

suggests a need to limit excessive discounting and focus on balanced pricing strategies to sustain profitability

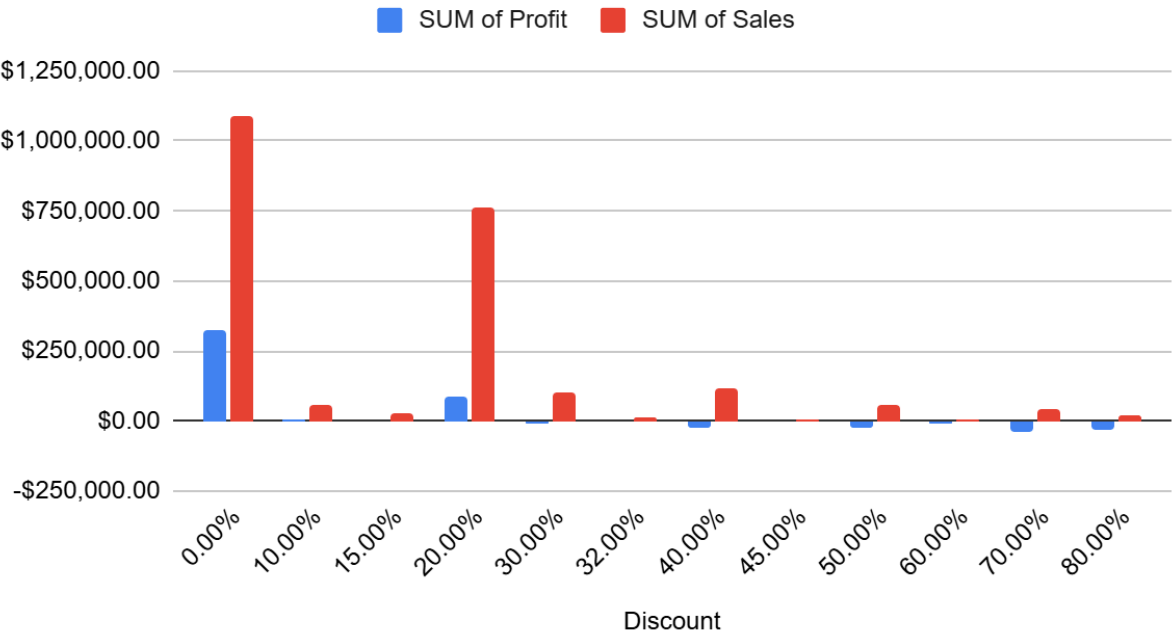
SUM of Profit vs. Discount



SUM of Sales vs. Discount



SUM of Profit and SUM of Sales



This phase involved exploratory analysis using filters in Google Sheets to identify early patterns and relationships between profit and key variables (e.g., discount rate, product category, shipping period, and region).

The goal was to gather preliminary observations before conducting structured analysis using SQL.

Top Products (Profit > \$1,000)

General Patterns

- **Total Transactions:** 42
- **Most Common Ship Mode:** Standard Class
- **Most Common Segment:** Consumer (~45%)
- **Most Common City:** New York City (~30%)
- **Most Common Category:** Technology (~64%)
- **Frequent Subcategories:** Binders (Office Supplies), Copiers and Machines (Technology)

Sales Overview

- **Max:** \$17,499.95 | **Min:** \$2,321.90
- **Average:** \$5,447.04 | **Median:** \$4,380.03 | **Mode:** \$4,899.93

Discount Overview

- **Range:** 0.00% – 40.00%
- **Average:** 5.71% | **Median:** 0.00% | **Mode:** 0.00%

Profit Overview

- **Range:** \$1,007.98 – \$8,399.98
- **Average:** \$2,211.58 | **Median:** \$1,469.83 | **Mode:** \$1,906.49

Profit Margin Overview

- **Range:** 13.33% – 50.00%
- **Average:** 40.76% | **Median:** 46.50% | **Mode:** 50.00%

Insight: Most high-profit orders belong to Technology products sold to Consumer segments via Standard shipping, typically with no or minimal discounts ($\leq 5\%$). This suggests that maintaining low discount rates in Technology may sustain profitability.

