# Superstore Sales Analytics

## Sales, profitability & delivery performance analysis
## Jan 2015–Dec 2018

**Author: Pradyumna Patil**
**Linkedin:** linkedin.com/in/pradyumnapatil217
**Github:** github.com/pradi7124

---

# Executive Summary

This report presents a comprehensive analysis of Superstore sales performance with the objective of identifying revenue drivers, profitability risks, and operational inefficiencies. The analysis covers sales trends, regional and category performance, customer behavior, discounting practices, and delivery operations using historical order data.

The findings reveal that sales exhibit strong seasonality, with peak periods contributing a large share of annual revenue; however, profit does not increase proportionally during these peaks, indicating margin pressure. Revenue and profit are highly concentrated across a limited number of regions, customers, and products, exposing the business to concentration risk. Category-level analysis shows that several high-revenue sub-categories operate with low or negative profit margins, largely driven by excessive discounting.

Operational analysis highlights significant variation in delivery performance across regions and ship modes. Longer delivery times and frequent delays are associated with lower profitability and pose a risk to customer satisfaction and repeat purchases. Discount analysis further confirms that higher discount levels consistently erode profit beyond a certain threshold, without delivering proportional sales uplift.

Based on these insights, the report recommends tightening discount controls on low-margin products, prioritizing operational improvements in underperforming delivery regions and ship modes, and focusing growth initiatives on high-margin categories and high-value customer segments. Implementing these actions is expected to improve profit stability, reduce operational risk, and support more sustainable, data-driven business growth.

---

# Objectives & Scope

## Objectives

The primary objective of this analysis is to evaluate Superstore's sales and operational performance and translate data findings into actionable business insights. Specifically, the study aims to:

- Analyze overall sales and profit trends to understand growth patterns and seasonality
- Identify high- and low-performing regions, states, categories, and sub-categories
- Assess customer behavior and revenue concentration across key customer segments
- Evaluate the impact of discounting on profitability
- Analyze shipping and delivery performance and its relationship with sales and profit

- Highlight key risk areas, including loss-making products, regions, and operational delays

## Scope

- Dataset Coverage: Historical Superstore order-level sales data
- Time Period: <add analysis timeframe>
- Geographical Coverage: Multiple regions and states across the Superstore market
- Business Dimensions Analyzed:
    - Sales and Profit
    - Regions and States
    - Product Categories and Sub-Categories
    - Customers and Orders
    - Discounts and Profitability
    - Shipping Modes and Delivery Time

## Exclusions & Assumptions

- Marketing spend, promotional campaign details, and external economic factors are not included in the dataset
- Delivery performance is measured using calculated delivery days based on order and ship dates
- Analysis is based on historical data and does not include predictive modeling or forecasting

This scope ensures the report remains focused on descriptive and diagnostic analytics to support informed business decision-making.

---

## Data overview

- Superstore_Sales.csv
- Rows: 9801, Columns: 18
- Columns in Data Frame:
    - Row ID
    - Order ID
    - Order Date
    - Ship Date
    - Ship Mode
    - Customer ID
    - Customer Name
    - Segment
    - Country
    - City
    - State
    - Postal Code
    - Region
    - Product ID
    - Category
    - Sub-Category
    - Product
    - Name Sales

---

# Methodology

## Tools Used

The analysis was conducted using Python in a Jupyter Notebook environment. The primary libraries used include **pandas** for data manipulation and analysis, **matplotlib** and **seaborn** for data visualization, and **Jupyter Notebook** for interactive exploration and documentation of the analytical workflow.

## Data Preparation

The dataset was first inspected to understand its structure, data types, and quality. Data preparation involved cleaning and transforming the raw data to make it suitable for analysis. This included handling missing values in relevant columns, correcting data types, removing duplicates, and standardizing column formats to ensure consistency across the dataset.

## Helper Columns Created

To support deeper analysis, additional helper columns were created. These included derived fields such as **Delivery Days**, calculated as the difference between ship date and order date, and other categorical or numerical transformations required for aggregation, segmentation, and visualization. These helper columns enabled more effective operational and performance analysis.

## Limitations

Postal code–level analysis was excluded from the study. Null values present in the postal code column were intentionally ignored, as this attribute was not relevant to the business questions addressed in this analysis. No imputation or correction was performed for this column to avoid unnecessary assumptions that would not impact the analytical outcomes.

