**Advanced Sales and Customer Insights on Superstore Dataset using R & Excel**

**Submitted by:**

Parasa Jahnavi

**Project Duration:**

July 17 – July 30, 2025

**Tools & Technologies Used:**

Microsoft Excel: Data cleaning, pivot analysis, basic insights.

R Programming: Data visualization (ggplot2), exploratory data analysis (EDA), clustering (k-means), time series analysis, basic statistical insights, and data cleaning (dplyr).

Packages in R: dplyr, ggplot2, lubridate, stats.

Independent Portfolio Project

**Project Overview & Objective**

Why I Chose This Project?

I chose the Superstore Sales dataset to apply my R and Excel skills in a realistic business scenario. My focus was on performing detailed exploratory data analysis, cleaning data, and building visual dashboards to uncover region-wise and category-wise trends. This reflects my technical proficiency and readiness for hands-on data analysis roles.

Though the dataset is widely used, my aim was to go beyond basic analysis and deliver:

Deeper storytelling

Advanced visualizations

Customer-focused insights

This project simulates everyday business challenges like managing sales, customers, regions, and profitability, making it the ideal choice to showcase both my skills and mindset.

**Project Goal**

The primary goal of this project was to analyze and clean sales data end-to-end using R and Excel, and extract actionable insights related to:

Variation of sales, profit, and discounts across regions, categories, and segments

Identifying loss-making areas and suggesting ways to reduce them

Exploring time-based patterns and seasonality in sales

Additionally, I wanted to simulate a real analyst workflow from organizing folders and cleaning messy Excel files to visualizing data and presenting findings in a professional report format.

**Key Questions I Aimed to Answer**

Which regions and product categories are most profitable?

Are discounts helping or hurting overall profits?

Do sales patterns change over months or seasons?

Can we identify customer segments through clustering?

What are the key issues a business manager should know from this dataset?

**Dataset Description**

1. Dataset Source

The dataset used for this project is the Sample Superstore Sales dataset a publicly available business dataset commonly used for data analysis practice. It is available on platforms like Kaggle, Tableau, and various educational portals.

2. Size of the Dataset

The dataset contains approximately 9,994 rows and 21 columns, representing individual order-level records.

3. Key Variables in the Dataset

Some of the most important columns include:

* Order ID – Unique identifier for each order
* Order Date – Date when the order was placed
* Ship Date – Date when the order was shipped
* Category – Product category (e.g., Furniture, Office Supplies, Technology)
* Sub-Category – More specific product type
* Sales – Total sale value
* Quantity – Number of units sold
* Discount – Discount applied to the order
* Profit – Profit earned from the sale
* Region, State, City – Geographic details
* Customer Segment – Type of customer (e.g., Consumer, Corporate, Home Office)

4. Time Period of the Data

The dataset spans over four years, typically from 2014 to 2017.

5. Initial Data Issues Identified

Upon exploring the dataset, the following issues were observed:

* Missing values: A few blank fields (especially in postal codes) were cleaned.
* Outliers: Identified in numerical fields such as Sales, Profit, and Discount. However, these were retained for analysis since they reflect real-world business variations like bulk purchases, high discounts, and loss-making orders.
* No duplicates were found.

**Data Cleaning & Preparation:**

Tools Used:

For data cleaning, I used both Excel and R. Initial data inspection and cleanup were done in Excel, followed by deeper structural and programmatic cleaning in R.

**Cleaning Tasks Performed**:

In Excel:

* Checked and removed unnecessary columns (e.g., helper column with value '1')
* Created basic pivot tables to understand Sales and Profit by Category

In R:

* Checked for Duplicates.
* Inspected data using str(), glimpse(), and dfSummary() for structure and type checks
* Removed a redundant last column that was unintentionally added.
* Checked for and confirmed the presence of duplicate entries using duplicated()
* Verified and handled missing values with colSums(is.na())
* Used summary() and table() to understand distributions of variables.

