Superstore Sales Data Analysis Documentation

Prepared by:

1. Mayar Saad Khalifa

2. Aml Mohamed

3. Mohammed Salama

4. Fady Makram

5. Ayman Elsayed Abdelhalim

Table of Contents

1. Introduction 3

Overview: 3

Objective: 3

Stakeholders: 3

2. Data Collection 4

Data Sources 4

Data Description 4

Data Quality Assessment: 4

3. Data Cleaning & Preprocessing 5

Python 5

DAX Queries 7

4. Exploratory Data Analysis (EDA) 9

✓ Overview Sales Analysis Report (January 2015 - December 2018) .9

✓ Customer Analysis 10

✓ Region Analysis 12

✓ Product Analysis 14

Data Visualization 16

5. Results & Insights 17

Key Findings: 17

Business Implications: 17

6. Recommendations & Next Steps 18

Actionable Insights: 18

Future Work: 19

SQL ERD and Mapping 20

7. Conclusion 24

1. Introduction

Overview:

This project involves analyzing a superstore sales dataset to uncover key trends, patterns, and actionable insights that can support business decision-making. The dataset includes information on customer orders, shipping methods, regions, product categories, and sales performance.

Objective:

- Identify sales trends over time.
- Discover top-performing product categories and regions.
- Detect customer segments contributing most to revenue.
- Highlight inefficiencies in shipping or sales processes.

Stakeholders:

- Business Analysts
- Sales and Marketing Teams
- Regional Managers
- Senior Management and Decision-Makers

2. Data Collection

Data Sources

The dataset was sourced from a cleaned internal Superstore sales report, originally exported from the company's retail transaction system.

Data Description

Key variables in the dataset include:

- Order and Shipping: Order ID, Order Date, Ship Date, Ship Mode
- Customer Info: Customer ID, Customer Name, Segment, City, State, Region
- Product Details: Product ID, Category, Sub-Category, Product Name
- Sales Data: Sales amount

Data Quality Assessment:

- No missing values found in critical fields
- Minor duplication in customer names (same customer ID)
- Consistent formatting across dates and categories

3. Data Cleaning & Preprocessing

Several steps were taken to clean the data and remove duplicates and unknown values as follows:

Python

1- Importing Required Libraries:

```
import pandas as pd
```

2. Reading Data from a CSV File:

Here, we read the data from the file 'superstore sale dataset.csv' and load it into a DataFrame using the following command:

```
df = pd.read_csv(r'D:\Data Analytics\superstore sale dataset.csv')
```

3. Describing the Displayed Data:

The following function was used to generate a statistical summary of the data:

```
df.describe( )
```

4. Displaying General Information About the Dataset:

To get a complete overview of the dataset, the following function was used:

```
df.info( )
```

From the output, it was found that the dataset contains 9,800 rows, including 11 rows with missing values in the 'Postal Code' column.

5. Displaying the First 20 Random Rows:

To verify that the data was successfully loaded, we randomly displayed 20 rows using the following function:

```
df.sample(n=20)
```

6. Reviewing 50% of the Dataset:

To achieve a higher level of data verification, we viewed 50% of the dataset’s content.

7. Replacing Missing Values with Fixed Values:

To facilitate data analysis, we replaced missing values in the 'Postal Code' column with a fixed value using the following command:

```
df['Postal Code'].fillna('Unknown', inplace=True)
```

8. We reviewed some insights and results, such as:

- **Total Customers: 793**
- **Total Sales: $2.26 million**
- **Total Orders: 4,922**
- **Top 10 Cities with the highest sales**

9. Saving the Cleaned Data:

We saved the cleaned data after ensuring all modifications were successfully applied.

10. Exporting the Cleaned Data to an Excel File:

After data cleaning, we exported the updated dataset to a new Excel file in the specified path using the following command:

```
df.to_excel(r'D:\Data Analytics\cleaned_superstore sales.xlsx', index=False)
```

This makes it easier to retrieve and use the data for further analysis in other programs such as Power BI.

