Superstore Retail Analytics – CRISP-DM Project Report

Prepared by: Ehsan Keikhavani

Contents

1. Business Understanding

Superstore is a multi-category retail business that sells Office Supplies, Furniture, and Technology products across several regions. Management is concerned that, despite strong sales, profit margins appear inconsistent across categories, locations, and customers. They also want to understand which customers are at risk of churn and how to focus marketing and operational efforts more effectively.

- Key business questions include:
- Which categories, sub-categories, and regions are most and least profitable?
- How seasonal are sales and when are the strongest and weakest months?
- Which customers drive the most value and which are at risk of churn?
- How can we segment customers to support targeted campaigns and retention strategies?

2. Data Understanding

The dataset contains transactional order-line records. Each row represents one product sold to a customer on a given date. Key fields include:

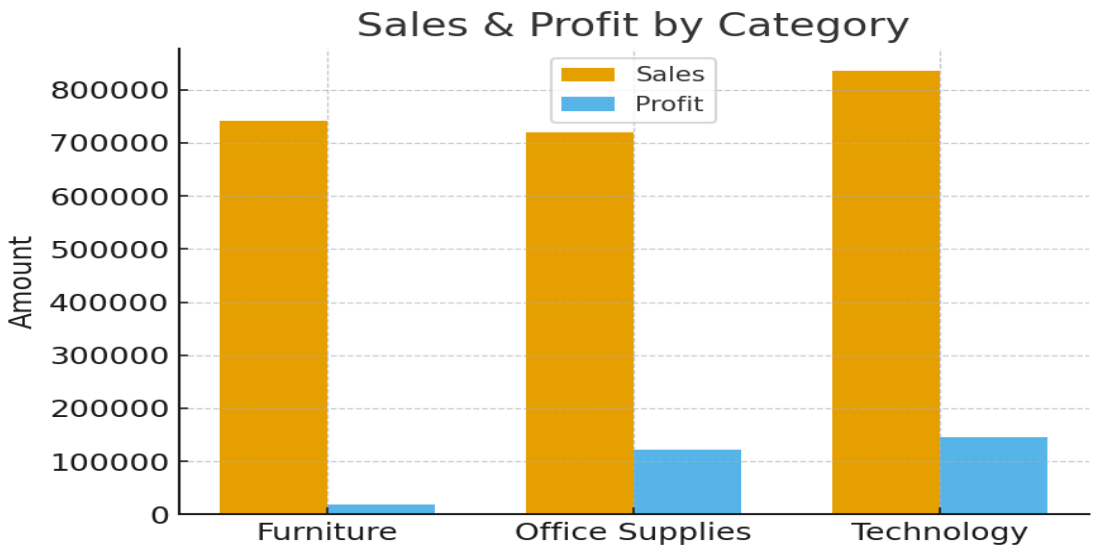
- **Order Date**
- **Customer ID**
- **Region, City**
- **Category, Sub-Category**
- **Sales, Quantity, Discount, Profit**

Initial checks confirmed that data types were appropriate (e.g. dates as datetime, numeric columns as floats/integers) and that there were no critical missing values that would prevent analysis.

To better understand business performance, exploratory data analysis (EDA) was carried out using aggregations and visualisations.

