Superstore Sales Analysis

Urvesh Mistry - K00302035

2025-04-25

Table of contents

# 1. Problem Statement

“Behind every purchase is a decision, behind every return is a story, and behind every product is a pattern waiting to be discovered”. In an era where consumers have higher expectations and loyalty is earned rather than taken for granted, companies want more than dashboards. They require clear understanding.  
  
This study connects the dots between sales performance, customer sentiment, and return behavior using a comprehensive Superstore dataset. By designing unique measures such Performance Index and Inventory Stress and then displaying them to show where we are making progress, where we are losing confidence, and where action is overdue, it transcends conventional KPIs.  
  
Every bar, scatter, and trend line in this report is a discussion starter for data-driven decisions that generate income and preserve loyalty.

# 2. Solution Summary

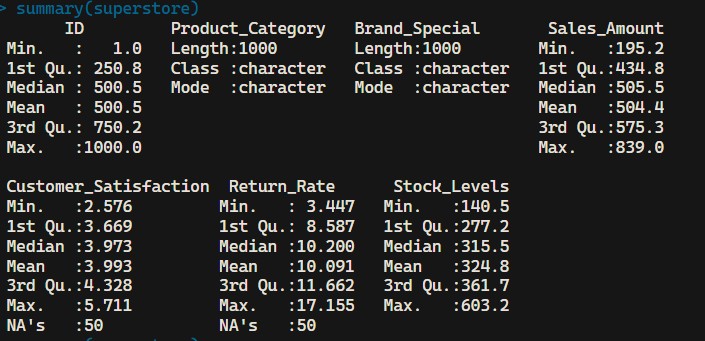
This analysis applied a blend of **Data Cleaning**, **Feature Engineering**, and **Visual Storytelling** to the Superstore data-set. which are as follows:

1. **Data Quality Improvement:** The data set was thoroughly evaluated to identify and rectify any missing, inconsistent, or incorrect data. Textual fields were cleaned and standardized, with errors and special characters removed. Numerical fields are formatted to maintain precision and uniformity throughout the dataset.
2. **Descriptive Analysis**: Explored the distribution of key metrics of data set through **summary statistics** including mean, median, mode, SD (standard deviation), and IQR for different variables
3. **Feature Engineering: Created new, insightful variables** to enhance analysis depth and improve the predictive power for business decision making.
4. **Visualization:** Created **visual representations** (e.g., histograms, bar charts, box plots, scatter plots and Dual axis chart) to understand the data distribution and relationships better.
5. **Conclusions**: Deciding via the **EDA** on customer satisfaction, return risk patterns, and inventory control helps to identify areas for improvement, increase profitability, and strengthen relations with customers.

# 3. Descriptive Statistics:-

### 3.0.1 Summary statistics

Overview of my dataset



Summary Statistics

The dataset consists of 1,000 observations and includes the following variables: Product category, Brand special, Sales amount, Customer satisfaction, Return rate, and Stock levels. Below is a detailed summary of each variable:

**Product\_Category**

* **Length**: 1,000
* **Class**: Character
* **Description**: Indicates the category of the product sold (e.g., Electronics, Furniture, Clothing, Groceries).

**Brand\_Special**

* **Length**: 1,000
* **Class**: Character
* **Description**: Represents the brand or outlet associated with the product.

**Sales\_Amount**

* **Minimum**: 195.2
* **1st Quartile**: 434.8
* **Median**: 505.5
* **3rd Quartile**: 575.3
* **Maximum**: 839.0
* **Class**: Numeric (2 decimal precision)
* **Description**: Total sales amount for each product line item.

**Customer\_Satisfaction**

* **Minimum**: 2.58
* **1st Quartile**: 3.67
* **Median**: 4.00
* **3rd Quartile**: 4.32
* **Maximum**: 5.71
* **Class**: Numeric (rounded to 2 decimals)
* **Description**: Satisfaction score given by the customer.

**Return\_Rate**

* **Minimum**: 3.44
* **1st Quartile**: 8.56
* **Median**: 10.20
* **3rd Quartile**: 11.61
* **Maximum**: 17.15
* **Class**: Numeric (%)
* **Description**: Percentage of products returned per line item.

**Stock\_Levels**

* **Minimum**: 140
* **1st Quartile**: 277
* **Median**: 315
* **3rd Quartile**: 362
* **Maximum**: 603
* **Class**: Integer
* **Description**: Reflects the number of units in stock at the time of transaction.