**Columns Created or Modified:**

* Converted Order Date and Ship Date to proper Date format.
* Created two new columns using the lubridate package.
* Order\_Year: extracted year from Order Date.
* Order\_Month: extracted labeled month from Order Date.

**Outlier Analysis**

As part of data exploration, an outlier detection step was performed using the Interquartile Range (IQR) method on key numerical variables — Sales, Profit, and Discount.

The analysis revealed:

* Sales outliers: 1,167 records
* Profit outliers: 1,881 records
* Discount outliers: 856 records
* These values represent extreme but realistic business cases, such as:
* High-value orders (e.g., bulk sales, premium products).
* High discounts during clearance or promotional events.
* Significant profits/losses from specific transactions.

**Decision:** Outliers were retained because:

All visualizations, insights, and statistical summaries appeared consistent and meaningful.

The distribution shapes and clustering logic remained interpretable.

Removing outliers might risk losing important business patterns and rare-event behavior.

While they can influence some metrics, they also represent valid extremes in business data and for this project, their inclusion offers a more complete picture.

**Special Considerations:**

* Ensured all date conversions were accurate to enable time-series analysis.
* Removed test or placeholder columns (like the one with a constant value '1') that could skew results.
* Standardized customer-related variables before clustering to avoid bias in K-Means algorithm.
* This careful preparation ensured that the dataset was clean, structured, and ready for advanced visualizations and segmentation analysis.

**Exploratory Data Analysis (EDA) & Visual Insights**

1. Region-wise Profit

Purpose: To understand which geographical regions generate the highest and lowest profits.

A graph of red bars

AI-generated content may be incorrect.

Fig:1

Insights:

The Western region leads in profit, showing strong market performance. The Southern region lags behind, indicating possible operational or sales issues. The Eastern region maintains steady profits, offering potential for strategic growth.

2. Segment-wise Profit

Purpose: To analyze which customer segments are most profitable.

A graph of a bar chart

AI-generated content may be incorrect.

Fig:2

Insights:

Corporate and Consumer segments drive the highest profits, while the Home Office segment lags behind, indicating an area worth investigating.

3. Monthly Sales Trend

Purpose: To identify seasonal sales patterns and peak months.