2.1 Sales and Profit by Category

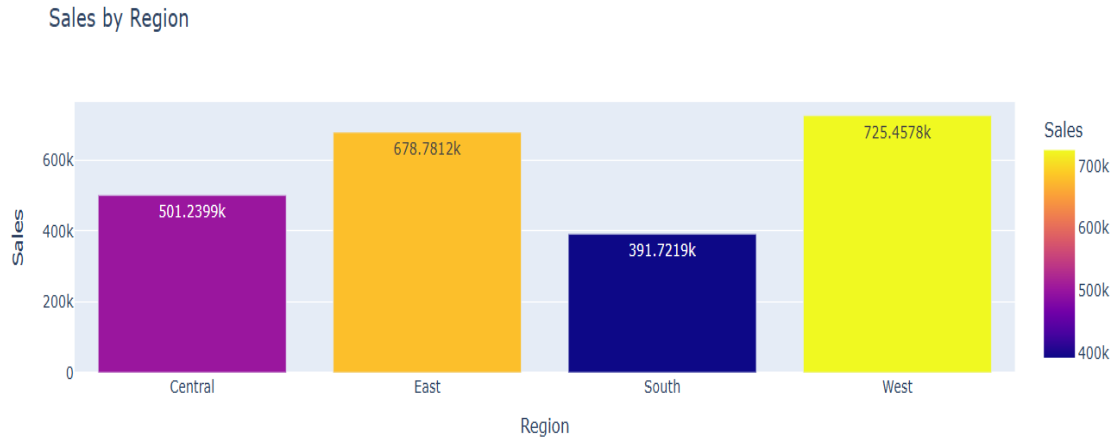
This chart compares total Sales and total Profit for each high-level product category. It shows where revenue is generated and whether that revenue translates into acceptable profit.



Technology and Office Supplies usually generate strong sales, while Furniture can be more volatile in terms of profit. This indicates that not all revenue growth automatically translates into profit and that margin management is important for certain categories.

2.2 Sales by Region

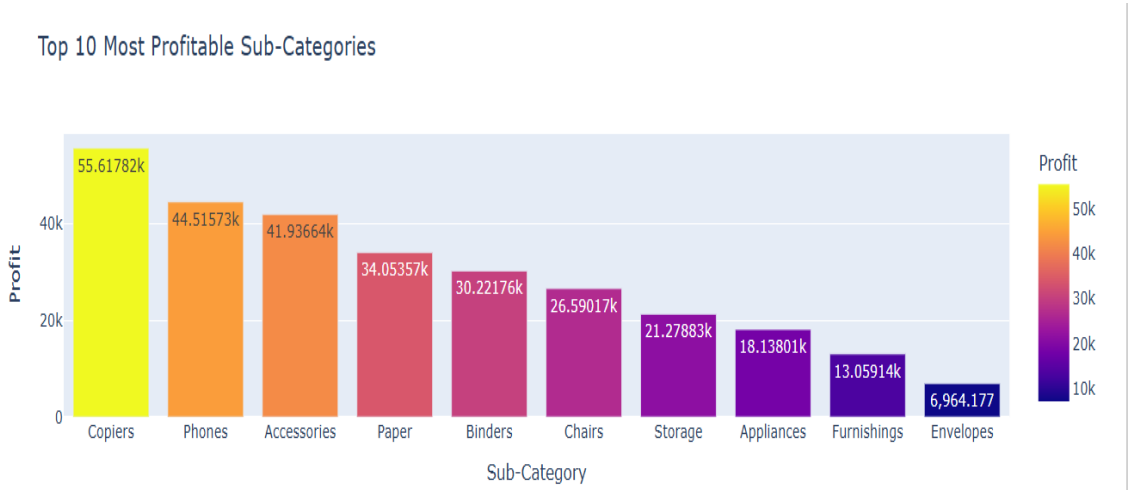
Total sales by region highlight which markets are most important to the business. Regions with significantly higher sales volumes indicate stronger market presence or demand.



The West and East regions contribute the most to total sales, while the South underperforms. Marketing and resource allocation should Favor high-performing regions, while deeper investigation is needed for the South.

2.3 Top 10 Most Profitable Sub-Categories

Ranking sub-categories by total profit identifies which specific product lines drive financial performance. These high-profit sub-categories can be prioritized for stock, promotions, and cross-selling opportunities.



Copiers, Phones, and Accessories are the highest drivers of profitability. Increasing ad spend, stock levels, and bundle promotions for these categories will maximize ROI.