Data summary

|  |  |
| --- | --- |
| Name | superstore |
| Number of rows | 1000 |
| Number of columns | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 2 |
| numeric | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Product\_Category | 0 | 1 | 8 | 11 | 0 | 4 | 0 |
| Brand\_Special | 0 | 1 | 8 | 10 | 0 | 4 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | 0 | 1.00 | 500.50 | 288.82 | 1.00 | 250.75 | 500.50 | 750.25 | 1000.00 | ▇▇▇▇▇ |
| Sales\_Amount | 0 | 1.00 | 504.41 | 100.76 | 195.21 | 434.85 | 505.53 | 575.35 | 839.04 | ▁▅▇▃▁ |
| Customer\_Satisfaction | 50 | 0.95 | 3.99 | 0.49 | 2.58 | 3.67 | 3.97 | 4.33 | 5.71 | ▁▆▇▃▁ |
| Return\_Rate | 50 | 0.95 | 10.09 | 2.27 | 3.45 | 8.59 | 10.20 | 11.66 | 17.16 | ▁▅▇▃▁ |
| Stock\_Levels | 0 | 1.00 | 324.84 | 71.53 | 140.54 | 277.20 | 315.48 | 361.65 | 603.23 | ▁▇▆▂▁ |

With **skim()** we can find a more detailed statistical analysis of each column with the count of missing data as shown in the figure

# 4. Data Cleaning:

Data cleaning in R is a crucial step in preparing raw data for analysis, ensuring it’s accurate, consistent, and structured correctly.

The Superstore data set, like many real-world datasets, contains inconsistencies, missing values, and formatting issues that can hinder analysis. By cleaning the data, it ensure that it is **accurate, standardized, and optimized** for tasks

### 4.0.1 Text Cleanup & Type Conversion

By refining **Product\_Category** and **Brand\_Special**, it ensured consistency and eliminated potential grouping errors, which is essential for reliable analysis. For original data before cleaning look image

During the data cleaning phase,

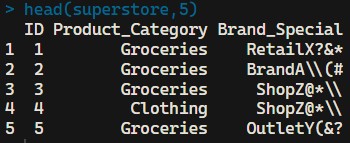
1. **Product\_Category Column Cleaning**

* Used str\_trim() to eliminate leading and trailing spaces, preventing mismatches (e.g., "Electronics " vs "Electronics").
* Converted the column to **character format** to ensure consistent handling in grouping, filtering, and visualization processes.

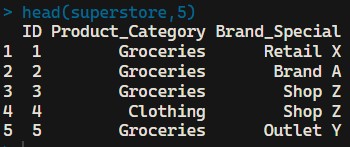
1. **Brand\_Special Column Cleaning**

* Applied str\_trim() to remove excess spaces, similar to Product\_Category.
* Used str\_replace\_all("[^[:alnum:] ]", "") to eliminate characters like %, &, or @, ensuring uniform brand naming.
* Inserted a space between lowercase-uppercase combinations (e.g., "RetailX" to "Retail X") to improve readability.

A clear formatted and consistently labeled brand column and Product category enabling accurate aggregation and ensuring that each brand and category is recognized correctly in grouping operations and visualization.



Text Cleanup & Type Conversion - Original

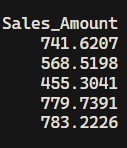


Text Cleanup & Type Conversion

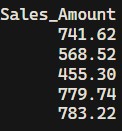
### 4.0.2 Sales\_Amount Column Cleaning

* Used as.numeric() to prevent errors caused by text-formatted numbers, ensuring proper calculations.
* Applied round(, 2) to maintain correct currency representation, avoiding unnecessary precision that could complicate financial summaries.

This refinement ensures that **aggregations, comparisons, and financial models** are reliable and error-free



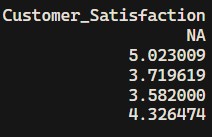
Sales amount - Original



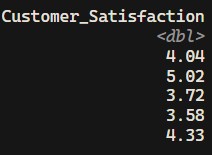
Sales amount

### 4.0.3 Customer\_Satisfaction Column Cleaning

* The dataset was grouped by **Product\_Category** and **Brand\_Special** to ensure that imputed values reflected relevant product-brand combinations.
* Within each group, the missing values (NA) were replaced by the median satisfaction score, maintaining the natural variation of customer feedback.
* Example: For "Groceries" & "Retail X", the known satisfaction scores **(3.65, 3.8, 4.87)** resulted in a median **(4.04)**, which was used to fill missing values.
* All satisfaction scores were rounded to **two decimal places** round(, 2) to maintain a consistent numeric format throughout the dataset.



