Indiana University, Bloomington

INFO-I 590 DATA VISUALIZATION

Final Report

Priyanka Prem Kumar
Department of Data Science, Indiana University Bloomington, 107 S Indiana Ave, Bloomington, Indiana, 47405, USA
Tel: (726) 239-9953
prpremk@iu.edu

Sumedh Ambapkar
Department of Data Science, Indiana University Bloomington, 107 S Indiana Ave, Bloomington, Indiana, 47405
Tel: (313) 230-8314
suambapk@iu.edu

Analyzing Retail Sales Performance Across the USA Regions

ABSTRACT

The Superstore Sales dataset is a structured, fictional dataset commonly used in data visualization and analytics training. It contains detailed transactional data from a global office supply company, filtered to represent sales activity within the USA region. The dataset includes information across key business dimensions such as orders, shipping, customers, products, sales revenue, discounts, and profitability. It serves as a representative model for analyzing retail operations across various geographic and customer segments.

The primary objective of this analysis is to evaluate business performance by identifying trends in sales, analyzing regional and segment-based profitability. The analysis combines statistical methods, Python's matplotlib,seatborn and plotly, Power BI, observable HQ's Dr.js visualizations, and basic predictive modeling to explore sales trends across orders, customers, products, and regions.

The results are expected to show significant variation in profitability across countries and customer segments. High discounts reduce profit margins in some product categories, while specific segments and subcategories drive revenue. Shipping modes and delivery timelines significantly impact customer satisfaction and efficiency. The analysis offers data-driven insights to guide pricing, promotion, regional focus, and supply chain improvements. It identifies strategic levers that can enhance financial and logistical performance across the USA retail operations.

INTRODUCTION

Motivation

In today's highly competitive, digitally driven retail landscape, businesses face increasing pressure to meet evolving customer expectations while maintaining operational efficiency. With the rise of omnichannel strategies, shortened delivery windows, and dynamic pricing pressures, retailers must embrace data-driven decision-making to remain agile and profitable. The availability of granular transactional data presents a unique opportunity to derive actionable insights that inform everything from inventory planning to customer segmentation and logistics (Waller & Fawcett, 2013).

This project is centered around the analysis of a retail sales dataset from a U.S.-based office supply company. The dataset contains over 9,994 individual transactions across 49 states, covering three major product categories (Furniture, Office Supplies, and Technology) and spanning more than 180 unique cities. It includes detailed records of customer orders, including order and shipping dates, product sub-categories, sales figures, discounts, profit margins, shipping modes, and customer segments (Consumer, Corporate, and Home Office). These dimensions allow for a rich, multidimensional exploration of the factors influencing sales performance.

Retail analytics, the process of applying data science to optimize retail operations, has emerged as a core capability for organizations seeking a competitive edge (Chen et al., 2012). For instance, analyzing shipping timelines reveals bottlenecks that directly affect customer satisfaction and repeat purchase behavior (Hao et al., 2020).

The dataset also enables exploration of customer segmentation, revealing how different groups (Consumer, Corporate, Home Office) behave in terms of purchasing habits, discount sensitivity, and profitability (Wedel & Kamakura, 2000). Similarly, regional trends can inform localized strategies, as regional preferences and challenges often vary significantly across geographies (Dholakia, 2012). While many existing dashboards offer only high-level summaries, this project takes a deeper dive, layering traditional KPIs with advanced visualizations such as contour plots, chord diagrams, and sunburst charts. These tools allow for fine-grained insight into patterns over time, cross-segment behaviors, and inefficiencies in fulfillment pipelines.

Ultimately, this project demonstrates how structured, visual data exploration can transform complex retail data into strategic insights, highlighting not only what is happening in the business, but why. In an industry where small inefficiencies can lead to significant losses, this kind of analysis provides a clear path toward smarter, faster, and more targeted decision-making.

Background

This project explores several key analytical themes that are commonly addressed in retail analytics: First, it examines product-level performance, breaking down sales,

discounts, and profits by product category and sub-category. This allows us to pinpoint which products contribute most to revenue and which may be underperforming due to heavy discounting or low margins. Understanding these dynamics is essential for optimizing pricing strategies and refining the product portfolio.

The dataset's temporal attributes, including order and shipping dates, enable the analysis of logistics and operational efficiency. By evaluating variations in delivery times across states and shipping modes, the project identifies potential supply chain bottlenecks or inefficiencies that could impact customer satisfaction and retention.

The analysis also leverages customer segmentation, focusing on three key groups: Consumer, Corporate, and Home Office. These segments are expected to show distinct purchasing behaviors and profitability profiles. For example, corporate clients may generate higher order volumes, while individual consumers may respond more to pricing and convenience.

In addition to applying standard techniques like aggregation and trend analysis, the project incorporates more nuanced visualizations such as contour plots, Pareto charts, and multi-level heatmaps. These allow for clearer insight into interactions between time, location, segment, and product performance, going beyond basic charts to support more strategic decision-making.

Finally, the inclusion of geographic data, with sales recorded at the state and regional level, allows for a spatial analysis of market performance. This reveals regional trends, preferences, and gaps, providing insights that can inform regional marketing, inventory distribution, and resource allocation strategies.

Existing work

The existing visualization is based on a subset of the Superstore Sales dataset, emphasizing transaction-level sales data from various regions of the United States. Each row represents a single product line item from a customer order, including attributes such as Ship Mode, Customer ID, Segment, Region, Product Category, Sales, Discount, and Profit. The dataset showcases diverse purchasing behaviors across customer segments (e.g., Consumer, Corporate, Home Office) and product types (Furniture, Office Supplies, Technology).

Below are some of the visualizations that can be considered as existing and showcase the sales of items.

Inline bar chart

The visualization shown in Figure 1 (Kizley Benedict,2019) is a combination of a KPI card and a mini bar chart, often referred to as a spark bar chart or inline bar chart. This type of chart is commonly used in executive dashboards to present key performance indicators (KPIs) along with a concise visual representation of trends over time.

In this case, the main focus is on "Total Sales", which is prominently displayed as a large numerical value ($109,456) (Figure 1a). Just below it, there's a percentage indicator (+12.6%) accompanied by an upward arrow, signifying a positive change in sales compared to a previous period (such as week-over-week or month-over-month). The color green is typically used to indicate favorable performance.

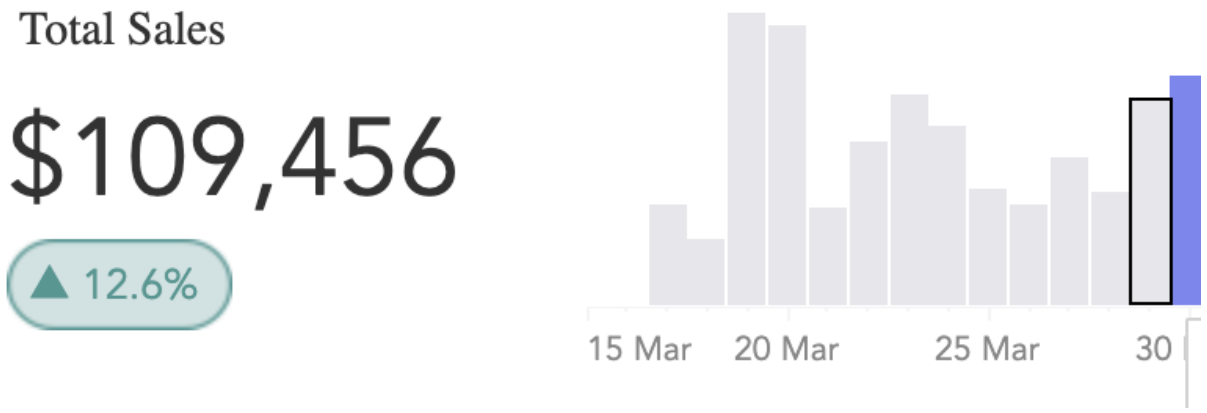

Figure 1: a) KPI card for Total Sales, b) Inline Barchart showing sales trends

To the right (Figure 1b), the mini bar chart shows daily sales trends from March 15 to March 30. Each vertical bar represents total sales for a specific day, with the bar for March 30 highlighted in blue, drawing attention to either the most recent data point or a significant sales spike. The rest of the bars are in lighter shades, providing historical context without overwhelming the viewer.

This chart type is effective for quick performance monitoring, allowing stakeholders to observe both the overall value and the recent trend at a single glance without needing to analyze a full-sized graph.

Heat Map

Consider Figure 2 as a sample graph, but the same can be drawn for the dataset that we are using for this project.

The current heat map visualization, which displays sales by time of day and day of the week, can be effectively modified to show sales by date instead of by hour. In our project, the x-axis would represent calendar dates (such as March 1st to March 31st), while the y-axis could represent categories like days of the week, customer segments, product types, or regions. Each cell in the heat map would then display the total sales for that particular date-category combination, with color intensity indicating the volume of sales, darker shades reflecting higher sales, and lighter shades indicating lower sales.

Sales	Gross Income	COGS	Orders	Rating

	10 am	11 am	12 pm	1 pm	2 pm	3 pm	4 pm	5 pm	6 pm	7 pm	8 pm
Sunday	891.2	106.6	1,811.6	1,916.2	1,656.0	364.5	325.4	1,029.2	2,183.3	2,802.9	383.4
Monday	931.7	1,374.1	1,643.2	735.0		1,660.9	832.6	710.5	1,037.5	418.5	1,077.5
Tuesday	575.3	666.1	1,772.9	1,736.6	1,043.5	1,799.3	1,505.8	2,004.1	1,598.8	3,633.2	1,275.9
Wednesday	1,027.0	1,800.0	990.5	1,042.0	1,087.5	1,692.8	1,079.1	132.6	354.0	3,127.6	1,184.2
Thursday	372.7	905.7	1,085.7	850.1	1,906.8	748.2	2,768.3	1,347.5	580.4	1,096.4	1,098.0
Friday	2,860.9	288.1	215.7	1,371.5	2,995.3	962.3	1,270.0	291.4	512.6	4,121.0	1,015.7
Saturday	1,826.4	1,607.8	2,639.6	4,003.7	2,792.8	1,022.4	1,387.1	3,301.3	2,335.7	4,549.1	305.0

Figure 2: Heat map for showcasing the total sales for that particular date-category

This format is particularly useful for identifying daily sales trends, spotting unusual spikes or drops in revenue, and analyzing the effectiveness of marketing campaigns or seasonal patterns over time. By visualizing data across a calendar timeline, businesses can more easily correlate sales performance with external events,

promotions, or operational changes. Tools like Tableau or Power BI allow easy implementation of this structure by using dates on the x-axis, a relevant grouping on the y-axis, and sales values mapped to color.

Choropleth Map

Again map below (Figure 3) is a sample map from random sales data, but the same can be showcased using the existing dataset. The visualization shown is a choropleth map of the United States, used to display sales performance by region. Each region is color-coded based on the level of total sales: dark blue represents regions with the lowest sales, olive green indicates regions with higher sales, and yellow denotes regions with average sales. This type of map is particularly effective for showing the geographic distribution of performance metrics, allowing users to quickly identify which regions are underperforming or excelling without needing to examine detailed numerical data.

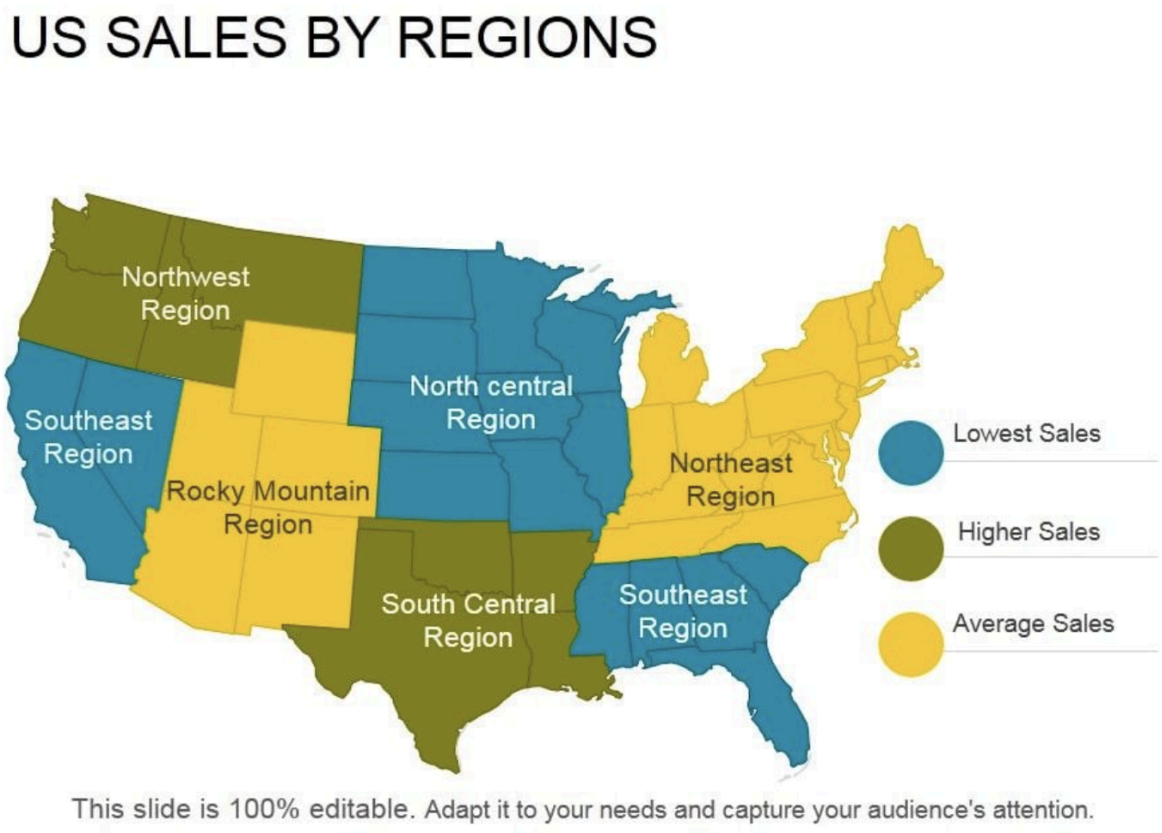

Figure 3: Choropleth graph to depict performance by region

When related to the previously shared sales dataset, this map visually reinforces the performance patterns seen in the raw data. For instance, the Southeast Region, which includes areas like Florida and North Carolina, is shaded in dark blue, aligning with data entries that showed lower profitability or losses, such as a high-value table sale with a significant negative profit margin. In contrast, the South Central Region, shaded in green, may correspond to stronger performance cases like high-profit sales in Texas. The Northeast and Rocky Mountain Regions, marked in yellow, likely reflect moderate activity with a mix of high and low transactions, contributing to an overall average sales level.

This choropleth map serves as a useful tool for executive summaries, regional sales planning, and market prioritization, providing a clear and intuitive way to interpret complex sales data across large geographic areas.

Line Chart

The Line Chart below can be used as a reference to analyze Sales Annually and Quarterly, as shown in Figure 4.

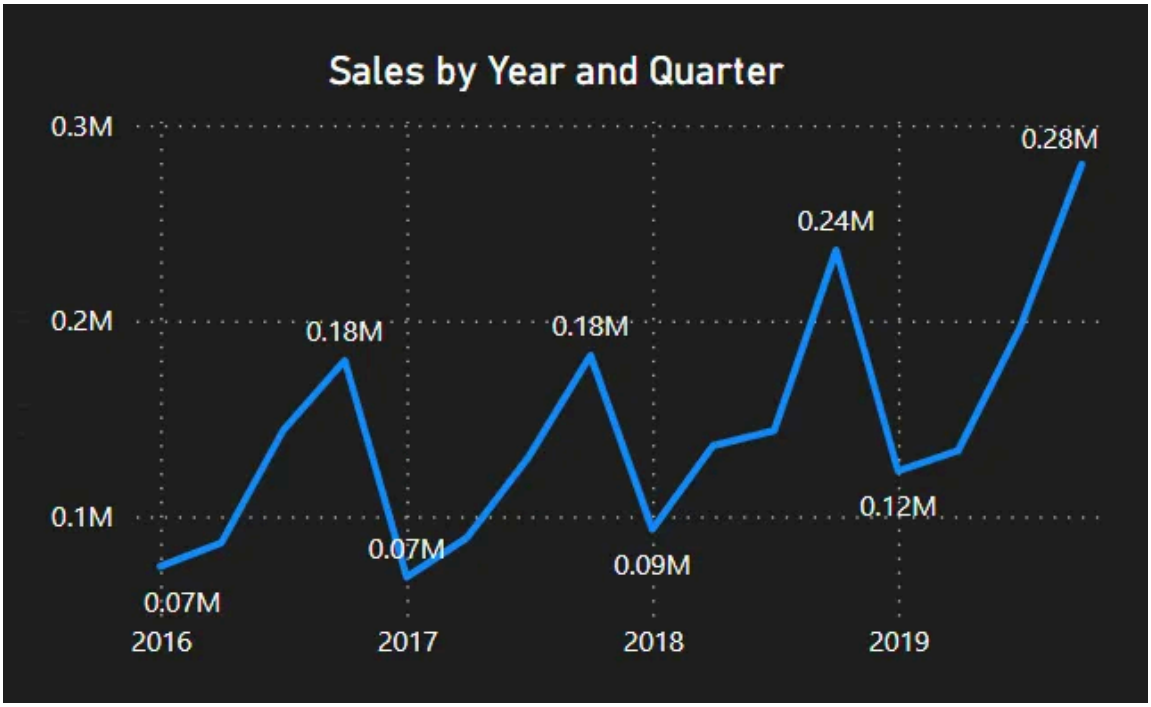

Figure 4: Line Graph to design Sales Annually and Quarterly

Analyzing the sales data from 2017 to 2019 through a yearly line graph (Knaflic, 2015) provides a comprehensive visual narrative of the Superstore's performance over these years. The graph reveals trends, patterns, and fluctuations in sales across the three years, allowing for a clear understanding of the company's growth trajectory or potential challenges. This visualization aids in offering valuable insights for strategic planning, resource allocation, and decision-making within the organization.

Bar Graphs

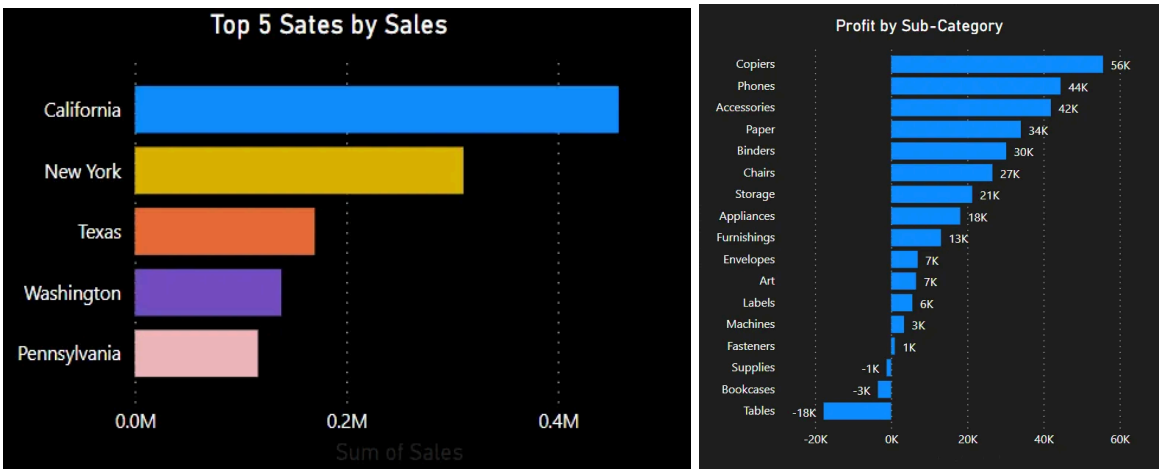

Figure 5: Bar Graphs to indicate the top 5 states and profits made by sub-categories

Displaying the revenue per state for all 49 states would not look nice, but limiting the horizontal bar chart to the top 5 states by revenue generated would leverage human perceptual abilities to compare lengths easily and accurately (Cleveland, 1984). The top 5 states and profits made by sub-categories that generated the most revenue between 2017 and 2019 is shown here.