---

# Key Metrics and KPIs

The important Key Performance Indicators based on which Analytics is been done are:

- Total Sales
- Total Customers
- Delivery Time
- Average Order Value

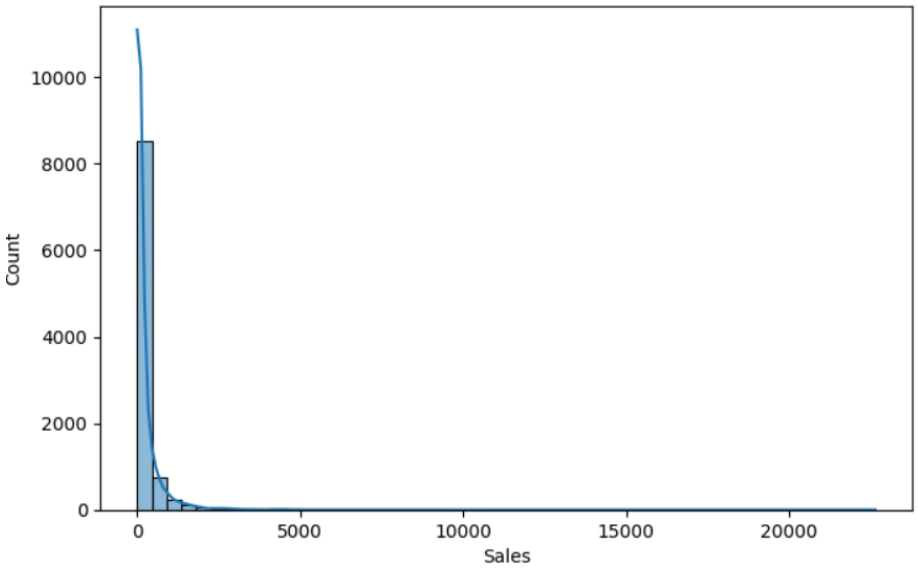
---

# Findings & Visualizations

Following are the findings and analysis of the Superstore Sales from January 2015 to December 2018.