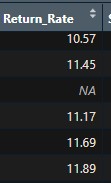
Customer satisfaction - Original



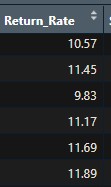
Customer satisfaction

### 4.0.4 Return\_Rate Column Cleaning

* The dataset was grouped by **Product\_Category** and **Brand\_Special** to ensure that imputed values reflected relevant product-brand combinations.
* Replaced NA values with the **median return rate** of their respective group (e.g., "Furniture & Brand A" → **9.83%** median imputed).



Return\_Rate Original



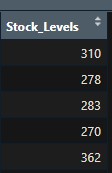
Return\_Rate

### 4.0.5 Stock\_Levels Column Cleaning

* The Stock\_Levels column represents the number of product units available in inventory at the time of sale. The stock is always measured in full units, the values were **rounded to the nearest whole number** using the round() function in R.
* There were **no missing values** in this column, so no imputation was needed.



stock levels - Original



stock levels

# 5. Feature Engineering:

As part of this analysis, three new features were engineered to deepen the understanding of product performance, customer behavior, and inventory efficiency. These derived variables — Satisfaction\_Return\_Ratio, Performance\_Index, and Inventory\_Stress — combine existing metrics to uncover relationships

### 5.0.1 Satisfaction\_Return\_Ratio

* This feature calculates the ratio between **Customer Satisfaction** and **Return Rate**
* It combines two key customer experience signals that how happy customers are and how often they return products.
* A high ratio indicates that customers are satisfied and not frequently returning the product a positive sign. Conversely, a lower ratio may highlight inconsistencies, such as high satisfaction scores paired with high return rates, which could point to quality issues, misleading product descriptions, or false-positive reviews.

### 5.0.2 Performance\_Index

* A composite metric measuring overall product performance by combining sales, satisfaction, and return behavior
* This single value reflects a more complete picture of a product’s performance by integrating revenue, experience, and risk.
* It’s especially useful for ranking or scoring products and brands in terms of effectiveness, and for identifying high-selling items that may still carry hidden risks due to high dissatisfaction or return volume.

### 5.0.3 Inventory\_Stress

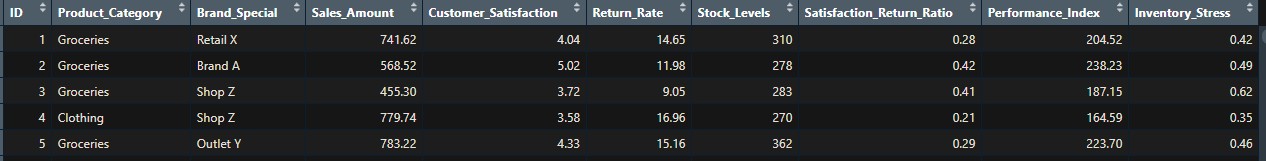
* Inventory\_Stress measures the strain on inventory by comparing how much stock is held versus how much is sold
* Calculated as Stock\_Levels ÷ Sales\_Amount, a high value indicates products that are heavily stocked but not selling efficiently, while a low value may indicate strong demand or potential understocking.
* This metric is valuable for supply chain and inventory teams, as it quickly highlights which products may require attention in terms of restocking, discounting, or supply adjustment.

# 6. Cleaning and Feature Engineering

* The displayed table represents a snapshot of the superstore dataset after undergoing a complete data cleaning and feature engineering process. It highlights a few initial rows with all key columns cleaned, standardized, and enriched to support insightful analysis.
* Categorical columns such as Product\_Category and Brand\_Special were cleaned by removing any leading or trailing white spaces and correcting inconsistent formatting
* Numerical columns including Sales\_Amount, Customer\_Satisfaction, Return\_Rate, and Stock\_Levels were standardized. Values were rounded to improve readability and reporting consistency.
* Additionally, missing values in customer satisfaction and return rate were imputed using the **median value** within each unique combination of product category and brand.
* Median imputation is used because it is robust to the outliers.
* As the original dataset is not perfectly normal so to maintain its data integrity without skewing outcomes. i have use median imputation.
* Three new columns were created through **Feature Engineering**:
* **Satisfaction\_Return\_Ratio** captures how well satisfaction aligns with return behavior.
* **Performance\_Index** aggregates sales, satisfaction, and return rate into a single performance score.
* **Inventory\_Stress** compares stock levels with actual sales, revealing whether products are overstocked or efficiently managed.