Donut Charts

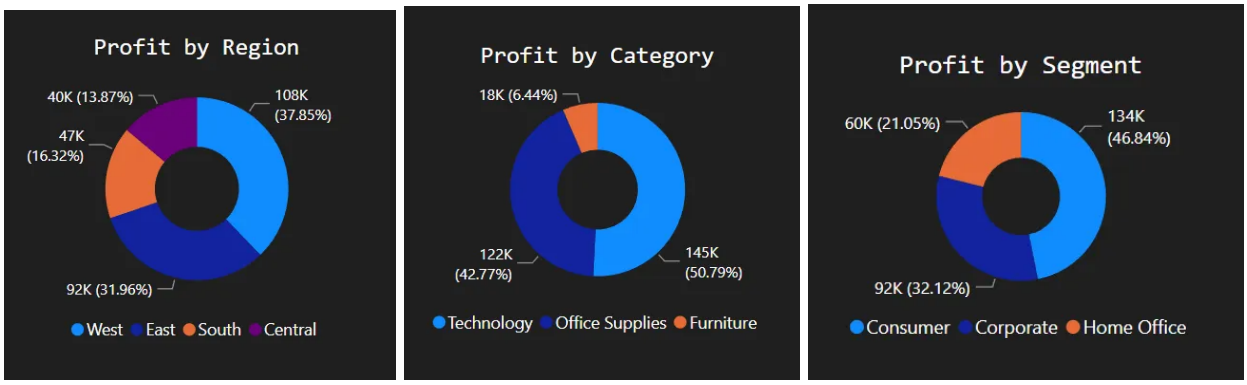

Figure 6: Donut charts to procure Sales by different categories

To identify products that significantly boost sales, presenting the data as a donut chart (Figure 6) is the most effective way to understand sales distribution across different categories (Tufte, 2001). This approach will aid in reducing stock-up costs, provide valuable insights for the company's marketing team for focused advertisements, and much more (Shmueli, 2017). From 2017 to 2019, the West region generated $108K, accounting for 37.85% of the store's total revenue during that period. Technology contributed $145K, while Consumer sales reached $134K, indicating where corporations should focus next based on these trends.

These existing visuals offer a strong foundation for exploring sales, regional trends, and customer behavior, and help guide the direction of our exploratory visualizations

Objectives

This project aims to apply descriptive and visual analytics to explore a retail sales dataset from a U.S.-based office supply company. The goal is to uncover actionable insights that can support strategic business decisions in pricing, marketing, logistics, and customer engagement.

Specifically, the objectives are to:

1. Investigate product performance by identifying high-performing and underperforming items based on sales and profit metrics.
2. Analyze the impact of discounting on profitability across product categories and sub-categories.

3. Evaluate operational efficiency by examining order and shipping timelines across different regions and shipping modes.
4. Segment customers (Consumer, Corporate, Home Office) to uncover behavioral patterns, purchasing trends, and differences in profitability.
5. Assess regional performance by analyzing geographic variations in sales and profit at the state and regional levels.

In undertaking this analysis, we expected to find that:

1. A small number of products and customers contribute disproportionately to overall revenue.
2. Certain categories or segments may consistently underperform due to excessive discounting or low margins.
3. Operational inefficiencies such as delayed shipping may vary by region or shipping mode.
4. Regional and segment-based trends would provide insights into where marketing or resource allocation can be optimized.

The project integrates multiple analytical perspectives, descriptive, diagnostic, and exploratory, to transform raw transactional data into meaningful business intelligence. Through visualizations and aggregation techniques, it aims to reveal the drivers behind sales performance and customer behavior. Ultimately, the insights derived can guide strategic decisions to improve profitability, efficiency, and customer satisfaction.

DATA AND METHODS

Analysis of Data:

This project employs descriptive and visual analytics techniques to extract insights from a transactional retail dataset of a U.S.-based office supply company. The methodology integrates data preprocessing, visualization, and interpretation, leveraging a combination of modern data visualization libraries and platforms. It is the publicly available Superstore Sales dataset, widely used for retail analytics projects and data visualization practices. This dataset contains detailed transaction-level data for an office supply company operating across the United

States. It is provided in a tabular format a CSV file, with each row representing an individual product line item from a customer order.

The dataset includes a rich variety of columns spanning both categorical and numerical data types. Key categorical variables include: Order ID, Customer ID, Segment, Region, State, City, Ship Mode, Product Category, Sub-Category, and Order Priority. The numerical and date/time columns include: Sales, Quantity, Discount, Profit, Order Date, and Ship Date. These diverse data types enable a multidimensional analysis of customer behavior, product performance, geographic trends, and operational logistics.

Data Preprocessing

Before conducting analysis, several preprocessing steps were performed to prepare the dataset for exploration and visualization:

1. Data Loading and Inspection: The dataset was imported from a CSV file and the first few rows were inspected to understand its structure.
2. Missing Value Check: A check for missing values was conducted to ensure data completeness. No major imputation or cleaning was needed, suggesting the dataset was well-structured.
3. Data Type Verification: Columns were reviewed to ensure correct data types, especially for date fields like Order Date, which were explicitly converted to datetime format for temporal analysis.
4. Feature Extraction:
    a. Order Month was extracted from Order Date to enable time-based grouping and trend analysis.
    b. Day of the Week was derived from the order date to analyze weekday sales patterns.
    c. Month Index and Segment Index were created using factorization for plotting advanced visualizations like contour plots.
5. Indexing: Order Date was set as the DataFrame index to support time series operations such as resampling for monthly trends.
6. Categorical Encoding (for plotting): Customer segments were factorized into numeric values to support grid-based interpolations in contour plots.

7. Aggregations and Groupings: Data was grouped by variables such as Category, Sub-Category, Region, Segment, and Customer ID to prepare for aggregated visual summaries and performance analysis.

These steps ensured that the dataset was clean, properly formatted, and analytically ready for both descriptive and visual analysis techniques.

Visualization Tools and Platforms

To ensure high interactivity and accessibility in our analysis and presentation, a combination of business intelligence platforms and advanced visualization tools will be employed:

1. Power BI: Used for building dynamic executive dashboards that summarize key performance indicators such as sales, profit, and regional performance in an accessible, user-friendly format.
2. Observable HQ: Enables browser-native, interactive visualizations using JavaScript-based frameworks like D3.js—ideal for embedding live data views and storytelling.
3. Python (Plotly, Seaborn, Matplotlib): Leveraged for in-depth exploratory data analysis, statistical visualization, and advanced plots such as contour and heatmaps. These tools provide flexibility for custom analysis and fine-grained control over visuals.

We iteratively prototyped visualizations in Python and Observable to test clarity, label density, color encoding, and user interaction. Feedback was used to refine bubble scaling, label filtering, and region-level drilldowns. This multi-tool approach ensures that the final deliverables are both analytically robust and visually engaging for diverse audience.

Planned Visualizations

Each visualization is purposefully selected to address key business questions related to performance, customer behavior, and operational efficiency. The techniques used include both exploratory and explanatory visuals to provide clarity at multiple levels:

1. Sunburst Chart
   To captures hierarchical relationships such as Category → Sub-category → Product or Region → State → City, offering an intuitive view of how sales or profits are distributed across business levels.
2. Chord Diagram
   To illustrates the relationships and flow between dimensions, for example, between customer segments and product categories, or regions and shipping modes, revealing interaction patterns and logistical dependencies.
3. Contour Plot
   To shows density and gradient patterns over two continuous variables. It is particularly useful to identify clusters or anomalies in sales vs. discount relationships or profit trends over time.
4. Bubble Chart
   To enable multivariate comparison at the product level: X-axis: Discount, Y-axis: Sales, Bubble Size: Profit, Color: Product Category. This visual simplifies spotting trade-offs between discounts and profitability.
5. Bar Charts & Heatmaps
   To compare profit margins and evaluate discount effectiveness across regions, segments, and product types. These visuals make ranking and hotspot identification easy.
6. Histograms and Time-Lag Plots
   To analyze distribution of shipping delays and highlight inconsistencies across shipping modes or geographic locations, helping uncover operational inefficiencies.
7. Choropleth Maps
   To eospatially map sales and profit figures by state or region, providing insight into regional demand, performance disparities, and potential logistical constraints.
8. Treemap
   To provides a space-efficient view of proportional data, especially useful for comparing sales or profit contributions by product, sub-category, or region. It helps quickly identify top-performing and underperforming areas through block size and color.

Unlike standard bar charts and KPIs, this method uncovers multi-variable relationships (e.g., discount vs. sales vs. profit) and hidden structure in categorical flows and time-based anomalies. The integration of D3 and Python tools enables a richer, interactive storytelling layer beyond what's possible in traditional BI tools alone.

Candidate Visualization Methods:

Bubble Chart

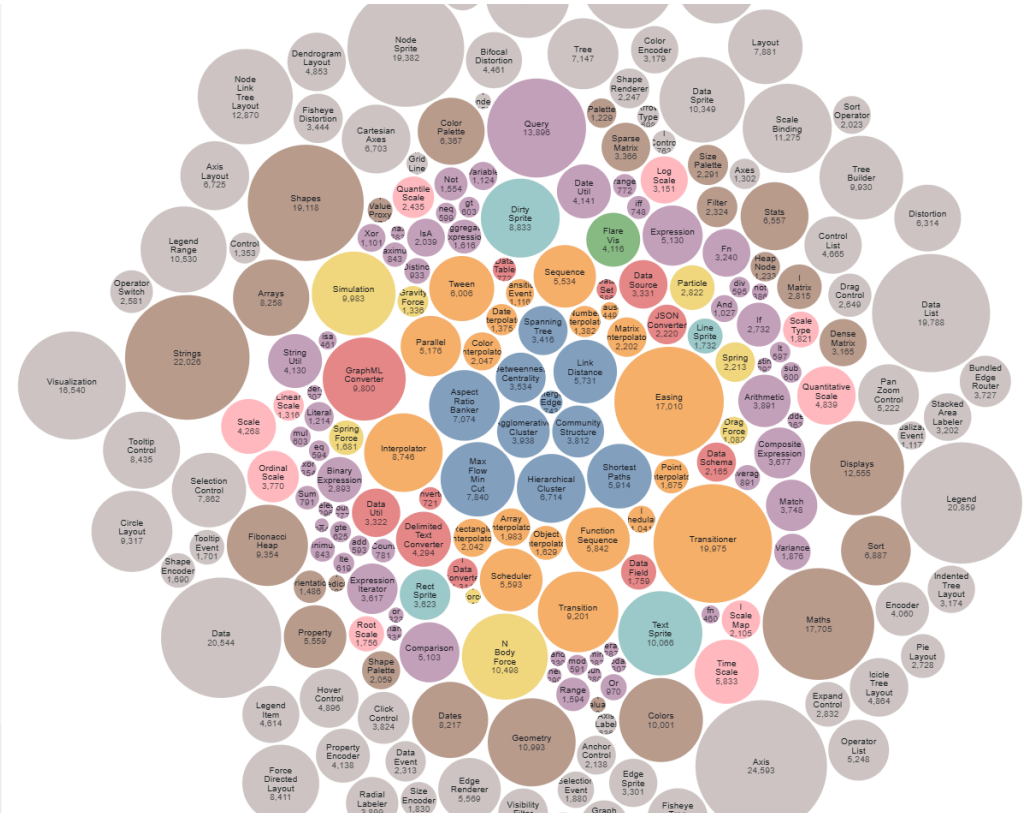

Figure 7: Bubble Chart for Sales vs Profit Visuals

The **Bubble Chart** (Figure 7) is selected as a candidate visualization due to its strength in representing multivariate data in an intuitive and compact form. By plotting Discount on the x-axis, Sales on the y-axis, bubble size for Profit, and color for Product Category, it provides a rich overview of product performance across multiple dimensions. This enables quick identification of items that generate high revenue but low profit, or those where discounts are driving losses. In comparison to basic scatter or bar charts, the bubble chart offers a deeper layer of

insight into trade-offs between sales volume and profitability. When applied to regional analysis, it further visualizes disparities in total sales (bubble size) and shipping mode (bubble color), helping detect delivery inefficiencies or underperforming regions (Ware, 2012).

Contour Graphs

Contour plots (Figure 8) are chosen over basic scatter or heatmaps due to their ability to represent continuous density variations without binning, offering smoother gradient transitions across the data space (Wilke, 2019). In the Discount vs. Profit Density analysis, contour plots highlight dense regions where moderate discounts align with higher profits, aiding in the identification of optimal pricing strategies—something difficult to achieve with traditional visuals. For Shipping Duration vs. Product characteristics, they reveal consistent delivery zones and outliers indicating delays. In geospatial applications, contour overlays on maps expose regional sales hotspots, enabling more targeted logistics and marketing. Compared to basic charts, contour plots excel at uncovering subtle trends and anomalies in continuous, multidimensional data.point representing a transaction can be overlaid with a contour map to highlight areas with high transaction density.

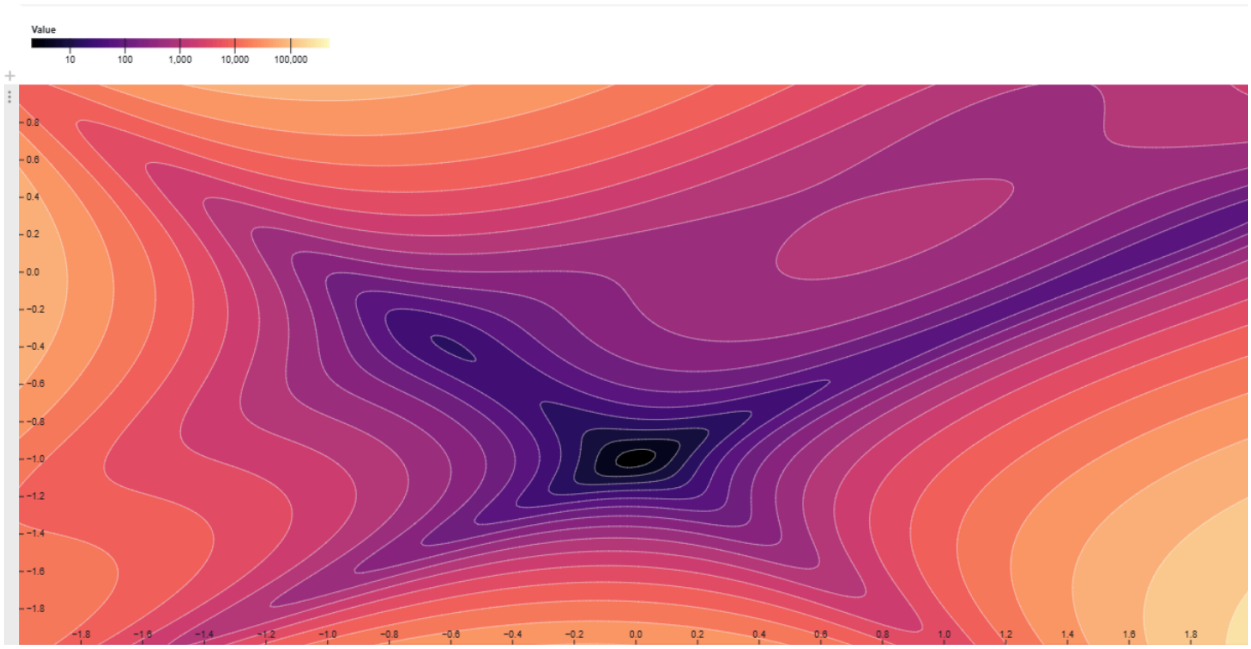

Figure 8: Contour Diagram for Discount vs. Profit Density analysis

Sunburst Chart:

The **Sunburst chart** (Figure 9) is chosen over stacked bar charts due to its ability to naturally represent multi-level hierarchies and scale efficiently with increasing category depth (Kirk, 2016). In this case, it visualizes relationships across Customer Segment → Order Priority → Shipping Mode, starting from the center outward. This layout makes it easier to detect behavioral patterns, such as which segments prioritize urgent orders and their preferred shipping methods. Unlike stacked bars, which become cluttered with more levels, the sunburst offers an intuitive, compact, and interactive way to explore nested categorical data for improved service planning and targeting.

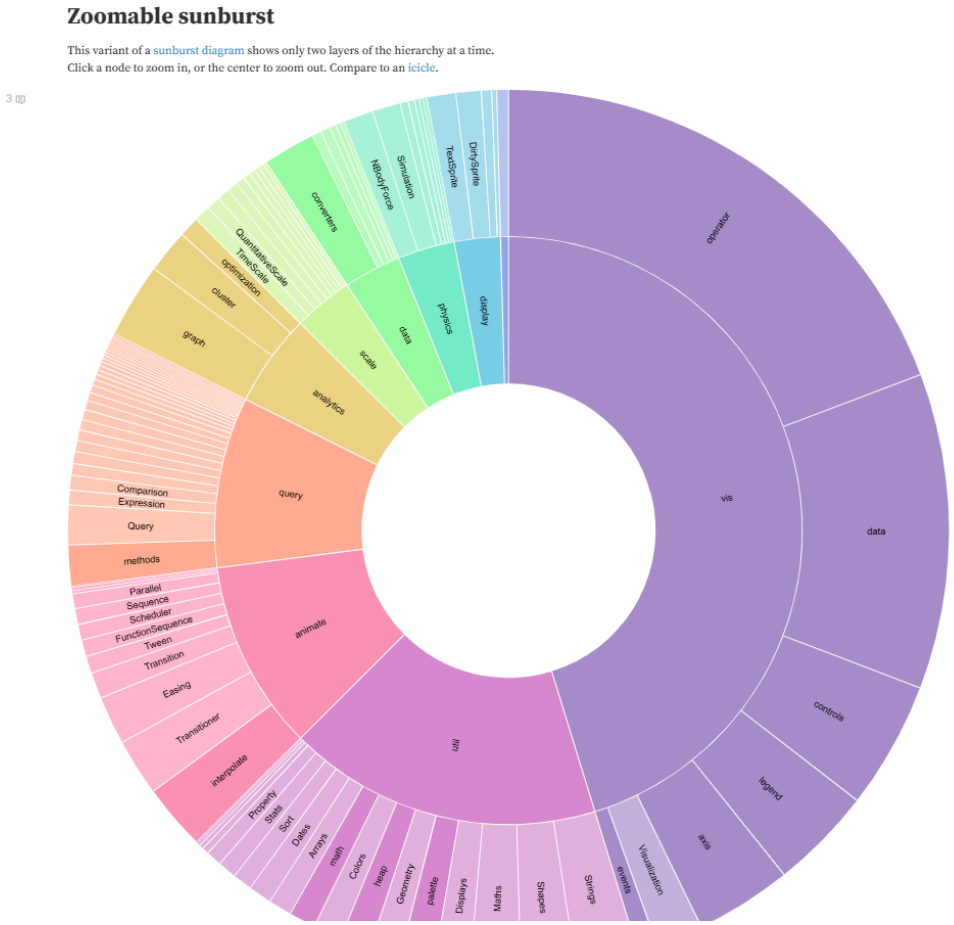

Figure 9: A Sunburst chart can be used to depict hierarchical relationships among customer segments, order priority, and shipping mode

Chord Diagram:

The Chord Diagram (Figure 10) is selected for its ability to visually encode interrelationships between categorical variables with clarity and efficiency, especially when compared to basic cross-tabulations or grouped bar charts (Krzywinski et al., 2009). It enables a compact summary of flows and connections, such as mapping Region to Product Category to highlight regional demand trends, or State to Shipping Mode to uncover delivery preferences and potential logistical bottlenecks. It also reveals Customer Segment to Product Category affinities, helping identify which segments contribute most to various product lines. This format allows for a holistic view of categorical interdependencies that would be cumbersome to analyze through disjointed charts.

