

Table of Contents

| | |
|----------------------------|----|
| Introduction | 03 |
| Background | 04 |
| Objective | 05 |
| Dataset Overview | 06 |
| Data Cleaning | 08 |
| Customer Behavior Analysis | 11 |
| Product Category Analysis | 23 |
| Forecasting | 30 |
| Dashboard Visualization | 37 |
| Recommendations | 41 |
| Conclusion | 42 |

Introduction

Executive Summary

This report provides a detailed analysis of the Superstore dataset, focusing on customer behavior and product performance. The dataset did not require any data modeling as the structure was already well-organized for analysis. We performed data cleaning, developed key questions to guide the analysis, and conducted SQL queries to address both customer and product-based inquiries. Our dashboards provided a clear visual representation of the insights gained from the analysis, with the third dashboard containing filters for user interaction. Finally, we used Excel's Forecast Sheet tool to predict future trends, allowing the business to plan for upcoming sales and demand.



Key takeaways:

Customer Analysis: focused on identifying the most valuable segments and understanding how different segments behave in terms of order size, repeat purchases, and shipping preferences.

Product Analysis: Revealed the top-performing product categories and sub-categories, with a focus on seasonal spikes and regional performance.

The forecasting models: predicted future growth, particularly for the Technology and Consumer segments, enabling data-driven decisions for resource allocation and marketing strategies.



Background

The Superstore dataset is a fictional retail dataset commonly used for data analysis. It contains detailed records of customer orders, shipping preferences, and product performance. The goal of this analysis is to explore customer behavior and product sales to drive revenue growth and improve overall business performance. No data modeling was required for this dataset, as it was already well-structured for direct analysis.





Objectives:

Customer Behavior

Our mission is to understand purchasing patterns to improve engagement, retention, and revenue.

Product Categories

Assess product sales performance to identify high-revenue categories and sub-categories, enabling data-driven decisions for optimizing marketing efforts and promotional strategies.

Dataset Overview

④ Order Information:

Order ID: Unique identifier for each transaction.

Order Date and Ship Date: Essential for time-based sales and shipping performance analysis.

Ship Mode: Customer-selected delivery method (Standard Class, Second Class, etc.).

④ Customer Information:

Customer ID: Unique identifier for each customer.

Customer Name: Key for segmentation and personalized marketing.

Segment: Customer type (Consumer, Corporate, or Home Office).

Country, City, State, Region: Geographic information for location-based analysis.

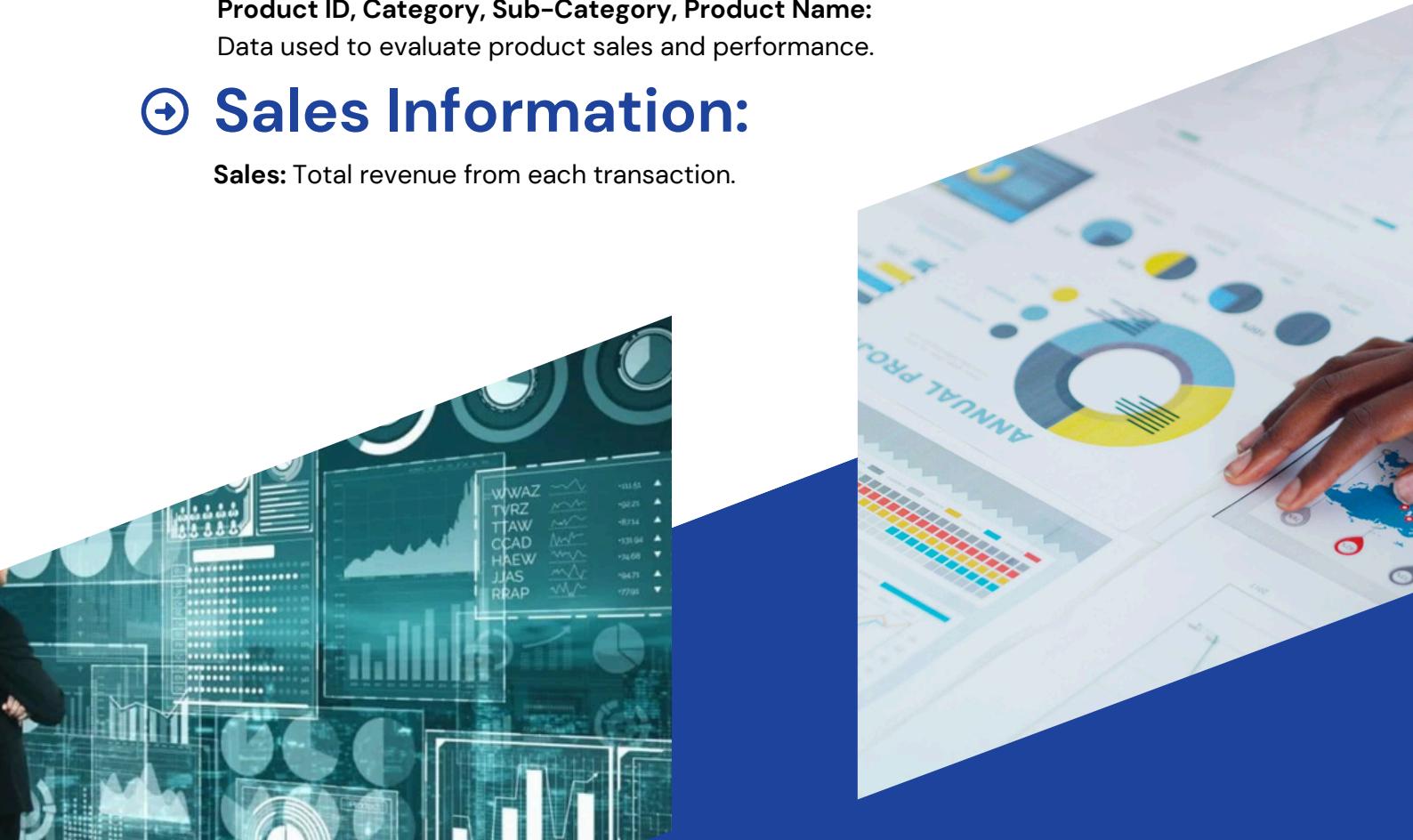
④ Product Information:

Product ID, Category, Sub-Category, Product Name:

Data used to evaluate product sales and performance.

④ Sales Information:

Sales: Total revenue from each transaction.



Our Approach



Understanding
Your Business

Creative
Execution

Data-Driven
Optimization

Continuous
Improvement

3. Data Cleaning

- Before conducting the analysis, we cleaned the dataset to ensure it was accurate and usable.

→ First: Finding Duplicates

- A check was performed to detect duplicates, and no duplicates were found. The dataset was already well-organized, as confirmed by the following Python code:

```
if df.duplicated().sum()>0:
    print('Duplicates are present')
else:
    print('No duplicates')
```

No duplicates

→ Second: Handling Null Values

- Problem:** Some customer records from Vermont were missing postal codes.
- Solution:** The mode (most frequent postal code) for Burlington, Vermont, was used to fill in the missing data:

```
burlington_data = df[df['City'] == 'Burlington'][['City', 'State', 'Postal Code']]
print(burlington_data)

      City      State  Postal Code
683  Burlington  North Carolina     27217.0
684  Burlington  North Carolina     27217.0
1008  Burlington       Iowa     52601.0
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9800 entries, 0 to 9799
Data columns (total 18 columns):
 #   Column            Non-Null Count  Dtype    
---  -- 
 0   Row ID             9800 non-null   int64    
 1   Order ID           9800 non-null   object   
 2   Order Date         9800 non-null   datetime64[ns]
 3   Ship Date          9800 non-null   datetime64[ns]
 4   Ship Mode          9800 non-null   object   
 5   Customer ID        9800 non-null   object   
 6   Customer Name      9800 non-null   object   
 7   Segment             9800 non-null   object   
 8   Country             9800 non-null   object   
 9   City                9800 non-null   object   
 10  State               9800 non-null   object   
 11  Postal Code        9789 non-null   float64  
 12  Region              9800 non-null   object   
 13  Product ID          9800 non-null   object   
 14  Category             9800 non-null   object   
 15  Sub-Category         9800 non-null   object   
 16  Product Name         9800 non-null   object   
 17  Sales                9800 non-null   float64  
dtypes: datetime64[ns](2), float64(2), int64(1), object(13)
memory usage: 1.3+ MB
```

```
burlington_data = df[df['City'] == 'Burlington'][['City', 'State', 'Postal Code']]
print(burlington_data)

      City      State  Postal Code
683  Burlington  North Carolina     27217.0
684  Burlington  North Carolina     27217.0
1008  Burlington       Iowa     52601.0
1038  Burlington  North Carolina     27217.0
1039  Burlington  North Carolina     27217.0
1393  Burlington  North Carolina     27217.0
2234  Burlington       Vermont      NaN
2928  Burlington  North Carolina     27217.0
5065  Burlington  North Carolina     27217.0
5066  Burlington  North Carolina     27217.0
5274  Burlington       Vermont      NaN
8317  Burlington  North Carolina     27217.0
8318  Burlington  North Carolina     27217.0
8410  Burlington  North Carolina     27217.0
8798  Burlington       Vermont      NaN
9146  Burlington       Vermont      NaN
9147  Burlington       Vermont      NaN
9148  Burlington       Vermont      NaN
9209  Burlington       Iowa      52601.0
9210  Burlington       Iowa      52601.0
9386  Burlington       Vermont      NaN
9387  Burlington       Vermont      NaN
9388  Burlington       Vermont      NaN
9389  Burlington       Vermont      NaN
9741  Burlington       Vermont      NaN
```

```
burlington_mode = df[df['City'] == 'Burlington'][['Postal Code']].mode()[0]
df.loc[(df['City'] == 'Burlington') & (df['State'] == 'Vermont'), 'Postal Code'] = burlington_mode
```

- **Note :** In this analysis, we used the **loc function** to update missing postal codes for Burlington, Vermont. First, we calculated the mode (most frequent value) of the postal codes for Burlington. Then, we applied loc to identify all rows where the city was Burlington and the state was Vermont, and replaced the missing values in the 'Postal Code' column with the mode. This ensured data consistency for accurate location-based analysis.