Average Products (Profit > \$0 & < \$1,000)

General Patterns

- **Total Transactions:** 8016
- **Most Common Ship Mode:** Standard Class (~58.89%)
- **Most Common Segment:** Consumer (~51.57%)
- **Most Common Category:** Office Supplies (~63.59%)
- **Frequent Subcategories:** Paper (~17.09%)
- **Least Frequent Subcategories:** Copiers (~0.69%), Machines (~0.72%)

Sales Overview

- **Max:** \$8,187.65 | **Min:** \$0.99
- **Average:** \$196.11 | **Median:** \$49.97 | **Mode:** \$12.96

Discount Overview

- **Range:** 0.00% – 40.00%
- **Average:** 8.12% | **Median:** 0.00% | **Mode:** 0.00%

Profit Overview

- **Range:** \$0.06 – \$944.99

- **Average:** \$43.62 | **Median:** \$13.19 | **Mode:** 6.2208

Profit Margin Overview

- **Range:** 1.00% – 50.00%
- **Average:** 29.37% | **Median:** 31.25% | **Mode:** 26.00%

Insight:

Average-profit orders are dominated by Office Supplies, particularly Paper, and are mostly sold to Consumer segments through Standard shipping. Discounts remain modest (average 8.1%, often zero), and profit margins hover around 30%. Notably, Copiers and Machines — previously leading subcategories in high-profit orders — appear among the least frequent here, suggesting that while these products can generate high profits, they are sold less frequently or in different conditions that affect profitability consistency.

Zero Profit Products (Profit = \$0)

General Patterns

- **Total Transactions:** 65
- **Most Common Ship Mode:** Standard Class (~70.77%)
- **Most Common Segment:** Consumer (~53.85%)
- **Most Common Category:** Furniture (~50.77%)
- **Frequent Subcategories:** Storage (~36.92%)
- **Least Frequent Category:** Technology (~4.62%)

Sales Overview

- **Max:** \$2,803.92 | **Min:** \$3.96
- **Average:** \$425.95 | **Median:** \$272.94 | **Mode:** 501.81

Discount Overview

- **Range:** 0.00% – 30.00%
- **Average:** 12.77% | **Median:** 20.00% | **Mode:** 0.00%

Profit Overview

- **Range:** \$0.00 – \$0.00
- **Average:** \$0.00 | **Median:** \$0.00 | **Mode:** \$0.00

Profit Margin Overview

- **Range:** 0.00% – 0.00%
- **Average:** 0.00% | **Median:** 0.00% | **Mode:** 0.00%

Insight:

Zero-profit transactions are mainly concentrated in the **Furniture category**, particularly **Storage** and **Chairs**. These orders are predominantly shipped via **Standard Class** to **Consumer customers**. Despite moderate sales values (average \$426), none generate profit, likely due to **high discount levels** (average 12.8%, median 20%) and **unaccounted shipping costs**, which may significantly reduce margins. Technology items are rarely found in this group, suggesting Furniture products are especially vulnerable to pricing and cost-related losses.

Losing Products

General Patterns

- **Total Transactions:** 1871
- **Most Common Ship Mode:** Standard Class (~62.69%)
- **Most Common Segment:** Consumer (~53.61%)
- **Most Common Category:** Office Supplies (~47.35%), Furniture (~38.16%)
- **Frequent Subcategories:** Binders (~32.76%)
- **Least Frequent Category:** Technology (~14.48%)

Sales Overview

- **Max:** \$22,638.48 | **Min:** \$0.44
- **Average:** \$250.51 | **Median:** \$71.09 | **Mode:** 3.168

Discount Overview

- **Range:** 10.00% – 80.00%
- **Average:** 48.09% | **Median:** 40.00% | **Mode:** 20.00%

Profit Overview

- **Range:** -\$6,599.98 – -\$0.09
- **Average:** -\$83.45 | **Median:** -\$18.09 | **Mode:** -6.237

Profit Margin Overview

- **Range:** -275.00% – -1.11%
- **Average:** -62.46% | **Median:** -38.33% | **Mode:** -73.33%

Insight:

Losing orders are dominated by **Office Supplies (47%)** and **Furniture (38%)**, mainly shipped via **Standard Class** to **Consumer customers**. **Binders** are the most frequent subcategory, appearing in nearly one-third of loss-making transactions. These orders are strongly associated with **heavy discounting** (average 48%, median 40%) and deeply negative **profit margins** (average -62%). The combination of high discounts, potential unrecorded costs (like shipping or handling), and low pricing power in office products likely drives these losses. Technology items are less affected, suggesting their higher margins offer better resilience against discount-driven losses.