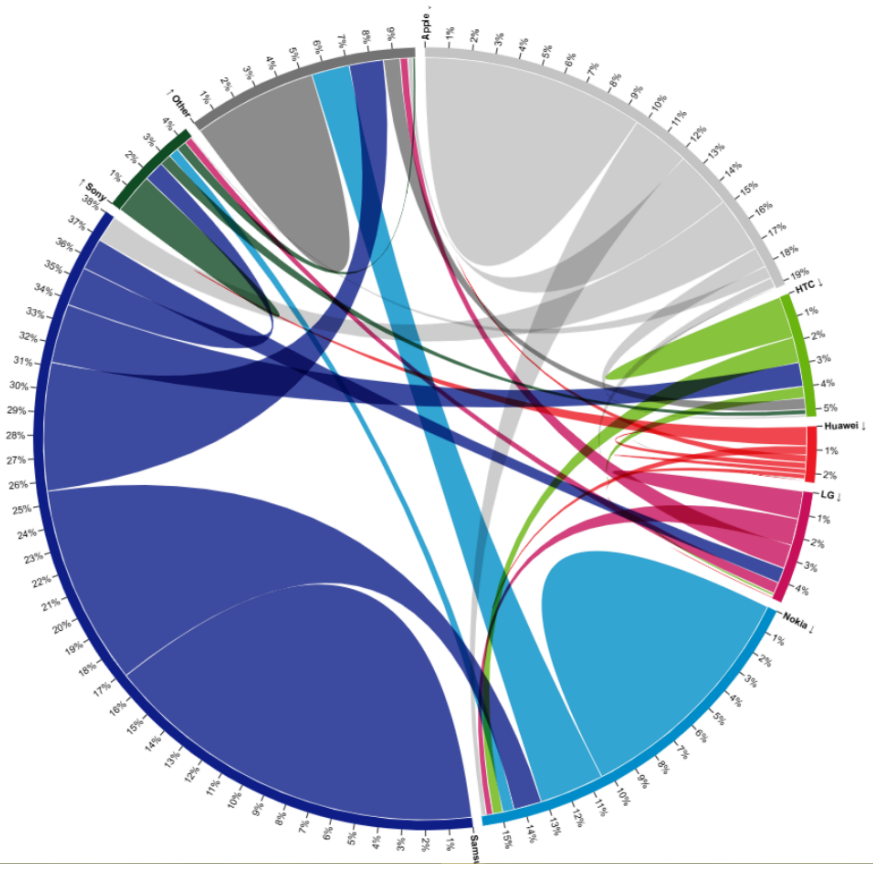

Figure 10: Chord Diagram to visualize inter-relationships or flows between categorical variables

Failed Experiments:

1. To identify high-performing and underperforming products based on sales and profit metrics, we initially chose pie charts as shown in figure 11, assuming they would be effective for displaying product-wise contributions.

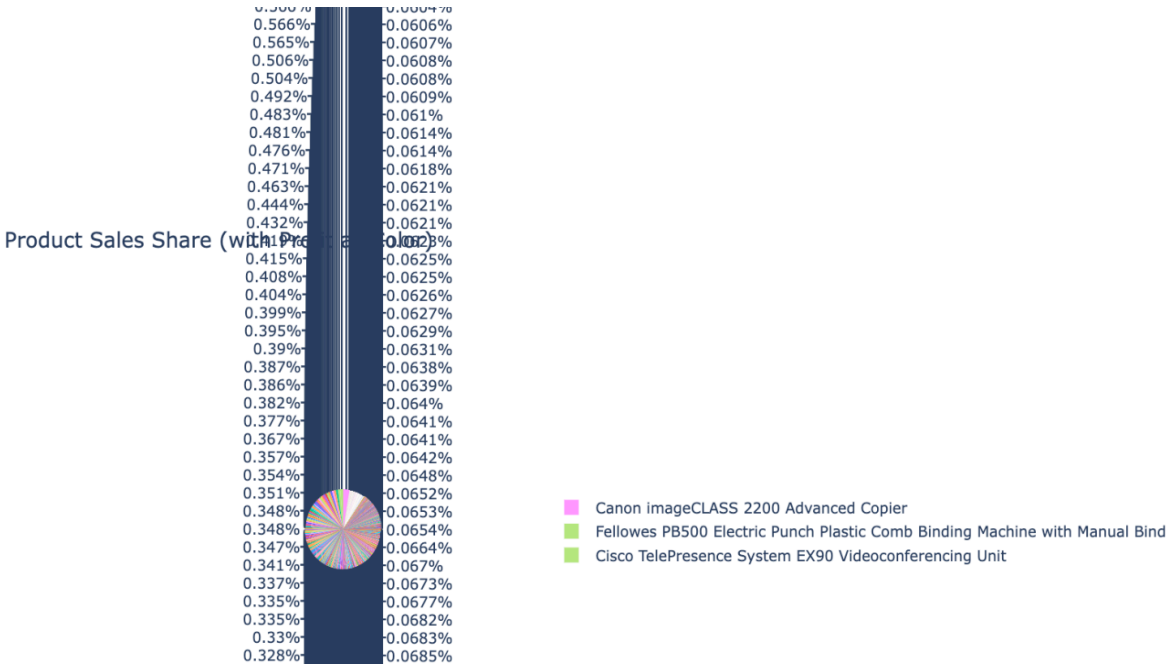

Figure 11: Piechart showing product sale share

However, this approach proved inadequate due to the limitations of pie charts in conveying detailed comparisons across multiple categories, especially when dealing with small variations or a large number of products. Below is the reason:

a. Wrong Chart Type: A pie chart is inappropriate for comparing two metrics like sales and profit simultaneously.
b. Pie charts work well for part-to-whole relationships but fail to show underperformance trends.
c. Misleading Color Encoding: Using profit as a color dimension in a pie chart confuses interpretation, since viewers assume the size represents value but color mapping lacks proper legend context.

d. No Clear High vs. Low Performance: There's no axis or logical ordering to distinguish high or low performance. Products with negative or zero profit (Product D, E) are not easily distinguishable.
e. No Combined Metric: There's no composite performance metric like "Profit Margin" or "Profit per Sale" that might help identify efficiency.

2. Next, to evaluate the impact of discounts on profitability across product categories, we initially used a line chart as shown in figure 12 a. However, this visualization proved ineffective, the reason being:
    a. Incorrect chart type for categorical data: Line charts imply continuous or time-sequential data, but product categories are discrete and unordered, so connecting them with lines is misleading.
    b. Shows only one metric: It displays average discount only, without including any profitability data, so the relationship between discount and profit is completely missing.
    c. No insight into impact: The chart does not show how discounts affect profitability, which is the key goal of the analysis.
    d. False sense of trends: Connecting category points with lines suggests a trend or progression between unrelated categories, which does not exist.
    e. No detection of outliers or extremes: It cannot highlight products or categories with unusual discount-profit behavior, which is critical for identifying underperformers or over-performers.

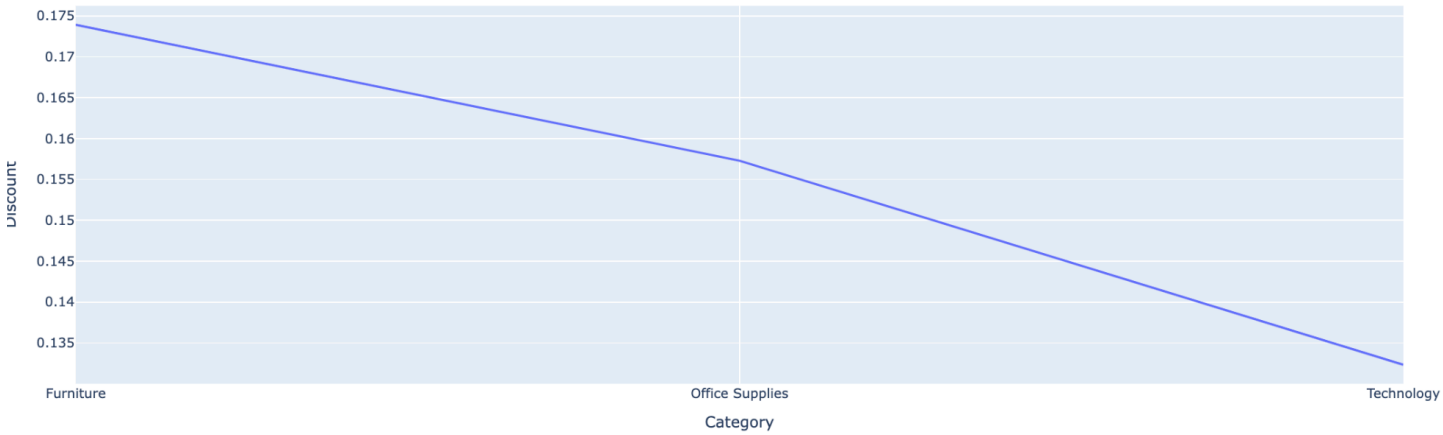

Figure 12 a: Line chart for showing average discount by category

3. Moving on, to analyze order and shipping timelines and assess operational efficiency, we initially considered using a scatter chart as shown infifure 12 b. However, this approach fell short. Scatter charts are effective for visualizing correlations between two continuous variables, but in this case, they:
    a. Ignored Shipping Timeline: It only shows order dates, completely ignoring shipping dates, which are critical to assess operational efficiency.
    a. There was no measure of Delay: Without comparing order and shipping times, you can't evaluate how fast shipments are fulfilled.
    b. No Duration Metric: There's no calculated metric like "shipping delay" or "time between order and ship" to analyze.
    c. Poor Use of Scatter Plot: Scatter plots need two related numeric/date variables to show correlation or delay — using only order dates and product names shows just when orders were placed.
    d. No Trend or Insight: Cannot spot slow or fast shipping products, bottlenecks, or peak load times.

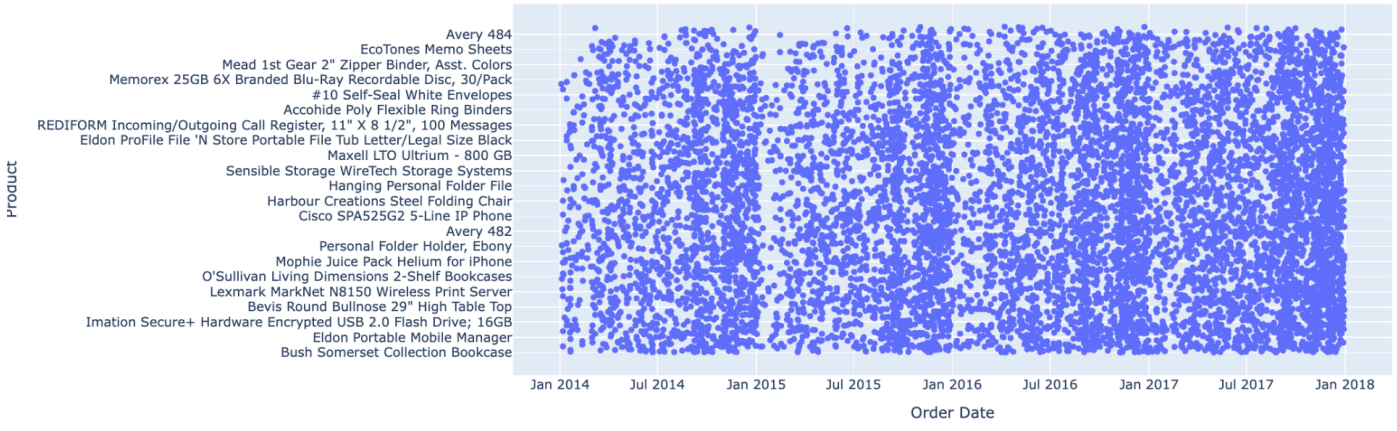

Figure 12 b: Scatterplot showing Order dates by product

4. Finally, for our last objective, to utilize descriptive and visual analytics to analyze a U.S. office supply retail business, we initially chose a sunburst chart as shown in figure 13. However, this proved to be an ineffective choice as.

a. Only visualizes one metric (Sales): The chart focuses only on sales and ignores other important metrics like profit, quantity sold, or discount, missing a multi-dimensional analysis.
b. No temporal context: It doesn't show trends over time (e.g., monthly or yearly performance), which is crucial for understanding seasonal or evolving business behavior.
c. No regional or geographic insight: The objective references a U.S. dataset, yet this chart completely ignores Region or State, losing geographic context.
d. Difficult to interpret proportionally: In Sunburst charts, it's hard to accurately compare segment sizes or understand the scale, especially with nested rings.
e. Doesn't highlight outliers or inefficiencies: The chart does not reveal underperforming products or categories with high sales but low profit, missing actionable insights.

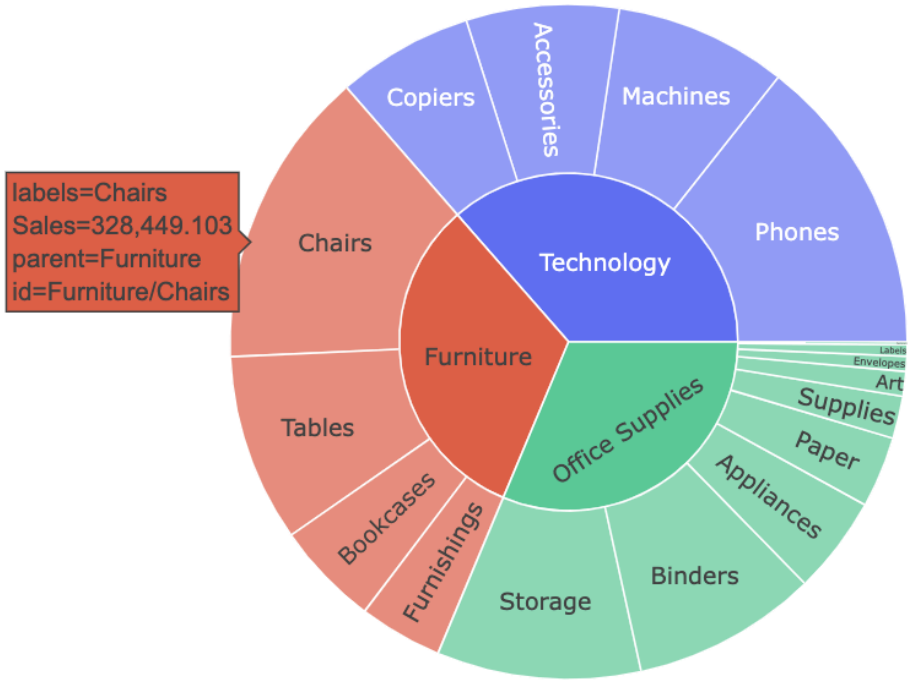

Figure 13: Sunburst chart by Category and Sub-Category

RESULTS AND INSIGHTS

Objective 1: Utilize descriptive and visual analytics to analyze a U.S. Office supply retail dataset.

We began our exploratory data analysis/ objective 1 analysis by loading the retail dataset and reviewing the first few rows to understand its structure. Summary statistics provided a quick overview of the central tendencies and spread of numerical features. We also checked for missing values and confirmed data types to ensure the dataset's readiness for analysis. To explore data distributions, we plotted histograms for all numerical columns, which revealed varying scales and skews in features like Sales and Profit. For categorical variables such as Ship Mode, Segment, State, and Category, count plots highlighted the distribution of entries, helping identify dominant categories.

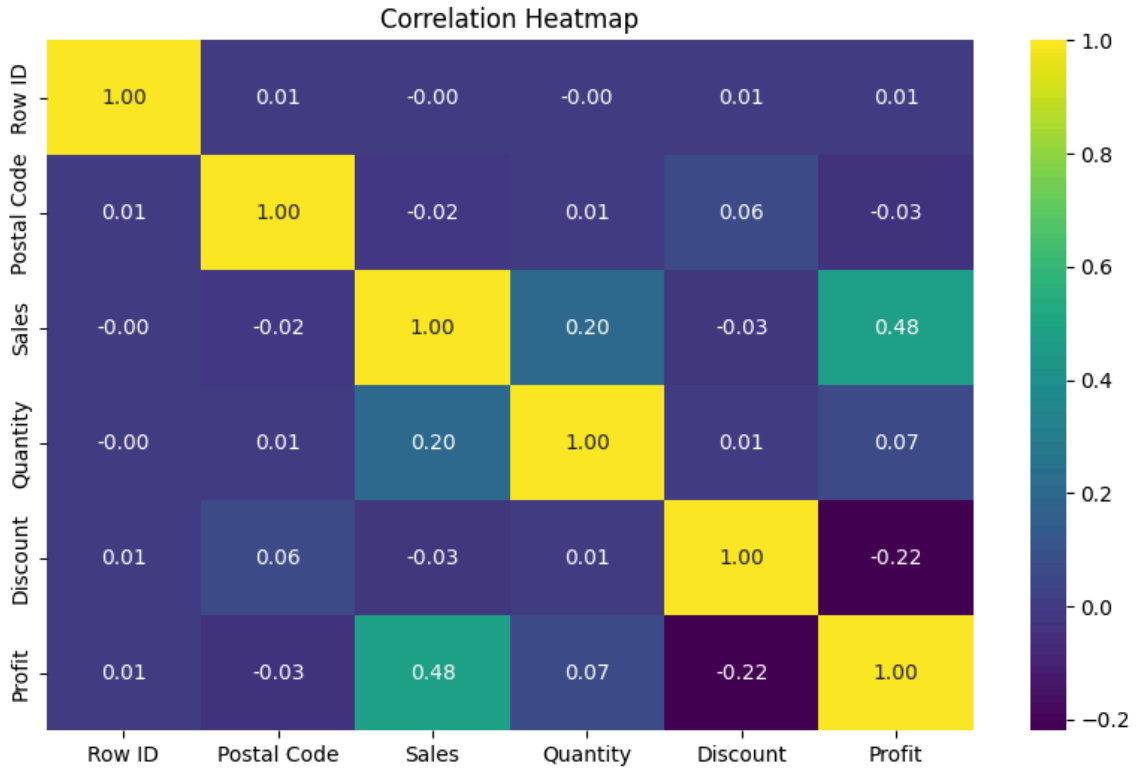

Figure 14: Correlation heatmap among numerical variables

A correlation heatmap as shown in figure 14 was generated to examine relationships between numerical variables. This revealed expected positive correlations between Sales and Profit, as well as weaker or negligible relationships among some other features. Additionally, we identified the top 10 customers by total sales using the bar chart as shown below in figure 15, highlighting key contributors to revenue. This helped in recognizing potential high-value customers for targeted strategies.

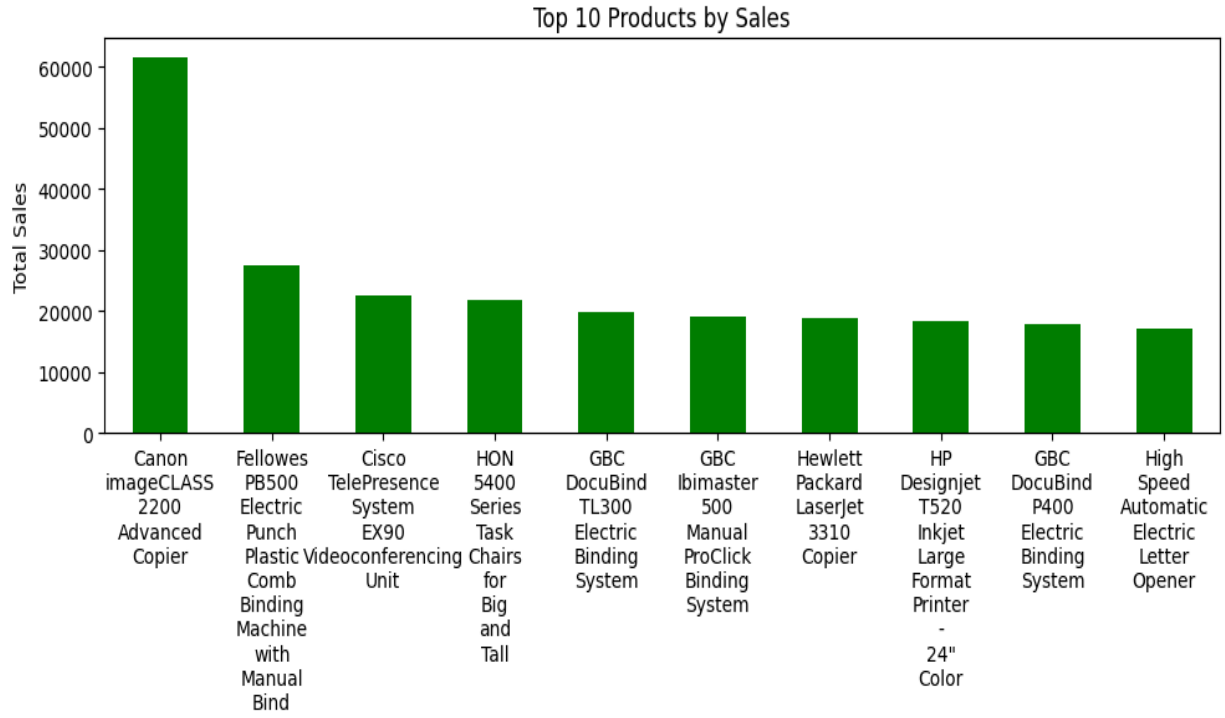

Figure 15: Vertical Bar chart showing top 10 customers by total sales

We analyzed sales by day of the week using a boxplot as shown in Figure 16, revealing variations in daily sales patterns. This helps identify high- and low-performing days, useful for planning promotions and operations. Temporal analysis included a monthly sales trend shown in Figure 17, which revealed seasonality and fluctuations over time, and a weekday analysis, showing how sales vary across days of the week.

