🛒 Supermarket Sales Performance: Exploratory Data Analysis (EDA) Report

Executive Summary

This report presents a detailed Exploratory Data Analysis (EDA) of a supermarket's transaction data spanning from 2015 to 2018. The analysis of 9,789 transactions reveals critical insights into sales performance, customer segmentation, regional dynamics, and product category trends. Key findings indicate that the **Consumer segment** in the **West and East regions** are the primary revenue drivers, the **Technology category** generates the highest sales, and clear **seasonal peaks** occur during the year-end holiday period. Strategic recommendations are provided to optimize marketing, inventory, logistics, and regional operations to capitalize on these insights and drive future growth.

---

1. Introduction & Data Overview

1.1 Project Objective

To analyze historical sales transaction data, uncover underlying patterns, relationships, and trends, and provide data-driven recommendations to enhance business strategy and operational efficiency.

1.2 Dataset Description

- **Source**: Kaggle - Supermarket EDA Dataset
- **Total Records**: 9,800 (9,789 after cleaning)
- **Features**: 18 columns encompassing order details, customer information, product details, and sales figures.
- **Period**: Transactions from 2015 to 2018.

1.3 Tools & Libraries

- **Python** for data manipulation and analysis.
- **Pandas & NumPy** for data cleaning and transformation.
- **Matplotlib, Seaborn, and Plotly** for data visualization.
- **Dython** for advanced correlation analysis.

---

2. Data Understanding & Cleaning

2.1 Initial Assessment

The initial dataset contained 9,800 rows and 18 columns. A preliminary check revealed:

- **Missing Values**: 11 missing entries in the `Postal Code` column.
- **Data Types**: `Order Date` and `Ship Date` were stored as objects (strings) instead of datetime objects.
- **Unique Values**: The dataset contained 793 unique customers, 1,861 unique products, and transactions from 49 states.

2.2 Data Cleaning Steps

1. **Handling Missing Values**: The 11 rows with missing `Postal Code` were dropped, resulting in a clean dataset of **9,789 records**.
2. **Data Type Conversion**: The `Order Date` and `Ship Date` columns were converted to the `datetime64[ns]` data type for time-series analysis.
3. **Feature Engineering**: New features were extracted from the dates for deeper analysis:
    - `Order-year`, `Order-Month`
    - `Ship-Year`, `Ship-Month`

2.3 Processed Data Snapshot

Row ID	Order ID	Order Date	Ship Mode	Customer Segment	Region	Category	Sub-Category	Sales	Order-Year
1	CA-2017-152156	2017-11-08	Second Class	Consumer	South	Furniture	Bookcases	261.96	2017
2	CA-2017-152156	2017-11-08	Second Class	Consumer	South	Furniture	Chairs	731.94	2017
3	CA-2017-138688	2017-06-12	Second Class	Corporate	West	Office Supplies	Labels	14.62	2017

3. Univariate Analysis

3.1 Analysis of Numerical Variables: Sales

The `Sales` variable is the primary KPI for this analysis.

Statistic	Value
Count	9,789
Mean	$230.77
Standard Deviation	$626.65
Minimum	$0.44
25th Percentile	$17.25
Median (50th)	$54.49
75th Percentile	$210.61
Maximum	**$22,638.48**