④ Third: Correcting Data Types

- The Order Date and Ship Date fields were converted to Datetime format to allow for accurate time-based analysis. Additionally, a Year column was created to analyze yearly trends:

```
df['Order Date'] = pd.to_datetime(df['Order Date'], format='%d/%m/%Y')
df['Ship Date'] = pd.to_datetime(df['Ship Date'], format='%d/%m/%Y')
```

⑤ Fourth: Outlier Detection

- Outliers were identified using boxplot analysis, and they represent exceptional transactions large sales. These outliers were analyzed further to understand their impact on business performance.

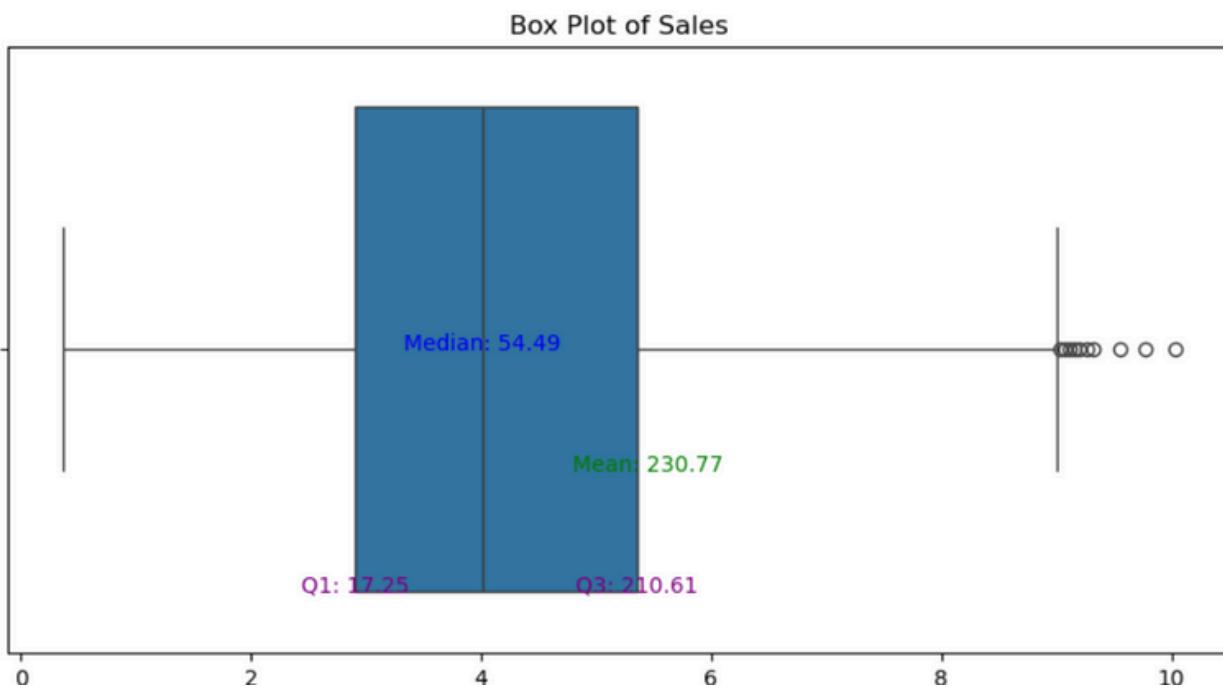
```
# Create the box plot
plt.figure(figsize=(10, 5))
sns.boxplot(x=log_sales)

# Add title and labels
plt.title("Box Plot of Sales")
#plt.xlabel("Log Sales")
#plt.ylabel("Frequency")

# Calculate the necessary statistics
min_val = np.min(df["Sales"])
max_val = np.max(df["Sales"])
q1 = df['Sales'].quantile(0.25)
q3 = df['Sales'].quantile(0.75)
mean_val = np.mean(df["Sales"])
median_val = np.median(df["Sales"])

plt.text(np.log1p(mean_val), 0.2, f'Mean: {mean_val:.2f}', horizontalalignment='center', color='green')
plt.text(np.log1p(median_val), 0, f'Median: {median_val:.2f}', horizontalalignment='center', color='blue')
plt.text(np.log1p(min_val), 0.6, f'Min: {min_val:.2f}', horizontalalignment='center', color='red')
plt.text(np.log1p(max_val), 0.6, f'Max: {max_val:.2f}', horizontalalignment='center', color='red')
plt.text(np.log1p(q1), 0.4, f'Q1: {q1:.2f}', horizontalalignment='center', color='purple')
plt.text(np.log1p(q3), 0.4, f'Q3: {q3:.2f}', horizontalalignment='center', color='purple')

plt.show()
```



- A boxplot was created for the 'Sales' column to identify outliers.

The data showed a right-skewed distribution, with outliers representing unusually large transactions, possibly from bulk orders or high-value products.

These outliers could skew the analysis by inflating average sales, leading to misinterpretations of customer behavior and product performance.

- **Possible handling methods include:**

- **Capping:** Limiting extreme values to reduce their impact.
- **Removing:** Excluding outliers from the dataset to focus on regular sales patterns.
- **Separate analysis:** Analyzing outliers separately to understand their impact on the business.

4. Analysis

- We divided the analysis into two main areas: Customer Behavior and Product Category Performance. For each area, key questions were defined to guide the SQL queries and analysis.

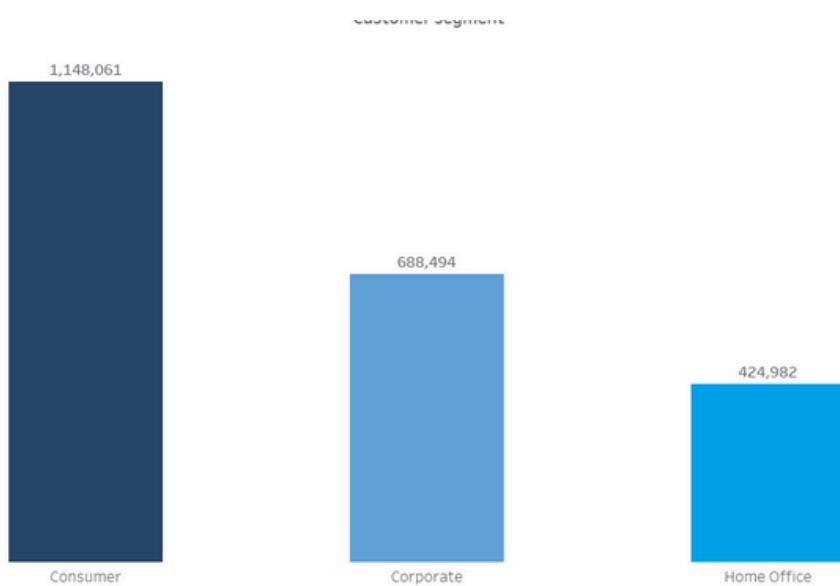
④ First: Customer Behavior Analysis

We focused on understanding how customer segments contribute to revenue, order sizes, and repeat purchases. The following questions were addressed:

1. Which customer segment contributes the most to revenue?

```
select Segment , round(sum(sales),2) Total_Sales  
from cleaned_data2  
group by segment  
order by Total_Sales Desc
```

| | Segment | Total_Sales |
|---|-------------|-------------|
| 1 | Consumer | 1148060.53 |
| 2 | Corporate | 688494.07 |
| 3 | Home Office | 424982.18 |



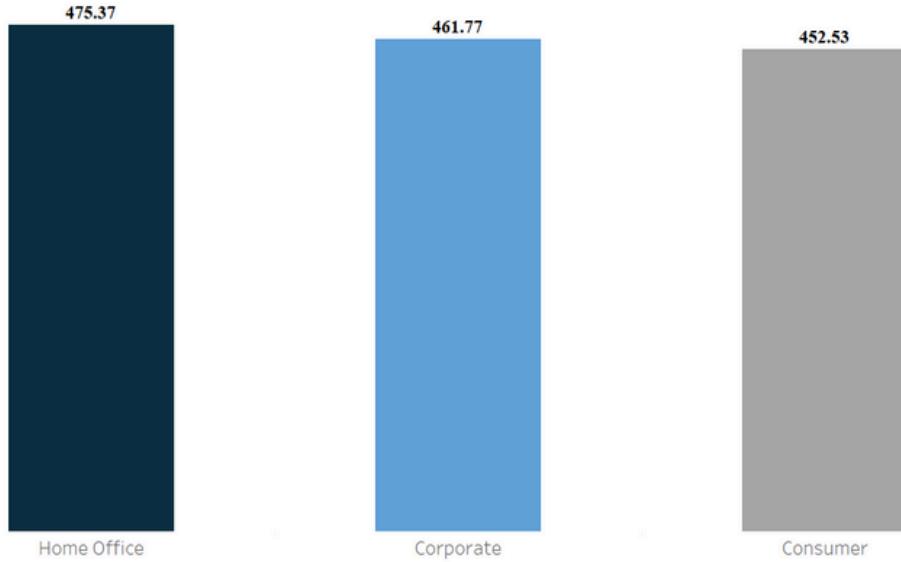
Result: The Consumer Segment contributed the most to revenue, with a total of \$1,148,060. This indicates that individual consumers drive a significant portion of sales, making them a key target for marketing efforts.

④First: Customer Behavior Analysis

What is the average order size by customer segment?

```
select Segment
      , round(Sum(sales)/COUNT(distinct [order id]),2) average_order_size
  from cleaned_data2
 group by segment
 order by average_order_size desc
```

| | Segment | average_order_size |
|---|-------------|--------------------|
| 1 | Home Office | 475.37 |
| 2 | Corporate | 461.77 |
| 3 | Consumer | 452.53 |



Result: The Home Office Segment had the highest average order size at \$475.37, indicating that businesses in this segment tend to place larger orders compared to other segments.