Original Dataset

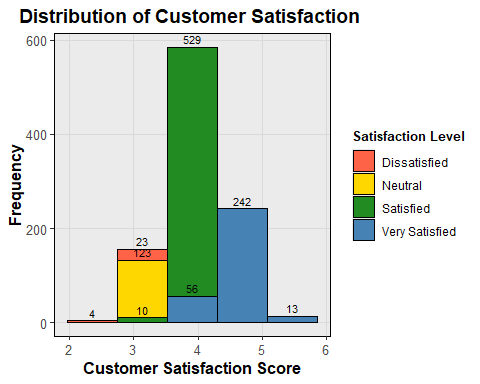


Cleaned and Feature Engineered Dataset

# 7. Exploratory Visualisations

#### 7.0.0.1 **Distribution of Customer Satisfaction by Ratings.**

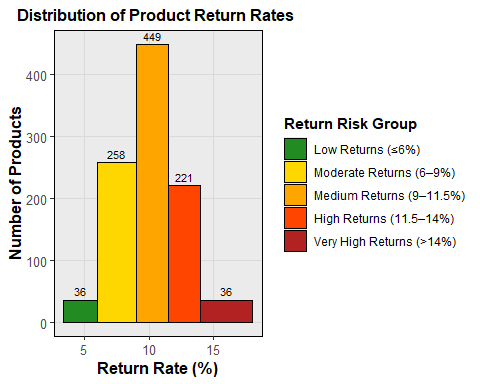
* To categorize the satisfaction scores into meaningful **ratings**. These thresholds are **business-friendly** and align with how satisfaction is often interpreted in real-world settings (e.g., 5-star ratings).
* It visualizes how customers are distributed across satisfaction levels, broken into 5 bins that represent levels from “Very Dissatisfied” to “Very Satisfied.”



* This chart serves as an effective starting point for evaluating the **overall tone of customer feedback**, highlighting areas of strength and signaling where improvement efforts can be targeted.
* By aggregating satisfaction into grouped ratings and visualizing it clearly, it supports **data-driven decisions** aimed at enhancing customer retention and brand rating
* From the plot, it’s evident that most customers fall into the *Satisfied* and *Very Satisfied* categories, suggesting a generally positive customer experience across the dataset.
* Very few responses fall into the *Very Dissatisfied* or *Dissatisfied* range, which is encouraging.
* However, the **moderate number of neutral responses** indicates there may be opportunities for product or service improvement to further push customers into higher satisfaction bands.

#### 7.0.0.2 **Return Rate Distribution**

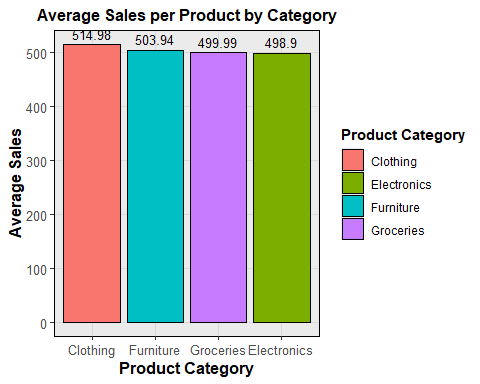
* This histogram illustrates the distribution of product return rates in the superstore dataset, grouped into clearly defined risk bands.
* Each product was classified into one of five **Return Risk Groups,** ranging from *Low Returns (≤6%)* to *Very High Returns (>14%)* , using logical thresholds that help differentiate levels of return behavior.



* The chart reveals that the majority of products fall into the *Medium Returns (9–11.5%)* group, with **449 occurrences**, followed by the *Moderate Returns (6–9%)* category at **258 occurrences**. This suggests that most products return at a manageable but non-negligible rate.
* However, there’s a noticeable cluster on the higher-risk end, **221 products** fall under the *High Returns (11.5–14%)* band.
* **36 products** in the *Very High Returns (>14%)* category could indicate serious quality issues or mismatches in customer expectations.
* On the other hand, only **36 products** show very low return rates (≤6%), revealing that very few items are consistently immune to returns.