DAX Queries

Average Order Value (AOV)

```
MEASURE Sheet1[Average Order Value(AOV)] = DIVIDE(Sheet1[Total Sales],Sheet1[Total Orders],0)
```

Average Sales Per Customer

```
MEASURE Sheet1[Average Sales Per Customer] = DIVIDE([Total Sales], [Total Customers])
```

Cities Served

```
MEASURE Sheet1[Cities Served] = DISTINCTCOUNT('Sheet1'[City])
```

Customers2015

```
MEASURE Sheet1[Customers2015] = CALCULATE(DISTINCTCOUNT(Sheet1[Customer ID]),
Sheet1[YearOnly] = 2015)
```

Customers2016

```
MEASURE Sheet1[Customers2016] = CALCULATE(DISTINCTCOUNT(Sheet1[Customer ID]),
Sheet1[YearOnly] = 2016)
```

Customers2017

```
MEASURE Sheet1[Customers2017] = CALCULATE(DISTINCTCOUNT(Sheet1[Customer ID]),
Sheet1[YearOnly] = 2017)
```

Customers2018

```
MEASURE Sheet1[Customers2018] = CALCULATE(DISTINCTCOUNT(Sheet1[Customer ID]),
Sheet1[YearOnly] = 2018)
```

Order Count

```
MEASURE Sheet1[Order Count] = DISTINCTCOUNT('Sheet1'[Order ID])
```

Orders 2015

```
MEASURE Sheet1[Orders 2015] = CALCULATE(DISTINCTCOUNT(Sheet1[Order ID]),
Sheet1[YearOnly] = 2015)
```

Orders 2016

```
MEASURE Sheet1[Orders 2016] = CALCULATE(DISTINCTCOUNT(Sheet1[Order ID]),
Sheet1[YearOnly] = 2016)
```

Orders 2017

```
MEASURE Sheet1[Orders 2017] = CALCULATE(DISTINCTCOUNT(Sheet1[Order ID]),
Sheet1[YearOnly] = 2017)
```

Orders 2018

MEASURE Sheet1[Orders 2018] = CALCULATE(DISTINCTCOUNT(Sheet1[Order ID]), Sheet1[YearOnly] = 2018)

product 2015

MEASURE Sheet1[product 2015] = CALCULATE(DISTINCTCOUNT(Sheet1[Product ID]), Sheet1[YearOnly] = 2015)

product 2016

MEASURE Sheet1[product 2016] = CALCULATE(DISTINCTCOUNT(Sheet1[Product ID]), Sheet1[YearOnly] = 2016)

product 2017

MEASURE Sheet1[product 2017] = CALCULATE(DISTINCTCOUNT(Sheet1[Product ID]), Sheet1[YearOnly] = 2017)

product 2018

MEASURE Sheet1[product 2018] = CALCULATE(DISTINCTCOUNT(Sheet1[Product ID]), Sheet1[YearOnly] = 2018)

sales2015

MEASURE Sheet1[sales2015] = CALCULATE(SUM(Sheet1[Sales]), Sheet1[YearOnly] = 2015)

sales2016

MEASURE Sheet1[sales2016] = CALCULATE(SUM(Sheet1[Sales]), Sheet1[YearOnly] = 2016)

sales2017

MEASURE Sheet1[sales2017] = CALCULATE(SUM(Sheet1[Sales]), Sheet1[YearOnly] = 2017)

sales2018

MEASURE Sheet1[sales2018] = CALCULATE(SUM(Sheet1[Sales]), Sheet1[YearOnly] = 2018)

Total Customers

MEASURE Sheet1[Total Customers] = DISTINCTCOUNT('Sheet1'[Customer ID])

4. Exploratory Data Analysis (EDA)

- ✓ **Overview Sales Analysis Report (January 2015 - December 2018)**

A. Key Performance Indicators (KPIs):

- Total Orders: 4,922
- Total Revenue: 2.26M
- Average Order Value: 459.48

B. Sales by Segment:

- Consumer: 53% (most significant contributor)
- Corporate: 30%
- Home Office: 17%

C. Sales by Category:

- Technology: 827.46K (36.59% of total revenue)
- Furniture: 705.42K (31.19%)
- Office Supplies: 728.6K (32.22%)

D. Sales Trend by Year and Month:

- 2015: Steady growth in sales over the year
- 2016: Significant peaks in the middle of the year and variability between months
- 2017: Ongoing volatility with high peaks and troughs
- 2018: Good performance in the first few months, with consistent weakening through to year-end

Seasonal trends and individual sales peaks suggest areas where further investigation would be valuable.