Phase 3 – SQL-Based Analysis and Insights

Objective: To extend the analysis using SQL Server for more advanced queries, aggregations, and data validation.

Tools: Microsoft SQL Server Management Studio (SSMS)

SQL Import Summary:

The dataset was successfully imported into SQL Server after converting all numeric and percentage fields in Google Sheets to numeric format and defining their SQL data types as

DECIMAL. This prevented previous conversion issues and preserved value accuracy across key metrics such as Sales, Profit, Discount, and Profit Margin.

The derived field “*Dates in Months*” was excluded from the SQL version of the dataset because it was imported as text and could not support chronological sorting or aggregation. Monthly and yearly time-based analyses will instead be generated dynamically within SQL using date functions for accuracy and flexibility.

Notes :

- General Statistics Results Similar to those in google sheets

Top Products

In Terms Of Product Per Unit :

1. Canon imageCLASS 2200 Advanced Copier
2. Canon imageCLASS MF7460 Monochrome Digital Laser Multifunction Copier
3. Zebra ZM400 Thermal Label Printer
4. Ativa V4110MDD Micro-Cut Shredder
5. 3D Systems Cube Printer, 2nd Generation, Magenta

In Terms Of Total Profit :

1. Canon imageCLASS 2200 Advanced Copier
2. Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind
3. Hewlett Packard LaserJet 3310 Copier
4. Canon PC1060 Personal Laser Copier
5. HP Designjet T520 Inkjet Large Format Printer - 24" Color

Top Products Insight :

Top by Profit Per Unit: High-margin office and technology equipment dominated, particularly Canon imageCLASS copiers, HP Designjet printers, and Ativa shredders.

Top by Total Profit: The same product family (copiers, printers, and binding machines) led in total profit, reinforcing their strong profitability and brand consistency.

Discount impact :

Discount_Range	Total_Sales	Total_Profit	Profit_Margin	Number_Of_Orders
0-10%	1142277.77	330017.09	0.336623	4892
10-20%	792152.87	91757.14	0.175591	3709

20-30%	103226.76	-10369.34	-0.114933	227
30-50%	195314.91	-48447.87	-0.296258	310
> 50%	64228.76	-76559.23	-1.139135	856

Profitability declines sharply beyond 20% discounts, turning negative after 30%.

Shipping Period Insights:

Optimal profitability was observed at shipping periods of **2–6 days**, with sharp deviations in longer deliveries (≥ 60 days) likely representing delayed or invalid records.

Annual Sales and Profit Trends (2011–2014):

Yearly aggregation revealed consistent growth in both sales and profit over the four-year period.

- 2011: Sales \$484K | Profit \$49K
- 2012: Sales \$471K | Profit \$62K
- 2013: Sales \$608K | Profit \$82K
- 2014: Sales \$734K | Profit \$94K

Insight : This upward trajectory suggests improved sales performance and profit efficiency across time, reflecting stronger market presence and operational refinement.

Discount vs. Profitability by Category:

A category-level analysis revealed a clear inverse relationship between discount levels and profit margins.

- *Technology* maintained the highest profitability (Avg. Margin 15.65%) with moderate discounts (~13.23%), confirming its strong value perception.
- *Office Supplies* achieved stable performance (Margin 13.83%) under moderate discounts (~15.73%).
- *Furniture* showed the weakest profitability (Margin 3.88%) under the heaviest discounting (~17.39%), indicating aggressive markdowns severely erode profit.

These findings highlight the need for category-specific discount policies, especially for Furniture.

Profit Distribution Analysis:

Quartile analysis revealed a highly right-skewed profit distribution:

- Q1: \$1.73 | Median: \$8.67 | Q3: \$29.36
This indicates that 75% of orders earn less than \$30 in profit, while a small fraction of high-profit orders drive the majority of total profitability.
The finding reinforces earlier observations that profitability is concentrated in a few high-value products and categories (notably Technology).