#### 7.0.0.3 Average Sales Per Product

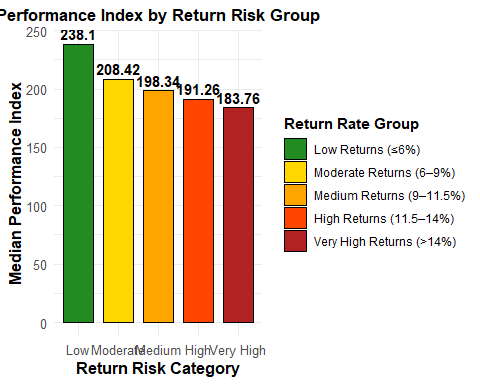
* **Bar chart that visualizes the average sales amount per product category** in the Superstore dataset.
* A custom color palette is defined using specific codes, assigning a unique and consistent color to each product category.
* The dataset is grouped by Product\_Category, and the average of the Sales\_Amount is calculated for each group using the summarise() function from the dplyr package.



* **Clothing** has the highest average sales per product, with approximately **$514.98**.
* **Furniture** follows with **$503.94**, then **Groceries** at **$499.99**.
* **Electronics** has the lowest average at **$498.90**.
* Although the differences between categories are relatively small, the chart clearly demonstrates that **Clothing tends to generate slightly higher revenue per product sold** compared to the others.
* This type of visualization is important in decision-making because it highlights **which product categories are more profitable on average**

#### 7.0.0.4 Performance Index by Return Rate Group

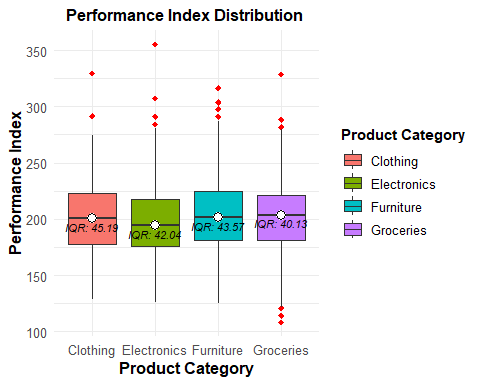
* The code first groups the Superstore dataset by return rate risk categories and calculates the **median performance index** for each group.
* It simplifies the return rate labels for cleaner plotting and enforces a business-intuitive order ie. (Low to Very High).



* Then, it visualizes these values in a **color-coded bar chart**, where each bar represents a return group and its corresponding median performance score.
* The resulting graph clearly shows that products with **lower return rates** tend to have **higher performance indexes**, which signals that high-performing products are also more stable

#### 7.0.0.5 Performance Index Distribution by Product Category

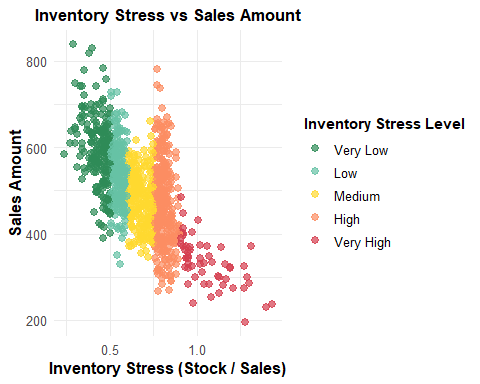
* Boxplot visualization that compares the **distribution of Performance Index** across different product categories in the Superstore dataset
* It first calculates the **interquartile range (IQR)** and **median** of the performance index for each category and stores the results in a summary table.



* The **middle 50%** of values (interquartile range or IQR)
* The **median** is highlighted with a **white dot** and a bold horizontal line inside the box.
* **Red diamonds** represent **outliers**
* The IQR value is displayed under each median, giving you an idea of how widely values vary in each category.