④ Important Note

④ Contrasting Sales Dynamics: High Total Sales in the Consumer Segment vs. Higher Average Order Size in the Home Office Segment

Consumer Segment: While the Consumer Segment has the highest total sales (\$1,148,060), this is driven by a large number of individual purchases. Consumers tend to place more frequent but smaller orders.

Home Office Segment: The Home Office Segment, on the other hand, has the highest average order size at \$475.37, indicating that although they place fewer orders, these are typically larger in value.

Relationship: The two segments complement each other. Consumers generate a high volume of transactions, while Home Office customers contribute significantly through larger, less frequent orders. This combination provides both steady revenue (from consumers) and high-value transactions (from Home Offices).

- **Insights and Opportunities:**

Marketing Strategy: Focus marketing efforts on both segments. Target consumers with promotions that encourage repeat purchases, while offering tailored incentives or loyalty programs to Home Offices to capitalize on their larger orders.

Product Bundles: For Home Office customers, offering bulk discounts or product bundles could further increase order sizes.

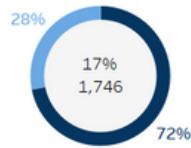
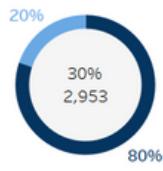
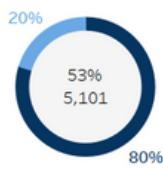
Customer Retention: Since Home Offices already place larger orders, implementing retention strategies such as personalized account management or corporate discount programs can enhance long-term business relationships.

④ First: Customer Behavior Analysis

What is the number of orders for each segment ?

```
select Segment, COUNT([Order ID]) as number_of_orders  
from cleaned_data2  
group by Segment
```

| | Segment | number_of_orders |
|---|-------------|------------------|
| 1 | Consumer | 5101 |
| 2 | Corporate | 2953 |
| 3 | Home Office | 1746 |



Result: The Home Office Segment had the highest average order size at \$475.37, indicating that businesses in this segment tend to place larger orders compared to other segments.

⌚Important Note

⌚ Understanding the Dynamics Between Consumer and Home Office Segments: Order Volume vs. Average Order Size

Finding: The Consumer Segment had the highest number of orders, indicating a strong volume of transactions, while the Home Office Segment had the highest average order size based on sales.

Difference: The Consumer Segment generates a large number of smaller transactions, reflecting frequent purchases by individual customers. In contrast, the Home Office Segment, while placing fewer orders, tends to make larger purchases, contributing to a higher average order size.

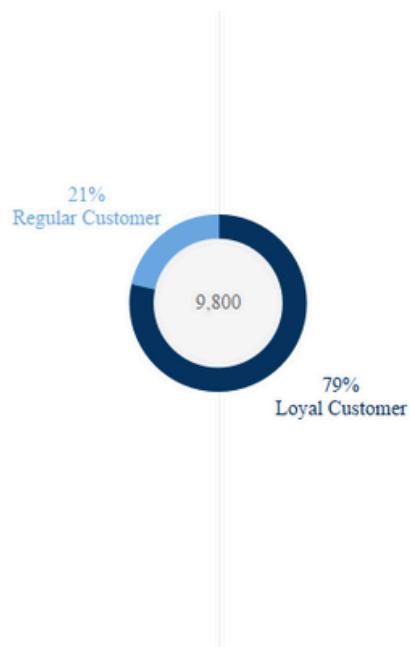
Relationship: These two segments are interconnected. The high order volume from consumers provides steady revenue, while the high average order size from Home Office customers contributes significant value per transaction. This combination enhances overall business performance.

Relevance: Understanding this relationship is crucial for effective marketing and sales strategies. While consumers drive consistent sales, targeting Home Office customers with tailored offerings can maximize revenue potential through their larger orders. Thus, analyzing both metrics is essential for informed decision-making in resource allocation and promotional efforts.

④ First: Customer Behavior Analysis

What is the Proportion of Loyal and Regular Customers Relative to Total Orders?

```
SELECT  
    [Customer Type],  
    COUNT(*) AS Customer_Count,  
    ROUND((COUNT(*) * 100.0 / (SELECT COUNT(*) FROM cleaned_data2 )), 2) AS Percentage  
FROM  
    cleaned_data2  
GROUP BY  
    [Customer Type]
```



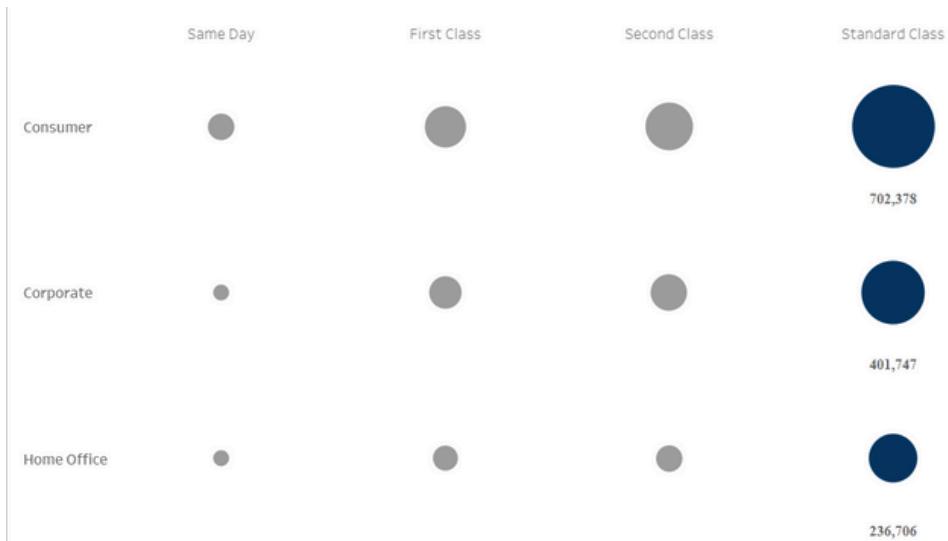
Result: The analysis revealed that the percentage of loyal customers is 79%, while regular customers account for 21%. This suggests that a significant majority of customers are repeat buyers, indicating strong customer loyalty within the Corporate Segment.

⊕ First: Customer Behavior Analysis

Do certain customer segments prefer specific shipping methods?

```
select Segment , [ship mode], count([order id]) order_count, sum(sales)
from cleaned_data2
group by Segment , [ship mode]
order by Segment , order_count desc
```

| | Segment | ship mode | order_count | total_sales |
|----|-------------|----------------|-------------|------------------|
| 1 | Consumer | Standard Class | 3031 | 702377.775271177 |
| 2 | Consumer | Second Class | 1003 | 230125.535160065 |
| 3 | Consumer | First Class | 755 | 158104.946645439 |
| 4 | Consumer | Same Day | 312 | 57452.2726843357 |
| 5 | Corporate | Standard Class | 1782 | 401747.405991375 |
| 6 | Corporate | Second Class | 589 | 139045.289665699 |
| 7 | Corporate | First Class | 468 | 102580.053216815 |
| 8 | Corporate | Same Day | 114 | 45121.3227184415 |
| 9 | Home Office | Standard Class | 1046 | 236706.125020981 |
| 10 | Home Office | Second Class | 310 | 80743.3529978991 |
| 11 | Home Office | First Class | 278 | 84887.256162405 |
| 12 | Home Office | Same Day | 112 | 22645.4428949356 |



Result: Standard Class was the preferred shipping method across all segments, but corporate clients showed a higher tendency to use First Class for faster delivery.

⊕ First: Customer Behavior Analysis

Do certain customer segments prefer specific shipping methods?

```
WITH MonthlySales AS (
  SELECT
    Segment,
    YEAR([Order Date]) AS OrderYear,
    MONTH([Order Date]) AS OrderMonth,
    SUM(Sales) AS TotalSales
  FROM
    cleaned_data2
  GROUP BY
    YEAR([Order Date]), MONTH([Order Date]), Segment
),
RankedSales AS (
  SELECT
    Segment,
    OrderYear,
    OrderMonth,
    TotalSales,
    RANK() OVER (PARTITION BY OrderYear, Segment ORDER BY TotalSales DESC) AS SalesRank
  FROM
    MonthlySales
)
```

```
SELECT
  Segment,
  OrderYear,
  OrderMonth,
  TotalSales
FROM
  RankedSales
WHERE
  SalesRank <= 3
ORDER BY
  OrderYear, Segment, TotalSales desc
```

Explanation of SQL Code

Purpose: This SQL code is designed to analyze monthly sales performance across different customer segments. Specifically, it identifies the top three segments with the highest total sales for each month and year.

How It Works:

1. MonthlySales Common Table Expression (CTE):

This section calculates the total sales for each segment, grouped by year and month.

Key Components:

YEAR([Order Date]) and MONTH([Order Date]): These functions extract the year and month from the order date, allowing for time-based aggregation.

SUM(Sales) AS TotalSales: This sums the sales figures for each combination of year, month, and segment.

The result is a summarized view of total sales per segment for each month.

2. RankedSales CTE:

This part ranks the total sales for each segment within each year.

Key Components:

RANK() OVER (PARTITION BY OrderYear, Segment ORDER BY TotalSales DESC) AS SalesRank: This function assigns a rank to each segment based on its total sales within the same year, ordering them from highest to lowest.

This ranking enables the identification of the top sales performers in each segment for every month.

3. Final Selection:

The final SELECT statement retrieves segments, years, months, and total sales from the RankedSales CTE.

The WHERE SalesRank <= 3 clause filters the results to include only the top three segments based on sales for each month and year.

The results are ordered by year, segment, and total sales in descending order to present the highest sales figures first.

