# " GLOBAL SUPERSTORE DATA ANALYSIS: INSIGHTS AND RECOMMENDATIONS"

BY

POLLAPU RUDRA SHIVA KUMAR

**Internship Period:**

**Organization: “NXTIVIA TECHNOLOGIES PRIVATE LIMITED”**

**Date: 31-03-2025**

# Internship Project Report on Global Superstore Project

## 1. Abstract

This report presents a comprehensive analysis of the Global Superstore dataset. The analysis focuses on understanding sales patterns, product performance, profitability, and shipping logistics. Key findings are presented with detailed interpretations, offering actionable insights and recommendations relevant to an internship project. This report aims to demonstrate analytical skills and provide realistic business recommendations

## 2. Executive Summary

This report presents a comprehensive analysis of the Global Superstore dataset, conducted during my internship, focusing on identifying key trends and opportunities for optimizing sales, profitability, and shipping costs. The analysis employed a variety of data analysis techniques, including data cleaning, exploratory data analysis (EDA), and visualization, to extract actionable insights from the extensive dataset.

The analysis revealed significant disparities in product performance, with "Apple Smart Phone, Full Size" emerging as the top-selling product, generating $86,935.79 in sales. However, this high sales volume was accompanied by concerning negative profit margins, indicating potential inefficiencies in pricing or cost management. Conversely, "Eureka Disposable Bags for Sanitaire Vebra Groomer I Upright Vac" recorded the lowest sales, at a mere $1.62, illustrating a long-tail distribution in sales performance, where a small number of products generate a large portion of the revenue.

Smartphones, including models from "Cisco," "Motorola," and "Nokia," were identified as key revenue drivers, alongside high-value office equipment such as the "Canon image CLASS 2200 Advanced Copier" and executive leather armchairs from brands like "Hon" and "Office Star." Notably, the "Canon image CLASS 2200 Advanced Copier" demonstrated the highest average profit, at $5040.0, highlighting the strong profitability of certain office equipment categories.

However, the analysis also highlighted significant variations in profitability across products. Certain items, such as the "Cubify CubeX 3D Printer Double Head Print," incurred substantial losses, with a negative profit of -$6599.98. This discrepancy underscores the need for a granular review of product-level profitability to identify and address loss-making items. The presence of the "Apple Smart Phone, Full Size" in both the top-selling and negative-profit lists is a critical point that demands further investigation.

Shipping cost analysis revealed significant variations in expenses based on product type and shipping mode. Smartphones and executive armchairs, both high-value and potentially bulky items, incurred the highest shipping expenses. "Same Day" shipping was identified as the most expensive shipping mode, with an average cost of $42.9, while "Standard Class" shipping was the most frequently used and cost-effective, averaging $20.0. The high utilization of "Standard Class" suggests a strong preference for cost-effective shipping options among customers.

Furthermore, a critical gap was identified in the analysis of average shipping costs by order priority. This metric is essential for understanding the impact of expedited shipping on overall costs and customer satisfaction. Therefore, further investigation is required to determine the average shipping costs associated with each order priority.

Key findings from this analysis include:

- A concentration of sales in a few high-performing products, indicating potential opportunities for targeted marketing and inventory management.
- Significant variations in product profitability, with some high-selling items generating losses, necessitating a review of pricing and cost structures.
- Substantial differences in shipping costs based on product and shipping mode, highlighting the need for optimized shipping strategies.
- The dominance of "Standard Class" shipping due to its lower cost, suggesting a customer preference for economical shipping options.
- The critical need to calculate the average shipping costs for each order priority to improve shipping cost analysis.

Based on these findings, recommendations include:

- Conducting a thorough review of the pricing and cost structures for high-selling products with negative profit margins, such as the "Apple Smart Phone, Full Size."
- Optimizing shipping strategies to reduce costs for high-value and bulky items, while maintaining customer satisfaction.
- Further analysing the relationship between product characteristics and shipping costs to identify potential cost-saving opportunities.
- Leveraging the popularity of "Standard Class" shipping while managing the higher costs of expedited shipping options.
- Calculating the average shipping costs for each order priority.