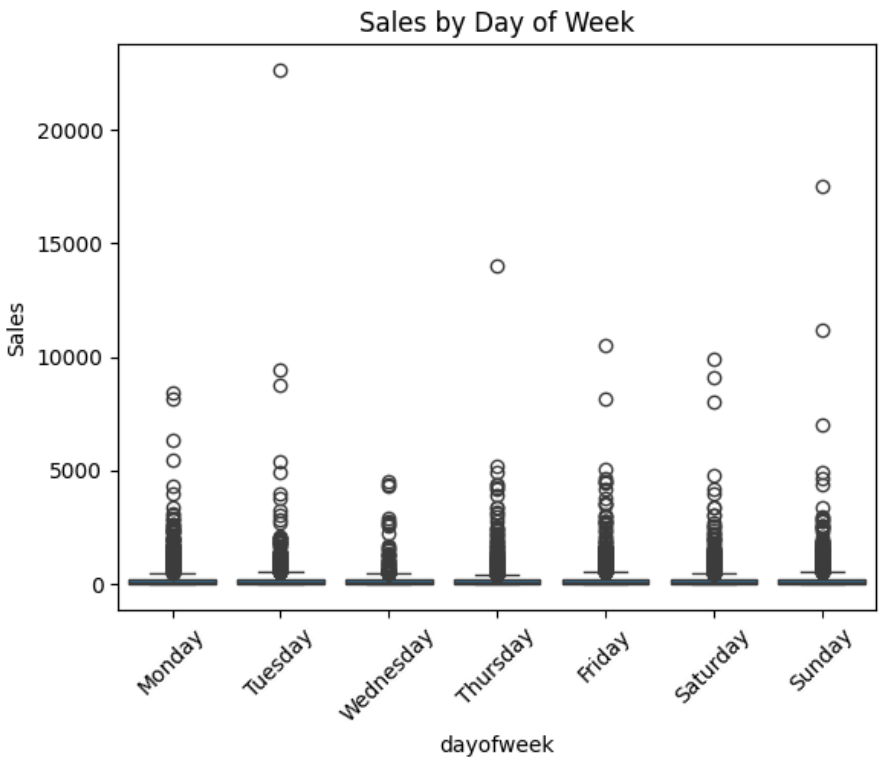

Figure 16: Boxplot showing sales by day of the week

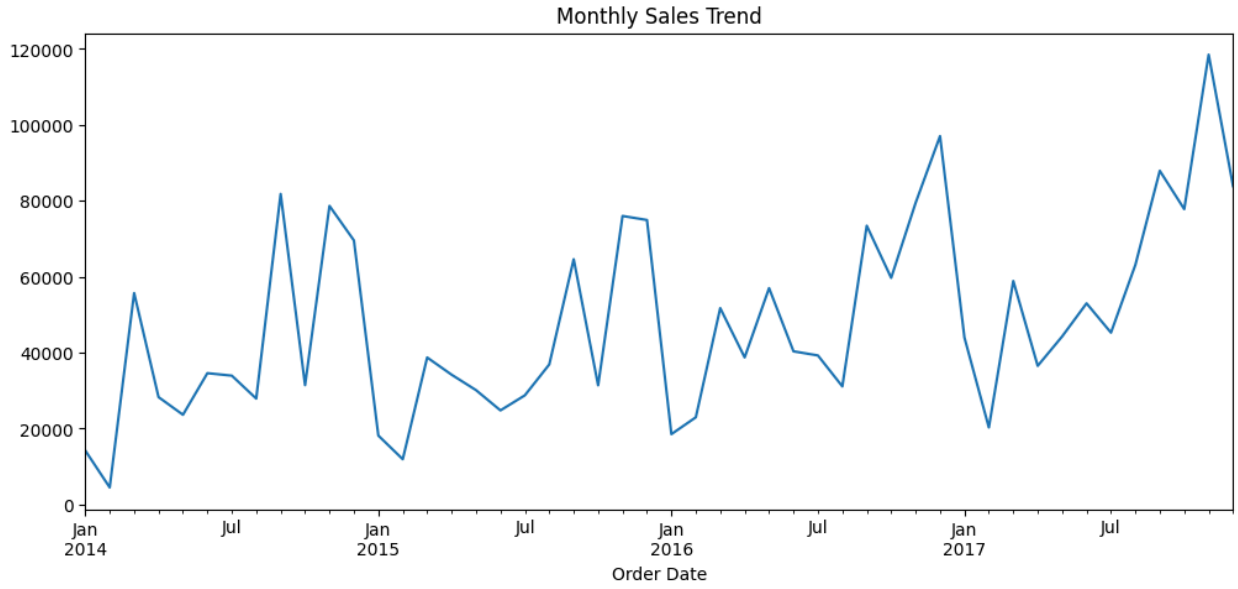

Figure 17: Line chart showing temporal analysis on monthly sales

We created a pairplot as shown in figure 18, on a random sample to visualize interactions and clusters among numerical features.

Figure 18: A pairplot to visualize interactions and clusters among numerical features.

A pareto analysis on Cumulative percentage of Total Sales Contribution by customers as demonstrated in figure 19, shows a small percentage of customers contribute disproportionately to total revenue, consistent with the 80/20 rule. This insight supports more targeted marketing and customer relationship strategies.

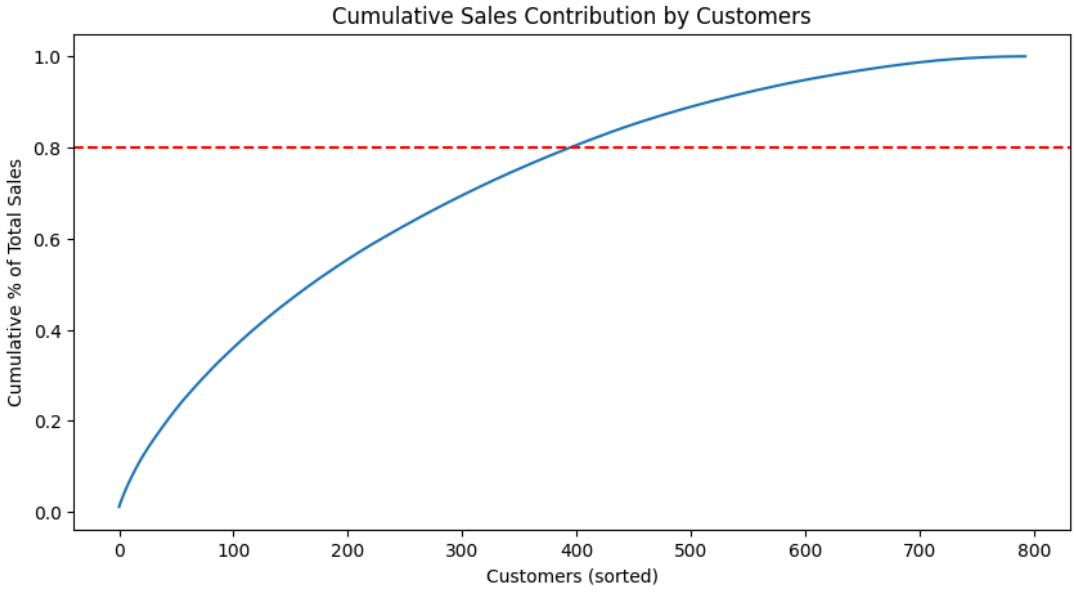

Figure 19: Cumulative percentage of Total Sales Contribution by customers

Histograms of Sales and Profit as shown in figure 20, showed both variables to be highly skewed, with most values clustered at the lower end, especially for Profit.

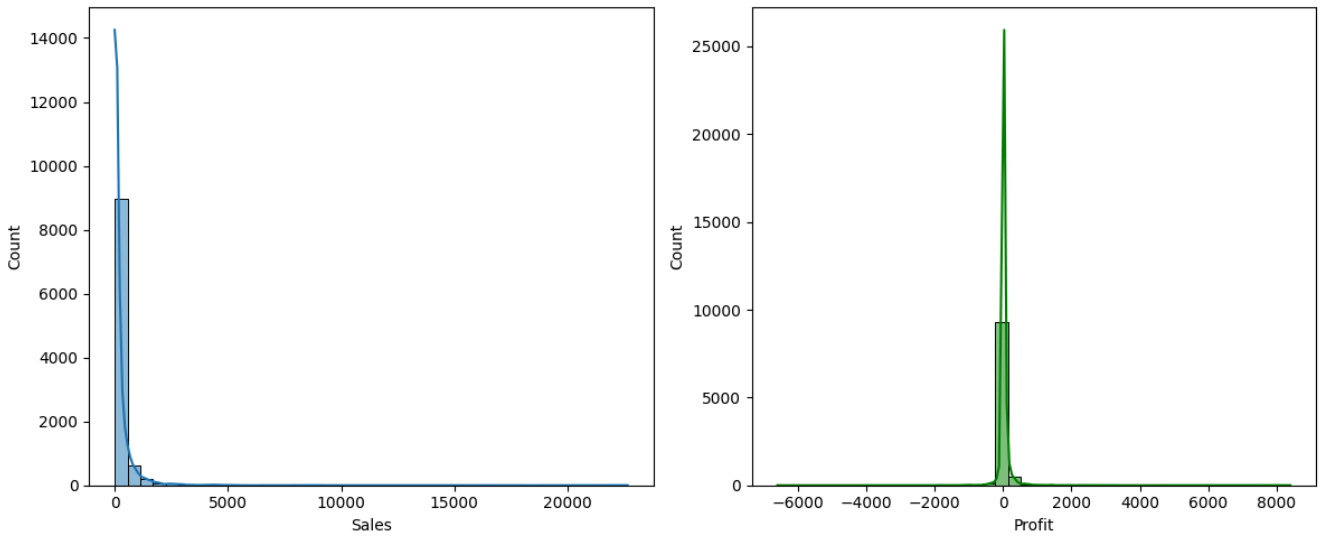

Figure 20: Histogram plot of Sales and Profit

We examined Sales and Profit by Region as shown in figure 21, where certain regions stood out in revenue generation but differed in profitability. Similarly, a category-wise breakdown showed that some categories, despite high sales, may contribute less to overall profit, highlighting areas for cost optimization or pricing adjustments.

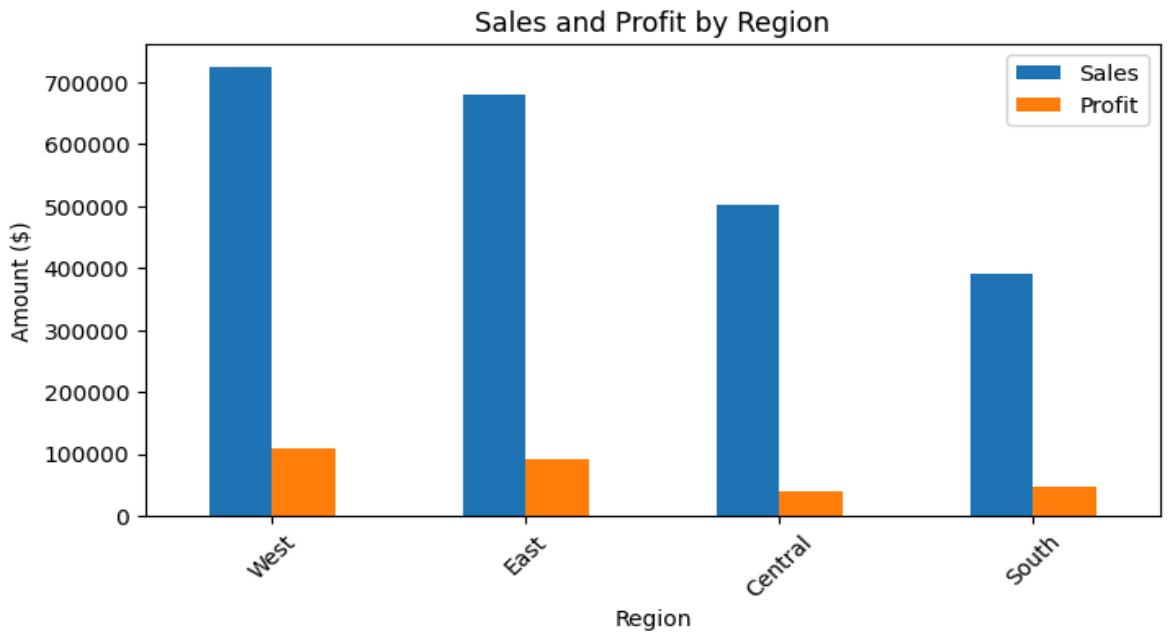

Figure 21: Grouped Barchart indicating Sales and Profit by Region

We further visualized total profit by Sub-Category using a horizontal bar chart as shown in Figure 22. This made it easy to spot which sub-categories are underperforming or potentially incurring losses, useful for guiding pricing or product focus decisions.

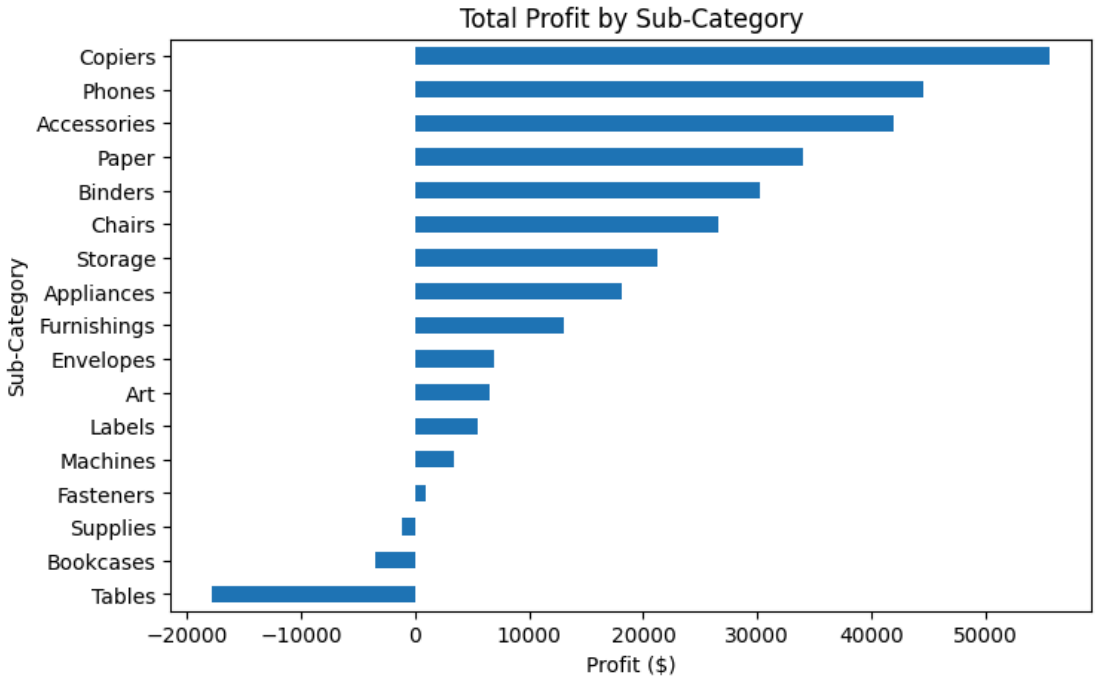

Figure 22: Horizontal Bar chart for visualizing Total Profit by Sub-Category

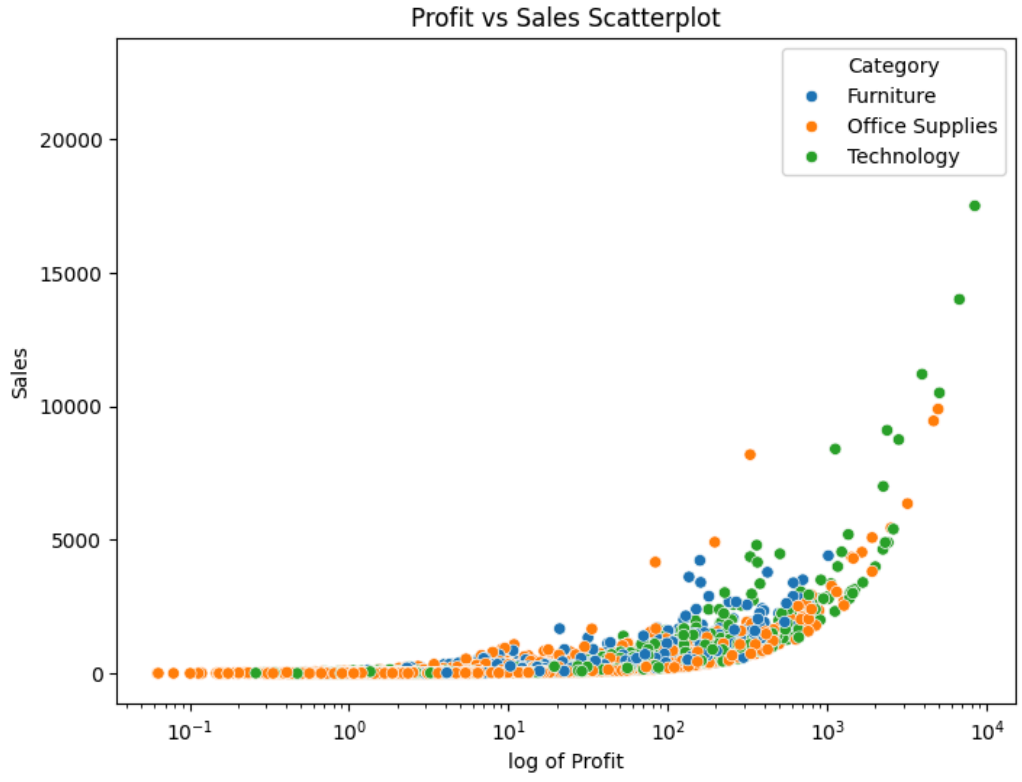

Figure 23: Scatterplot of Log of Profit vs Sales, hue by Category

A scatterplot of Profit vs Sales, colored by Category as shown in figure 23, illustrated profitability patterns at different sales levels. The log scale on Profit made it easier to observe the spread across a wide range of values. To evaluate high-level product group performance, we aggregated data by Category and Sub-Category, summing up Sales, Profit, and Discount. This allowed us to identify which product segments contribute most to revenue and profitability. The results helped highlight top-performing combinations and those with lower or even negative profit margins.

By creating a pivot table and visualizing it with a heatmap, we were able to compare profit levels across customer segments (e.g., Consumer, Corporate, Home Office) in each region as shown in Figure 24. This helped reveal where certain segments are more profitable and where additional marketing effort may be needed to improve performance.