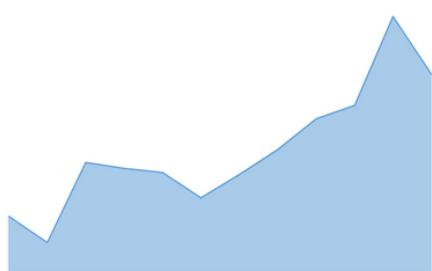
Conclusion: This SQL code provides a clear view of the best-performing segments in terms of sales on a monthly basis, enabling targeted analysis and decision-making for marketing strategies and resource allocation.

| | Segment | OrderYear | OrderMonth | TotalSales |
|----|-------------|-----------|------------|------------------|
| 1 | Consumer | 2015 | 9 | 59536.8264935017 |
| 2 | Consumer | 2015 | 11 | 46228.4978436232 |
| 3 | Consumer | 2015 | 12 | 35979.4242483377 |
| 4 | Corporate | 2015 | 11 | 20900.2307229042 |
| 5 | Corporate | 2015 | 12 | 20573.5560419559 |
| 6 | Corporate | 2015 | 9 | 15507.7449470162 |
| 7 | Home Office | 2015 | 3 | 32384.3184428215 |
| 8 | Home Office | 2015 | 12 | 11614.0779948235 |
| 9 | Home Office | 2015 | 11 | 10778.9319643974 |
| 10 | Consumer | 2016 | 12 | 39561.97108078 |
| 11 | Consumer | 2016 | 9 | 38734.313282609 |
| 12 | Consumer | 2016 | 11 | 32423.6688239574 |
| 13 | Corporate | 2016 | 11 | 27299.6443556547 |
| 14 | Corporate | 2016 | 12 | 15929.90117383 |
| 15 | Corporate | 2016 | 9 | 15268.2089176178 |
| 16 | Home Office | 2016 | 12 | 19051.7292666435 |
| 17 | Home Office | 2016 | 11 | 15526.0859971046 |
| 18 | Home Office | 2016 | 9 | 9131.08407568932 |
| 19 | Consumer | 2017 | 12 | 49294.9892009497 |
| 20 | Consumer | 2017 | 11 | 41104.392608285 |
| 21 | Consumer | 2017 | 9 | 38495.127603054 |
| 22 | Corporate | 2017 | 12 | 35335.6858446598 |
| 23 | Corporate | 2017 | 10 | 32130.5444011688 |
| 24 | Corporate | 2017 | 11 | 26493.1799249649 |
| 25 | Home Office | 2017 | 10 | 17339.598212719 |
| 26 | Home Office | 2017 | 9 | 12895.4509665966 |

| | | | | |
|----|-------------|------|----|------------------|
| 27 | Home Office | 2017 | 3 | 12644.4112565517 |
| 28 | Consumer | 2018 | 11 | 49566.4846957922 |
| 29 | Consumer | 2018 | 12 | 49433.5257019997 |
| 30 | Consumer | 2018 | 9 | 43759.9027414322 |
| 31 | Corporate | 2018 | 11 | 44357.9861465693 |
| 32 | Corporate | 2018 | 8 | 25678.4441123009 |
| 33 | Corporate | 2018 | 10 | 24877.1679894924 |
| 34 | Home Office | 2018 | 10 | 29705.5148932934 |
| 35 | Home Office | 2018 | 11 | 24013.6838222742 |
| 36 | Home Office | 2018 | 9 | 19183.8259179592 |



Consumer Segment



Corporate Segment



Home Office Segment

Result: We identified seasonal spikes in sales, particularly during the back-to-school period in September and the holiday season in November and December. These patterns were consistent across all segments but were particularly strong in the Consumer Segment.

⊕ First: Customer Behavior Analysis

Which sub-categories have the highest sales among each customer segment?

```
WITH RankedSubCategories AS (
    SELECT
        Segment,
        [Sub-Category],
        SUM(Sales) AS Total_Sales,
        ROW_NUMBER() OVER (PARTITION BY Segment ORDER BY SUM(Sales) DESC) AS Rank
    FROM cleaned_data2
    GROUP BY
        Segment,
        [Sub-Category]
)
SELECT
    Segment,
    [Sub-Category],
    Total_Sales
FROM
    RankedSubCategories
WHERE
    Rank = 1;
```

| | Segment | Sub-Category | Total_Sales |
|---|-------------|--------------|------------------|
| 1 | Consumer | Chairs | 171174.095386505 |
| 2 | Corporate | Chairs | 95203.1516876221 |
| 3 | Home Office | Phones | 68209.1501820087 |

Purpose: This SQL code is designed to analyze sales data to identify the highest-performing sub-categories within each customer segment based on total sales.

How It Works:

1. RankedSubCategories Common Table Expression (CTE):

This section aggregates sales data by segment and sub-category to calculate total sales.

Key Components:

SUM(Sales) AS Total_Sales: This function computes the total sales for each combination of segment and sub-category.

ROW_NUMBER() OVER (PARTITION BY Segment ORDER BY SUM(Sales) DESC) AS Rank: This function assigns a rank to each sub-category within its segment based on total sales, ordering from highest to lowest.

The result is a summarized view of total sales for each sub-category within each segment, along with a ranking that indicates the top sub-category.

2. Final Selection:

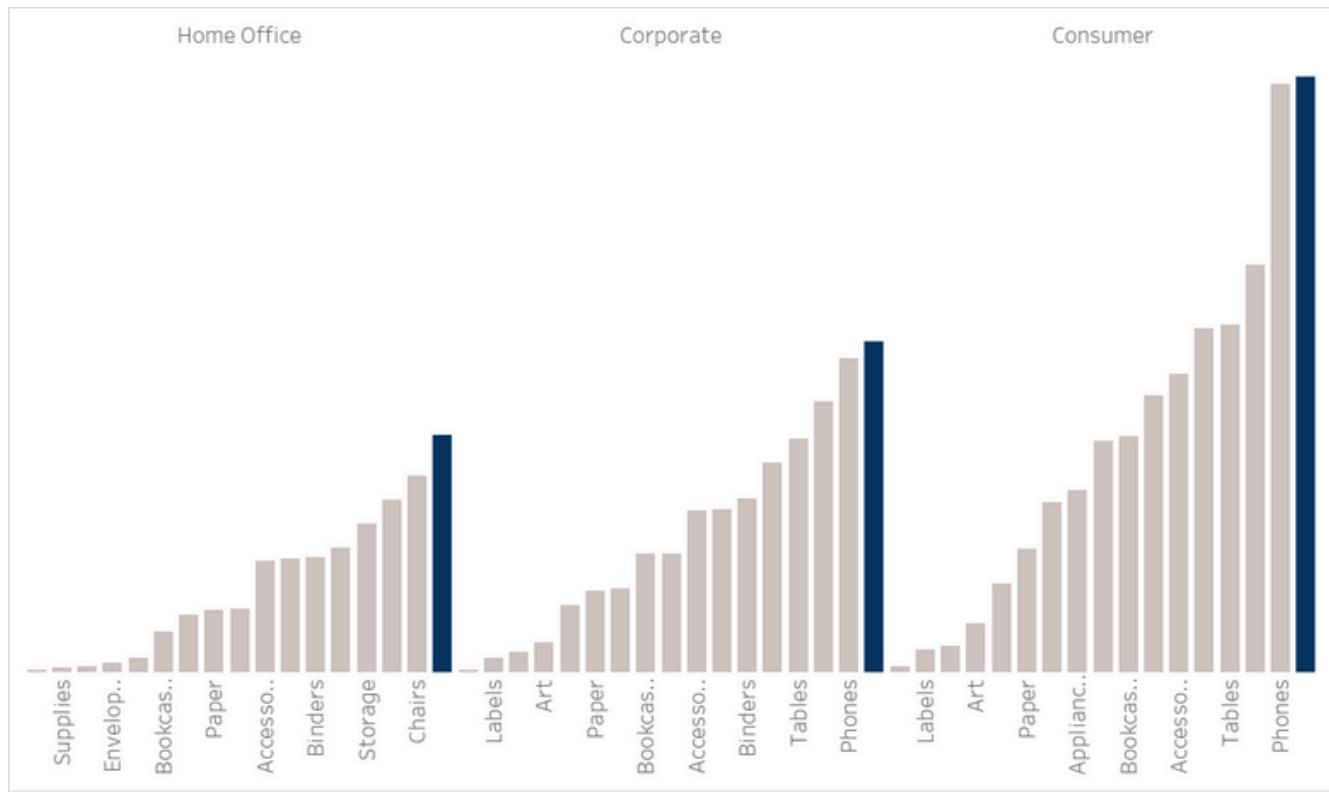
The final SELECT statement retrieves the segment, sub-category, and total sales from the RankedSubCategories CTE.

The WHERE Rank = 1 clause filters the results to include only the top sub-category for each segment based on total sales.

Results:

The analysis reveals the following top sub-categories by total sales:

- In the Consumer Segment, the highest sub-category is Chairs, with total sales of **\$171,174**.
- In the Corporate Segment, the highest sub-category is also Chairs, with total sales of **\$95,203**.
- In the Home Office Segment, the highest sub-category is Phones, with total sales of **\$68,209**.



Results:

The analysis reveals the following top sub-categories by total sales:

- In the Consumer Segment, the highest sub-category is Chairs, with total sales of **\$171,174**.
- In the Corporate Segment, the highest sub-category is also Chairs, with total sales of **\$95,203**.
- In the Home Office Segment, the highest sub-category is Phones, with total sales of **\$68,209**

④ Second: Product Analysis

which product categories are the most profitable?

```
select category
      , round(sum(sales),2) Total_Sales
from cleaned_data2
group by category
order by Total_Sales desc
```

| | category | Total_Sales |
|---|-----------------|-------------|
| 1 | Technology | 827455.87 |
| 2 | Furniture | 728658.57 |
| 3 | Office Supplies | 705422.33 |



Results: Technology was the most profitable product category, followed by Furniture.

④ Second: Product Analysis

what are the best selling sub categories within each product category

```
] WITH RankedSales AS (
SELECT Category, [Sub-Category], ROUND(SUM(sales), 2) AS total_sales,
ROW_NUMBER() OVER (PARTITION BY Category ORDER BY SUM(sales) DESC) AS sales_rank
FROM cleaned_data2
GROUP BY Category, [Sub-Category])
SELECT Category, [Sub-Category], total_sales
FROM RankedSales
WHERE sales_rank = 1
ORDER BY total_sales DESC
```

| | Category | Sub-Category | total_sales |
|---|-----------------|--------------|-------------|
| 1 | Technology | Phones | 327782.45 |
| 2 | Furniture | Chairs | 322822.73 |
| 3 | Office Supplies | Storage | 219343.39 |

Purpose: This SQL code is designed to identify the highest-performing sub-categories within each product category based on total sales.