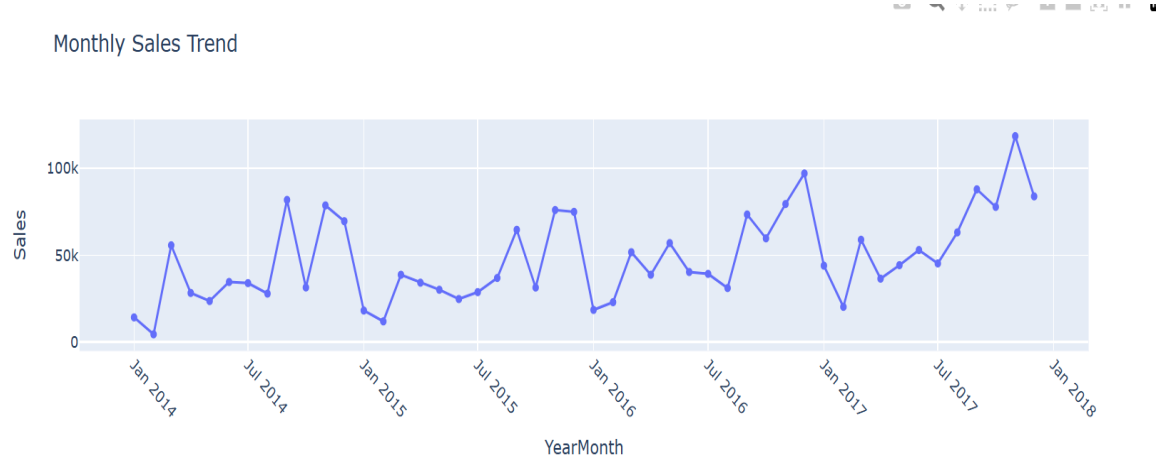
2.4 City-Level Profitability

Profit by city shows that profitability is not evenly distributed geographically. Some cities generate very high profit, while others operate at or below break-even. This supports more granular, city-specific strategy instead of a single national approach.

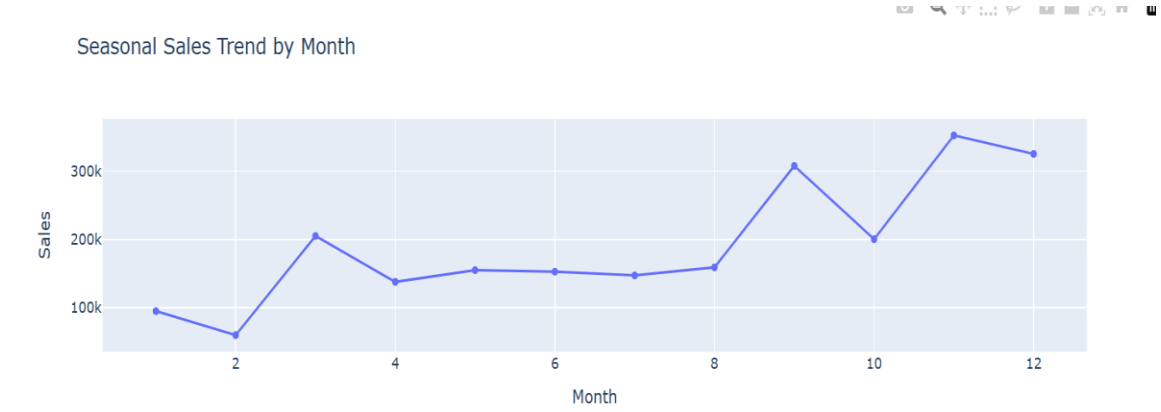
Cities like New York, Los Angeles, and Seattle drive the majority of profits, while Philadelphia, Houston, and San Antonio generate the largest losses. A pricing or discount review is required for loss-making cities.

2.5 Time and Seasonality Effects

Monthly sales trends and within-year seasonality provide insight into peak demand periods. Strong sales peaks towards the end of the year and weaker performance in early months suggest classic retail seasonality. Major peaks occur near the end of each year, suggesting holiday-season campaigns significantly boost revenue.