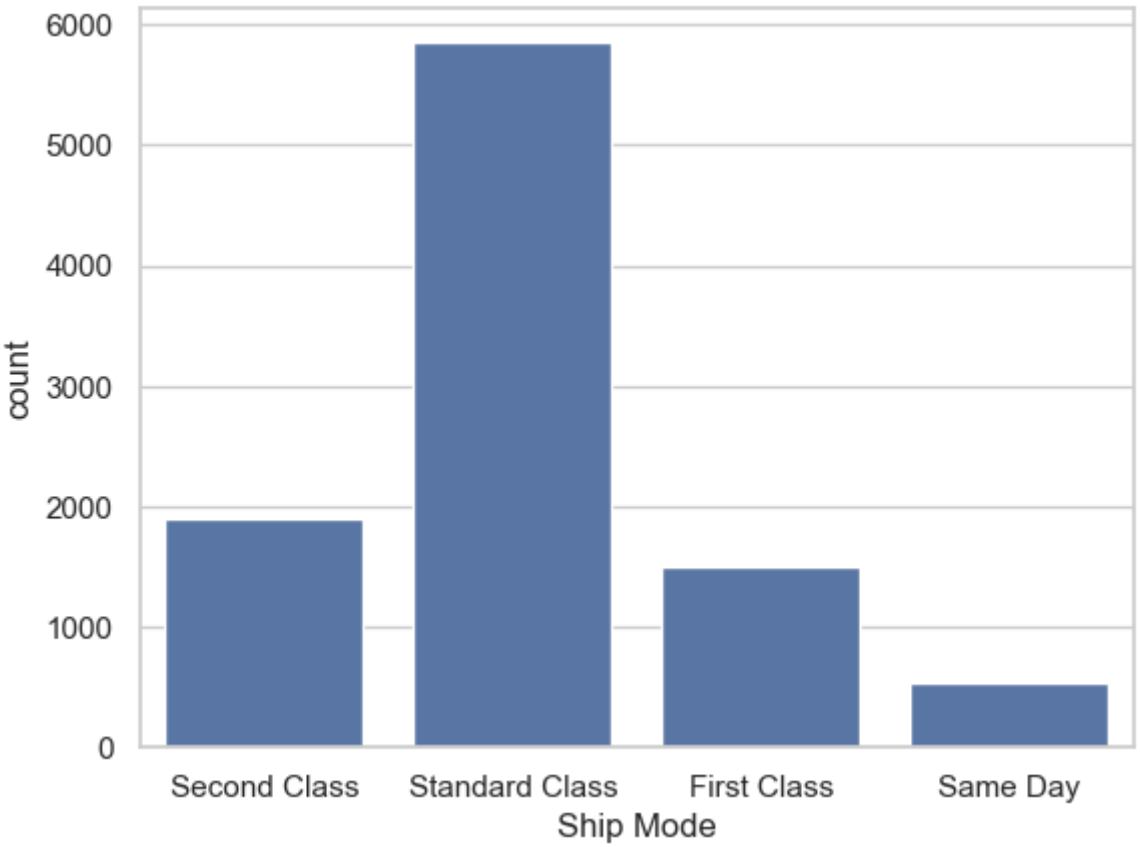
Initial Insight: The significant difference between the mean and median, coupled with a very high maximum value, suggests a right-skewed distribution with potential outliers.

3.2 Analysis of Categorical Variables

A. Shipping Modes

The supermarket offers four shipping modes. **Standard Class** is the most frequently used option by a significant margin, indicating it is the default or most economical choice for customers.

```
Standard Class     5859
Second Class       1943
First Class        1530
Same Day            457
```



B. Customer Segments

The customer base is divided into three segments. **Consumers** are the largest segment, forming the backbone of the business.

```
Consumer      5101
Corporate     3025
Home Office   1663
```

C. Geographic Distribution (Top States & Cities)

- **States**: California is the top state by number of orders, followed by New York and Texas.

- **Cities**: New York City is the leading city, with Los Angeles a distant second. This highlights key urban centers as primary markets.