How It Works:

1. RankedSales Common Table Expression (CTE):

This section aggregates sales data by sub-category and category to compute total sales for each sub-category.

Key Components:

ROUND(SUM(sales), 2) AS total_sales: This function calculates the total sales for each sub-category and rounds the result to two decimal places for better readability.

ROW_NUMBER() OVER (PARTITION BY Category ORDER BY SUM(sales) DESC) AS sales_rank: This function assigns a rank to each sub-category within its respective category, ordering them from highest to lowest total sales.

The result is a detailed view of total sales for each sub-category within each category, along with a ranking indicating the top sub-category.

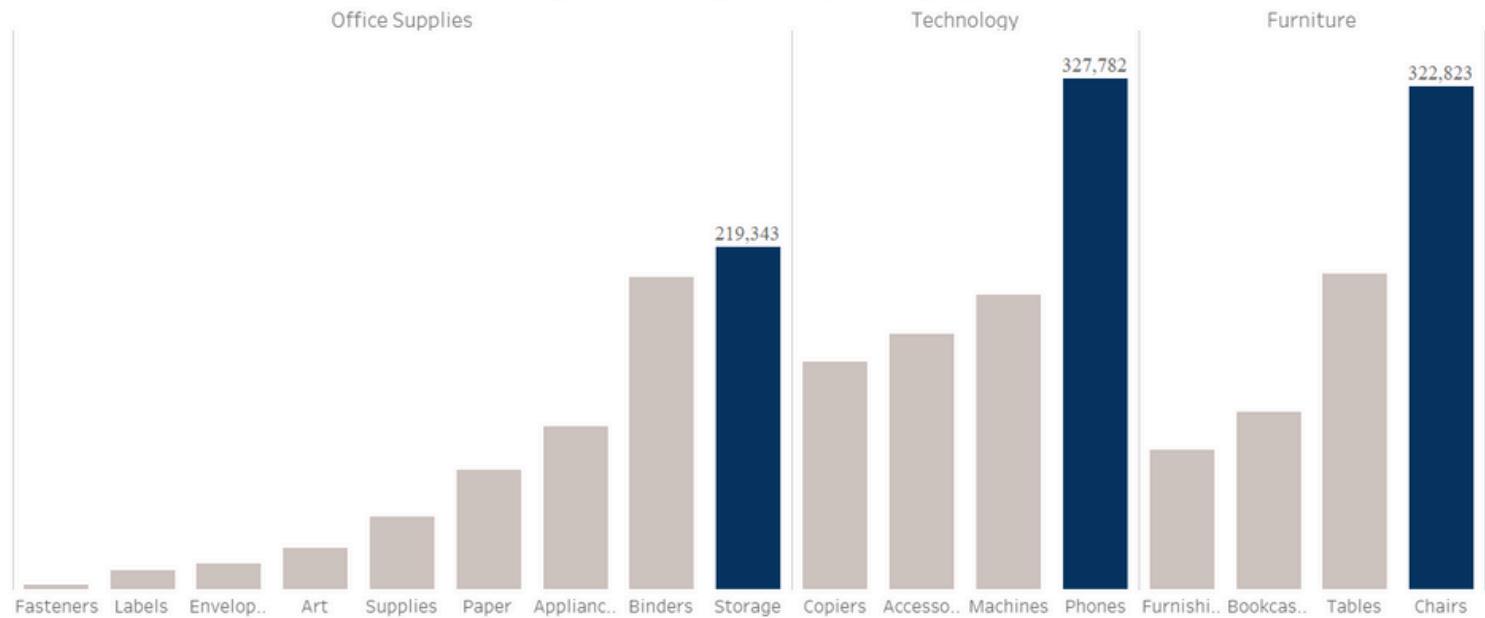
2. Final Selection:

The final SELECT statement retrieves the sub-category, category, and total sales from the RankedSales CTE. The WHERE sales_rank = 1 clause filters the results to include only the top sub-category for each product category based on total sales.

The results are ordered by total sales in descending order to present the highest sales figures first.

Results: Technology was the most profitable product category, followed by Furniture.

Top Sub-Categories by Category



Results:

The analysis reveals the following top sub-categories by total sales:

In the Technology category, the highest sub-category is Phones, with total sales of **\$327,782.44**.

In the Furniture category, the highest sub-category is Chairs, with total sales of **\$322,822.73**.

In the Office Supplies category, the highest sub-category is Storage, with total sales of **\$219,343.39**.

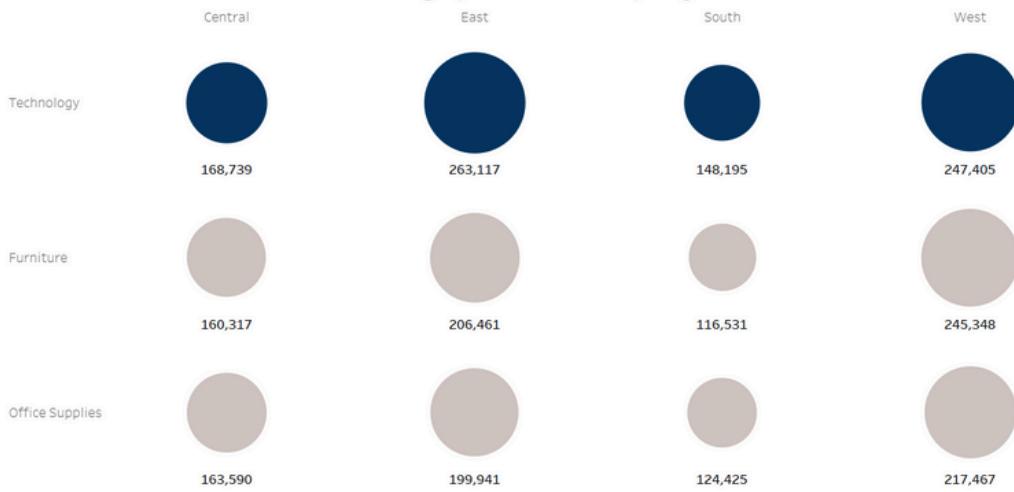
④ Second: Product Analysis

How do product categories perform across different regions?

```
select Region
    ,Category
    ,round(sum(sales),2) Total_Sales
    ,DENSE_RANK() over(partition by Region order by round(sum(sales),2) desc) as Sales_Rank
from cleaned_data2
group by Region
    ,Category
order by Region , Sales_Rank
```

| | Region | Category | Total_Sales | Sales_Rank |
|----|---------|-----------------|-------------|------------|
| 1 | Central | Technology | 168739.21 | 1 |
| 2 | Central | Office Supplies | 163590.24 | 2 |
| 3 | Central | Furniture | 160317.46 | 3 |
| 4 | East | Technology | 263116.53 | 1 |
| 5 | East | Furniture | 206461.39 | 2 |
| 6 | East | Office Supplies | 199940.81 | 3 |
| 7 | South | Technology | 148195.21 | 1 |
| 8 | South | Office Supplies | 124424.77 | 2 |
| 9 | South | Furniture | 116531.48 | 3 |
| 10 | West | Technology | 247404.93 | 1 |
| 11 | West | Furniture | 245348.25 | 2 |
| 12 | West | Office Supplies | 217466.51 | 3 |

Category Performance by Region



Result: Technology performed best in all regions, but Furniture showed stronger performance in the West and Central regions.

④ Second: Product Analysis

Are there any notable sales spikes during specific months for certain products?

```

WITH MonthlySales AS (
    SELECT
        Category,
        YEAR([Order Date]) AS OrderYear,
        MONTH([Order Date]) AS OrderMonth,
        SUM(Sales) AS TotalSales
    FROM
        cleaned_data2
    GROUP BY
        YEAR([Order Date]), MONTH([Order Date]), Category
),
RankedSales AS (
    SELECT
        Category,
        OrderYear,
        OrderMonth,
        TotalSales,
        RANK() OVER (PARTITION BY OrderYear, Category ORDER BY TotalSales DESC) AS SalesRank
    FROM
        MonthlySales
)

```

```

SELECT
    Category,
    OrderYear,
    OrderMonth,
    TotalSales
FROM
    RankedSales
WHERE
    SalesRank <= 3
ORDER BY
    OrderYear, Category, TotalSales DESC;

```

Purpose: This SQL query is used to analyze the top-performing categories by sales within each year, identifying the months with the highest sales for each category.

1. MonthlySales Common Table Expression (CTE):

This part of the query aggregates the total sales for each category on a monthly basis. It groups the data by category, year, and month, and calculates the sum of sales.

Key Components:

YEAR([Order Date]) AS OrderYear and MONTH([Order Date]) AS OrderMonth: These functions extract the year and month from the order date.

SUM(Sales) AS TotalSales: This function calculates the total sales for each category per month.

The data is grouped by year, month, and category to provide a monthly sales breakdown..

2. RankedSales Common Table Expression (CTE):

This section ranks the sales figures for each category within each year. The ranking helps identify the months with the highest sales within a specific year.

Key Components:

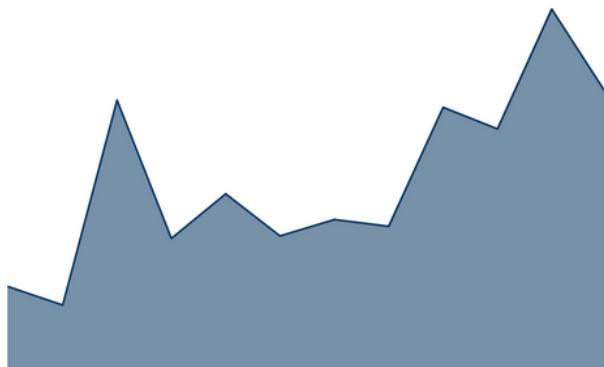
RANK() OVER (PARTITION BY OrderYear, Category ORDER BY TotalSales DESC): This ranking function assigns a rank to each month within the same year and category, based on total sales in descending order. The month with the highest sales receives a rank of 1.