This analysis provides a solid foundation for strategic decision-making, aiming to improve sales, profitability, and operational efficiency within the Global Superstore.

## 3. Introduction

The Global Superstore dataset represents a comprehensive record of sales transactions for a multinational retail company, encompassing a wide array of product categories, customer segments, and geographical markets. This dataset, rich in detail and complexity, provides a unique opportunity to delve into the dynamics of global retail operations and extract actionable insights for business improvement. The data spans various regions, including the United States, Europe, Asia Pacific, and Latin America, and encompasses a diverse product portfolio ranging from office supplies and technology to furniture.

This internship, focused on data analysis within the context of the Global Superstore dataset, was designed to provide a practical learning experience in leveraging data to address real-world business challenges. The primary objective was to thoroughly explore the dataset to identify key trends, patterns, and anomalies that could inform strategic decision-making. The analysis aimed to answer critical business questions, including:

- How can we improve overall profitability across different product categories and geographical markets?
- What are the key drivers of customer behaviour, and how can we leverage this information to enhance customer satisfaction and loyalty?
- How can we optimize shipping strategies to minimize costs and improve delivery efficiency?
- What are the high performing products, and what are the low performing products?
- What is the effect of discounts on profits?
- What is the effect of shipping modes on shipping costs?

These questions were chosen to provide a holistic view of the company's performance and to identify areas for potential improvement.

The scope of this internship was broad, encompassing the entire data analysis pipeline, from data cleaning and preparation to exploratory data analysis (EDA) and visualization. Specific tasks performed during the internship included:

- **Data Cleaning and Preprocessing:** This involved identifying and addressing inconsistencies, missing values, and outliers within the dataset. SQL queries were used to manipulate and clean the data.
- **Data Transformation and Feature Engineering:** Creating new calculated columns to derive meaningful metrics, such as profit margins and sales per customer, to facilitate deeper analysis.
- **Exploratory Data Analysis (EDA):** Conducting in-depth analyses of sales, profitability, customer behaviour, and shipping costs using statistical methods and visualizations.
- **Geographical Analysis:** Mapping sales and profitability across different regions to identify geographical trends and patterns.
- **Product Analysis:** Determining the best and worst performing products in terms of sales and profits.
- **Shipping Analysis:** Analysing the effects of shipping modes and priorities on shipping costs.
- **Customer Segmentation:** Identifying distinct customer segments based on purchasing behaviour and demographic characteristics.
- **Reporting and Visualization:** Creating comprehensive reports and visualizations to communicate findings and recommendations to stakeholders.

The internship provided hands-on experience in applying data analysis techniques to a large and complex dataset, fostering the development of critical thinking and problem-solving skills.

To effectively perform these tasks, a combination of tools and technologies was utilized.

- **PostgreSQL and MySQL:** These relational database management systems were instrumental in data cleaning and analysis. SQL queries were used to extract, transform, and load data, as well as to perform complex data aggregations and filtering.
- **Pandas (Python):** The Pandas library, a powerful tool for data manipulation and analysis in Python, was used to process, clean, and transform the dataset. Pandas was also used to create new data frames, and to export data to other formats.
- **Power BI:** This business intelligence tool was used to create interactive dashboards and visualizations to communicate findings effectively. Power BI allowed for the creation of dynamic charts, graphs, and maps, enabling stakeholders to explore the data and gain insights.

The use of these tools facilitated a comprehensive and efficient analysis of the Global Superstore dataset, enabling the extraction of actionable insights that could inform strategic decision-making. Through this internship, I gained valuable experience in applying data analysis techniques to a real-world business scenario, enhancing my skills in data manipulation, analysis, and visualization.