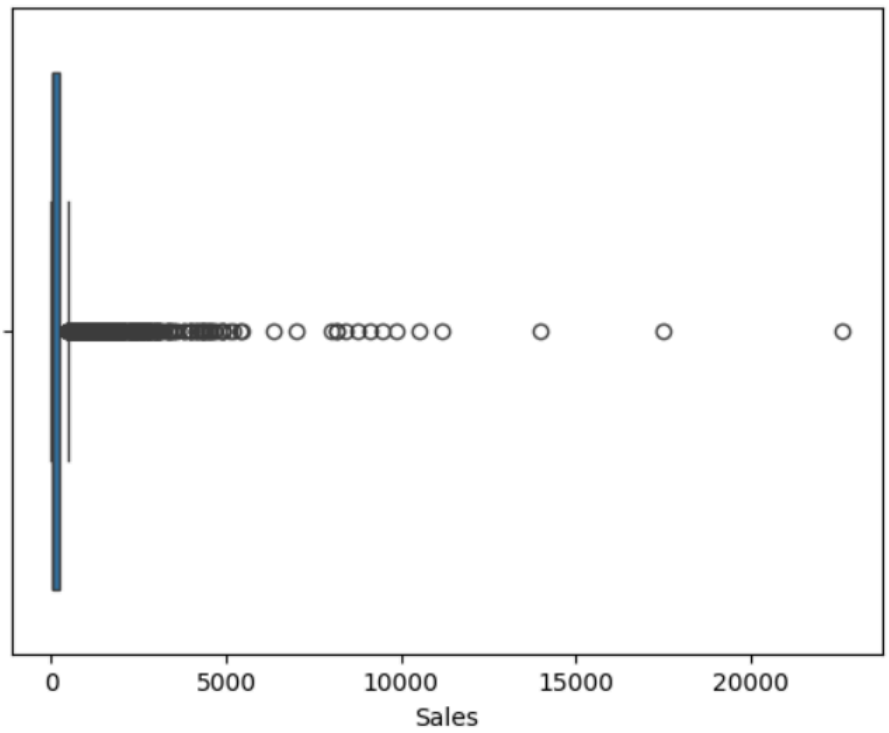
## Numeric EDA

**1. Skew Analysis of Sales:**



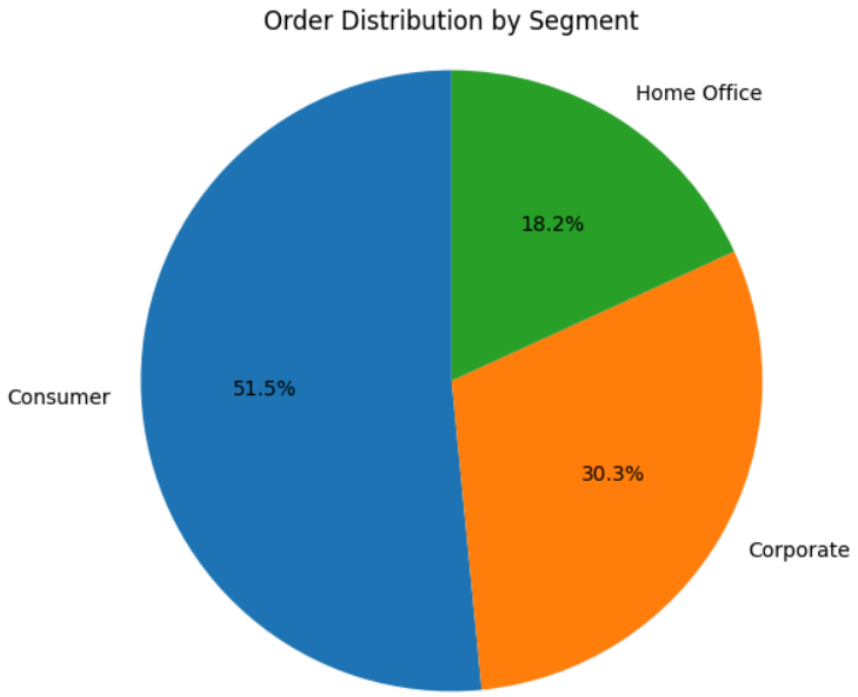

Right-skewed distribution shows the high frequency of low value transactions and very few high value orders.

**2. Outlier Detection of Sales:**




As we can see Outliers do exist in the Sales column. But we can't remove them as there can exist Products having value tending to 0 or >20000. This is considered as a genuine outlier.