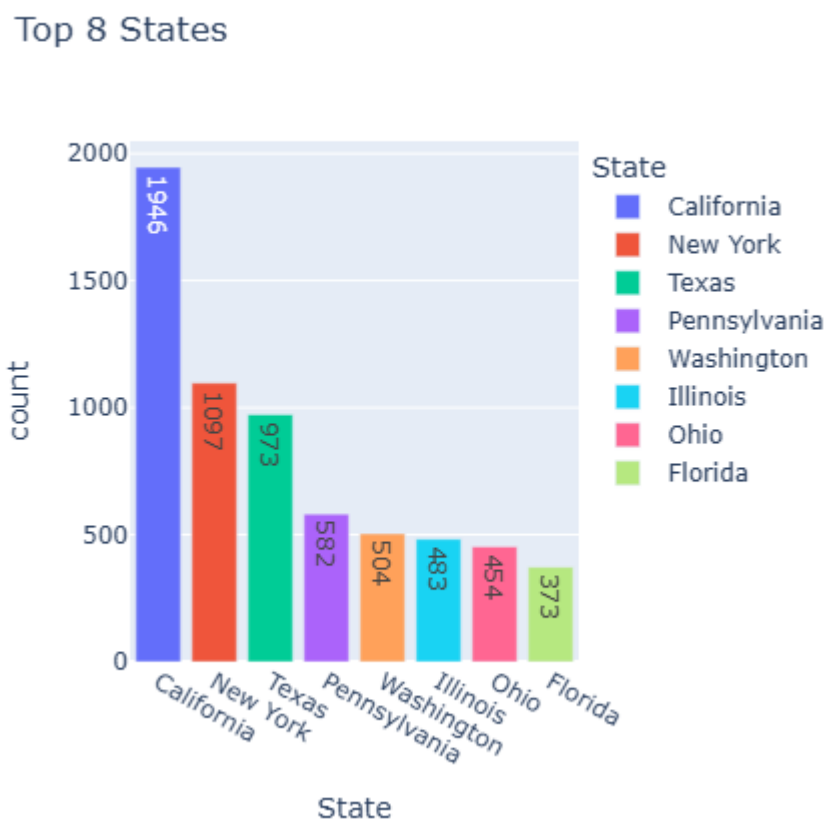


Figure: Top 8 States by Number of Orders

D. Product Categories and Sub-Categories

- **Categories**: **Office Supplies** is the most frequently sold category, followed by Furniture and Technology.
- **Sub-Categories**: **Binders** are the top sub-category, suggesting high volume in everyday office needs.

Figure: Frequency of Orders across Regions, Categories, and Sub-Categories

---

4. Outlier Analysis

4.1 Identification

A boxplot of the `Sales` variable confirmed the presence of numerous extreme outliers.

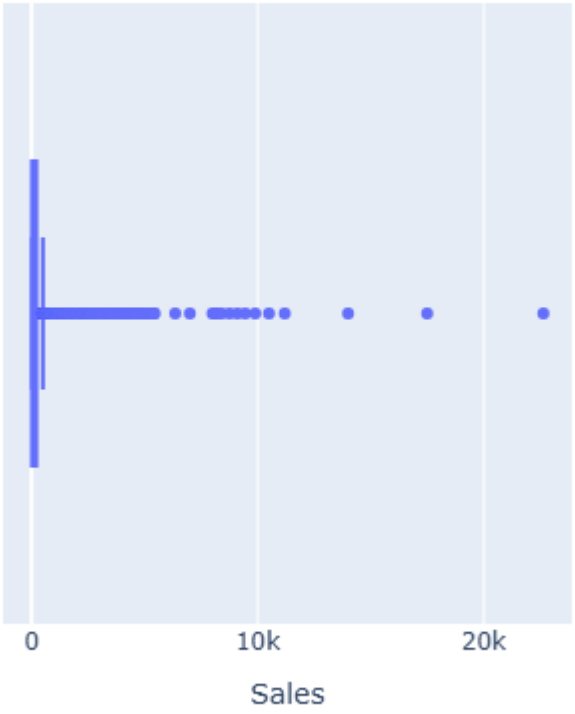


Figure: Boxplot revealing extreme outliers in the Sales data.

4.2 Investigation & Decision

- **Findings**: 1,288 outlier records were detected. These records had a mean sale of $1,267, far above the overall mean of $230. The maximum outlier was $22,638.
- **Conclusion**: These outliers represent genuine high-value transactions (e.g., bulk corporate orders, expensive technology purchases) rather than data entry errors.
- **Action**: **Outliers were retained** in the dataset as they contain valuable business information about high-value sales opportunities.

---

5. Multivariate Analysis

5.1 Sales by Segment and Region

This analysis reveals which customer segments contribute the most revenue in different geographical areas.

Key Insight: The **Consumer segment generates the highest total sales in every region**, with the West and East regions being the largest revenue generators.