3. Final Selection:

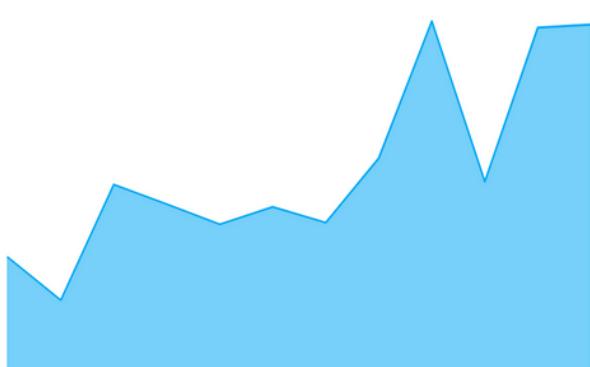
The final SELECT statement retrieves the category, year, month, and total sales from the RankedSales CTE. The WHERE SalesRank <= 3 clause ensures that only the top three months with the highest sales for each category within each year are selected.

The results are ordered by year, category, and total sales in descending order to present the highest sales figures first.

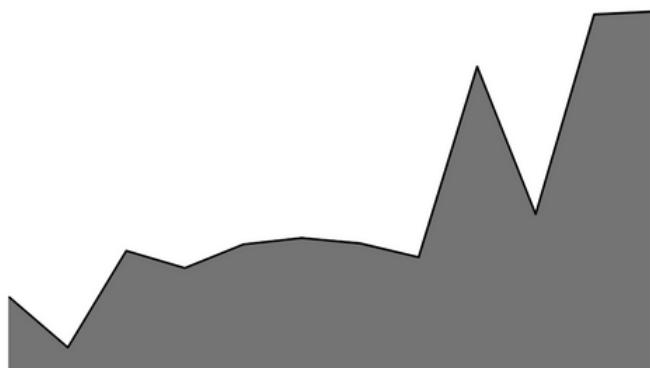
| | Category | OrderYear | OrderMonth | TotalSales |
|----|-----------------|-----------|------------|------------------|
| 1 | Furniture | 2015 | 12 | 30637.3423743248 |
| 2 | Furniture | 2015 | 9 | 23816.4809741974 |
| 3 | Furniture | 2015 | 11 | 21471.0407242775 |
| 4 | Office Supplies | 2015 | 9 | 27423.2981255651 |
| 5 | Office Supplies | 2015 | 11 | 26363.1959513426 |
| 6 | Office Supplies | 2015 | 12 | 16956.4919661283 |
| 7 | Technology | 2015 | 3 | 32359.9744105339 |
| 8 | Technology | 2015 | 9 | 30383.7482056618 |
| 9 | Technology | 2015 | 11 | 30073.4238553047 |
| 10 | Furniture | 2016 | 11 | 30197.5003643036 |
| 11 | Furniture | 2016 | 9 | 25085.4230618477 |
| 12 | Furniture | 2016 | 12 | 22812.2513747215 |
| 13 | Office Supplies | 2016 | 11 | 21178.2980004549 |
| 14 | Office Supplies | 2016 | 9 | 19031.0550745726 |
| 15 | Office Supplies | 2016 | 12 | 16099.3219621181 |
| 16 | Technology | 2016 | 12 | 35632.0281844139 |
| 17 | Technology | 2016 | 11 | 23873.6008119583 |
| 18 | Technology | 2016 | 9 | 19017.1281394959 |
| 19 | Furniture | 2017 | 12 | 36604.9647674561 |
| 20 | Furniture | 2017 | 11 | 31783.6286377907 |
| 21 | Furniture | 2017 | 9 | 26755.4747447968 |
| 22 | Office Supplies | 2017 | 12 | 37332.9382256269 |
| 23 | Office Supplies | 2017 | 9 | 23013.4018874168 |
| 24 | Office Supplies | 2017 | 11 | 20141.8080967665 |
| 25 | Technology | 2017 | 10 | 31533.3734617233 |
| 26 | Technology | 2017 | 5 | 28832.6903800964 |



Technology



Office Supplies



Furniture

September: Sales peak in September, likely due to the back-to-school season, indicating a demand for products such as technology and office supplies during this time.

November and December: Sales spike in both November and December, reflecting the holiday shopping season. This suggests that consumers are more likely to purchase high-value items, especially in categories like technology and furniture, during this period.

④ Optimizing Sales During Key Seasonal Peaks: September, November, and December

Maximizing September Sales: Focus on promoting back-to-school deals in September, particularly for technology and office supplies. Offering discounts or bundling products for students and professionals can help capture this seasonal demand.

Leveraging November and December: For November and December, launch holiday promotions and marketing campaigns that highlight high-value items like technology and furniture. Consider special holiday discounts, gift bundles, and targeted ads to capitalize on increased consumer spending during the holiday season.

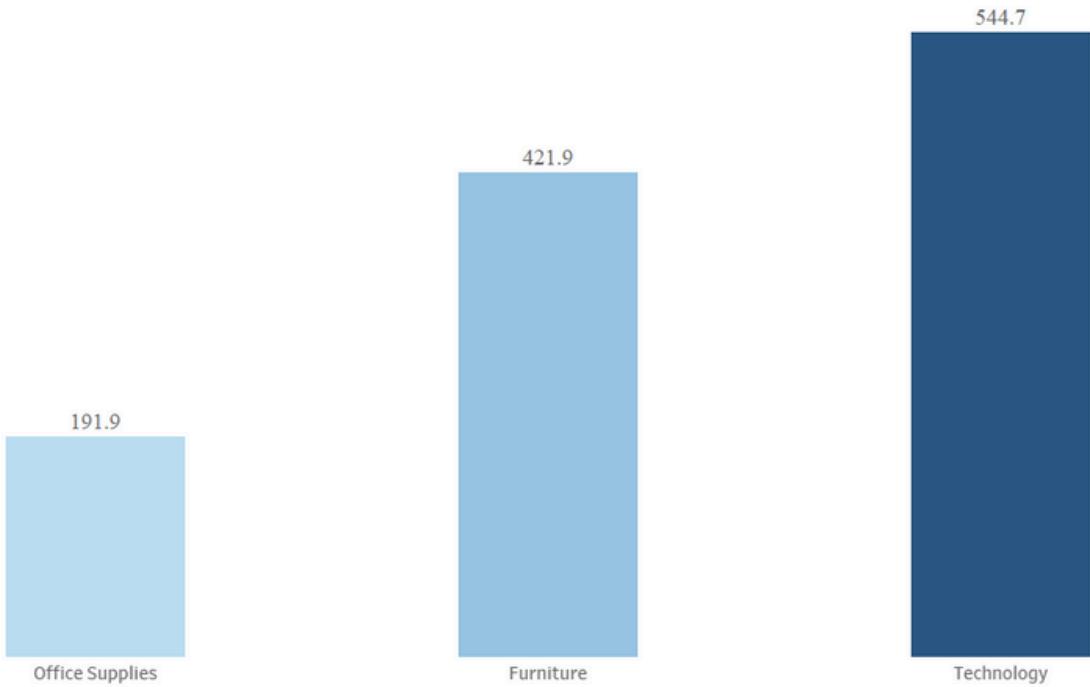


④ Second: Product Analysis

How do average sales per order vary by product category?

```
SELECT Category, ROUND(SUM(sales) / COUNT(DISTINCT [Order ID]), 2) AS avg_sales_per_order
FROM cleaned_data2
GROUP BY Category
ORDER BY avg_sales_per_order DESC
```

| | Category | avg_sales_per_order |
|---|-----------------|---------------------|
| 1 | Technology | 544.74 |
| 2 | Furniture | 421.92 |
| 3 | Office Supplies | 191.9 |



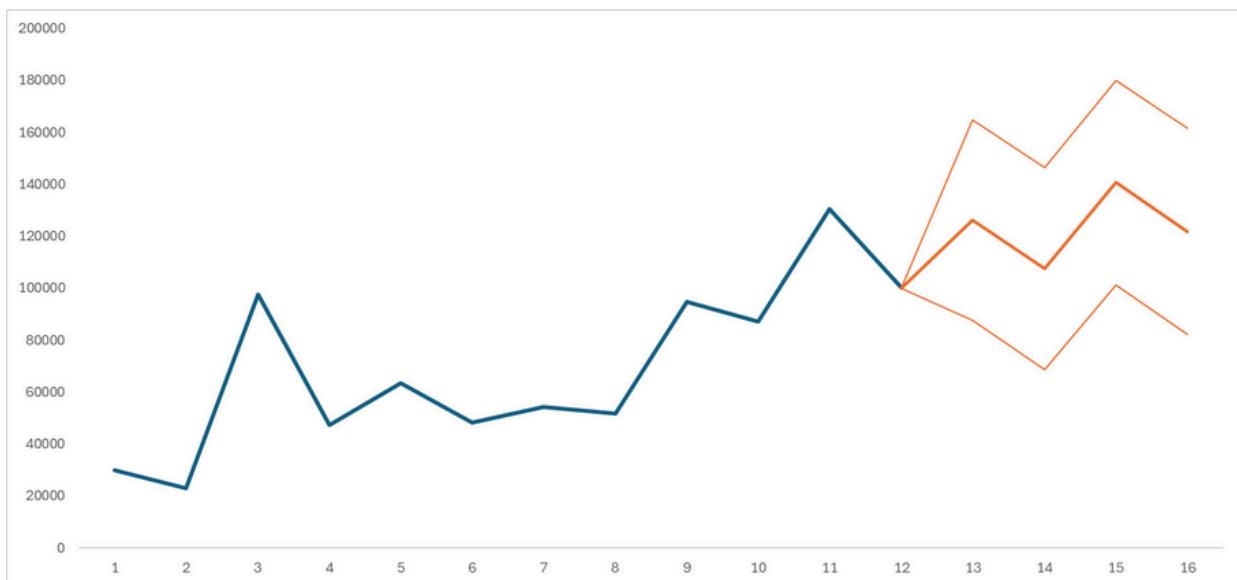
Result: Technology had the highest average sales per order, followed by Furniture, reflecting the higher price points of items in these categories.