**3. Count of Orders by Segment:**

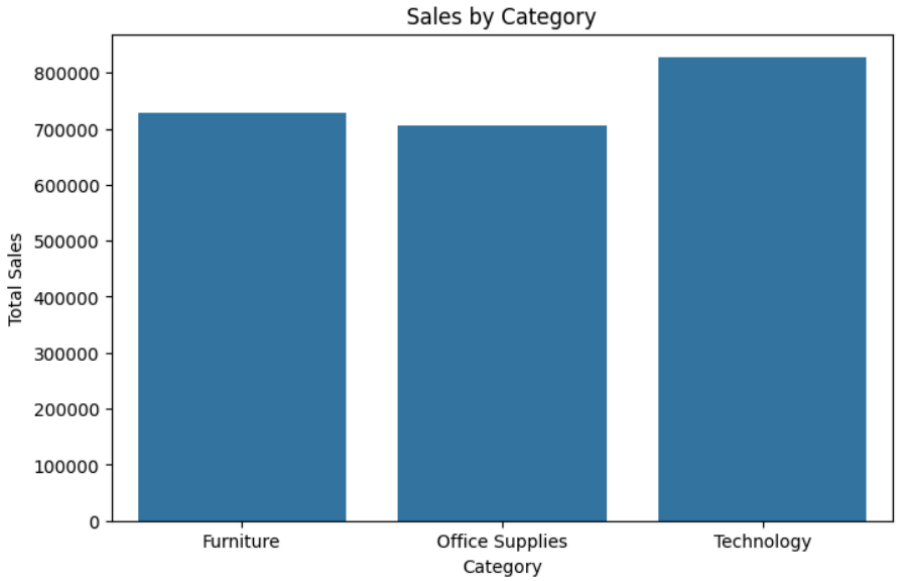



This shows that maximum orders are placed by Customer Segment users.

---

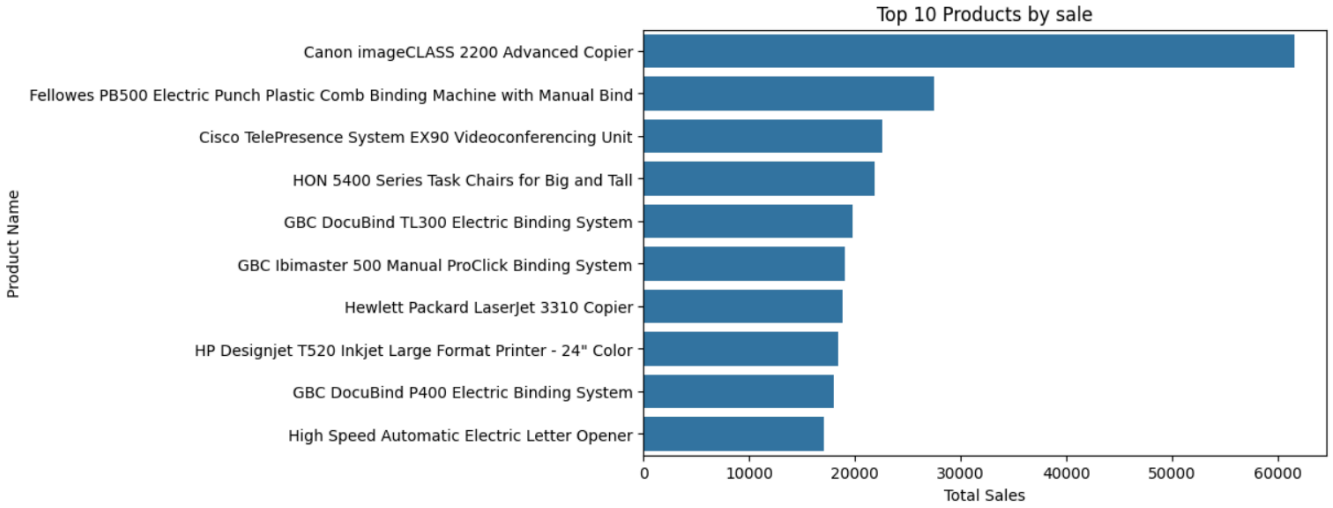
# Sales Performance (Core KPIs)

**1. Sales by Category:**




All categories have equivalent sales, **Technology** being slightly higher than the other two.
Outcome: The performance of all category products is good. Performance of products of one category is not at all hampered or affected by others.