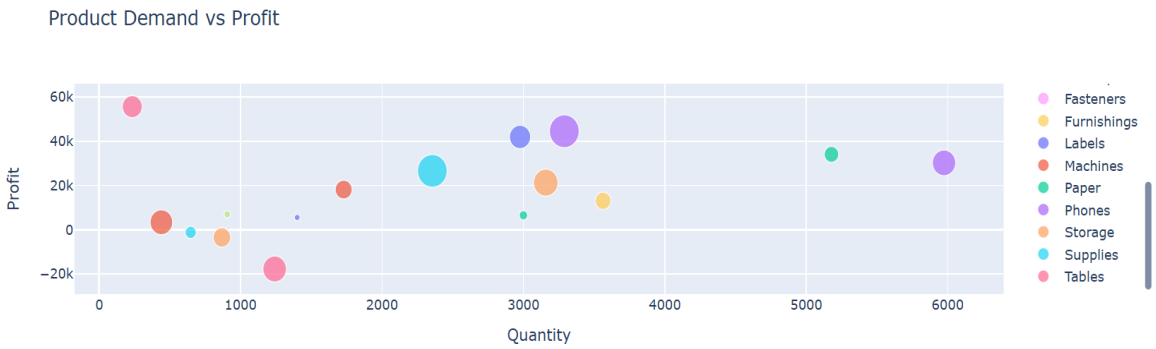
Result Interpretation:



The monthly trend typically shows clear peaks toward the end of each year and weaker performance in the early months. This is consistent with retail seasonality and highlights the need to prepare for strong demand in Q4 and manage costs more carefully during slower periods.

2.6 Product Demand versus Profit

Plotting total quantity sold against total profit (with bubble size proportional to sales) shows that some product groups generate high profit from relatively low volumes, while others sell in high quantity but deliver modest margins. This confirms that demand alone does not guarantee profitability.



Some high-demand products generate losses (e.g., Tables, Bookcases), indicating pricing or discounting issues. Copiers and Phones deliver strong profits despite lower volume and should remain priority products.

3. Data Preparation

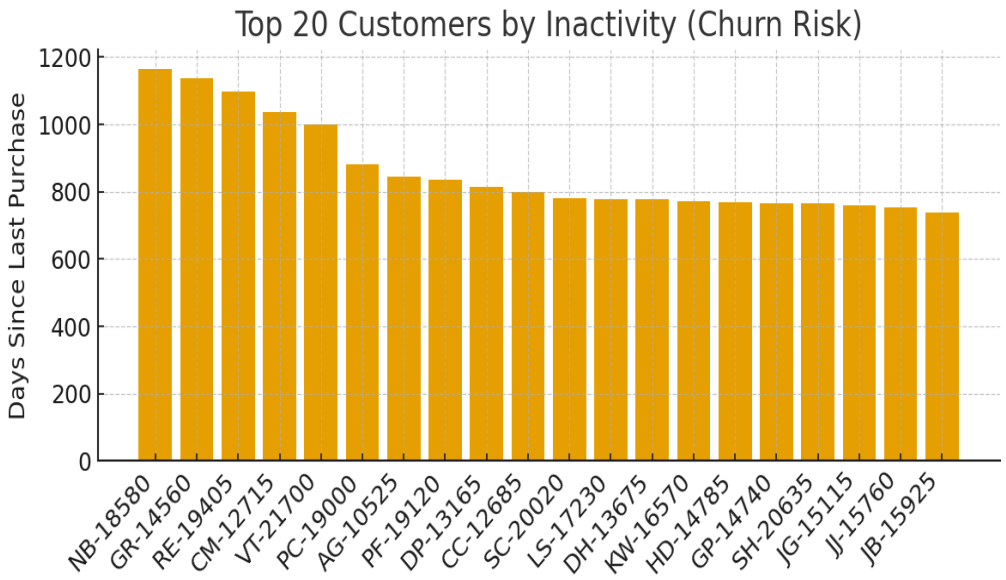
To move from transactional data to customer-level insights, the data was aggregated by Customer ID to create behavioral features similar to an RFM framework:

- Recency – days since the customer's most recent purchase (lower means more recent activity).
- Frequency – number of unique orders placed by the customer.
- Monetary – total revenue (Sales) generated by the customer.
- Avg Discount – average discount applied across that customer's orders.