E. Sales by Customer:

- Top Customers:
  - Tamara Chand: Total Revenue 19,052.22; Orders 5
  - Sean Miller: Total Revenue 25,043.05; Orders 5
  - Raymond Buch: Total Revenue 15,117.34; Orders 6

F. Key Findings:

-Technology drives the highest revenue and growth.

-The West region is the most lucrative market.

-Office Supplies, though less profitable, exhibit steady performance.

-Loyal repeat customers significantly impact sales volume.

G. Business Implications:

-Reallocate marketing efforts toward Technology products.

-Focus on the West region for scaling operations.

-Create loyalty programs for high-value customers.

✓ Customer Analysis

- **Total Customers:** 793
- **Total Orders:** 4,922
- **Total Sales:** $2.26M
- **Average Sales per Customer:** $2.85K

Top Performing Cities by Sales

City	Total Sales

New York	$0.25M
Los Angeles	$0.17M
Seattle	$0.12M
San Francisco	$0.11M
Philadelphia	$0.11M
Others	$0.06M - $0.02M

- Top Customers by Number of Orders
  - Emily Phan: 17 orders
  - Chloris Kavanaugh, Joel Eaton, Patrick Gardner, Zuschuss Donatelli: 13 orders each
  - Other key customers: 12 - 11 orders

- **Top Customers by Total Sales**
  - Sean Miller: $25K
  - Tamara Chand: $19K
  - Raymond Buch, Tom Ashbrook: $15K each
  - Adrian Barton, Ken Lonsdale, Sanjit Chandrasekhar: $14K each
  - Other top spenders: $13K - $11K

- **Customer Distribution by Category**
  - Office Supplies: 787 customers
  - Furniture: 705 customers
  - Technology: 684 customers
  - Combined coverage: 86.9%

- **Data Coverage Period**

From March 1, 2015 to December 30, 2018

✓ Region Analysis

- **Cities Served:** 529 unique cities were included in the dataset.
- **Order Count:** A total of 4,922 orders were processed.
- **Regional Sales Performance:**

  -The West region leads with the highest sales (31.40% of total)

  -East region follows at 29.60%

  -Central region contributes 21.76%

  -South region has the lowest at 17.21%

- **Order Distribution:**

  -Total order count is 4,922 across all regions

  -West region has the highest proportion of orders (25.24%)

  -Central region follows at 27.81%

  -East region accounts for 23.49%

  -South region has 16.46% of orders

- **Category Performance by Region:**

  -Technology category is performing well across regions

  -Office Supplies shows consistent sales

  -Furniture appears to have lower sales compared to other categories

  -Central region shows strong performance in Technology

  Sales Trends (2015-2018):

  -All regions show positive growth trends

  -West region maintains leadership throughout the period

  -East region shows the steepest growth, especially from 2017-2018

-Central and South regions show more modest but steady growth

- **Geographic Coverage:**

  -The business serves 529 cities across all regions

- **Sales by Region**

  The West region led with 31.40% of total sales, followed by East (29.60%), Central (21.78%), and South (17.21%).

  Order count also showed dominance of West (32.24%) and East (27.81%).

- **Sales by Category and Region**

  Technology category saw the highest revenue in several regions, especially Central and West.

  Furniture showed strong performance in Central and West regions.

  Office Supplies had consistent presence across all regions with relatively lower sales.

- **Trend Over Time**

  From 2015 to 2018, West and East regions showed consistent growth.

  South region experienced a dip in 2016 but recovered by 2018.

  Central remained steady but showed the lowest growth among all regions.