#### 7.0.0.6 Inventory Stress vs Sales Amount

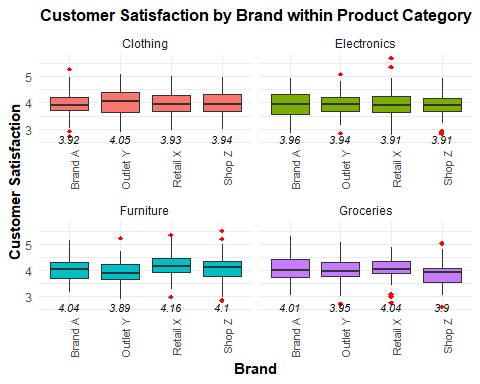
* Scatterplot visualizes the **relationship between inventory stress and sales performance** across all products.
* This code creates a new column Inventory\_Stress\_Group that **categorizes the Inventory\_Stress ratio into 5 meaningful bands**.



* These groups help to segment products based on how stressed their inventory is (i.e., how much stock they hold in relation to their sales performance).
* The **x-axis** shows how much inventory a product has relative to its sales (lower is better).
* The **y-axis** shows total sales.
* Products with low inventory stress tend (green/teal) to achieve higher sales, while those with high stress (especially in the red categories) tend to perform poorly in terms of revenue.

#### 7.0.0.7 Customer Satisfaction by Different Brands Within Category

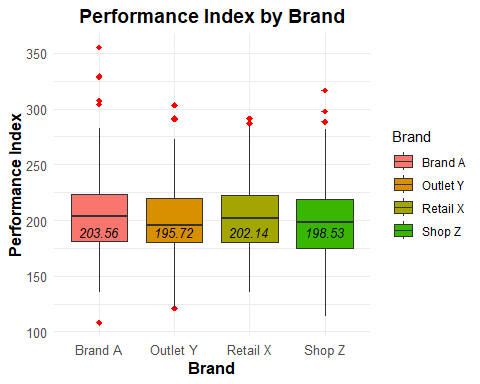
* A **faceted boxplot** to visualize **customer satisfaction by brand** within each Product\_Category from the Superstore dataset
* it calculates the **median customer satisfaction score** for every combination of product category and brand, storing it in a summary table.
* Then, using ggplot2, it creates a panelled (faceted) chart that displays boxplots of satisfaction scores for each brand, segmented by product category.



* Each **box** represents the distribution of satisfaction ratings for a brand.
* The **red diamonds** highlight outlier scores that either have high or low values.
* The **black italic numbers** below each box are the median scores for each brand,
* In **Furniture**, “Retail X” stands out with the highest median score (4.16), suggesting strong brand satisfaction.
* **Outlet Y** underperforms in both **Electronics** and **Furniture**, with visibly lower medians.

#### 7.0.0.8 Performance Index by Brand

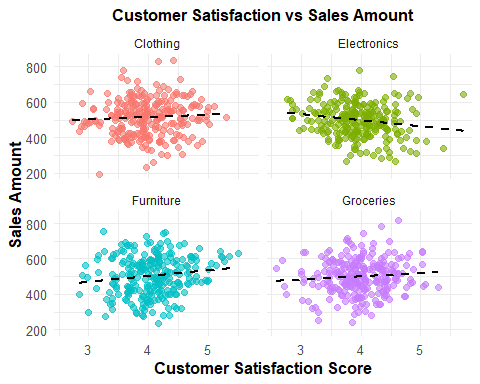
* This R script creates a **boxplot visualization** to compare the **Performance Index** distribution across different brands in the Superstore dataset.
* It starts by calculating the median performance score for each brand, which is then used as a reference point in the visualization.



* The plot itself displays each brand’s performance distribution, highlighting central tendency and spread through boxplot statistics (Q1, Q3, and median), while visually flagging **outliers** in red.
* A clearly labeled **median value is displayed below each box**, enhancing interpretability.
* The chart helps to compare brand performance, spot inconsistencies, and identify brands with high variability or outlier behaviors

#### 7.0.0.9 Customer Satisfaction vs. Sales Amount (Facet Chart)

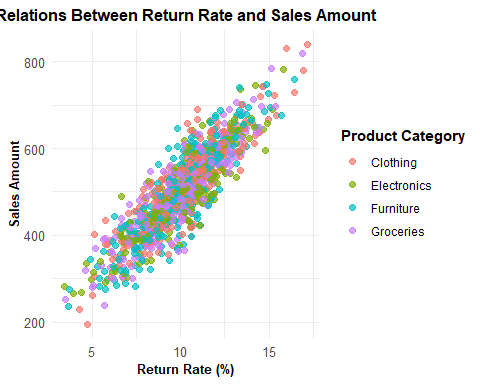
* A **faceted scatter plot** that analyzes the relationship between **Customer Satisfaction and Sales Amount** across different product categories in the Superstore dataset.
* Using ggplot2, the script maps individual product records as points and colors them by product category.