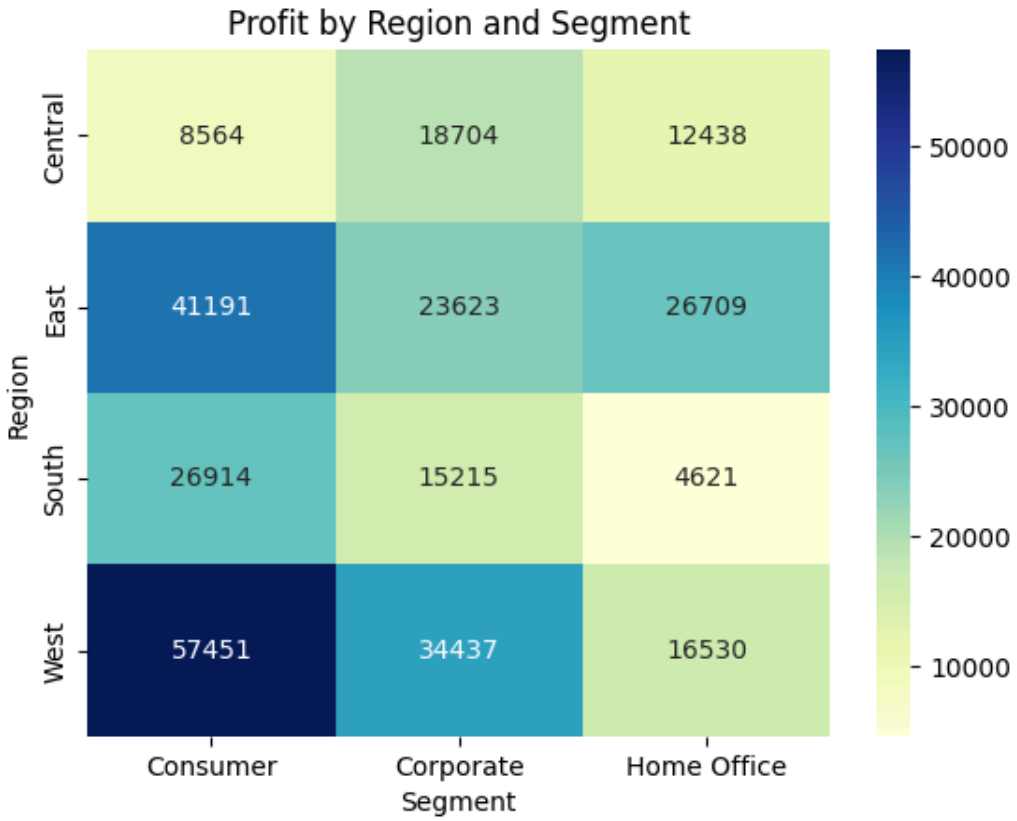

Figure 24: Heatmap of Profit by Region and Segment

The contour plot shown in figure 25 visualizes sales density over time across customer segments. The X-axis represents the order month, while the Y-axis corresponds to customer segments (Consumer, Corporate, Home Office). Using interpolated sales values, the plot reveals temporal trends and segment-specific performance patterns, helping identify sales peaks, seasonal behavior, or segment-driven surges over time.

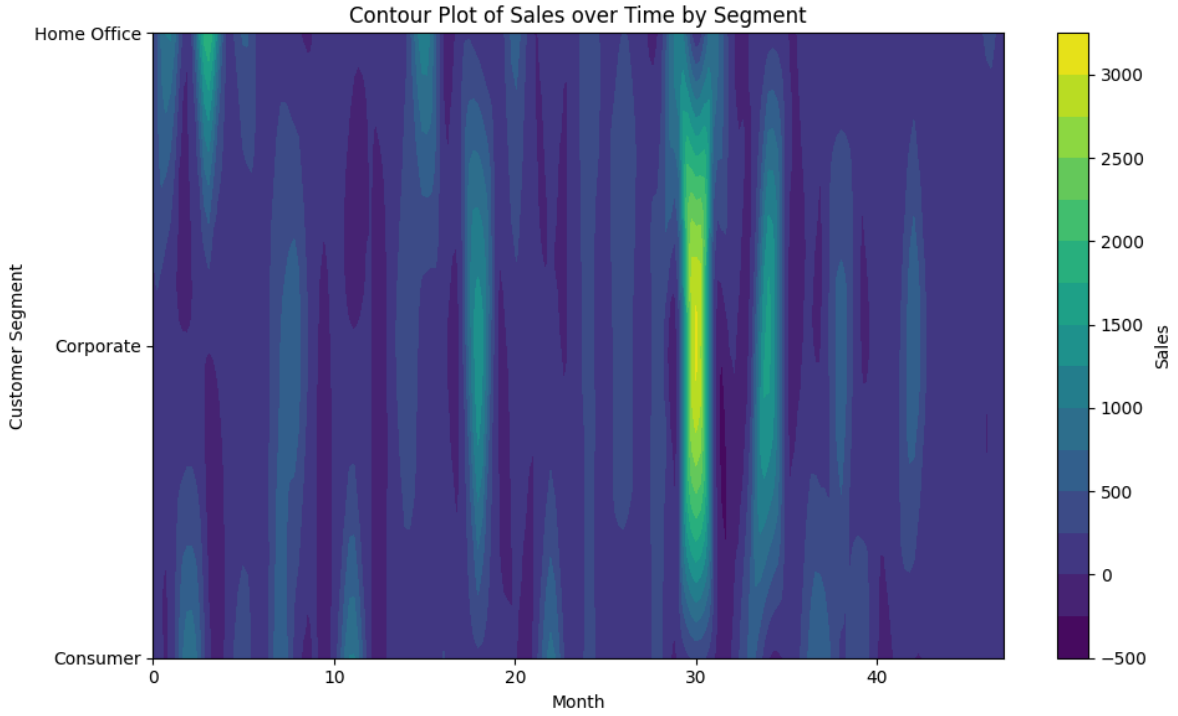

Figure 25: Contour plot visualizing sales density over time by customer segments

The Exploratory Data Analysis provided a comprehensive overview of the dataset, uncovering key patterns and trends across time, geography, customer segments, and product categories. Through visualizations and aggregations, we identified top-performing products, regions, and customers, along with areas of low profitability and potential improvement. Temporal and categorical insights, including sales trends, day-of-week behavior, and segment-wise performance, offered valuable direction for strategic decision-making. Overall, this analyis laid a strong analytical foundation for targeted actions and deeper business understanding.

Objective 2: Identifying High-Performing and Underperforming Products Using Sales and Profit Metrics

The objective of this report is to identify high-performing and underperforming products based on their sales and profit metrics using the Superstore sales dataset. This analysis was conducted in Power BI by aggregating total sales and total profit for each product and classifying them into performance categories using quartile-based logic. A new summarized table, ProductPerformance, was created using DAX to compute these aggregates, and a calculated column titled Performance Category was added to classify products as "High-Performing," "Underperforming," "Low Margin," or "Moderate." A scatterchart as shown in figure 26 was used to visualize product performance, with total sales on the X-axis and total profit on the Y-axis. Products were color-coded based on their performance category, and product names were displayed directly on the chart to allow quick identification without hovering.

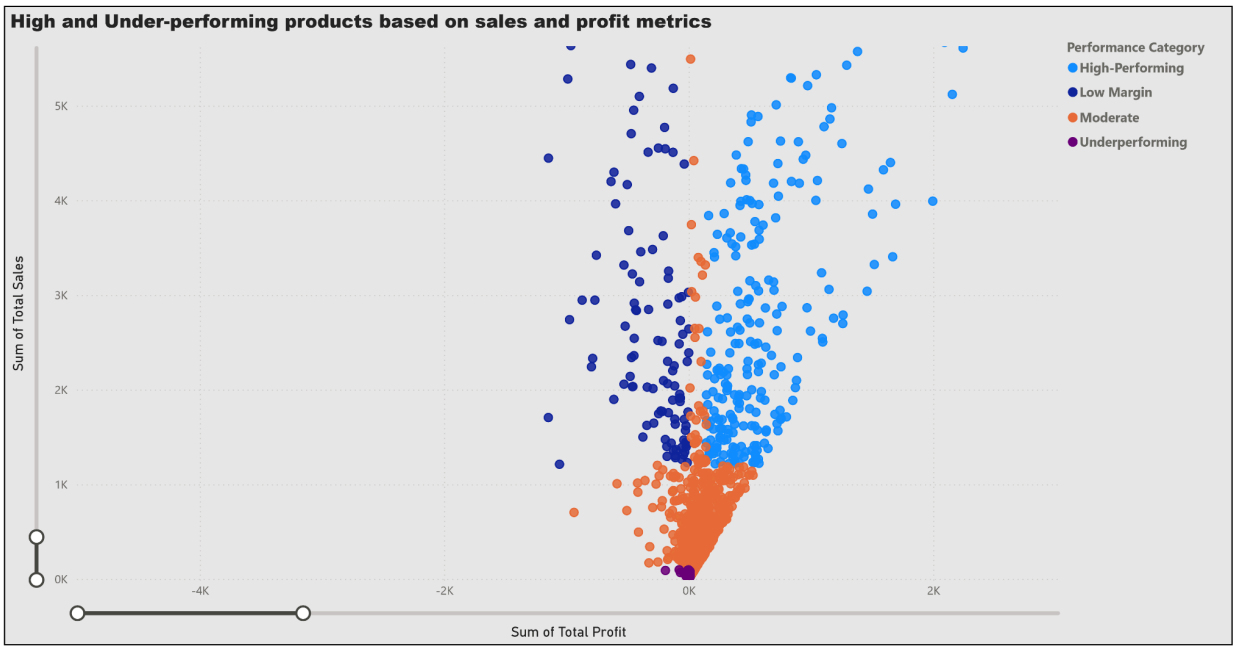

Figure 26: Scatterplot visualizing product performance, with total sales and total profit, products were color-coded based on their performance category

High-performing products as shown in figure 27 were defined as those with both total sales and profit above the 75th percentile. These products contribute significantly to both revenue and profitability, making them strategic assets in the company's portfolio. For instance, the Canon imageCLASS 2200 Advanced Copier generated $61,599 in sales and over $25,199 in profit, making it the most lucrative item in the dataset. Other high performers included the Fellowes PB500 Electric Punch Binding Machine, which earned over $27,000 in sales with a healthy profit margin of $7,753, and the HP Designjet T520 Large Format Printer and 3D Systems Cube Printer, both of which achieved strong financial performance.

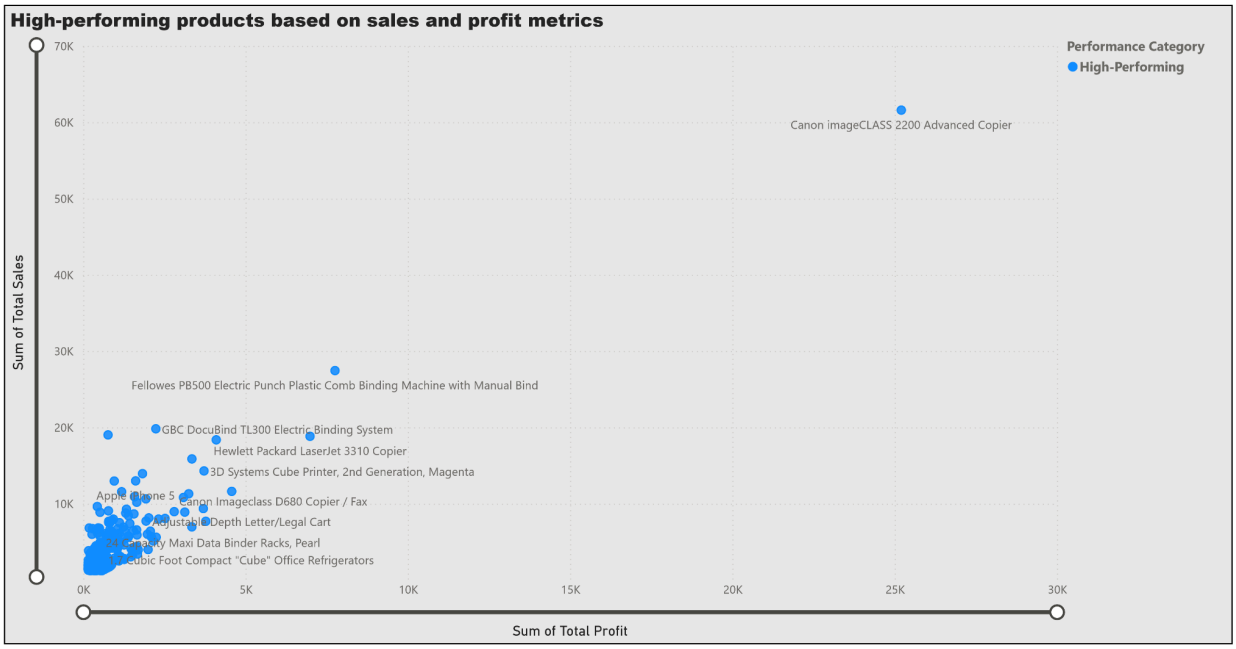

Figure 27: High Performing products

Underperforming products as shown in figure 28, were those with both sales and profit below the 25th percentile. These items not only failed to drive revenue but also incurred losses, making them potential candidates for discontinuation or strategic reevaluation. For example, the Bush Westfield Collection Bookcases recorded sales under $100 and a net loss of nearly $191. Similarly, office accessories such as Eldon Cherry Finish Desk Accessories, GBC Plasticlear Binding Covers, and Insertable Tab Post Binder Dividers experienced minimal sales and negative profits, likely due to low demand or uncompetitive pricing.

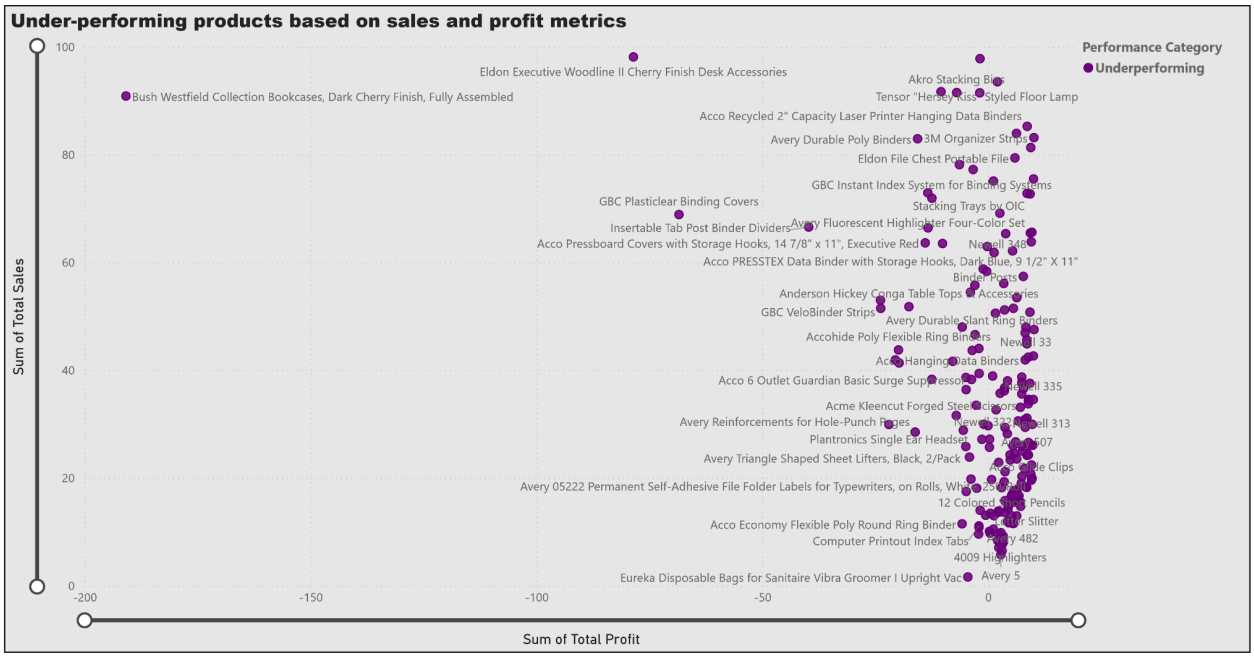

Figure 28: Under-Performing products

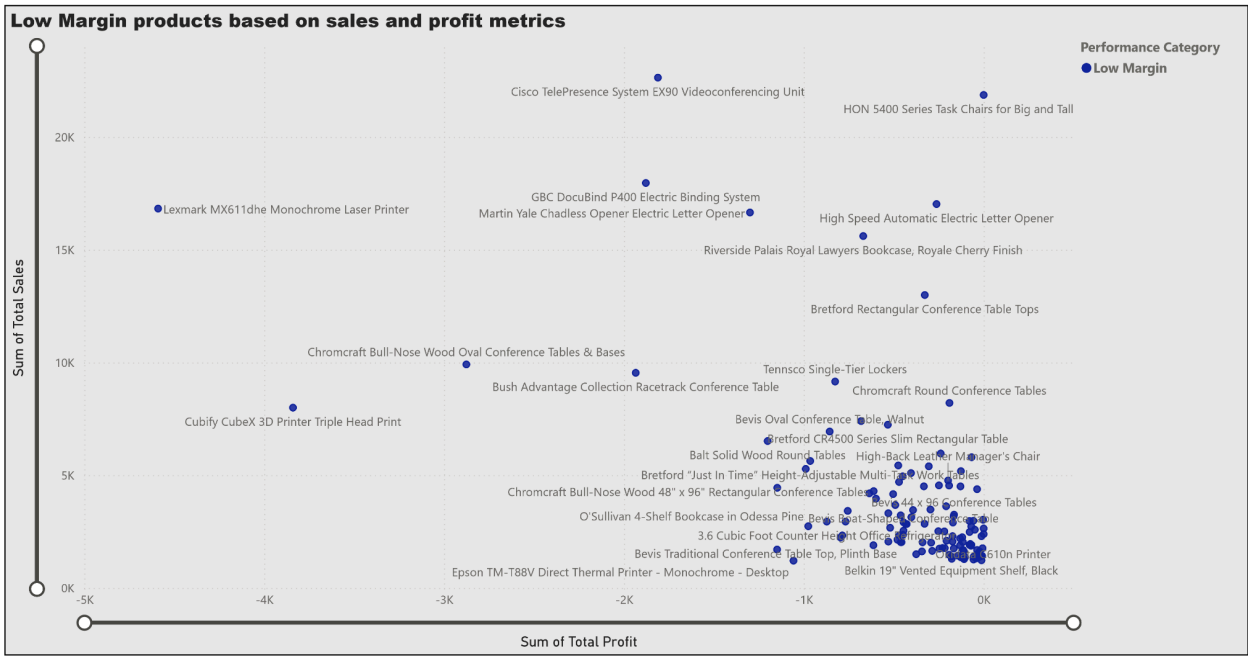

Figure 29: Low margin products

Low-margin products as shown in figure 29 are those that had relatively high sales but fell in the bottom 25% of the profit distribution. These items may appear successful at first glance due to high sales volumes, but their tight or negative margins pose long-term risks to profitability. Examples include certain office accessories or tech items like Plantronics Single Ear Headsets and the Texas Instruments TI-15 Calculator, which may be burdened by high fulfillment costs or excessive discounting. A visual representation—either a filtered scatter plot or a bar chart—can be inserted here to showcase products with high revenue but unsatisfactory profits, ideally in a different color like yellow or orange.

By classifying products based on sales and profit metrics, we uncovered clear performance patterns. High-performing items largely fall under office equipment and printing solutions, indicating strong demand for functional, high-tech products. In contrast, low-tech or general office supplies often underperform, suggesting a need for SKU rationalization or pricing review. An interactive scatter chart enabled stakeholders to easily explore these trends and derive actionable insights.

High performers should be prioritized in marketing, logistics, and pricing strategies. Underperformers may benefit from bundling, cost reductions, or phase-out. Low-margin products require targeted pricing reviews or supplier renegotiations. Overall, integrating sales and profit data into a performance classification model in Power BI provides a nuanced, data-driven view of product-level performance. It supports more informed decisions across marketing, inventory, and pricing, enhancing portfolio profitability and efficiency through visually supported analysis

The **scatterplot** was perfect for this objective because it clearly showed the relationship between sales and profit for each product. Plotting total sales on the X-axis and total profit on the Y-axis it allowed easy identification of products that perform well or poorly in both areas. The color-coding of products based on performance categories made it simple to see which items were high-performing, underperforming, or low-margin at a glance. Displaying product names directly on the chart helped stakeholders quickly recognize specific products without extra interaction. Additionally, the scatterplot highlighted important trade-offs, such as products with high sales but low profits, enabling better strategic decisions.