④ Third: Forecasting

For forecasting, we used Excel's Forecast Sheet tool to predict future sales trends. The tool allowed us to generate time-series forecasts with confidence bounds.

Technology Forecast:

| Timeline | Technology | Forecast | Lower Confidence Bound | Upper Confidence Bound |
|----------|------------|-----------|------------------------|------------------------|
| 1 | 29774.93 | | | |
| 2 | 23084.84 | | | |
| 3 | 97500.39 | | | |
| 4 | 47285.6 | | | |
| 5 | 63450.19 | | | |
| 6 | 48180.72 | | | |
| 7 | 54131.39 | | | |
| 8 | 51673.26 | | | |
| 9 | 94854.05 | | | |
| 10 | 87031.95 | | | |
| 11 | 130497.2 | | | |
| 12 | 99991.38 | 99991.38 | 99991.38 | 99991.38 |
| 13 | | 126222.05 | 87585.19 | 164858.92 |
| 14 | | 107528.76 | 68581.56 | 146475.97 |
| 15 | | 140582.66 | 101317.84 | 179847.49 |
| 16 | | 121889.37 | 82309.47 | 161469.27 |



④ Third: Forecasting

The forecasting chart and the accompanying table provide insights into the future sales performance of the Technology segment. The data predicts a general upward trend in technology sales, with the highest forecasted value occurring in March (140,582 \$). This steady growth indicates increasing demand for technology products over time, with March being a particularly strong month for sales.

First: Technology Forecasting

Key Insights from the Forecasting Chart:

1. Steady Growth Pattern:

- The blue line in the chart shows a consistent increase in sales over the forecast period, highlighting that demand for technology products is expected to rise steadily.
- This gradual growth aligns with technological advancements and the increasing need for tech products in both the consumer and business sectors.

2. March as a Peak Month:

- According to both the chart and the table, March is predicted to be the peak month for technology sales, with a forecasted value of 140,582.
- This could be attributed to multiple factors, such as corporate budget cycles, tech refreshes, and new product launches around this time.

3. Confidence Bounds:

- The orange lines in the chart represent the confidence intervals, showing the range within which the actual sales values are expected to fall.
- For example, in March, the lower bound of the prediction is 101,317 units, and the upper bound is 179,847 units. This indicates a level of uncertainty, which widens over time, suggesting that while March is likely to be a strong sales month, external factors could impact the exact numbers.

④ Third: Forecasting

4. Increasing Uncertainty:

- As shown in both the chart and the table, the confidence intervals widen as the forecast extends into the future, particularly from March onwards. This suggests that predictions become less certain over time due to market fluctuations, external factors, or unpredictable demand changes.

5. Actionable Recommendations:

Marketing & Sales Strategy for March:

- Given the peak in March sales, the company should prepare targeted marketing campaigns and ensure sufficient inventory is available during this period. Promotional efforts can be aligned with product launches or seasonal demand spikes.

Managing Uncertainty:

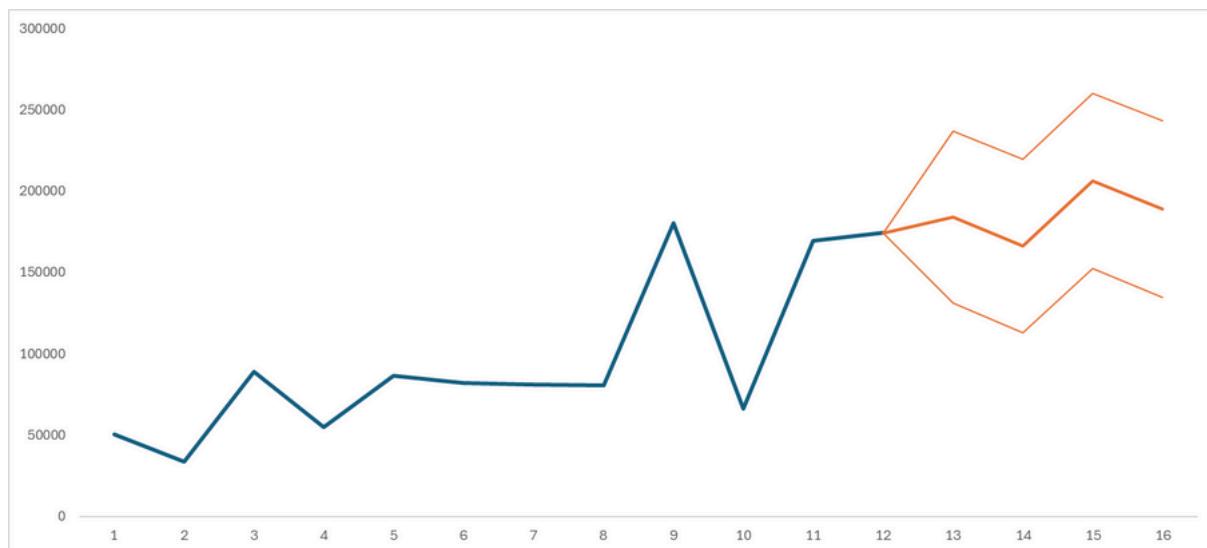
- Since the forecast shows increasing uncertainty in the later months, it is crucial to stay agile with inventory management and promotional tactics. Continuous monitoring of market trends and demand will help mitigate risks associated with this unpredictability.
- By leveraging these insights, the company can capitalize on the forecasted demand, particularly in high-sales periods like March, while preparing for potential fluctuations in the market as indicated by the widening confidence intervals.

④ Third: Forecasting

For forecasting, we used Excel's Forecast Sheet tool to predict future sales trends. The tool allowed us to generate time-series forecasts with confidence bounds.

Technology Forecast:

| Timeline | consum | Forecast | Lower Confidence Bound | Upper Confidence Bound |
|----------|----------|----------|------------------------|------------------------|
| 1 | 50273.9 | | | |
| 2 | 33446.04 | | | |
| 3 | 89174.71 | | | |
| 4 | 54774.92 | | | |
| 5 | 86406.5 | | | |
| 6 | 81911.22 | | | |
| 7 | 81008.54 | | | |
| 8 | 80591.67 | | | |
| 9 | 180526.2 | | | |
| 10 | 66353.9 | | | |
| 11 | 169323 | | | |
| 12 | 174269.9 | 174269.9 | 174269.91 | 174269.91 |
| 13 | | 183965.6 | 130988.96 | 236942.25 |
| 14 | | 166370.6 | 112968.39 | 219772.72 |
| 15 | | 206399.1 | 152561.44 | 260236.78 |
| 16 | | 188804.1 | 134534.38 | 243073.74 |



④ Third: Forecasting

The forecasting chart and the data table give us valuable insights into the expected consumption patterns over time for the consumer segment. The forecast shows a significant variation in future consumption with a consistent upward trajectory, particularly from December onwards. This rise in consumption is reflected in both the forecasted values and the increasing confidence intervals.

Second: Consumer Forecasting

Key Insights from the Forecasting Chart:

1. Rising Trend in Consumption:

- The initial data reveals fluctuations in consumer consumption over time. However, starting from the 12th timeline, we see a substantial jump in consumption, indicated by the blue line in the chart. This rise reflects growing consumer demand in the upcoming months.
- Insight: This pattern suggests that consumer demand is expected to accelerate, signaling that now might be the time to prepare for higher inventory levels or marketing campaigns targeted at these peak periods.

2. Confidence Intervals (Uncertainty):

- The orange lines on the chart represent the confidence bounds, displaying the upper and lower ranges for the forecast. As the timeline progresses, these intervals widen, indicating increased uncertainty regarding future consumption levels.

④ Third: Forecasting

- For example, in the 15th timeline mark, the lower bound is approximately 152,561 units, while the upper bound extends to 260,236 units. This wide range suggests considerable variability in the expected outcome.
- Actionable Insight: To mitigate risks from this uncertainty, businesses should employ flexible strategies, allowing for rapid response to fluctuations, particularly in periods of volatile demand.

3. Key Consumption Peaks:

- From the table, we can identify that the 15th timeline has one of the highest forecasted consumptions, reaching approximately 206,399 units. This spike suggests that this period is expected to have the highest consumer demand, which could be influenced by seasonal trends, promotions, or external factors like economic growth.
- Insight: It's vital to ensure that stock levels and supply chain operations are optimized to meet this potential surge in demand.

4. Gradual Widening of Forecast Range:

- The forecast range widens after the 12th timeline, suggesting increasing uncertainty about future consumption patterns. This uncertainty could be due to changes in market conditions, consumer preferences, or unforeseen economic factors.
- Actionable Insight: Companies should monitor real-time data and market trends to adjust to potential disruptions. A robust strategy for forecasting updates will allow for greater adaptability and more precise future projections.

④ Third: Forecasting

Actionable Recommendations:

1. Inventory & Sales Strategy:

- Given the spike in consumption forecasted from the 12th to the 16th timeline, it is crucial to align inventory management and sales efforts accordingly. Ensuring ample stock and supply chain readiness will allow you to meet this predicted surge in demand without facing stockouts or lost sales opportunities.

2. Prepare for Demand Variability:

- With the confidence bounds widening over time, a flexible approach to inventory management is essential. Consider implementing dynamic pricing strategies or promotional activities to stimulate demand during periods of uncertainty.