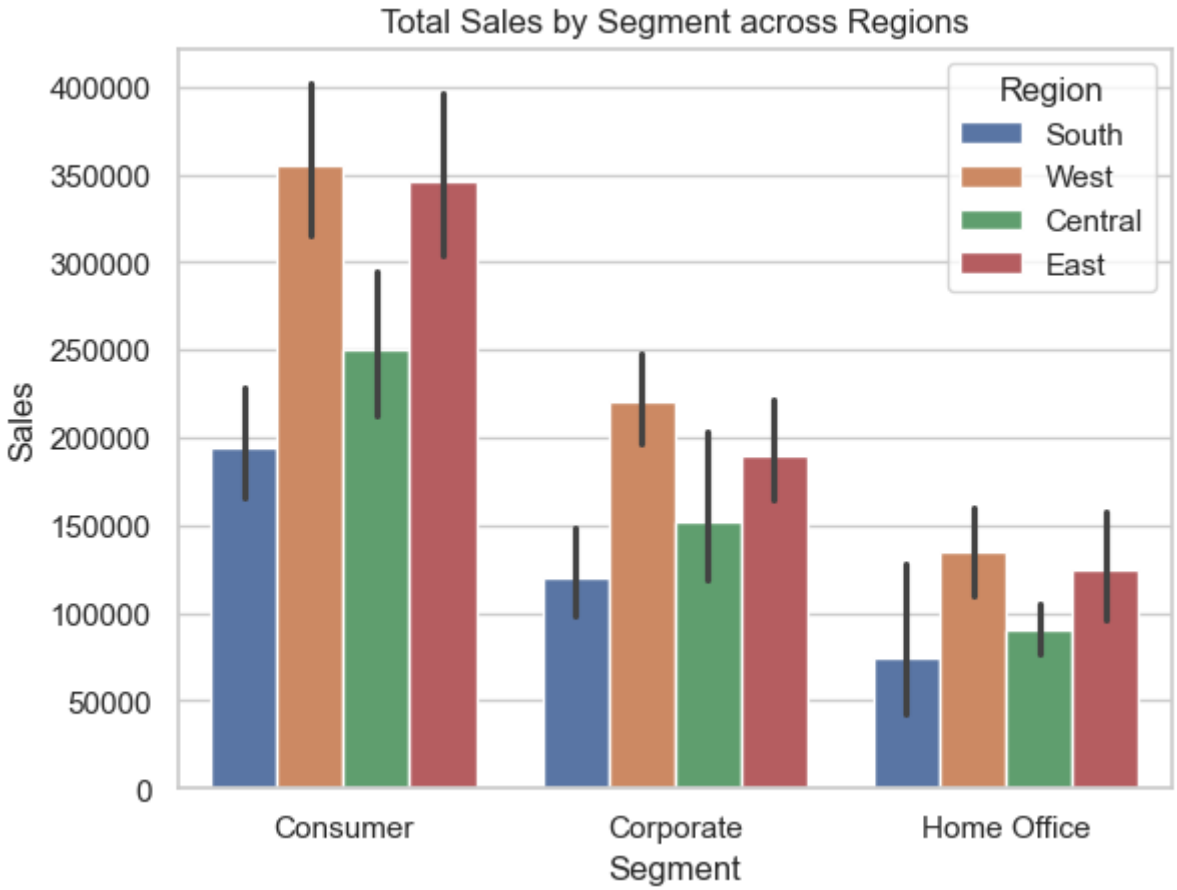


Figure A: Total Sales by Customer Segment across Regions.

5.2 Sales by Category and Ship Mode

This analysis shows how different product categories perform across various shipping options.

Key Insight: The **Technology category generates the highest total sales**, despite not being the most frequently ordered. Notably, while "Same Day" shipping is used less frequently, it is associated with high sales in the Technology category, suggesting customers are willing to pay a premium for fast delivery of high-value items.