✓ Product Analysis

1. Key Performance Indicators (KPIs)

- Total Orders: 4,922 — Represents the total number of customer transactions.
- Total Products: 1,861 — Reflects the diversity and volume of items in the store's catalog.
- Total Sales: $2.26M — Indicates total revenue generated over the analyzed period.

These KPIs serve as immediate indicators of scale and business health.

2. Key Insights

- Top-selling products account for a disproportionate share of sales.
- Poor performers are consistent across years — suggests low demand or visibility.
- All product categories contribute meaningfully — risk is balanced.
- Year-over-year order growth shows business momentum — worth capitalizing on.

3. Business Questions Answered

A- Which products generate the most/least revenue?

Top Revenue Products:

Products like Canon image CLASS multifunction printers lead in total sales, exceeding $62,000. These products have consistent demand and likely high customer satisfaction or business relevance.

Lowest Revenue Products:

Items such as Avery 5-Tab Index Dividers and similar low-visibility office supplies generate less than $6,000 in total sales. These are possibly niche or outdated products with limited.

B- What is the distribution of sales by category?

The three main product categories contribute relatively evenly to total sales:

Technology: ~36.6% of total revenue

Office Supplies: ~32.2%

Furniture: ~31.2%

Interpretation: The balanced distribution suggests a healthy business model that isn't overly reliant on a single category. It also indicates room to explore growth in all three verticals without major risk.

C- How have customer orders trended over time?

Trend: There's a consistent increase in the number of orders from 2015 through 2018.

Notable Peaks: Significant spikes occur toward the end of 2017 and again in late 2018, possibly aligning with seasonal promotions or fiscal year-end purchases.

Conclusion: The trend shows steady customer growth and improving market traction — useful signals for forecasting and capacity planning.

D- Are there specific cities or years where performance spikes?

The dashboard allows slicing by city and year. When this is applied:

Years: 2018 shows the highest number of orders across all years.

Cities: Performance varies, but larger metropolitan areas (e.g., New York, San Francisco) tend to outperform smaller cities.

Insight: Geographic segmentation can reveal high-potential markets, guiding localized promotions or inventory planning.

E- How can we optimize our catalog based on sales contribution?

Top-selling products: Should be prioritized in marketing, restocking, and potentially bundled with slower movers to enhance their visibility.

Low performers: Consider clearance, repackaging, or retiring completely to free up catalog space and reduce inventory waste.

Category-level optimization: Invest more in trending categories like Technology while keeping an eye on stable revenue from Furniture and Office Supplies.

Data Visualization

- Time series plot of sales over months
- Bar charts for category and sub-category performance
- Heatmaps showing regional sales distribution
- **Top 10 Selling Products (Bar Chart)**
- A horizontal bar chart ranking products by their total sales value.
  Insight: Canon imageCLASS generates over $62K, leading all products.
  Interpretation: These are the most valuable SKUs. Marketing, inventory, and pricing strategy should prioritize these winners.
- **Bottom 10 Selling Products (Bar Chart)**
- A similar chart showing products with the lowest sales performance.
  Insight: Several products yield under $6K in total revenue.
  Interpretation: These are underperformers — review for discontinuation, repackaging, or promotional rescue.
- **Total Sales by Category (Pie Chart)**
- Displays percentage contribution of each category:
  - Technology (36.6%)
  - Office Supplies (32.2%)
  - Furniture (31.2%)
  Insight: Revenue is well-distributed across categories.

Interpretation: Business risk is diversified — this balance allows for cross-category strategies and resilience against market shifts.

- **Orders Over Time (Line Chart)**
- Line chart capturing order volume from 2015 to 2018.
  Insight: Steady growth with notable peaks in 2017 and 2018.
  Interpretation: Identifies seasonal demand or successful promotional windows — useful for forecasting.
- **Filters and Slicers**
- Interactive filters allow slicing data by:
  - Year (2015 to 2018)
  - City (Multiple, e.g., Akron, Albuquerque, etc.)
  Utility: Enables localized or time-specific performance assessments to tailor decision-making to context.