## 4. Data Description and Preparation

### 1. Data Source and Initial Overview

The dataset used for this analysis was sourced from Kaggle.com, a platform widely recognized for its rich repository of datasets and resources for data science and machine learning. Specifically, the "Global Superstore" dataset was downloaded from Kaggle, providing a comprehensive view of sales transactions for a multinational retail company. This dataset is a popular choice for data analysis projects due to its real-world relevance and the variety of analytical opportunities it presents.

The dataset, initially provided in a CSV (Comma Separated Values) format, contained a substantial number of rows and columns, representing individual sales transactions and associated details. The initial examination revealed a complex structure, encompassing a wide range of variables related to orders, customers, products, and shipping.

### 2. Column Descriptions:

- **Row ID:** A unique identifier for each row in the dataset, serving as a primary key for individual records.
- **Order ID:** A unique identifier for each order placed by a customer, allowing for the grouping of related transactions.
- **Order Date:** The date when the order was placed, providing a time-based dimension for trend analysis.
- **Ship Date:** The date when the order was shipped, enabling the calculation of delivery times.
- **Ship Mode:** The method of shipping used for the order (e.g., Standard Class, Same Day, Second Class, First Class).
- **Customer ID:** A unique identifier for each customer, facilitating customer segmentation and analysis.
- **Customer Name:** The full name of the customer.
- **Segment:** The customer segment (e.g., Consumer, Corporate, Home Office).

- **City:** The city where the customer is located.
- **State:** The state where the customer is located.
- **Country:** The country where the customer is located.
- **Postal Code:** The postal code of the customer's location.
- **Market:** The market segment (e.g., US, EU, APAC, LATAM).
- **Region:** The geographical region (e.g., Central, East, South, West).
- **Product ID:** A unique identifier for each product.
- **Category:** The broad category of the product (e.g., Furniture, Office Supplies, Technology).
- **Sub-Category:** The specific sub-category of the product (e.g., Chairs, Binders, Phones).
- **Product Name:** The name of the product.
- **Sales:** The total sales amount for the product in the order.
- **Quantity:** The number of units of the product sold in the order.
- **Discount:** The discount applied to the product in the order.
- **Profit:** The profit generated from the product in the order.
- **Shipping Cost:** The cost of shipping the order.
- **Order Priority:** The priority of the order shipment.

3. **Data Cleaning and Handling Missing Values**

The initial data exploration revealed several data quality issues that required attention. Primarily, the dataset contained missing values in the "Postal Code" column, which were handled by either filling the missing values with the most frequently occurring postal code within the same city and state, or by removing the rows where this was not possible.

The data cleaning process was performed using a combination of PostgreSQL and MySQL for database-level manipulations, and the Pandas library in Python for more granular data transformations. SQL queries were used to identify and handle missing values, while Pandas was employed to standardize data formats and perform complex data transformations.

4. **Data Type Conversions:**

- The "Order Date" and "Ship Date" columns were converted from string format to datetime objects using the to_datetime() function in Pandas. This conversion allowed for time-based analysis and the calculation of delivery times.
- Numerical columns such as "Sales," "Quantity," "Discount," "Profit," and "Shipping Cost" were verified to ensure they were in numerical formats (float or integer). Any inconsistencies or non-numeric values were addressed.

**Handling Duplicates:**

- The dataset was checked for duplicate rows using the duplicated () function in Pandas. Duplicate rows were identified and removed to ensure data integrity and prevent skewed analysis results.

5. **Outlier Detection and Management**

Outlier detection was a crucial step in the data preparation process. Outliers were identified using a combination of visual inspection (box plots, scatter plots) and statistical methods (e.g., z-scores, interquartile range (IQR)). Significant outliers were found in the "Sales," "Profit," and "Shipping Cost" columns.

- For outliers in the "Sales" and "Profit" columns, a decision was made to retain them, as they could represent high-value transactions or significant losses that are relevant to the analysis.
- For the shipping cost outliers, they were kept as well because they are representative of the same day shipping, and other expedited shipping modes.

6. **Data Transformation and Feature Engineering**

To enhance the analytical capabilities of the dataset, several new calculated columns were created using Pandas.