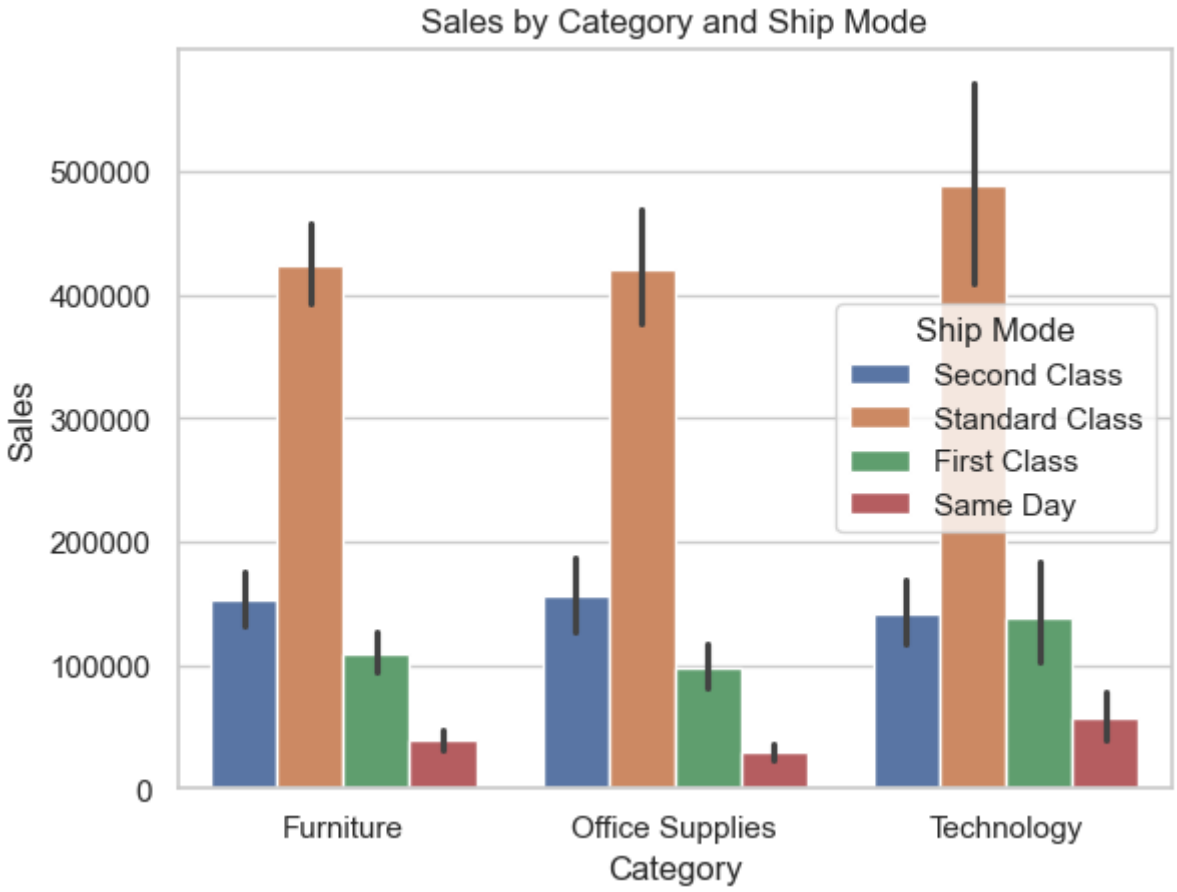


Figure B: Sales by Product Category and Shipping Mode.

5.3 Correlation Analysis

A heatmap of numerical variables shows strong positive correlations between time-based features.

Key Insight: The very high correlation (0.99) between `Order-year` and `Ship-year`, and between `Order-Month` and `Ship-Month` (0.91), indicates efficient logistics and confirms that sales are heavily influenced by temporal trends.