Outliers & Anomalies

- Few extremely high sales values identified in the "Furniture" category
- Sales spikes around holiday periods

5. Results & Insights

Key Findings:

- Office Supplies had the highest volume of orders, but Furniture generated higher revenue per order.
- Western and Southern regions had the highest sales figures.
- Corporate segment contributed more consistently to high-value sales.

Business Implications:

- Consider focusing more marketing efforts on the Furniture category.
- Optimize shipping logistics in high-volume regions to reduce costs.

Limitations

- Dataset does not include cost or profit data, limiting profitability analysis.
- Regional factors (like local holidays or promotions) not included.

6. Recommendations & Next Steps

Actionable Insights:

- Increase promotional campaigns for high-margin product categories.
- Target the Corporate segment with personalized offers.
- Review shipping efficiency in the Western region.
- Prioritize customer retention strategies for high-value clients like Sean Miller and Tamara Chand.
- Encourage repeat purchases from frequent buyers by offering exclusive deals and loyalty programs.
- Analyze low-performing cities and implement localized marketing campaigns to boost sales.
- Leverage the large customer base in the Office Supplies category by launching targeted promotions and tailored product lines.
- Prioritize end-of-year campaigns for Technology to align with peak demand.
- Target high-order customers with personalized offers and rewards.

- Investigate underperformance in the South region and develop improvement plans.

- Question and answer:

  Are the customers with the highest number of orders also the ones with the highest purchase value?

  The answer is no. The dashboard reveals that the customer Sean Miller has the highest total purchase value, amounting to 25k, even though he is not the customer with the most orders—he only made 5 orders.

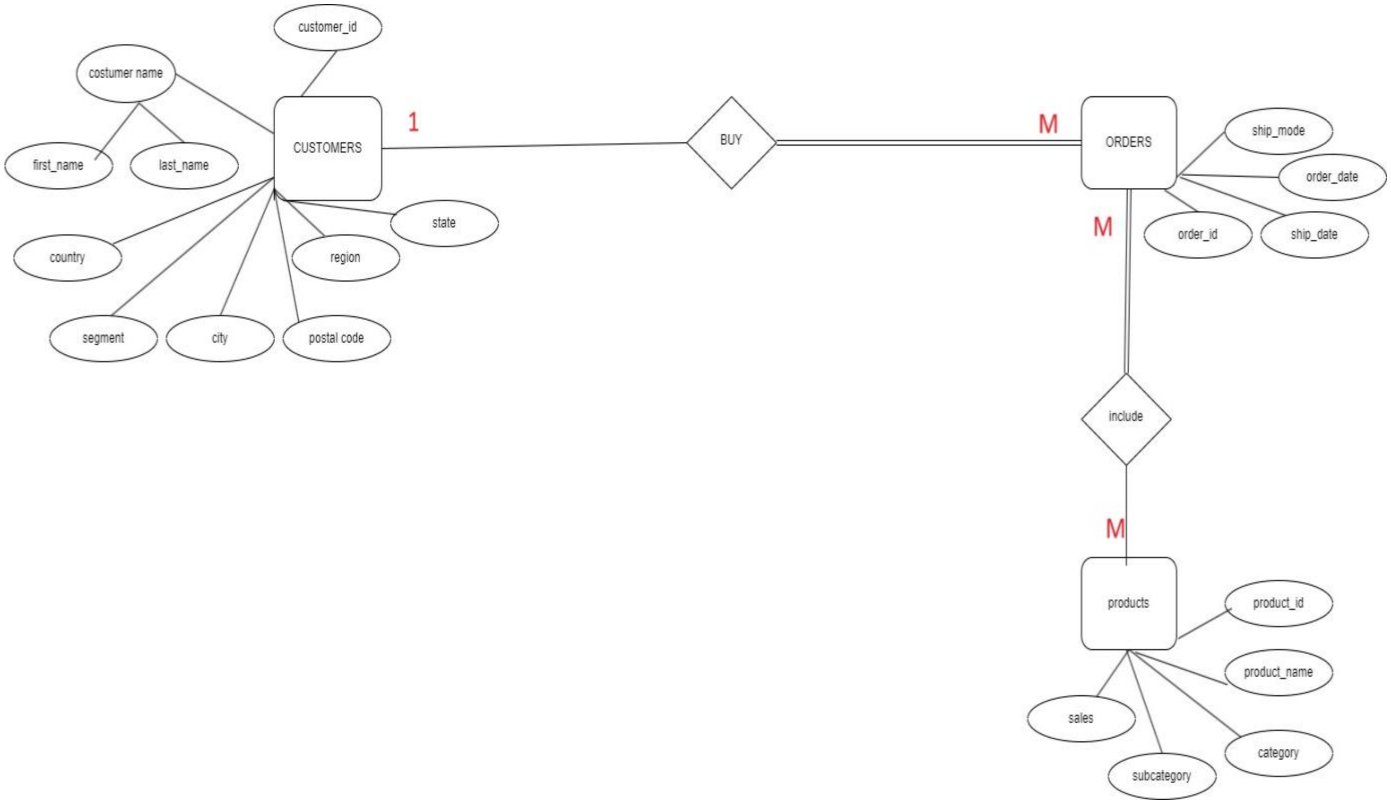
  On the other hand, the customer Emilly Phan placed the highest number of orders, totaling 17, but her total purchase value is only 5k.