Overall, this visualization made it easy to explore and understand complex performance data clearly and intuitively.

In the **Sunburst chart** provided in figure 30, this analysis is visually represented through a hierarchical breakdown of product categories and subcategories, with color gradients indicating profitability. Products in green shades are high performers, showing strong profit margins, while those in red indicate losses or low profitability.

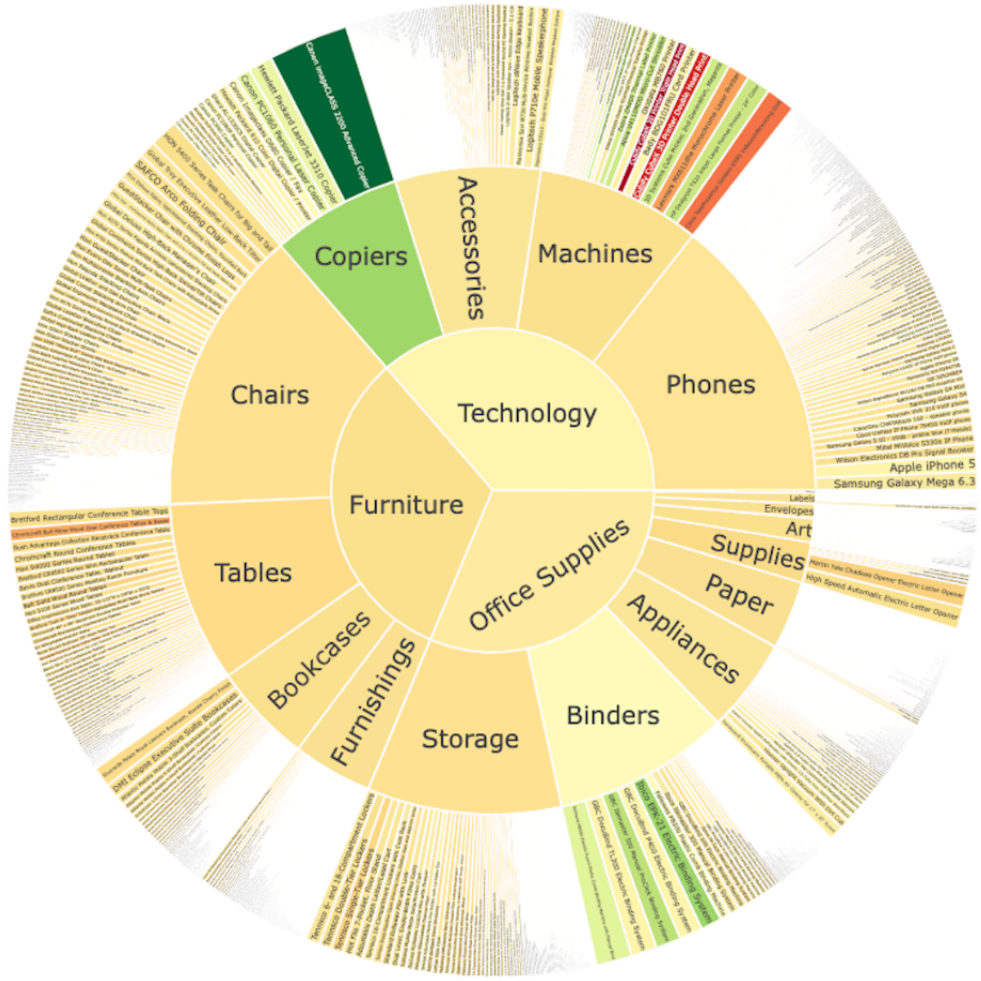

Figure 30: Sunburst Chart to visualize Sales and Profit analysis by Product Hierarchy

High-performing products, such as those in the Technology category—particularly Phones and Accessories—stand out due to their consistent sales and high profit margins. These products are likely well-aligned with market demand, priced effectively, and efficiently managed in terms of cost. Their success suggests that the company should consider investing more in these areas, possibly expanding the product range or increasing marketing efforts to capitalize on their popularity.

On the other hand, underperforming products, such as certain items in the Furniture category—like Tables or Bookcases—are marked in red, indicating negative profit margins. These products may suffer from high production or shipping costs, low sales volume, or pricing issues. Identifying these weak spots allows the company to take corrective actions, such as reevaluating pricing strategies, improving supply chain efficiency, or even discontinuing unprofitable items.

Overall, this Sunburst chart analysis empowers decision-makers to allocate resources more effectively, optimize their product portfolio, and enhance overall profitability by focusing on what works and addressing what doesn't

Improvement Plans for Enhancing Product-Level Profitability

1. Conduct margin audits for low-performing and low-margin products to identify cost drivers or excessive discounting.
2. Review pricing strategies and apply dynamic pricing models for frequently discounted but high-volume items.
3. Work with procurement to renegotiate supplier contracts for high-sales but low-profit products to improve margins.
4. Use time-series sales data to adjust inventory planning and avoid overstocking underperforming products.
5. Monitor product return rates and customer reviews to understand qualitative reasons behind low product performance.

Objective 3: Evaluate the impact of discounts on profitability across product categories.

The objective of this analysis is to evaluate how discounts influence profitability across various product sub-categories. Understanding this relationship is critical for

identifying pricing inefficiencies, uncovering areas of margin erosion, and guiding discounting strategies that align with profit goals. The analysis was conducted using ObservableHQ, leveraging its reactive programming capabilities to create an interactive and user-driven experience.

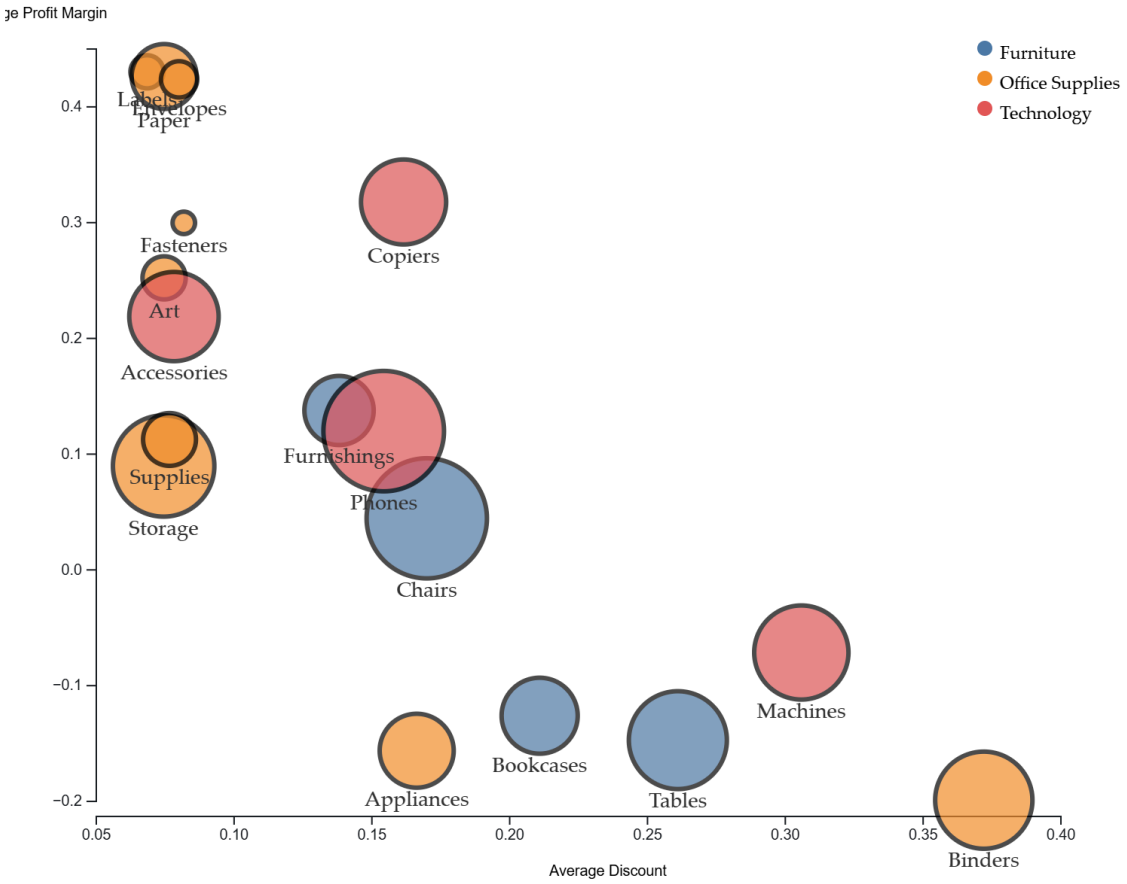

Figure 31: Bubble chart visualizing the relationship between average discount and average profit margin, with bubble size representing total sales and color indicating category

The analysis began by loading a Superstore sales dataset using FileAttachment or d3.csv() and aggregating it by sub-category using d3.rollups() to compute key metrics: average discount, average profit, profit margin, total sales, and record count. These summaries were formatted using .map() and enriched with category labels for clarity. An interactive dropdown (Inputs.select()) allowed users to filter data by category or view all sub-categories, with a reactive variable (filteredData)

updating visuals accordingly. A bubble chart as shown in figure 31 built with Observable HQ platform visualized the relationship between average discount (x-axis) and average profit margin (y-axis), with bubble size representing total sales and color indicating category. Tooltips, labels, and visual enhancements (like black outlines) made the chart interactive and easy to interpret.

Within the Furniture category, sub-categories like Furnishings, Chairs, Bookcases, and Tables exhibit distinct patterns in discounting, profitability, and sales volume. Furnishings is the most profitable, with the highest average profit margin and the lowest discount rate, though it lags in sales volume—highlighting growth potential through targeted promotion. Chairs lead in total sales and maintain a moderate but positive profit margin, indicating a healthy balance between revenue and profitability. In contrast, Bookcases and Tables show negative profit margins despite moderate sales and some of the highest average discounts, suggesting that over-discounting or high costs may be undermining profitability and require pricing or cost adjustments. Overall, the Furniture category illustrates the trade-offs between sales, discounting, and profitability—positioning Furnishings as a premium, profitable niche, Chairs as a high-volume, moderately profitable driver, and Bookcases and Tables as potential risks requiring corrective action to improve margin efficiency.

The Office Supplies category shows varied performance across sub-categories. Labels, Envelopes, and Paper stand out with high average profit margins (~0.4) and low discounts, reflecting efficient pricing and strong profitability—making them ideal for bundling or promotion. In contrast, Appliances and Binders are concerning due to high sales volumes paired with negative profit margins and steep discounts—especially Binders, with a 0.38 discount rate and -0.2 margin, making it a top candidate for pricing review. Supplies also generate high sales with a modest positive margin, indicating a more balanced strategy. Sub-categories like Fasteners, Art, and Storage show moderate margins and low discounts but low sales, suggesting untapped potential for targeted marketing or bundling. Overall, Office Supplies reveals a clear split between high-margin essentials, unprofitable high-volume items, and moderate performers with growth potential—offering clear opportunities for pricing optimization and promotion strategy.

The Technology category shows varied sub-category performance in profitability and discounting. Phones lead in sales but have thin margins (~12%) and moderate discounts (~15%), indicating strong demand but limited profitability. Copiers, though not top in sales, yield the highest profit margin (~32%) with moderate discounts (~17%), making them ideal for premium-focused strategies. Accessories offer a strong balance—solid margins (~22%) and low discounts (~8%)—suitable for bundling or growth initiatives. In contrast, Machines show high discounts (31%) and negative margins (-1%), making them unprofitable despite decent sales and a clear candidate for pricing or cost structure review. Overall, Technology mixes high-volume, low-margin items (Phones), high-margin niches (Copiers), and discount-driven loss leaders (Machines), providing a strong foundation for targeted pricing and promotional strategies.

The **Bubble chart** was ideal for analyzing this objective because it allowed for a clear and intuitive comparison of multiple key metrics at once. Each bubble represented a product sub-category, with its position showing the relationship between average discount (x-axis) and average profit margin (y-axis), while the size of the bubble indicated total sales volume. This multi-dimensional view made it easy to identify which sub-categories were highly profitable, over-discounted, or driving large volumes of sales. The use of color to distinguish broader product categories further enhanced readability and helped highlight performance differences within and across categories. Additionally, the chart supported interactive filtering, allowing users to focus on a specific category while retaining the ability to explore overall trends. Overall, the bubble chart provided a powerful way to spot patterns, outliers, and strategic opportunities in a single, unified visual.

The graph shown is a **Violin plot (or swarm plot)** in figure 32 that illustrates the distribution of profit across three major product categories—Furniture, Office Supplies, and Technology—with each data point representing an individual product transaction. The x-axis lists the product categories, while the y-axis shows the profit value associated with each transaction. The data points are colored based on the discount percentage applied, as indicated by the legend on the right.

In the Furniture category, profit values are generally clustered near the lower range with relatively limited variation. There are a few negative profit instances,

especially for higher discount values (such as 0.5 and 0.6), suggesting that heavy discounts may not be sustainable for profitability in this category. However, the category overall appears more stable, with fewer extreme profit outliers, indicating a more predictable margin structure.

The Office Supplies category displays a more varied profit distribution. A significant number of profitable transactions are observed even at zero or low discount levels (0.0, 0.1, 0.15). Nonetheless, some sharp dips into negative profit are visible, particularly when higher discounts (like 0.4 and above) are applied. This indicates that while office supplies can be sold profitably, discounts beyond a certain threshold tend to erode profitability quickly.

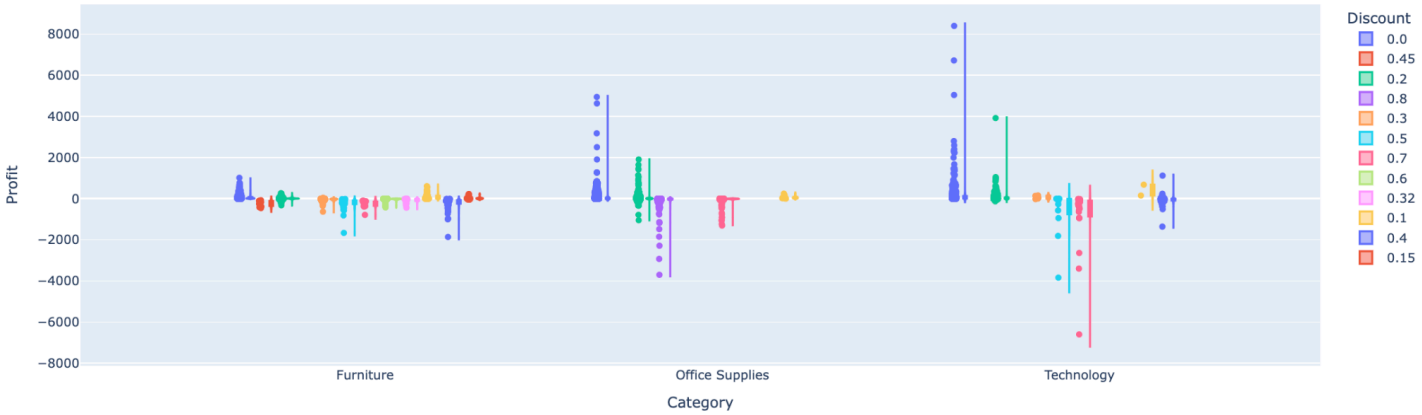

Figure 32: Violinplot indicating profit distribution per Category, by discount

The Technology category exhibits the widest spread of profit values, ranging from highly profitable sales to large losses. There is a noticeable pattern where many of the most negative profit outcomes are tied to higher discount levels (such as 0.7, 0.5, and 0.8), clearly indicating that aggressive discounting in the technology category leads to significant financial losses. On the other hand, some of the highest profit outliers also come from this category when no or low discounts are applied, suggesting potential for high margin products if discounting is controlled.

In summary, the violin plot reveals that discounting has a measurable and often negative impact on profit, especially in categories like Technology and Office Supplies. Lower or no discounts generally correlate with positive profits, while higher discounts often lead to losses, particularly in Technology. These insights can

help businesses tailor their pricing and discount strategies more effectively by identifying underperforming segments and reinforcing investment in high-performing products.

Improvement Plans for Optimizing Discount Strategies and Profitability

1. Limit discounts on loss-making items like Binders and Machines. Adjust pricing strategies using discount–profit trend analysis.
2. Bundle high-margin, low-sales products (e.g., Furnishings, Accessories) to boost volume. Set discount caps to protect margins in sensitive categories.
3. Investigate cost issues in low-margin sub-categories (e.g., Machines, Bookcases). Promote profitable, low-discount items like Copiers and Labels.
4. Use margin alerts to flag risky discounting in real time. Update visuals quarterly to track discount-policy impact.

Objective 4: Analyze order and shipping timelines to assess operational efficiency.

The primary objective of this analysis is to assess operational efficiency by evaluating the shipping timelines across different product categories and geographic regions. By examining the delay between the order date and the shipping date, we can uncover patterns in performance, identify bottlenecks, and pinpoint regions or categories that require process improvement. A heatmap visualization serves as an effective tool to present this multi-dimensional data compactly, highlighting variations in average shipping delays.

The dataset used for this analysis contains historical order-level information, including fields such as Order Date, Ship Date, Region, and Category. The data is assumed to be loaded in an Observable cell named data. Each record represents an individual order and contains textual date fields that need to be parsed into JavaScript Date objects for analysis.

To calculate the shipping delay, the following steps were taken:

1. Iterate over each row in the dataset.
2. Convert the Order Date and Ship Date strings into valid Date objects.
3. Subtract the order date from the shipping date to calculate the delay in days.

4. Filter out invalid or negative delay values to ensure data quality.

Once valid delays were extracted, we used the d3.rollup() function to compute the average shipping delay for each combination of Region and Category.

The heatmap visualization is a valuable decision-support tool for improving shipping efficiency and product availability across regions. By showing average shipping delays for each region–product category combination, it offers a clear, immediate view of operational performance. Business teams can quickly spot regions with persistent delays in specific categories—such as Technology in the South—prompting targeted investigations into local suppliers, logistics partners, or inventory levels. Low-delay cells indicate strong performance and can serve as operational benchmarks. The heatmap also highlights potential gaps in product availability; empty or gray cells may suggest unmet demand or strategic blind spots, allowing the business to assess whether these are due to supply constraints, low demand, or distribution choices. Overall, this visualization bridges data with action, enabling operations, sales, and logistics teams to align efforts that boost customer satisfaction, streamline inventory, and reduce delivery times across regions.

Figure 33: Heatmap showing average shipping delay by region and category

Each cell in the heatmap as shown in figure 33, represents the average shipping delay (in days) for a specific combination of region and product category. The numeric value is easy to interpret at a glance, and the color intensity gives a visual cue; darker shades indicate higher delays, while lighter shades suggest better performance. For example, in the central region, Office Supplies has the highest delay (4.1 days), which stands out due to the darkest color. All other categories and regions hover around 3.9–4.0 days, indicating relative consistency. This allows teams to instantly pinpoint operational bottlenecks or regions/categories where logistics need improvement.

From a business standpoint, decisions should focus on balancing regional supply chain efficiencies, improving inventory allocation strategies, and streamlining distribution processes. For instance, the company might consider increasing fulfillment resources in the Central region, revisiting vendor performance in Office Supplies, or implementing predictive demand planning tools. Additionally, since delays are relatively uniform in other regions, leadership can explore benchmarking best practices from those areas to standardize performance nationwide.

The **heatmap** is a powerful visualization tool that not only highlights surface-level issues like average shipping delays across regions and product categories but also enables deeper, more granular analysis. It facilitates exploration of patterns by shipping mode (e.g., standard vs. express), delivery performance of top-selling SKUs, and the impact of seasonal trends or external disruptions such as weather or labor shortages. By localizing operational bottlenecks and enabling cell-by-cell comparisons, the heatmap supports targeted, data-driven decisions. Its structured grid format allows for clear visualization of exact values across two categorical variables, making it easy to spot high or low performers and track changes over time. This makes it ideal for both diagnostics and continuous monitoring, driving improvements in logistics, supply chain resilience, and overall efficiency.