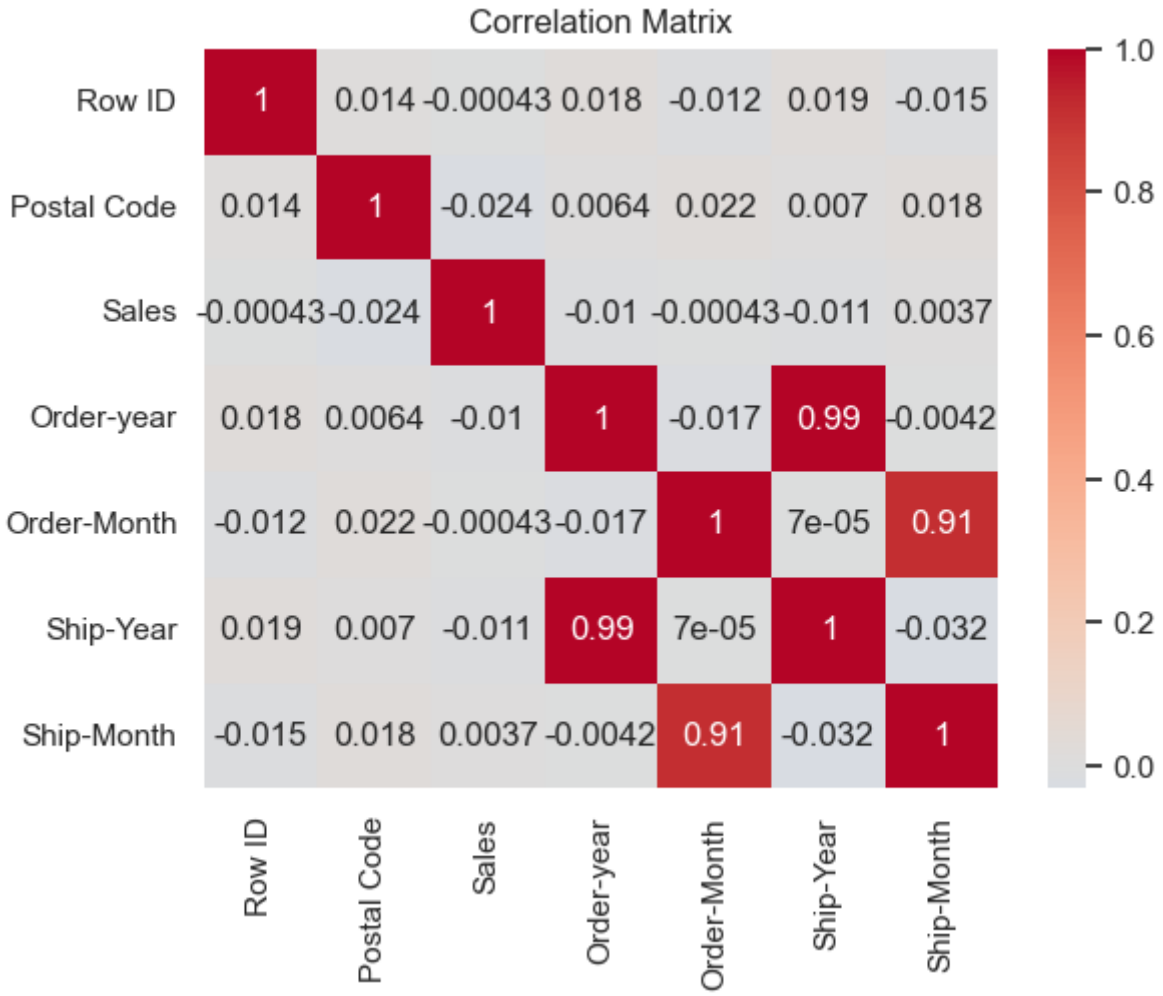


Figure C: Correlation Matrix of Numerical Features.

5.4 Time Series Analysis: Monthly Sales Trend

A line chart of monthly total sales reveals clear patterns over time.

Key Insight: There are consistent and pronounced **sales peaks in November and December of each year**, aligning with holiday shopping seasons (Black Friday, Cyber Monday, Christmas). This pattern strengthens over the years, indicating growing seasonal demand.