Order dates were converted to datetime, and a snapshot date was defined as the day after the latest order in the data. Recency was then calculated as the difference between the snapshot date and the customer's last order date. These features provide a compact view of recency, loyalty, value, and discount sensitivity for each customer.

Churn Risk – Inactive Customers

Using Recency, the 20 customers with the longest time since their last purchase were identified as the highest churn risk. These customers are prime targets for win-back campaigns or personalized offers, especially if they were historically high-value customers. Recency is a simple but powerful metric to prioritise retention efforts.



4. Modelling – Customer Segmentation with K-Means

The Recency, Frequency, Monetary, and Avg Discount features were **standardised** so that each variable had mean 0 and standard deviation 1. This step ensures that all features contribute comparably when distance-based algorithms are used.

K-Means clustering was then applied to the scaled data to discover groups of customers with similar purchasing behaviour.

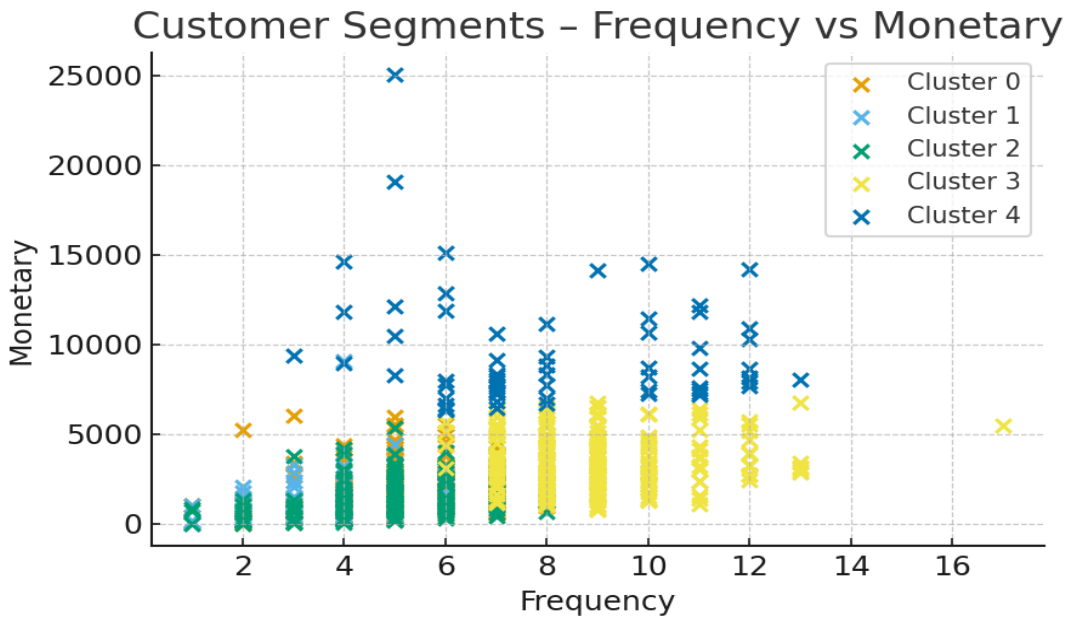
Different values of k (number of clusters) were evaluated using:

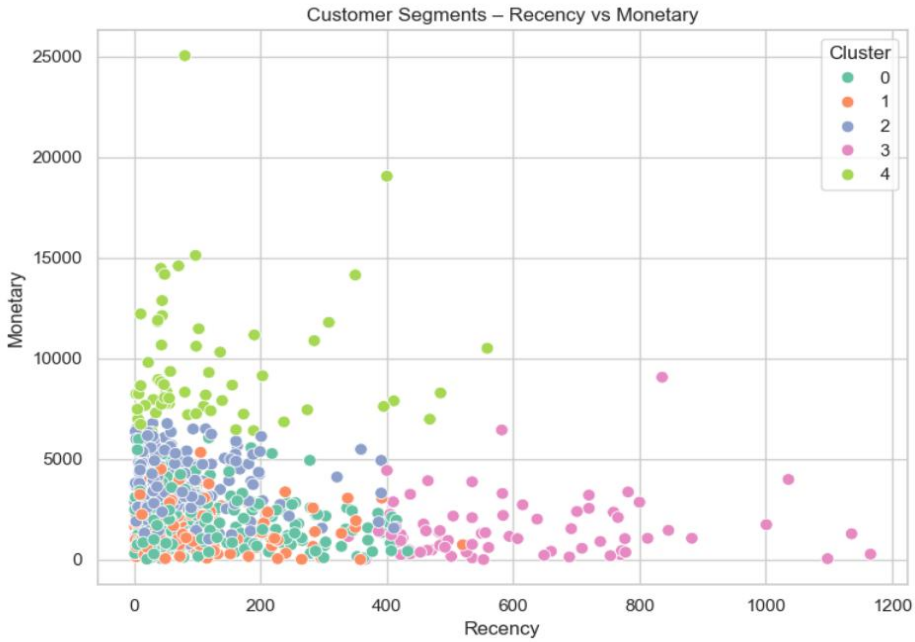
- **Elbow method** (inertia) – to see where adding more clusters stops giving big improvement
- **Silhouette score** – to measure how well-separated and compact the clusters are

A solution with a small number of clusters that still captured clear behavioural differences was selected so that the segments would be interpretable for business stakeholders (e.g. k = 5 in this case).

Visualizing Customer Segments

The scatter plot of Frequency versus Monetary, colored by cluster, shows distinct customer groups—for example, high-frequency high-spend customers, low-frequency low-spend customers, and intermediate groups. This visual check confirms that clusters capture meaningful behavioral differences.





This scatterplot compares customer recency (days since last purchase) with total spending. Each colour represents a customer cluster from the K-Means model.

- high-frequency, high-monetary customers (loyal, high-value segments)
- low-frequency, low-monetary customers (low-value or casual buyers)
- intermediate groups that sit between these extremes

This visual inspection confirms that clusters represent meaningful differences in behaviour.

Takeaway:
Customers who recently purchased tend to have higher monetary value. Customers with long recency and low spending may be at risk of churn and should be re-engaged with retention marketing.

5. Evaluation & Segment Interpretation

Cluster-level averages for Recency, Frequency, Monetary, and AvgDiscount were used to interpret the nature of each segment. For example:

- Segments with **low Recency, high Frequency and high Monetary** represent **loyal, high-value customers**.
- Segments with **high Recency, low Frequency and low Monetary** are more likely to be **at-risk or low-value customers**.
- Comparing **AvgDiscount** across clusters shows whether certain segments rely heavily on promotions.

In this dataset, high-value segments are not necessarily those receiving the deepest discounts, suggesting that loyalty is not driven solely by price cuts.

6. Deployment & Business Recommendations

The clustered customer dataset was merged back into the full transactional data and can be exported as a CSV file. This enriched dataset can be used directly in Tableau to create interactive dashboards that filter KPIs by segment, region, category, or time period.

Key recommended actions:

- **Prioritise high-frequency, high-monetary segments** with tailored retention and loyalty initiatives.
- **Target high-Recency customers** for win-back campaigns, especially those who were historically high spenders.
- **Review discount policies** for categories where sales are high but profit is weak, to protect margins without losing key customers.
- **Plan around seasonality**, ensuring stock, staffing, and marketing are aligned with peak and off-peak periods.

Overall, this project shows how applying the CRISP-DM framework to a retail dataset can turn raw transactions into customer-centric insights and practical recommendations for the business.