On the other hand, a **Contour plot** offers a more fluid and continuous view of the data. Instead of fixed blocks, it uses smooth color gradients and contour bands to show how values change gradually across dimensions. This makes it better for spotting underlying trends, transitions, and zones of similar intensity, such as delay

build-ups or shifting hotspots over time. While heatmaps focus on discrete clarity, contour plots excel at revealing patterns and movement within the data.

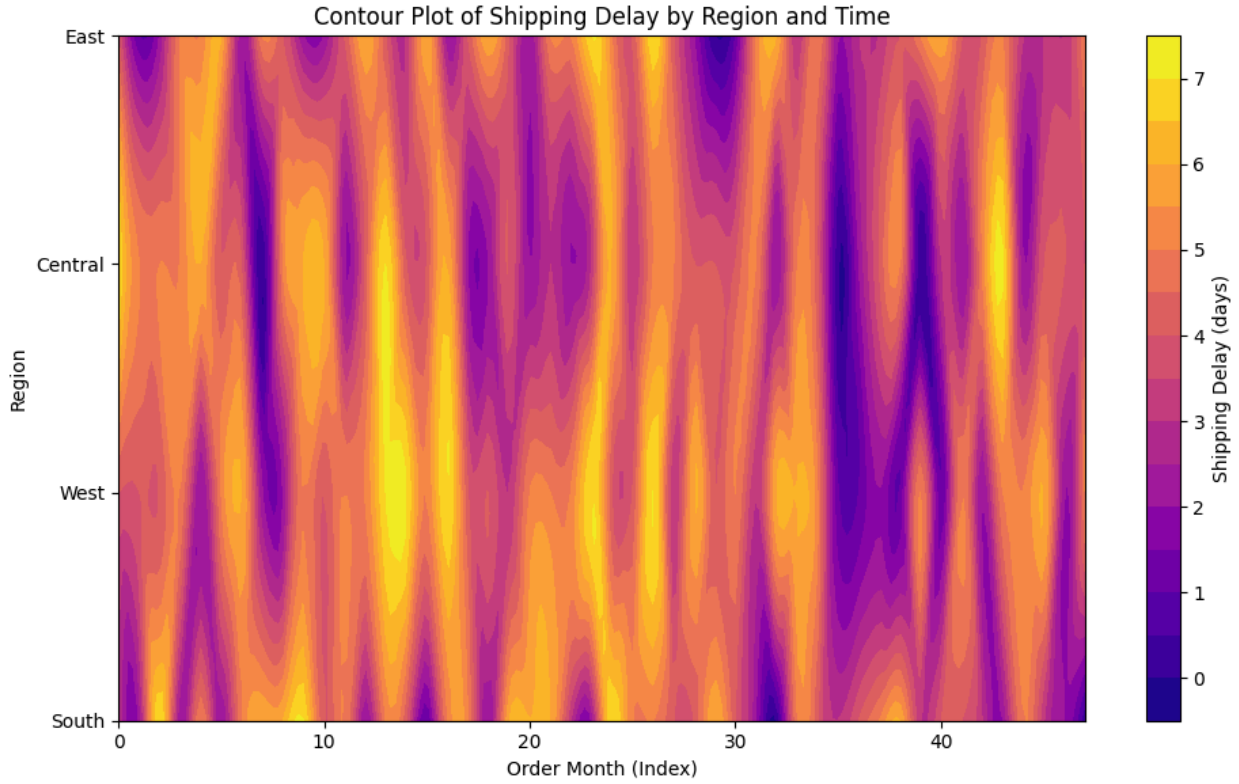

Figure 34: Contour plot showing shipping delay by region and time

This contour plot shows in figure 34, how shipping delays varied over time across different regions: South, West, Central, and East. The x-axis represents time in months (indexed numerically), while the y-axis lists the regions. The color gradient reflects the average shipping delay in days, with brighter colors (yellow) indicating longer delays and darker shades (purple) showing shorter delays.

From the plot, it's clear that shipping delays fluctuated over time in all regions, with noticeable spikes in delay at several points, particularly in the Central and West regions. These areas experienced more frequent and intense peaks, suggesting possible seasonal congestion or logistical issues. In contrast, the East and South regions showed more consistent performance, though occasional delays still occurred. Overall, the plot helps identify which regions faced the most significant

delays and when those delays happened, offering valuable insight for improving delivery operations and planning.

The **contour plot** is great for interpolating between values. This produces smooth transitions and continuous lines that reveal subtle trends and inflection points. Contour lines (or bands of color) group regions by delay levels, e.g., areas with delays between 3–5 days. This allows you to easily see zones of similar delay intensity, not just exact values. While a heat map might just show a jump in color from one cell to the next, a contour plot reveals how sharply or gradually delays increase or decrease, giving more nuance to how performance evolved. If you think of delays "spreading" across time and region like a wave or weather pattern, the contour plot gives a fluid, topographic view, ideal for this kind of interpretation. In the contour plot, banding effects make it easier to identify periodic peaks or clustered disruptions that might be harder to distinguish in a heat map's fixed cells.

Improvement Plans

1. Audit Central region logistics operations, check for inventory shortages, processing delays, or transportation issues.
2. Evaluate supplier lead times for Office Supplies and consider dual sourcing to avoid bottlenecks.
3. Implement regional performance tracking with alerts for deviations in delivery times.
4. Introduce automation in fulfillment centers to speed up order picking and packing.
5. Rebalance warehouse inventory across regions based on demand forecasts and delivery patterns.
6. Enhance carrier performance SLAs and consider alternative logistics providers in underperforming regions.
7. Regularly update the heatmap dashboard to track post-implementation performance improvements.

Objective 5: Segment customers (Consumer, Corporate, Home Office) to understand behavior and purchasing trends.

To visualize the relationships between customer segments and product categories using a chord diagram, we begin by aggregating the data. Using D3's d3.rollup function, the sales values are grouped by Segment and Category, with the total sales calculated for each pair. This results in a nested structure that is then flattened into an array of objects, where each object represents a directional connection from a segment (source) to a category (target) with an associated sales value.

Next, a unique list of nodes is extracted from the source and target fields. These nodes represent the distinct entities (segments and categories) that will be shown as arcs around the chord diagram. A square matrix is then constructed to define the relationships between all node pairs. Each cell in this matrix represents the sales value flowing from one node to another. If no relationship exists between two nodes, the corresponding matrix entry is set to zero. This matrix structure is required by D3's chord layout generator.

The chord diagram is then created using D3's chord, arc, and ribbon generators. Arcs are drawn for each node around a circle, and ribbons connect the nodes to illustrate the magnitude of sales flowing between them. The width of each ribbon corresponds to the sales volume, and a color scale is used to differentiate the nodes. Labels and tooltips are added to enhance interpretability, showing detailed information on hover.

The chord diagram as shown in figure 35, visualizes the relationship between customer segments and product categories through arcs and ribbons. The arcs around the circle represent the different nodes. Customer segments, such as Corporate and Home Office, are positioned on the left, while product categories like Furniture, Office Supplies, and Technology are on the right. The width of each arc corresponds to the total sales associated with that segment or category, with wider arcs indicating higher sales volumes. For instance, the Consumer segment, shown as a blue arc, is the widest, signifying it contributes the most to overall sales. Inside the circle, curved ribbons illustrate the flow of sales from each segment to each product category. The thickness of these ribbons reflects the volume of sales between the connected nodes; the thicker the ribbon, the greater the sales. Additionally, ribbons are colored based on the target category, allowing

for easier visual tracking of how each segment's revenue is distributed across product categories.

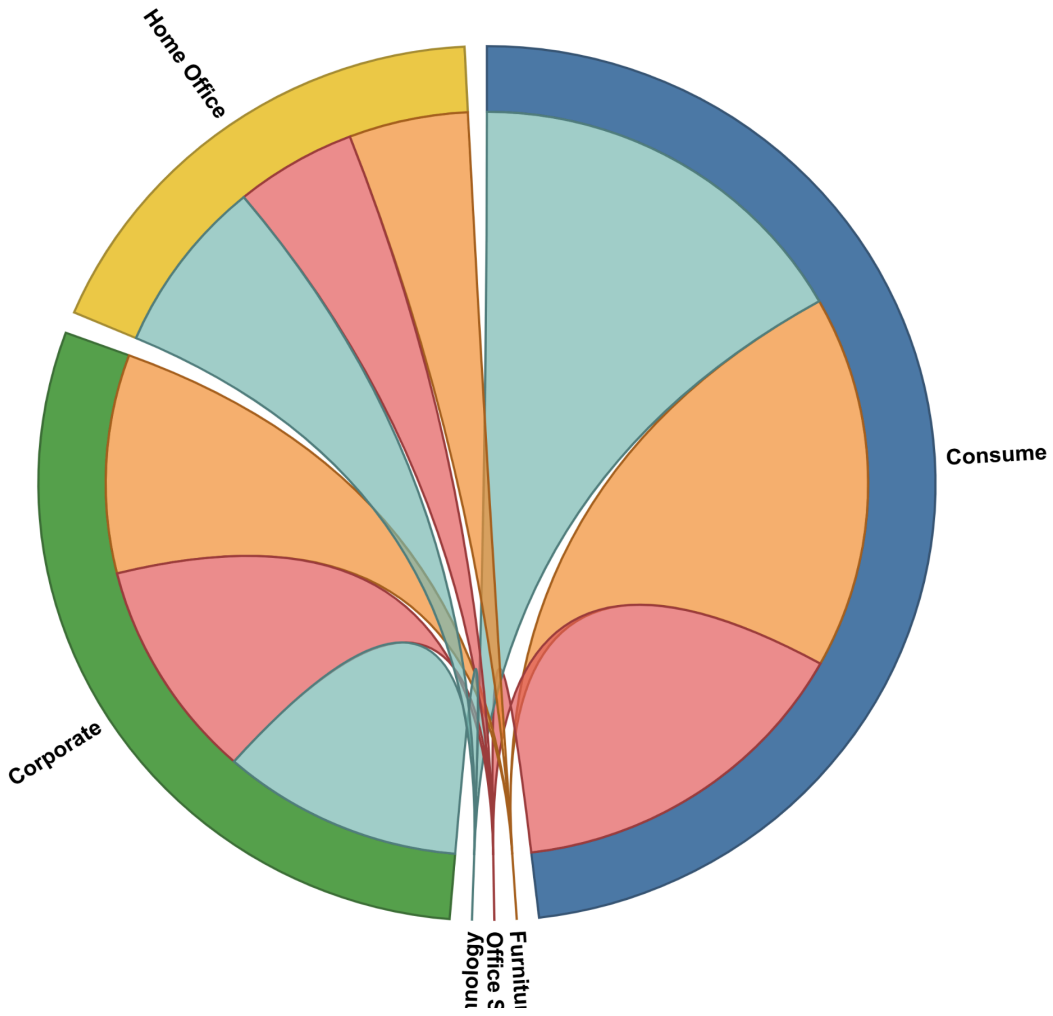

Figure 35: Chord diagram visualizing the relationship between customer segments and product categories through arcs and ribbons

The chord diagram effectively illustrates how sales are distributed from customer segments to product categories by combining both overall volumes and relational flows in a single visual. The Consumer segment contributes significantly to all three categories, Technology, Office Supplies, and Furniture, with a particularly strong emphasis on Technology, as indicated by both the wide arc and thick ribbons leading to that category. The Corporate segment also shows strong contributions across all categories, though at a slightly lower level than Consumer,

while Home Office has the smallest arc, reflecting its relatively lower total sales. On the category side, Technology receives the largest share of sales across all segments, followed by Office Supplies and then Furniture. The widths of the arcs represent total sales per segment or category, while the curved ribbons between them visually convey how those sales are shared. This makes the chord diagram an intuitive and powerful tool for understanding not just the size of each group, but also the distribution and strength of connections between segments and product categories.

The **chord diagram** was chosen to illustrate this data because it is particularly effective for visualizing interrelationships and flow between two categorical dimensions, in this case, customer segments and product categories. Unlike standard bar or pie charts that show totals independently, the chord diagram shows both the magnitude of sales per group and the directional flow between groups in a single, compact visual. This allows for immediate insights into which segments are driving sales in which categories, and how strongly each segment contributes to different product lines. The arcs provide a sense of total volume per node, while the ribbons emphasize the connections and proportions of those totals, enabling a more intuitive understanding of how sales are distributed and shared across the business.

We have used two visualizations, a stacked area chart and a stacked barchart, created in Power BI and Observable HQ respectively, to analyze purchasing trends over time by segment. These charts were chosen because they complement each other and reveal different aspects of sales behavior.

The stacked bar chart in figure 36, illustrates quarterly sales trends from 2014 to 2017, with each bar representing total sales for a specific quarter. Within each bar, the sales are divided by customer segments, Consumer, Corporate, and Home Office, using different colors. This allows for easy comparison of how each segment contributes to total sales across different quarters. The chart highlights seasonal patterns, such as consistently higher sales in Q4, and shows that the Consumer segment contributes the largest share overall.

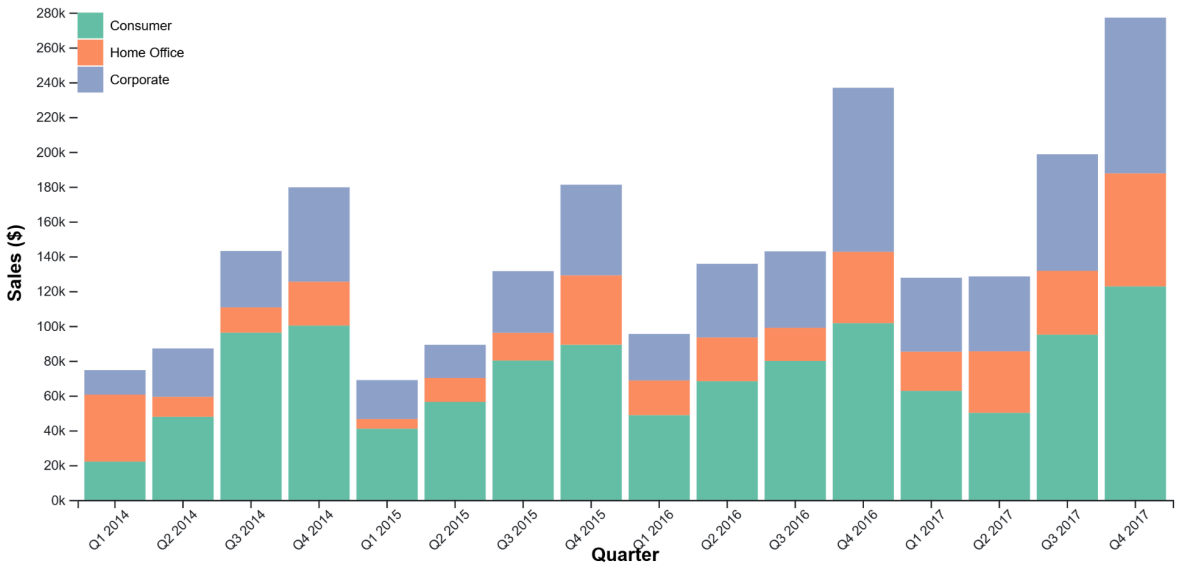

Figure 36: Stacked bar chart illustrates quarterly sales trends from 2014 to 2017

To ensure that the quarters are displayed in the correct chronological order, a custom sort column was created. This new column, often named something like "Order_Sort," was calculated by combining the year and quarter as a numeric value (e.g., 201401 for Q1 2014). Power BI was then instructed to sort the "Order_Quarter" labels using this numeric column. This step is crucial because without it, Power BI would sort the quarters alphabetically (e.g., Q1 2014 before Q2 2013), leading to an inaccurate time sequence in the visualization.

This **stacked bar chart** breaks down sales by individual quarters, giving a clearer picture of quarterly differences in absolute sales values. It's especially useful for identifying spikes or dips in specific quarters and observing seasonal patterns. The stacked format allows side-by-side comparison of how each segment performs within each quarter. For eg, Sales typically peak in Q4 each year, and Q4 2017 shows the highest overall sales, with all three segments contributing substantially, especially the Consumer segment.

The stacked area chart in figure 37 displays quarterly sales trends over time from Q1 2014 to Q4 2017, with sales divided by customer segments: Consumer (blue), Corporate (orange), and Home Office (gray). Each area represents the sales

contribution of a segment, and the combined height of all areas at each point represents the total sales in that quarter.

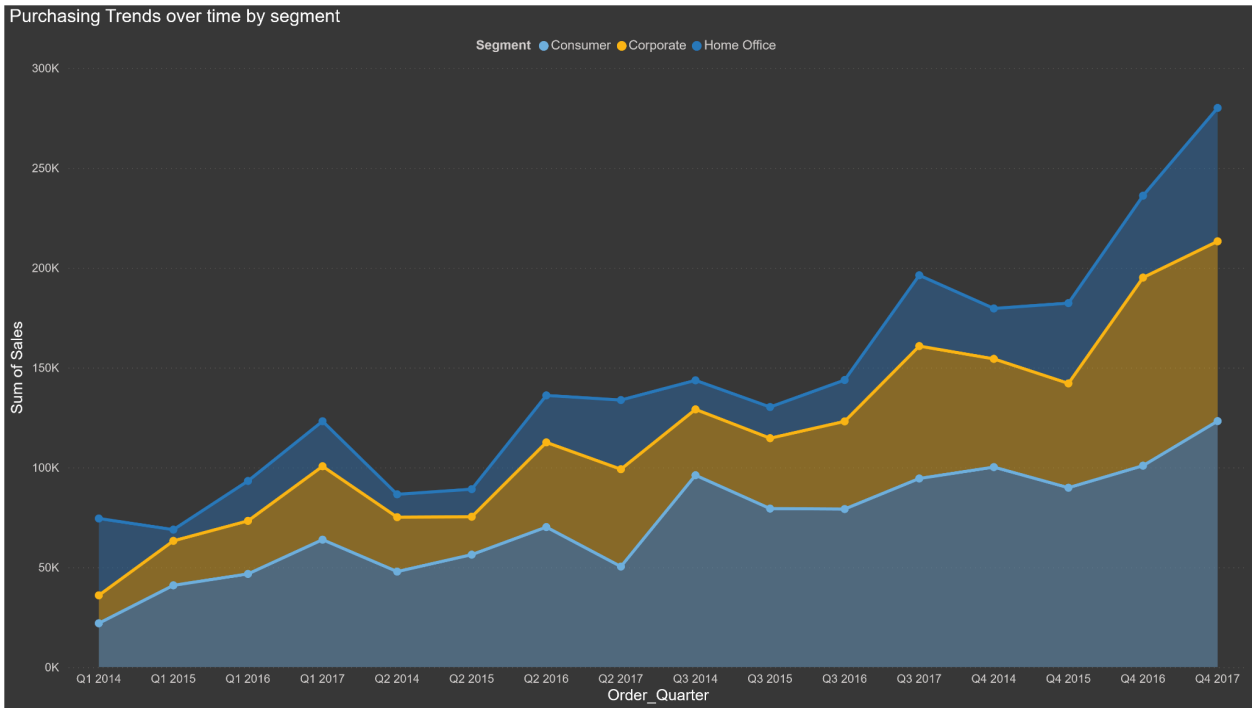

Figure 37: Stacked area chart displaying quarterly sales trends over time from Q1 2014 to Q4 2017, divided by customer segments

The chart shows a clear upward trend in total sales over the four-year period. The Consumer segment consistently contributes the largest share of sales, evident from the prominent blue area at the top. Corporate and Home Office also show steady growth, with Corporate gradually increasing its share over time. Sales peaks are visible in several quarters, especially Q4 periods, indicating seasonal spikes. The visual helps track both the growth in overall sales and how each segment's share changes over time.