  This difference is due to several factors, the most important being the variation in product value, not just the number of products ordered.

Future Work:

- Integrate profit data for deeper analysis.
- Analyze customer behavior over time using longitudinal data.
- Incorporate external factors such as economic indicators.
- Make database using SQL as following we showing data ERD and Mapping.

SQL ERD and Mapping



Based on this ERD

Types of Entities:

CUSTOMERS: Strong entity .

ORDERS: Strong entity .

PRODUCTS: Strong entity .

Types of Attributes:

For each entity, the attributes can be classified as follows:

CUSTOMERS:

Customer_id: Primary Key Attribute, Simple Attribute.

Customer_name: Composite Attribute composed of first_name and last_name.

Country, Segment ,City, Postal_code : Simple Attribute.

ORDERS:

Order_id: Primary Key Attribute ,Simple Attribute. .

Order_date, Ship_date , Ship_mode : Simple Attribute.

PRODUCTS:

Product_id: Primary Key Attribute, Simple Attribute.

Product_name, Category , Subcategory, Sales: Simple Attribute.

1. **Relationships**:

Relationship between CUSTOMERS and ORDERS :

Degree of Relationship: Binary Relationship, involving two entities (CUSTOMERS and ORDERS).

Cardinality: One-to-Many. One customer (CUSTOMERS) can place multiple orders (ORDERS). Each order (ORDERS) is placed by one customer (CUSTOMERS).

Participation:

Total Participation for ORDERS: Every order must be associated with a customer (an order cannot exist without a customer).

Partial Participation for CUSTOMERS: Not every customer may have placed an order (there might be registered customers who haven’t made any purchases yet).

Relationship between ORDERS and PRODUCTS (include):

Degree of Relationship: Binary Relationship, involving two entities (ORDERS and PRODUCTS).

Cardinality: Many-to-Many. One order (ORDERS) can include multiple products (PRODUCTS). One product (PRODUCTS) can be included in multiple orders (ORDERS).

Participation:

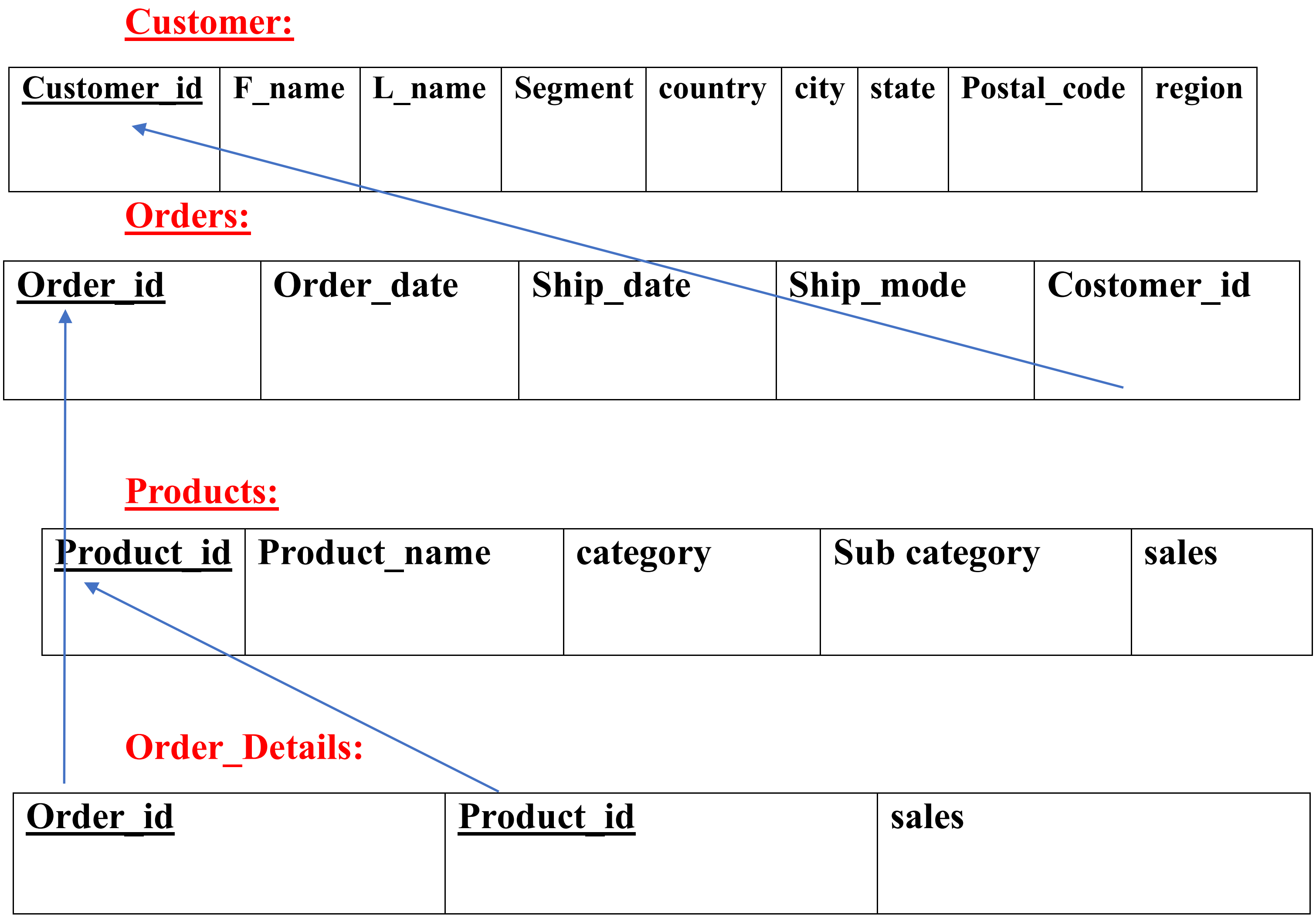
The relationship between Orders and Products is usually a many-to-many relationship because:

A single order can contain multiple products.

And a single product can be part of multiple different orders.

However, in relational databases, a direct many-to-many relationship cannot be represented. Therefore, an associative entity is used, often called Order_Details.

This entity contains foreign keys that link Orders and Products, along with additional attributes such as: sales



Customer:

Customer_id	F_name	L_name	Segment	country	city	state	Postal_code	region

Orders:

Order_id	Order_date	Ship_date	Ship_mode	Costomer_id

Products:

Product_id	Product_name	category	Sub category	sales

Order_Details:

Order_id	Product_id	sales

7. Conclusion

This project successfully explored sales patterns and customer segmentation in a superstore dataset. The insights gained can guide marketing and operational decisions to boost sales and improve customer satisfaction.