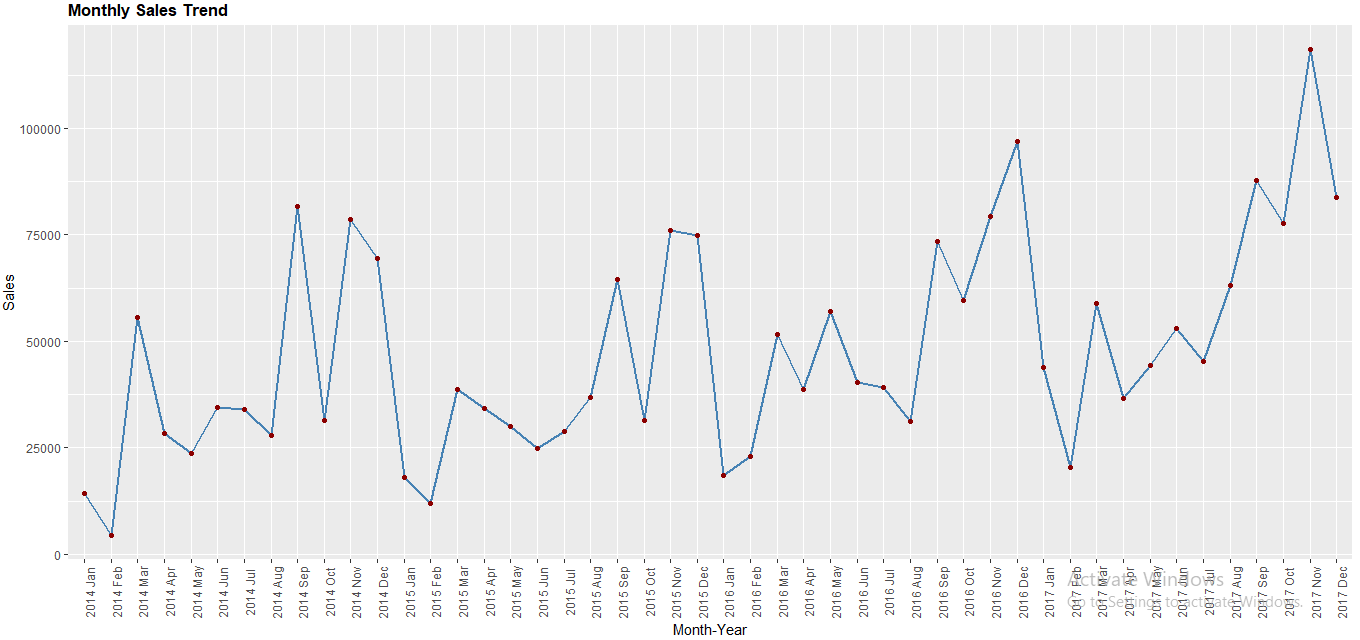


Fig:3

Insights:

Sales consistently peak in November–December, showing strong year-end demand.

January–February sees a drop, indicating post-holiday slowdown.

Mid-year months show moderate sales ideal for strategic promotions.

4. Discount vs Profit Relationship

Purpose: To understand how discounting impacts profitability.

A graph with blue and orange lines

AI-generated content may be incorrect.

Fig:4

Insights:

The plot reveals a sharp drop in profits as discounts increase, suggesting that aggressive discounting does not guarantee higher profitability. In fact, most transactions with heavy discounts result in negative or negligible profits, indicating that discounting beyond a threshold can be detrimental to revenue goals. Strategic discounting, therefore, must be balanced with profit margins to avoid value erosion.

5. Quantity vs Sales

Purpose: To study whether higher quantity sold leads to higher sales value.

A screen shot of a graph

AI-generated content may be incorrect.

Fig:5

Insights:

Most products cluster at lower sales and quantity levels, but a few outliers achieve extremely high sales. This confirms a skewed distribution, emphasizing the importance of identifying high-performing products early.

6. Top 10 Products by Sales

Purpose: To identify best-selling products based on total sales.

A graph with blue bars

AI-generated content may be incorrect.

Fig:6

Insights:

A small set of standout products drives most of the sales, indicating a Pareto-like 80/20 distribution. This helps pinpoint exact products for inventory prioritization and promotional targeting.

7. Profit by Sub-Category

Purpose: To identify profitable vs loss-making sub-categories.

A graph of a profit

AI-generated content may be incorrect.

Fig:7

Insights:

Profitability varies drastically across sub-categories.

Tables, Bookcases, and Supplies show steep losses, flagging them as areas needing urgent cost or pricing review. In contrast, Copiers and Phones generate the highest profits, signaling strong market demand and high margins.This contrast highlights where to cut losses and where to amplify investment..

8. Sales vs Profit by Region

Purpose: To compare sales and profit across regions side by side.

A graph of sales and profit

AI-generated content may be incorrect.

Fig:8

Insights:

Despite similar sales volumes across regions, profit margins vary sharply. The Southern region shows strong sales but low profit, pointing to high operational costs or pricing inefficiencies.

In contrast, the Western region excels in both sales and profit, making it a model for strategic replication in underperforming areas.

This reveals a clear opportunity to optimize costs and pricing in the South to boost overall profitability.

9. Customer Segmentation using K-Means

Purpose: To group customers based on similar purchasing behavior.

Clustering Methodology:

To segment customers, we aggregated the dataset by Customer ID and calculated Total Sales, Total Profit, and Total Quantity key indicators of customer value and activity. These were standardized using z-score normalization to ensure equal contribution.

We applied K-Means clustering with k = 3, chosen for practical interpretability to identify high-, medium-, and low-value customer groups. Though elbow method was not applied, visual separation in the clusters (Sales vs. Profit) suggests meaningful segmentation.