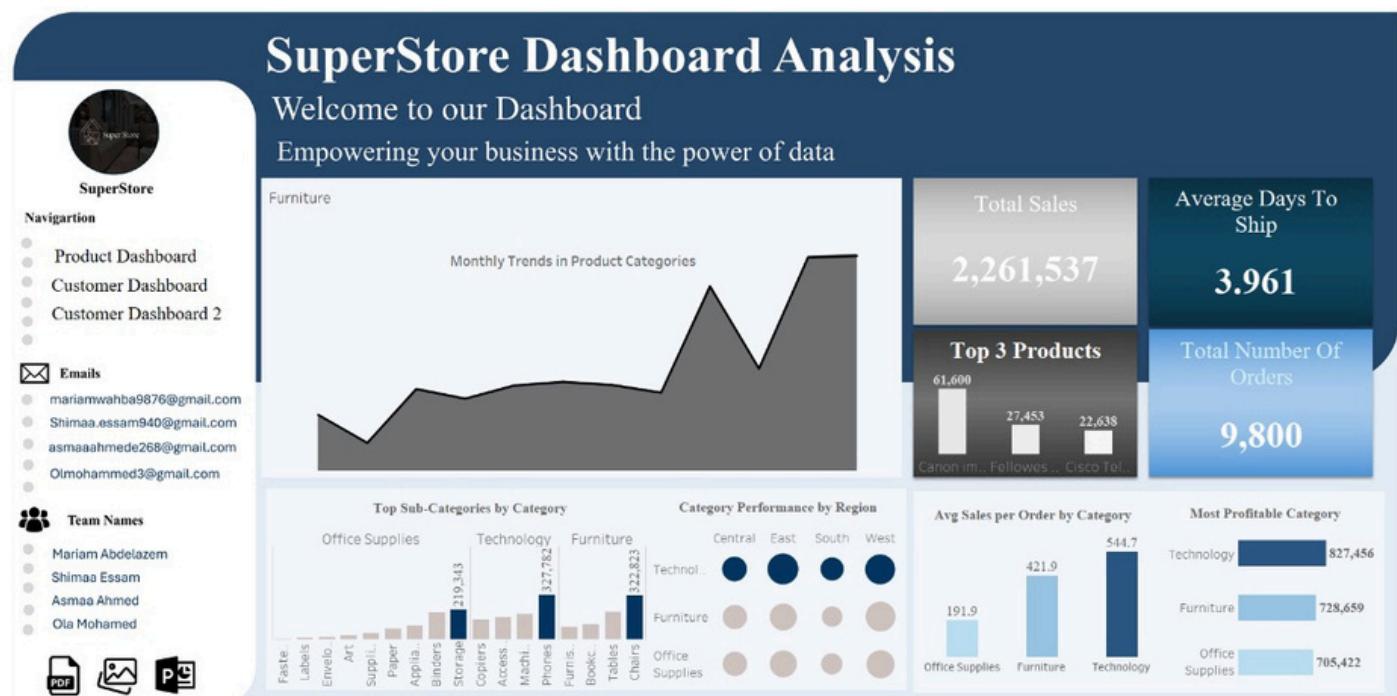
3. Agility in Marketing:

- Focus on targeted marketing initiatives in high-demand periods, particularly during the 12th to 16th timeline, when consumer activity is expected to peak. Customizing marketing campaigns during this time can maximize sales and customer engagement.

④ Fourth : Dashboard Visualization

- We created three interactive dashboards using Tableau to visualize the insights gained from the analysis. These dashboards provided stakeholders with an intuitive way to explore customer and product data:

1. Product Category Dashboard:



- This dashboard provides a comprehensive overview of sales performance across different product categories, including Technology, Furniture, and Office Supplies. It visualizes total sales, average sales per order, and trends across various regions.

- The chart on "Monthly Trends in Product Categories" allows users to observe patterns and spikes in sales, helping identify periods of high demand. The "Top 3 Products" section highlights the best-performing items, offering insight into which specific products drive the most revenue.
- The "Most Profitable Categories" bar chart reveals which categories contribute most to overall profitability, with Technology leading, followed by Furniture and Office Supplies. This visual insight can guide inventory and marketing strategies. The dashboard lacks interactive filters but provides a detailed, high-level view of product performance and trends.

2. Customer Behavior Dashboard:



- This dashboard focuses on segmenting customers based on purchase behavior, order size, and preferences. It aims to identify the most profitable customer segments and understand their buying patterns.

- It includes visualizations on "Top 3 Customers," showing the customers who bring the most revenue, alongside a breakdown of "Orders by Customer Segment," providing a snapshot of how different segments (Consumer, Corporate, and Home Office) contribute to total sales.
- The "Shipping Preferences by Customer Segment" chart shows preferred shipping methods, indicating that Corporate clients often choose faster, premium options, while Consumers may opt for standard delivery. Additionally, a pie chart on "Customer Types: Regular vs. Loyal" shows the proportion of loyal customers, emphasizing the importance of customer retention.

The dashboard provides high-level insights without filters, enabling businesses to quickly identify opportunities for targeted marketing and promotions.

3. Customer Records Dashboard:

Superstore Dashboard Analysis

SuperStore
Navigation

- Product Dashboard
- Customer Dashboard
- Customer Dashboard 2

- Emails**
- mariamwahba9876@gmail.com
- Shimaa.essam940@gmail.com
- asmaaahmed268@gmail.com
- Olmohammed3@gmail.com

- Team Names**
- Mariam Abdelazem
- Shimaa Essam
- Asmaa Ahmed
- Ola Mohamed

Customer Name
All

Customer Type
All

Category
All

Sub-Category
All

Region
All

Segment
All

| Order ID | Customer Details | Customer Type | Sales Data |
|----------|--|--------------------------------------|-----------------------------|
| AA-10315 | Alex Avila New York New York City | Regular Customer total orders: 11 | Consumer sales: 5,564 |
| AA-10375 | Allen Arnold California Rancho Cucamonga | Loyal Customer total orders: 15 | Consumer sales: 1,056 |
| AA-10480 | Andrew Allen Connecticut Middletown | Regular Customer total orders: 12 | Consumer sales: 1,791 |
| AA-10645 | Anna Andreadil Oklahoma Oklahoma City | Loyal Customer total orders: 18 | Consumer sales: 5,087 |
| AB-10015 | Aaron Bergman Oklahoma Oklahoma City | Regular Customer total orders: 6 | Consumer sales: 886 |
| AB-10060 | Adam Bellavance Pennsylvania Philadelphia | Loyal Customer total orders: 18 | Home Office sales: 7,756 |
| AB-10105 | Adrian Barton New York New York City | Loyal Customer total orders: 20 | Consumer sales: 14,474 |
| AB-10150 | Almee Bixby New York Long Beach | Regular Customer total orders: 12 | Consumer sales: 967 |

- Unlike the other two, this dashboard includes interactive filters, enabling a more in-depth exploration of customer data. Users can filter by customer name, region, customer type, product category, and segment, providing the flexibility to drill down into individual behaviors and purchase histories.
- The detailed customer list shows information like customer name, location, order history, and sales amounts. This helps identify high-value customers and analyze their purchasing habits, making it easier to customize marketing strategies and improve customer service.
- By filtering the data, users can also study specific product categories or customer segments to see which products resonate most with different audiences. This enables a deeper understanding of trends and patterns, helping in strategic decision-making around inventory, promotions, and customer engagement.

These dashboards collectively give Superstore a holistic view of its sales performance, customer behavior, and key areas for strategic focus, supporting the business in making data-driven decisions.

Recommendations

→ Focus on Profitable Categories

The Technology and Furniture categories consistently performed well across all regions. By increasing inventory and marketing efforts for these categories, Superstore can capitalize on their high profitability.

→ Capitalize on Seasonal Sales Spikes

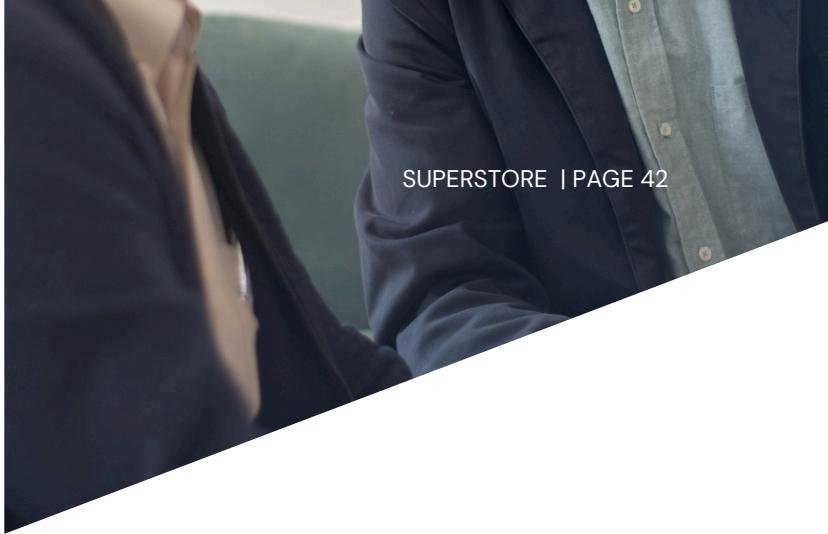
The data shows clear sales spikes in September, November, and December, driven by back-to-school shopping and holiday promotions. To maximize sales during these periods, Superstore should offer targeted discounts and marketing campaigns focused on high-demand products like technology and office furniture.

→ Enhance Customer Retention:

The analysis revealed that Corporate and Home Office segments are more likely to place repeat orders and larger orders on average. By developing loyalty programs or offering targeted promotions, Superstore can encourage these segments to continue making repeat purchases.

→ Optimize Shipping Strategies

The Consumer and Corporate segments show distinct shipping preferences, with Corporate clients favoring faster shipping methods. Tailoring shipping options based on customer segment preferences can improve satisfaction and streamline logistics.



Conclusion

- This report provides a comprehensive analysis of the Superstore dataset, focusing on customer behavior, product performance, and sales forecasting. By understanding key customer segments and optimizing product categories, Superstore can make informed decisions that drive revenue growth and improve customer satisfaction.
- The dashboards and forecasting models developed in this project offer powerful tools for future decision-making, helping to predict demand, manage inventory, and focus marketing efforts on the most profitable areas of the business. By implementing the recommended strategies, Superstore can continue to grow and enhance its operations.



by :

- Mariam abdalazym
- Shimaa Essam
- Asmaa Ahmeed
- Ola Mohammed