Conclusions & Recommendations

- The **Consumer** segment drives most revenue but has lower profitability.
- **Corporate** performs steadily in both.
- **Home Office**, while smallest in size, achieves the best margin efficiency.
- **West California and East New York** are **top performers** both in sales and profit, indicating strong pricing and product mix strategies.
- **Central Texas and East Pennsylvania** rank high in sales but show negative profits, suggesting high discounting, high costs, or unprofitable product categories in these regions.
- **West Washington, Central Michigan, South Virginia** achieve a good balance between sales and profitability.
- Top Sales: Phones, Chairs, Storage, Tables, Binders
Top Profit: Copiers, Phones, Accessories, Paper, Binders
Top Sold Quantities: Binders, Paper, Furnishings, Phones, Storage
Lowest Sales: Supplies, Art, Envelopes, Labels, Fasteners
Lowest Profit: Machines, Fasteners, Supplies, Bookcases, Tables
Lowest Quantities: Envelopes, Bookcases, Supplies, Machines, Copiers

- Some subcategories such as *Tables* and *Machines* achieve strong sales volumes but generate poor or even negative profits, while others like *Copiers* and *Accessories* achieve high profitability despite lower sales.
- This analysis highlights that sales volume doesn't necessarily translate to profitability. Future focus should prioritize high-margin subcategories and evaluate pricing or discount strategies in low-performing but high-selling categories.
- Higher discounts significantly erode profitability, while low to moderate discounts (especially around 10–20%) seem optimal for maintaining both sales momentum and profit margins. This suggests a need to limit excessive discounting and focus on balanced pricing strategies to
- Most high-profit orders belong to Technology products sold to Consumer segments via Standard shipping, typically with no or minimal discounts ($\leq 5\%$). This suggests that maintaining low discount rates in Technology may sustain profitability.
- Average-profit orders are dominated by Office Supplies, particularly Paper, and are mostly sold to Consumer segments through Standard shipping. Discounts remain modest (average 8.1%, often zero), and profit margins hover around 30%. Notably, Copiers and Machines — previously leading subcategories in high-profit orders — appear among the least frequent here, suggesting that while these products can generate high profits, they are sold less frequently or in different conditions that affect profitability consistency.
- Zero-profit transactions are mainly concentrated in the **Furniture category**, particularly **Storage** and **Chairs**. These orders are predominantly shipped via **Standard Class** to **Consumer customers**. Despite moderate sales values (average \$426), none generate profit, likely due to **high discount levels** (average 12.8%, median 20%) and **unaccounted shipping costs**, which may significantly reduce margins. Technology items are rarely found in this group, suggesting Furniture products are especially vulnerable to pricing and cost-related losses.
- Losing orders are dominated by **Office Supplies (47%)** and **Furniture (38%)**, mainly shipped via **Standard Class** to **Consumer customers**. **Binders** are the most frequent subcategory, appearing in nearly one-third of loss-making transactions. These orders are strongly associated with **heavy discounting** (average 48%, median 40%) and deeply negative **profit margins** (average -62%). The combination of high discounts, potential

unrecorded costs (like shipping or handling), and low pricing power in office products likely drives these losses. Technology items are less affected, suggesting their higher margins offer better resilience against discount-driven losses.

- Over the four-year period, total sales and profits display a clear upward trajectory, with the highest levels reached in late 2014. Seasonal peaks consistently occur in **October–December**, reflecting strong end-of-year demand. Early 2011 shows profit instability, where high sales months (e.g., March and July 2011) generated minimal or negative profits, possibly due to heavy discounting or cost inefficiencies. From 2013 onward, profit performance improves and becomes more consistent, suggesting more effective pricing and discount strategies.
- **High-value products** like copiers, printers, and binding machines consistently drive both per-unit and total profits, confirming them as strategic focus items.
- **Discounts above 20 %** should be tightly controlled — they rapidly erode profit and flip margins negative.
- **Efficient shipping (2–6 days)** correlates with higher profitability, indicating that operational speed plays a measurable role in financial outcomes.
- The SQL validation confirms that prior findings from Google Sheets were robust, reinforcing confidence in the dataset and the analytical pipeline.
- Technology consistently delivers superior profit margins with moderate discounts, while Furniture suffers from excessive discounting. Category-specific pricing and discount strategies could substantially improve total profitability.
- Profit generation is highly concentrated — most transactions yield minimal profit, emphasizing the importance of focusing on high-margin products and reducing low-profit, high-discount sales.