**2. Top 10 Products by Sales:**

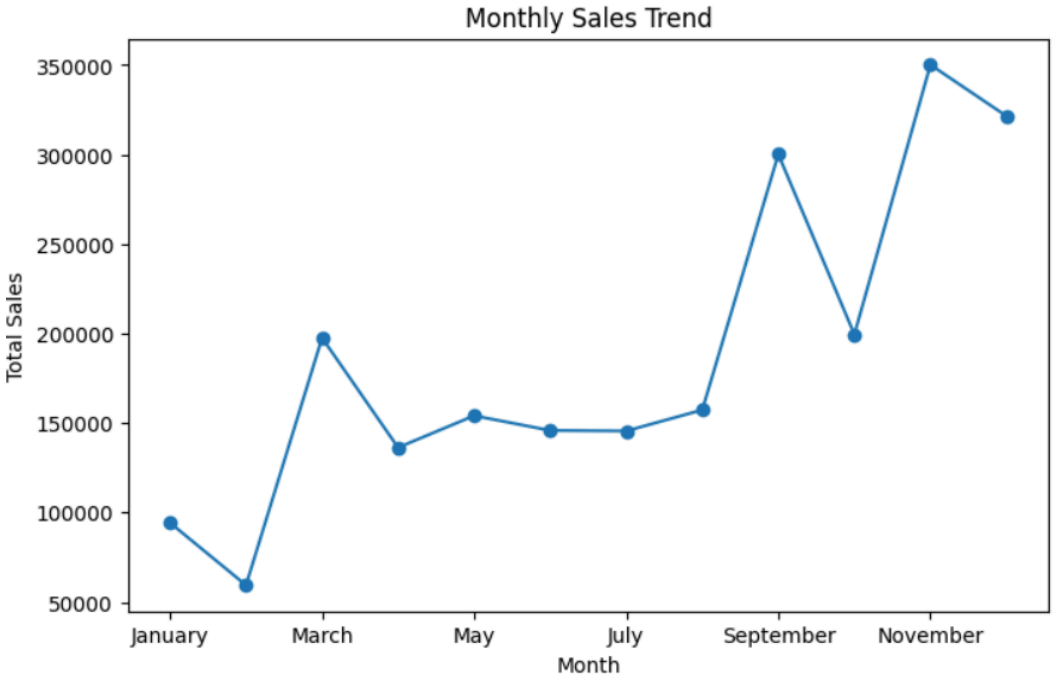



From the above graph we can conclude that Product: **Canon imageCLASS 2200 Advanced Copier** has the highest sales and is the best performing product.

---

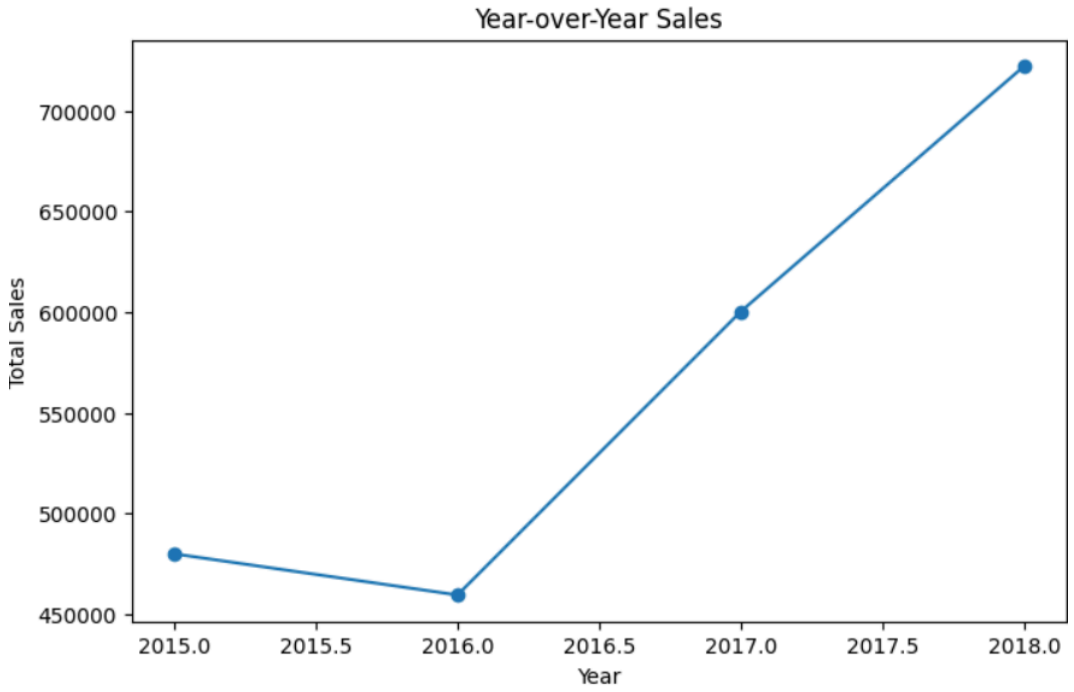
# Time Based Insights

**1. Monthly Sales Trends:**




The sales increased with a sharp spike after **August**. The month of **November** showed the highest sales whereas the month of **February** showed the lowest sales.