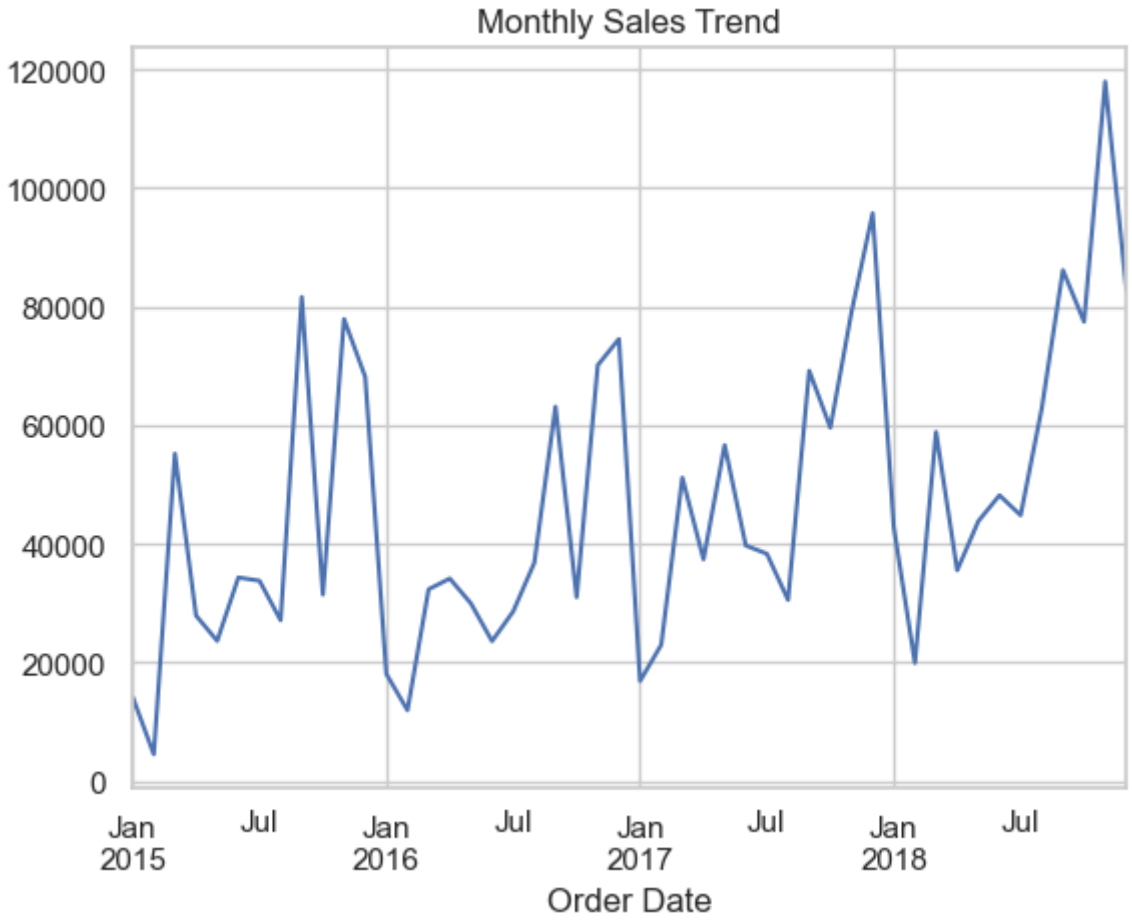


Figure D: Monthly Sales Trend showing clear seasonal peaks.

6. Insights & Strategic Recommendations

Insight Area	Key Finding	Recommendation
1. High-Value Segments	The **Consumer segment** dominates sales, especially in the **West and East**.	**Launch targeted marketing campaigns** (e.g., loyalty programs, personalized ads) focused on the Consumer segment in these high-performing regions.
2. Regional Opportunities	The **South and Central regions** underperform compared to the West and East.	**Implement regional promotions** and evaluate supply chain efficiency in these areas. Consider partnerships with local businesses to boost brand presence.
3. Category Performance	The **Technology category** is the top revenue generator, not the most frequent.	**Increase inventory and marketing investment in Technology**. Explore product bundling (e.g., laptops with accessories) to increase average order value.
4. Shipping Strategy	**Same Day shipping** is linked to high-value Technology sales.	**Promote express shipping options** for high-value categories. Offer free Same Day shipping on orders above a specific threshold to incentivize larger purchases.

Insight Area	Key Finding	Recommendation
5. Seasonality	Clear, strong **sales peaks in Nov-Dec**.	**Begin holiday planning in Q3**. Secure inventory, especially for Technology and Furniture, and run pre-holiday marketing campaigns to capture early demand.
6. Forecasting	Sales are **highly correlated with time**.	**Develop time-series forecasting models** to predict demand accurately. Use these models for optimized staffing, inventory management, and budget allocation.

7. Conclusion

This EDA has successfully transformed raw transaction data into actionable business intelligence. The analysis underscores the critical importance of the **Consumer segment**, the **Technology product category**, and strategic **regional focus** for driving growth. Furthermore, the identification of strong seasonal trends provides a reliable foundation for future planning and forecasting.

By implementing the recommended strategies, the supermarket can optimize its operations, enhance customer targeting, improve logistics, and ultimately, maximize profitability and market share.