A graph with many colored dots

AI-generated content may be incorrect.

Fig:9

Insights:

Customer segmentation reveals three distinct clusters with varying profit contributions. The green cluster demonstrates high sales with balanced profits, suggesting loyal and high-value customers ideal for retention strategies. The pink cluster shows potential for growth but requires targeted marketing or better pricing to improve profitability. The blue cluster, though smaller, contains low-profit or loss-making customers, indicating a need for cost control or revised engagement tactics.

This clustering helps in tailoring strategies for each group to maximize overall business impact.

**Key Findings**

This project provided actionable insights into Superstore's sales and profitability patterns through in-depth exploratory data analysis. Below are the major findings derived from the visual outputs:

* Western region emerged as the most profitable and consistent performer across all metrics. It also maintained high sales, making it a benchmark region for replication.
* Southern region, despite good sales figures, showed the lowest profit margins, signaling a need to optimize costs or revise pricing strategies.
* Corporate and Consumer segments were the most profitable, while the Home Office segment underperformed. This suggests focusing marketing efforts on Corporate and Consumer buyers.
* Profitability is severely impacted by high discounts. Transactions with higher discounts showed negative or negligible profit, highlighting that discounting strategies must be more selective and margin-aware.
* Top-performing products follow a clear 80/20 rule — a small set of items drives most of the revenue. Identifying and promoting these high-performing products is crucial.
* Certain sub-categories like Tables, Bookcases, and Supplies consistently incurred losses. Immediate action is needed in pricing, sourcing, or inventory for these items. On the other hand, Phones and Copiers showed strong profitability and should be prioritized in future strategy.
* Seasonal spikes in sales were clearly observed in November and December, making them key months for marketing and inventory alignment. Conversely, Q1 months like January and February need cost-saving or promotional tactics.
* Customer segmentation using K-Means clustering revealed three clear segments:
* High-value loyal customers (green cluster)
* Moderate profit customers with potential (pink cluster)
* Low-profit or loss-making customers (blue cluster) needing re-engagement or review
* Even when sales volume appears similar across regions, profit outcomes differ, emphasizing that not just sales but profit per region or product must be considered in decision-making.

**Conclusion & Recommendations**

The Sales Performance Analysis of the Superstore dataset revealed several critical insights into the company’s strengths and operational gaps. While specific product categories like Chairs, Phones, and Binders consistently contribute to high profits, other segments such as Tables, Bookcases, and Supplies showed recurring losses, especially in regions like the South and Central. Additionally, frequent high discounts seem to reduce profit margins rather than boost sales volume.

Based on the analysis, the following strategic recommendations are suggested:

1. Reassess Product Portfolio

Unprofitable products like Tables and Bookcases should be either repositioned with a better value proposition or phased out gradually. Consider bundling them with high-selling products or offering value-added services like free installation.

2. Regional Sales Strategy

Focus aggressive marketing and logistics improvement efforts in profitable regions such as the West and East. For low-performing regions, explore hyper-local customization adapting product choices based on demographic demand.

3. Rethink Discount Strategy

Instead of frequent blanket discounts, implement data-driven smart discounting target specific customer segments, special occasions, or bundle offers to retain profit margins while maintaining customer interest.

4. Invest in Technology & Forecasting

Leverage predictive analytics and machine learning to forecast high-demand seasons and optimize inventory. This minimizes overstocking of poor-performing items and improves supply chain efficiency.

5. Strengthen Customer Segmentation

Focus more on Corporate and Home Office segments that showed better returns. Create personalized marketing campaigns using customer behavior insights to improve loyalty and repeat purchases.

6. Optimize Time-Sensitive Sales

Time series analysis indicates periodic spikes in sales. Utilize this to launch seasonal campaigns with targeted products and dynamic pricing to maximize revenue during peak periods.

This analysis not only offers clarity on operational inefficiencies but also demonstrates how combining Excel and R programming can provide a powerful foundation for data-driven business decisions.