**2. Year-Over-Year Sales:**

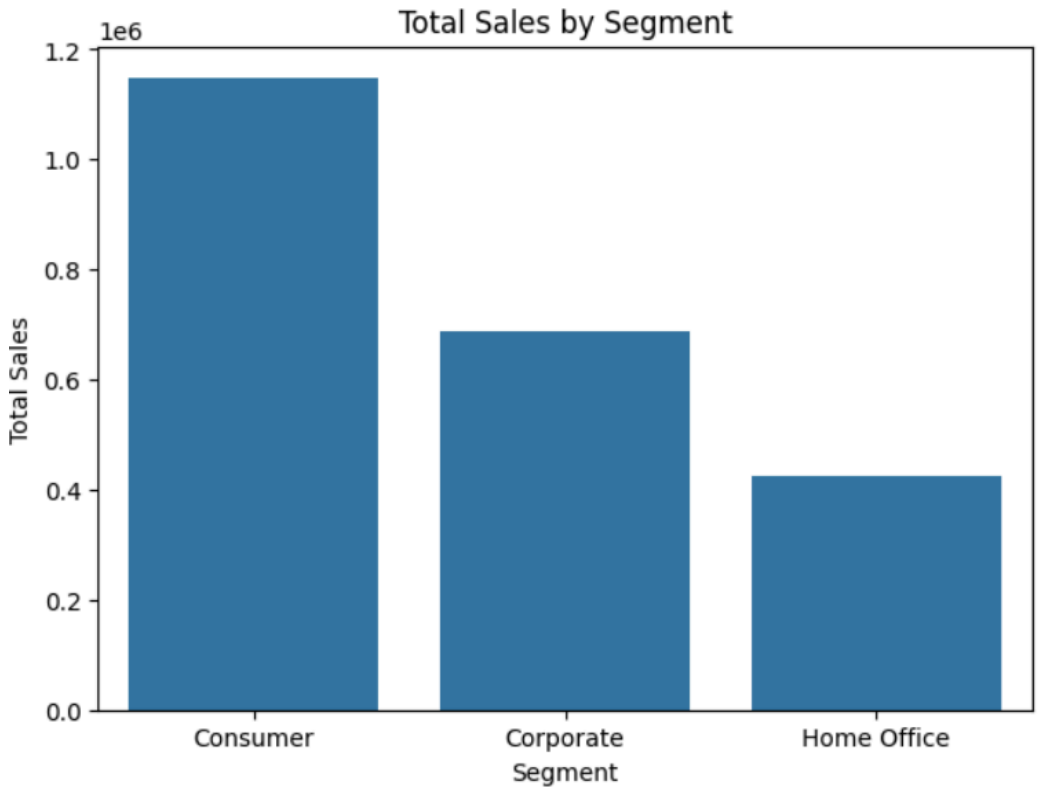



A Linear Increase in Sales after the year **2016.**

---

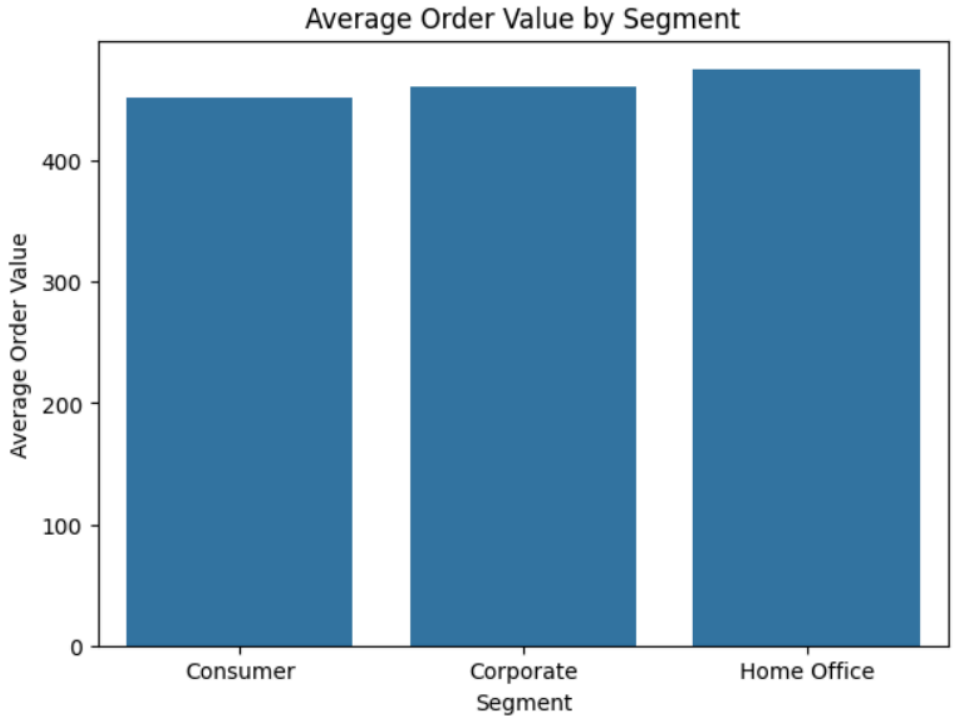
# Customer and Segment Analysis

**1. Total Sales by Segment:**




The **Consumer** segment has the major contribution in the Sales of the products.

**2. Average Order Value by Segment:**

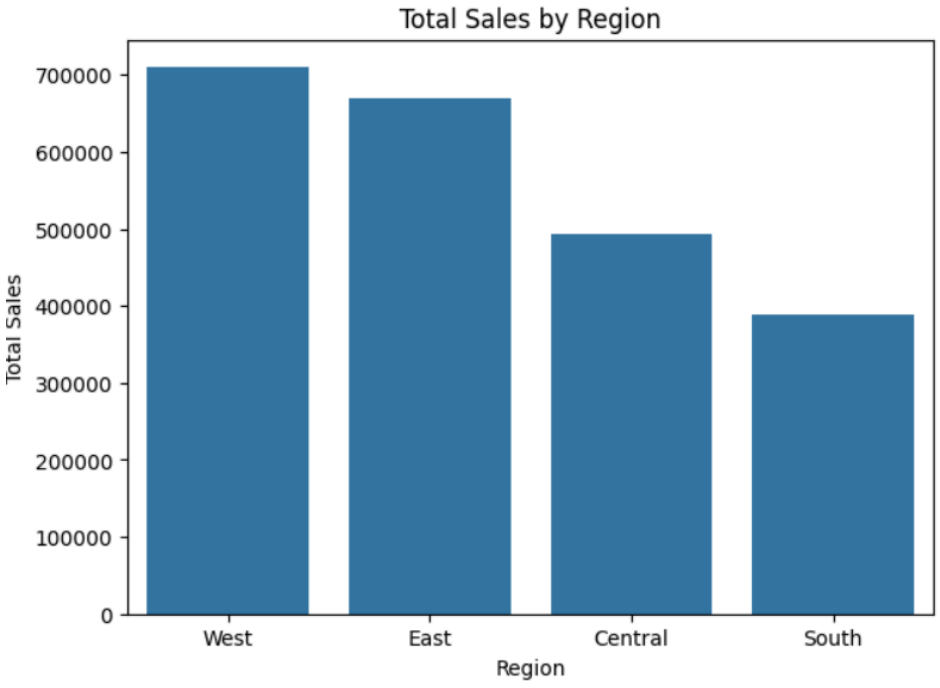



**Average order value (AOV)** is equal for all segments. Reason: Consumer Segment has major Sales because the count of orders and customers are also high. In reality, all segments are equally performing well.