Answers to Project Questions & Data Improvement Recommendations:

Q1. What factors drive sales for a retail store?

Answer:

Sales are primarily driven by **Technology** and **Office Supplies** categories, especially high-demand items such as phones, binders, and storage products. The **Consumer segment** generates the highest total sales, and regions such as **West California** and **East New York** consistently lead in sales performance.

Q2. What factors negatively affect sales or profitability?

Answer:

Excessive **discounting (above 20%)**, long **shipping periods**, and **high-cost furniture items** (Tables, Bookcases) are strongly associated with low or negative profit margins. Some regions, notably **Central Texas** and **East Pennsylvania**, also show high sales but negative profits, indicating inefficient pricing or high costs.

Q3. What are the top winning products?

Answer:

High-margin **Technology products** dominate the list:

Canon imageCLASS copiers, HP printers, and Fellowes binding machines deliver both high total profit and high profit per unit. These products combine strong pricing power with manageable discounting.

Q4. What are the periods of peaks and falls of sales?

Answer:

Sales and profits steadily increase from **2011 to 2014**, with seasonal peaks during **October–December** each year (end-of-year demand). Early 2011 showed unstable profits, but from 2013 onward performance became consistently positive.

Q5. Which factors contribute to higher profits versus just higher sales volume?

Answer:

High profits are achieved through:

- **Moderate discounting ($\leq 20\%$)**
 - **Efficient shipping (2–6 days)**
 - **Low operational costs** (likely for Technology items)
 - **High-value, lower-volume products**
In contrast, categories like Furniture achieve high sales but low or negative profit due to heavy discounting and costs.
-

Q6. Which Segment represents the biggest Sales Share and which represents the biggest Profit Share?

Answer:

- **Consumer Segment:** Largest share of total sales (~50%), but moderate profitability.
 - **Corporate Segment:** Steady in both sales and profit.
 - **Home Office Segment:** Smallest in sales but highest profit margin (~14%), showing superior efficiency.
-

Q7. What is the recommended discount range?

Answer:

Optimal discounts are **below 20%**, with the **0–10% range** yielding the best overall profitability (average margin ~33%). Beyond **20%**, profit margins decline sharply and turn negative past **30%**.

Q8. Is there a relationship between the number of sold items and profitability?

Answer:

No direct positive relationship. Many high-selling subcategories (e.g., Chairs, Tables, Storage) generate **low or negative profit**, while less frequent sales (e.g., Copiers, Machines) generate **high profit per unit**. Profitability is therefore **more influenced by pricing and cost structure than volume**.

Q9. Which Regions deliver the best profitability?

Answer:

West California and **East New York** perform best in both sales and profit.

Central Texas and **East Pennsylvania** have high sales but negative profits, suggesting inefficient cost management or discounting.

Q10. How do shipping times affect profitability?

Answer:

Orders shipped within **2–6 days** show the highest average profits. Longer shipping times (≥ 60 days) or invalid delivery records correlate with profit losses, suggesting delayed or costly logistics reduce overall profitability.

Q11. How can we improve our data collection system to get better understanding and performance?

Answer:

The analysis revealed several areas where the dataset could be enhanced to improve accuracy, insight depth, and future analytical performance:

- **Include Shipping Cost Data:**
Adding a shipping cost column would enable more precise profit calculations and help identify whether long shipping periods or logistics costs contribute to low or negative margins.
- **Ensure Date Field Validation and Consistency:**
During exploration, inconsistent date formats and invalid delivery timelines were identified. Implementing standardized date validation (e.g., enforcing **YYYY-MM-DD** format and verifying that ship dates always follow order dates) will improve time-based analysis accuracy.
- **Enhance Data Type Validation:**
Import issues revealed that percentage and numeric columns were sometimes stored as text. Stronger data entry validation rules and automated checks (e.g., ensuring numeric types for discount, sales, and profit fields) will reduce future conversion errors.

- **Optional Additions:**

Including fields like *product cost*, *marketing expense*, or *shipping method cost* would support deeper profitability modeling and ROI analysis.

These improvements will strengthen future data-driven decision-making by ensuring cleaner, more complete, and analysis-ready datasets.