* The chart includes a **dashed linear regression line** in each panel to highlight potential trends.
* The graph visually separates data by category using **faceting**, allowing targeted interpretation. For example, in categories like **Furniture and Clothing**, there appears to be a mild **positive relationship**, where higher satisfaction aligns with greater sales
* In contrast, **Electronics** shows little to no correlation, suggesting other factors might drive revenue in that category.

#### 7.0.0.10 Scatterplot between Return\_Rate and Sales\_Amount

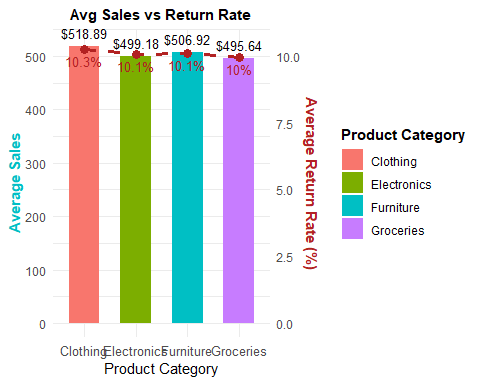
* A **scatterplot** that explores the relationship between Return\_Rate and Sales\_Amount across various product categories.
* Each point in the plot represents a product or transaction, with its color determined by its Product\_Category
* By applying transparency and minimal theming, the chart remains readable even with complex data.



* The graph reveals a **positive linear relationship** between return rates and sales amounts.
* As sales grow, return rates typically increase which are likely due to larger volumes leading to higher chances of returns.

#### 7.0.0.11 **Average Sales Revenue**

* A **dual-axis comparison** of **average sales** and **average return rate** across each product category using a combination of bar and line plots
* First, it computes the median sales and return rate per category and maps this in a bar chart for sales and a dashed line plot for return rate.
* It uses a **dual axis** system to represent both metrics clearly and formats the return rate using red indicators and percentage labels.



* The resulting graph provides categories like **Clothing** show the highest sales, they also carry **slightly higher return rates**, helping analysts **weigh profitability against risk**.

# 8. Conclusion from Exploratory Data Analysis

This comprehensive exploratory analysis of the Superstore dataset, which included data cleaning, feature engineering, and insightful visualizations, uncovered several key patterns and relationships across sales performance, return behavior, customer satisfaction, inventory efficiency, and brand dynamics. These findings provide actionable direction for business improvements across product, operations, and marketing strategies.

**Customer Satisfaction Distribution:**  
Most customers are satisfied or very satisfied, but a noticeable neutral segment reveals room for improvement.

**Return Rate Distribution:**  
Return rates are mostly moderate to high, suggesting a need for better product quality or customer alignment.

**Average Sales by Product Category:**  
Clothing generates the highest average sales, but return rates must be factored in to assess true profitability.

**Performance Index by Return Rate Group:**  
Lower return rates strongly correlate with higher product performance, while performance drops as return risk increases.

**Performance Index by Product Category:**  
Clothing and Furniture show high performance variability, while Electronics is more stable but underperforms overall.

**Inventory Stress vs Sales:**  
Higher inventory stress is linked to lower sales, indicating overstocking or weak demand.

**Customer Satisfaction by Brand:**  
Brand satisfaction varies widely within categories; “Retail X” performs well, while “Outlet Y” underperforms.

**Performance Index by Brand:**  
Brand A performs best overall, while others show inconsistency and outliers that may affect reliability.

**Satisfaction vs Sales:**  
Some categories (like Furniture) show that happier customers spend more, but this isn’t true across the board.

**Return Rate vs Sales:**  
Higher sales often bring higher returns, reinforcing the need to manage returns without hurting volume.

**Sales vs Return Rate (Dual Axis):**  
Clothing leads in sales but also has high return rates, requiring balance to protect margins.

# 9. Strategic Recommendations

* Reduce stock for underperforming items to free up space and improve sales efficiency.
* Introduce clearer expectations and better product information, especially in Clothing.
* Support consistent brands like “Retail X” and reevaluate those with weaker performance.
* Target “neutral” customers with follow-ups and incentives to turn them into loyal buyers.
* Use brand dashboards to guide promotions, pricing, and inventory planning.