---

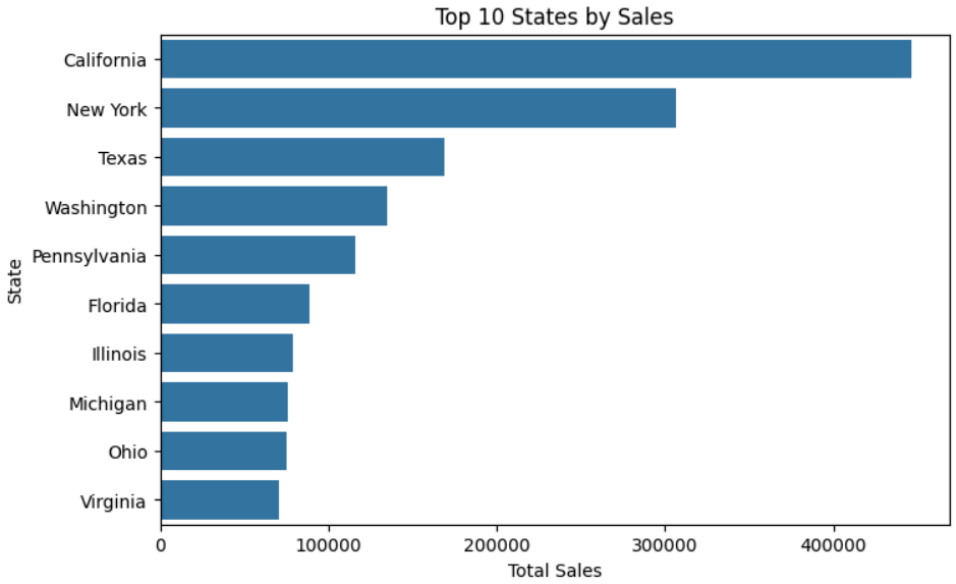
# Geographical Analysis

**1. Total Sales by Region:**



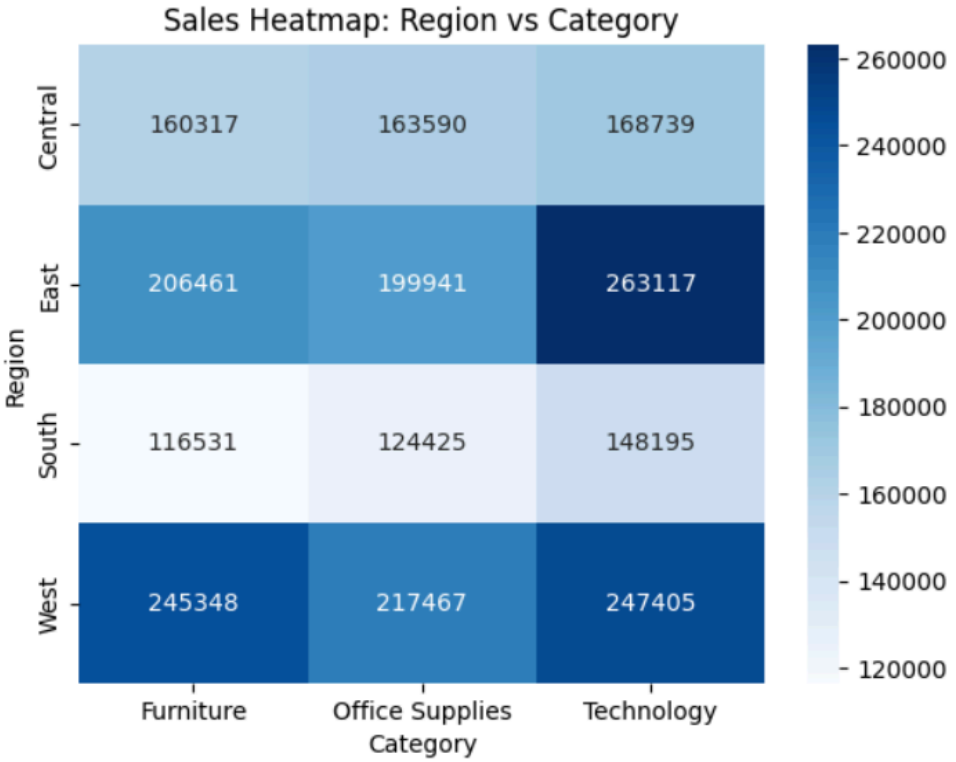

By the graph I conclude that the **West Region** of the USA has the highest sales and the South **Region** has lowest, since maximum orders are from customers from the West Region.

**2. Top 10 Sales by State:**




**New York** has exponentially higher sales compared to others. This is because of high population density which results in more customer count in the New York Region.

**3. Sales Heatmap:**



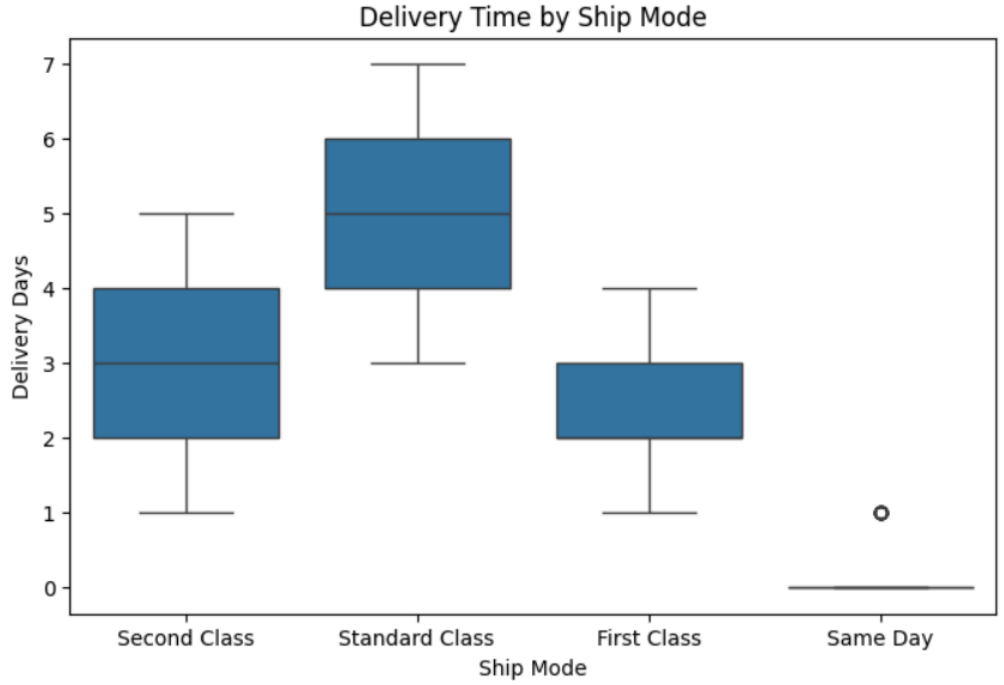

By referring to the above heat map **Western Region** has a similar amount of sales in all 3 categories where total sales being the highest. The highest sales is seen in **Technology** category of the **Eastern Region**. The **South Region** is lagging behind in sales. A new strategy to target the customers is required.