This **stacked area chart** is excellent for showing the evolution of sales trends over time. It reveals how the total sales volume grows or changes across quarters and allows you to compare the relative contributions of each segment, Consumer, Corporate, and Home Office, at every point in time. Because the areas are stacked, it also provides a clear view of the overall upward or downward trend, while still highlighting how much each segment contributes to the whole. For eg, in Q3 2014,

the Consumer segment had significantly higher sales than Corporate and Home Office, and this pattern of dominance continues over time with visible fluctuations.

Improvement Plans to Deepen Customer Segmentation and Optimize Engagement

1. Launch targeted promotions for Consumer segment on high-performing categories like Technology.
2. Tailor marketing strategies for Corporate and Home Office to boost engagement in underrepresented categories.
3. Use chord diagram insights to refine cross-selling opportunities based on strong segment–category links. Analyze seasonal patterns from stacked visuals to optimize campaign timing (e.g., Q4 peaks).
4. Monitor segment-specific growth and adjust inventory planning accordingly. Enhance personalization in sales outreach using segment-specific trend data.
5. Continue tracking quarterly performance with updated bar and area charts.

Objective 6: Examine regional sales patterns across U.S. states and regions to uncover geographic insights.

The **choropleth map** as shown in figure 38, presents a clear visualization of sales distribution across the United States by state. The data reveals significant regional disparities, with California leading by a substantial margin at $457,688 in sales, followed by Texas and New York, which record $170,188 and $310,876 respectively. These three states stand out as major sales hubs, as indicated by the darkest blue shading, demonstrating strong market demand and possibly more established sales infrastructure. Other states with relatively high sales figures include Washington ($138,641), Pennsylvania ($116,512), Florida ($89,474), Michigan ($76,270), Illinois ($80,166), and Ohio ($78,258), all showing moderate to strong performance compared to the national landscape. In contrast, a considerable number of states reflect much lower sales volumes, as indicated by the lighter shades of blue or near-white color. States like North Dakota ($920), South Dakota ($1,316), Kansas ($2,914), West Virginia ($1,210), and Maine ($1,271) exhibit very low sales numbers, suggesting under-penetrated markets or less effective sales outreach. Additionally, Alaska and Hawaii report zero sales, which may point to distribution challenges or lower market prioritization. Several

other states, primarily in the central and northern regions, fall into the low-to-moderate sales range, indicating room for growth and increased market engagement.

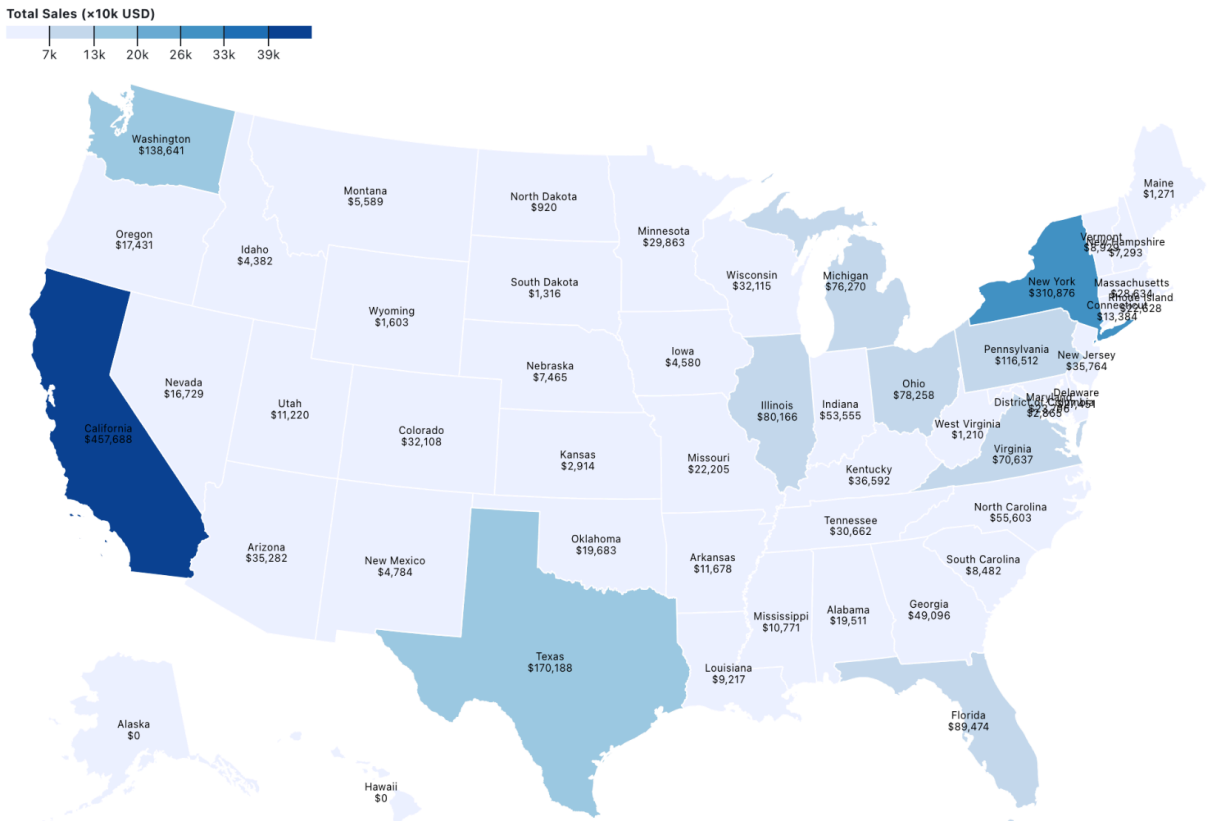

Figure 38: Choropleth map representing sales distribution across the United States by state

This uneven sales distribution points to potential strategic opportunities for growth by focusing efforts on the underperforming states. Expanding distribution channels in these regions could enhance product availability, while targeted marketing campaigns tailored to local preferences might increase customer awareness and engagement. Conducting market research in low-sales states can also help uncover specific barriers such as product fit, pricing sensitivity, or competitive dynamics that limit sales. Furthermore, incentivizing and empowering local sales teams with resources and rewards could drive improved performance. Engagement with communities through sponsorships and partnerships could also foster brand loyalty in these lesser-performing markets.

Overall, the map's visual insights emphasize the importance of a regionally differentiated strategy to optimize sales growth nationwide. While major states like California and New York remain critical to overall revenue, there is considerable untapped potential in many other states that, if addressed strategically, can contribute significantly to long-term business expansion and market share growth. By prioritizing investment and resources in these lower-performing areas, companies can build a more balanced and resilient sales network across the United States.

The **treemap** provided is a graphical representation of data distributed across U.S. states, where each state is depicted as a rectangle. The size of each rectangle corresponds to a numerical value, likely representing financial figures such as revenue, expenditure, or investment. This method of visualization allows for an immediate understanding of how values compare across different categories, in this case, states.

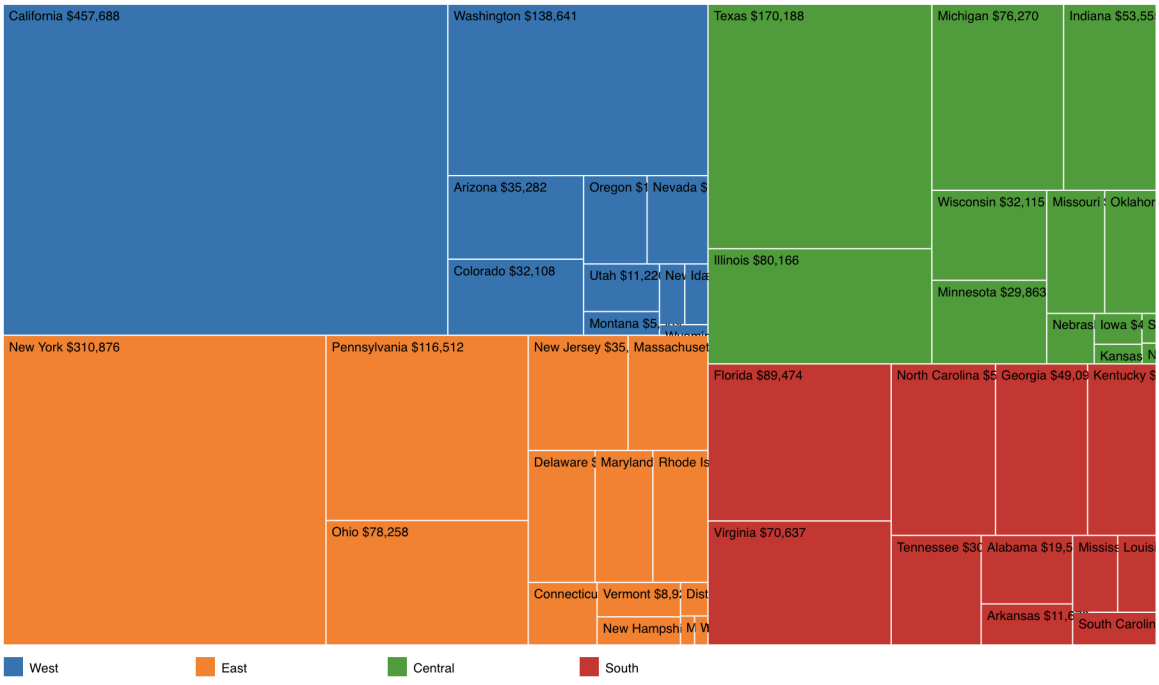

Figure 39: Treemap of Sales Performance by U.S. State and Region

From the treemap shown in figure 39, it's evident that California holds the largest value at $457,688, making it the most dominant state in this dataset. New York

follows with $310,876, and Texas with $170,188, indicating that these states are either contributing or receiving significantly more than others. Washington and Pennsylvania also show substantial values, suggesting their importance in the overall distribution.

In contrast, Midwestern states like Indiana ($53,555), Ohio ($78,258), and Illinois ($80,166) have moderate values, while states such as Wisconsin ($32,115) and Minnesota ($29,863) are on the lower end. Some states have values that are partially obscured or too small to read, which may indicate either lower figures or limitations in the visualization's resolution.

Overall, the treemap effectively highlights disparities and concentrations in the data, making it a useful tool for identifying regional trends and outliers.

Improvement Plan for Increasing Sales Across Underperforming States

1. Market Research: Analyze low-sales states (e.g., ND, SD, KS, WV, ME) to identify barriers like customer preferences, pricing, competition, and product fit.
2. Distribution & Accessibility: Expand reach in under-penetrated markets through local partnerships and assess warehousing options to improve availability and delivery speed.
3. Targeted Marketing: Launch region-specific campaigns using local messaging, channels, and geo-targeted digital ads; optimize with data-driven insights.
4. Sales Enablement: Support local sales teams with training, tools, and incentives; hire regionally knowledgeable talent to boost effectiveness.
5. Community Engagement: Build brand trust via local events, sponsorships, and influencer partnerships; respond actively to regional feedback.
6. Performance Monitoring: Track sales by state with KPIs and dashboards; adapt strategies quickly and review progress through quarterly regional reviews.

DISCUSSION

The analysis showcases a deeply data-driven understanding of U.S. sales performance across product categories, regions, and customer segments. Notably, high-performing products, largely in Technology and Office Equipment (e.g. copiers, printers), consistently generate strong sales and profitability, whereas several Office Supplies sub-categories (e.g. Binders, Appliances) and furniture items (e.g. bookcases, tables) demonstrate negative or marginal margins. This aligns with Pareto rule findings, where a small share of SKUs drives most profit (Han, Pei, & Kamber, 2011).

Discount strategy emerges as a critical lever: bubble-chart and violin-plot visualizations clearly show that aggressive discounts, especially in Technology and Office Supplies, erode margins, occasionally resulting in negative profitability. In contrast, sub-categories like Furnishings and Labels with lower discount rates yield healthy margins, suggesting an opportunity to segment pricing and tailor promotions more carefully.

Operationally, shipping delay analysis reveals notable efficiency challenges, particularly in the Central region's Office Supplies category where average delays exceed 4 days. While most regions maintained relatively uniform performance, the Central–Office Supplies combination stands out as a potential logistics bottleneck. Contour and heatmap plots illustrate how delays shift over time and across categories, revealing seasonal spikes especially in Central and West regions.

Segment-wise analysis using chord diagrams and stacked visualizations uncovers that the Consumer segment contributes the most in terms of volume, especially in Technology, while Corporate and Home Office segments, though smaller, generate meaningful revenue and show different product affinities. This segmentation allows strategic targeting via promotions or cross-selling, with chord flows highlighting ideal category alignment.

Geographically, sales concentration is heavily skewed: California, New York, and Texas dominate revenue, while many central and northern states underperform. Choropleth and treemap visuals highlight this imbalance, suggesting untapped

potential in under-penetrated markets, particularly those with near-zero sales (e.g., ND, SD, WV, ME).

These insights collectively point toward several business-critical opportunities: optimizing discount policies, pruning low-profit SKUs, improving logistics operations in specific regions, enhancing customer segmentation strategies, and expanding geographic reach. The multi-visual, interactive analytic framework also proves effective: scatterplots, contour maps, chord diagrams, and heatmaps work together to reveal complex patterns that wouldn't be apparent in any single visualization.

LIMITATIONS

While the analysis provides actionable insights, several limitations should be acknowledged. First, the dataset lacks detailed customer demographic and behavioral variables, which restricts deeper segmentation beyond broad categories like "Consumer" or "Corporate." Second, the analysis does not include data on returns, refunds, or customer satisfaction, factors that could influence profitability and operational metrics. Third, geographic insights are limited to state-level granularity; more localized data (e.g., city or ZIP code level) could enhance market targeting strategies. Lastly, the analysis assumes clean and accurate transaction logging; any underlying data entry errors or missing timestamps (e.g., for shipping delays) may introduce bias into conclusions. Future analyses could integrate external datasets (e.g., competitor pricing, macroeconomic indicators) and temporal comparisons across multiple years for a more robust, context-rich understanding.

CONCLUSION

This study illustrates how a structured, multi-dimensional exploration of transactional retail data can yield actionable insights across product performance, pricing strategy, logistics, segmentation, and regional deployment. By combining statistical summaries with advanced visualizations, the analysis provides clarity on which products and segments drive profitability and where inefficiencies lie.

Key conclusions include:

1. **Product-level clarity**: High performers are mainly high-tech office machines and copiers, while certain accessories and general supplies underperform or carry low margins.
2. **Discount impact**: High discount rates correlate strongly with margin erosion, especially where demand is steady but profits are slim (e.g. Binders, Machines).
3. **Operational disparities**: Shipping delays, particularly in the Central region and within certain categories, suggest process inefficiencies affecting fulfillment performance.
4. **Customer segmentation**: The Consumer segment dominates in volume, while Corporate and Home Office segments offer opportunities for targeted expansions.
5. **Geographic imbalance**: Market concentration in a few states reveals both strength and opportunity—the latter in under-served markets.

This project not only aimed to derive insights from data but also to communicate those insights effectively through a wide array of static and interactive visualizations (Yau, 2013), including bubble charts, contour plots, chord diagrams, sunburst charts, and heatmaps. These tools enabled the discovery of relationships between discounting and profit erosion, regional shipping delays, and product performance across customer segments (Wickham & Grolemund, 2016). For example, contour plots captured operational inefficiencies in logistics, while chord diagrams helped decode interdependent trends in customer behavior and regional sales.

Together, these findings provide a strong foundation for business leaders to refine pricing, product mix, and operational strategies. The integrated analytics approach—combining Tableau or Power BI dashboards with embedded D3 visualizations—delivers both high-level summaries and granular insights, enhancing decision-making agility(Few, 2009). These methods contributed to a robust, visually-driven dashboard and a detailed report that equips stakeholders to make more informed decisions in marketing, inventory planning, and fulfillment. Ultimately, this analysis serves as both a strategic guide and a proof of concept for how advanced data visualization can transform transactional data into meaningful business intelligence.

FUTURE WORK

To build upon the current analysis, several next steps can deepen insight (Sivarajah et al., 2017) and drive continuous improvement:

1. **Predictive Modeling:** Develop regression or machine-learning models to forecast sales and profit at SKU and state levels. Including discount as a predictor, along with temporal seasonality and regional indicators, would support dynamic pricing and inventory planning.
2. **A/B Testing of Pricing Strategies:** Implement controlled discount experiments (e.g. test vs control segments) to empirically measure the impact of different discount levels on both volume and profit, validating or refining the observed correlations.
3. **Customer Lifetime Value (CLV) Modeling:** Extend the dataset to transactional history and recency/frequency metrics to calculate CLV per segment. Inform CRM strategies—e.g. focus retention campaigns on high-value customers, bundle offers for mid-value consumers.
4. **Operational Root Cause Analysis:** Drill down into logistics: examine order processing, warehouse throughput, carrier performance, and regional supply routes to uncover root causes behind observed shipping delays. Correlate delay data with return rates and customer satisfaction for end-to-end insights.
5. **Market Expansion Assessment:** For states with low current sales, pair sales data with demographic, economic, and competitive market data to evaluate feasibility and design tailored entry strategies—especially via local partnerships or digital outreach.
6. **Real-Time Dashboard Implementation:** Transition static reports into live BI dashboards that auto-refresh with new orders. Include alerting for negative margins, unusually long shipping delays, or discount thresholds. Enable stakeholders to interactively filter by region, segment, and time.
7. **Sentiment and Returns Integration:** Incorporate customer review scores and return data to better understand quality issues tied to poor profit performance. Visualize these alongside product-level profit and discounts to spot correlations.

By continuing to integrate predictive analytics, experimentation, and operational data, the company could move from descriptive insights to strategic optimization, enabling smarter promotions, tighter logistics, and more profitable customer relationships in a dynamic retail environment.

REFERENCES

1. Waller, M. A., & Fawcett, S. E. (2013). *Data Science, Predictive Analytics, and Big Data: A Revolution that will Transform Supply Chain Design and Management*. Journal of Business Logistics, 34(2), 77–84. https://doi.org/10.1111/jbl.12010
2. Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). *Business Intelligence and Analytics: From Big Data to Big Impact*. MIS Quarterly, 36(4), 1165–1188. https://doi.org/10.2307/41703503
3. Hao, S., Wang, Y., & Yan, L. (2020). *Online Retailing and Consumer Satisfaction: The Impact of Logistics Performance*. Journal of Retailing and Consumer Services, 54, 102035. https://doi.org/10.1016/j.jretconser.2019.102035
4. Dholakia, R. R. (2012). *Retailing in the 21st Century: Current and Future Trends*. Springer.
5. Wedel, M., & Kamakura, W. A. (2000). *Market Segmentation: Conceptual and Methodological Foundations*. Springer.
6. Kizley Benedict, Superstore Visualization https://public.tableau.com/app/profile/kizley.benedict/viz/SupermarketAnalysis_17059336212200/KPIOverview
7. Superstore Sample Dataset – Tableau Public https://public.tableau.com/app/learn/sample-data
8. Build a Superstore Dashboard – Tableau Tutorial https://public.tableau.com/app/profile/mc85/viz/example_store/Categoryistogram
9. Tableau Public Gallery – Superstore Dashboards https://public.tableau.com/app/profile/p.padham/viz/SuperstoreDashboard_16709573699130/SuperstoreDashboard

10. Data School – How to Analyze Superstore Data in Tableauhttps://public.tableau.com/app/profile/mirandali/viz/AnalyzeSuperstore/Overview
11. Tufte, E. R. (2001). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
12. Knaflic, C. N. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Wiley.
13. Cleveland, W. S., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387), 531–554.
14. Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. Wiley.
15. Ware, C. (2012). *Information visualization: Perception for design* (3rd ed.). Morgan Kaufmann.
16. Wilke, C. O. (2019). *Fundamentals of data visualization: A primer on making informative and compelling figures*. O'Reilly Media.
17. Kirk, A. (2016). *Data visualisation: A handbook for data-driven design*. SAGE Publications.
18. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). *Circos: An information aesthetic for comparative genomics. Genome Research*, 19(9), 1639–1645. https://doi.org/10.1101/gr.092759.109
19. Few, S. (2009). *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press.
20. Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
21. Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 70*, 263–286.
22. Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.
23. Yau, N. (2013). *Data Points: Visualization That Means Something*. Wiley.

****