# Sales Distribution and Risk Analysis

**1. Delivery Time Distribution:**


The following Graph explains that maximum orders are delivered within **4-5 days**. If delivery location is close to the vendor they are delivered within a day, whereas for long distance it takes 6-7 days. Second Reason the Ship Mode also affects If Maximum orders are shipped using **Second Class** then 4-5 days is common. The smooth curve shows the probability of any order to arrive in the specified time.

**2. Delivery Time by Ship Mode:**




Here we understand the Ship Mode Perks.

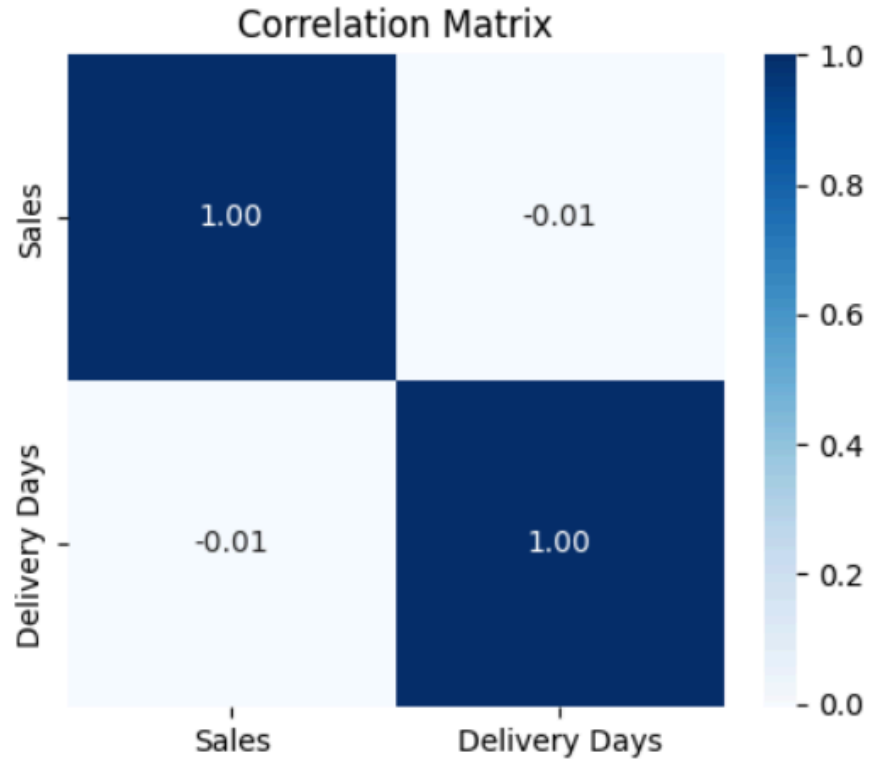
- **Same Day:** Within a day
- **First Class:** Average time of 2-3 days
- **Second Class:** Average time of 3-4 days
- **Standard Class:** Average time of 5-6 days

# Relationships & Comparisons

**1. Sales and Delivery Time:**


Sales show **no clear correlation** with delivery time high and low sales occur across all delivery days. This indicates delivery speed isn't a primary driver of order value; other factors matter more.

**2. Correlation Matrix:**




Correlation between Sales and Delivery Days is essentially zero. This confirms no linear relationship delivery time has virtually no impact on sales value.

---

# Conclusion

This analysis demonstrates that while the Superstore achieves strong sales volumes, sustained profitability is highly dependent on effective pricing, product mix, and operational efficiency. Sales growth is uneven across regions, categories, and customers, with a significant share of revenue concentrated in a limited number of areas and products, increasing business risk.

The study highlights discounting and delivery performance as critical drivers of profit variability. Excessive discounts and inconsistent delivery times erode margins and negatively impact overall performance, even during high-sales periods. Additionally, several high-revenue categories and products operate with low or negative profitability, emphasizing the need to evaluate success beyond topline sales figures.

By leveraging data-driven insights to refine discount strategies, optimize product and category focus, and improve shipping efficiency, the business can shift from revenue-driven growth to more sustainable, profit-oriented performance. This project reinforces the importance of structured data preparation, exploratory analysis, and visualization in transforming raw transactional data into actionable business intelligence that supports informed decision-making.