- **Profit Margin:** This column was calculated by dividing the "Profit" by the "Sales" and multiplying by 100, providing a percentage-based measure of profitability.
- **Sales per Customer:** This column was calculated by grouping the data by "Customer ID" and summing the "Sales," then dividing by the number of orders per customer, providing insight into average customer spending.

7. **Filtering, Sorting, and Grouping Data**

The data was extensively filtered, sorted, and grouped to facilitate various analytical tasks.

- Filtering was used to isolate specific subsets of the data, such as orders from a particular region or product category.
- Sorting was used to arrange the data in ascending or descending order based on specific columns, such as "Sales" or "Profit."
- Grouping was used to aggregate data based on categorical variables, such as "Category," "Segment," or "Ship Mode." For example, the groupby() function in Pandas was used to calculate total sales and profit by product category, region, and customer segment.

These data preparation steps were crucial for ensuring the accuracy and reliability of the subsequent analysis. The cleaned and transformed dataset provided a solid foundation for extracting meaningful insights and generating actionable recommendations.

## 5. Exploratory Data Analysis (EDA)

**Exploratory Data Analysis (EDA):**

- **Overall Sales and Profit:**
  - Total sales and profit.
  - Trends over time (using the "Order Date" column).
  - Sales and profit by market and region.
- **Customer Analysis:**
  - Customer segmentation (using the "Segment" column).
  - Top customers by sales and profit.
  - Customer purchase patterns.
  - Analysis of customer distribution by city, state, and country.
- **Product Analysis:**
  - Sales and profit by category and sub-category.
  - Top-selling products.
  - Analysis of discounts and their impact on sales and profit.
- **Shipping Analysis:**
  - Analysis of shipping modes and their impact on delivery time and cost.

    - Shipping cost analysis by region and market.
    - Analysis of Shipping priority.
- **Geographical Analysis:**
    - Map visualizations of sales and profit by country, state, and city.
    - Regional performance comparisons.
- **Key Metrics:**
    - Calculate and analyse key performance indicators (KPIs) such as profit margin, average order value, customer lifetime value (if possible with the data).
- **Visualization:**
    - Include charts and graphs to illustrate your findings (e.g., bar charts, line charts, pie charts, scatter plots, maps).
    - Use appropriate tools for visualization (e.g., Matplotlib, Seaborn, Tableau, Power BI).

## 6. Key Observations and Insights

Based on the analysis of the data extracted from the images, some key observations and insights include:

- **Sales Performance**: Sales vary significantly across different product categories. Smartphones are a high-performing category, but not all smartphone products are profitable.
- **Profitability**: Profit margins vary substantially across products. Some high-selling products generate losses, indicating potential issues with pricing or cost management.
- **Shipping Costs**: Shipping costs are a significant expense, particularly for certain products and shipping modes.
- **Order Fulfilment**: "Standard Class" is the most frequently used shipping mode, while "Same Day" is the least.

### 5. Detailed Analysis

The following sections provide a detailed analysis of the data extracted from the images.

### 6.1 Sales Analysis

### 6.1.1 Total Sales by Product

- "Apple Smart Phone, Full Size" is the top-selling product with $86,935.79 in sales.
- Smartphones ("Cisco," "Motorola," and "Nokia") are among the top 5 sellers.
- "Canon image CLASS 2200 Advanced Copier" is also a high-selling product.
- Executive leather armchairs from "Hon," "Office Star," and "Harbour Creations" are also among the top sellers.
- The total sales for all products in the dataset is $12,642,507.25
- Apple Smart Phone, Full Size" is the top-selling product.

**Interpretation**: Smartphones are a key sales driver, but other products like the Canon copier and executive armchairs also contribute significantly to revenue.

### 6.1.2 Lowest Sales by Product

The image "Screenshot 2025-03-31 202706.png" shows the product with the lowest sales:

- "Eureka Disposable Bags for Sanitaire Vebra Groomer I Upright Vac" has the lowest sales with $1.62

**Interpretation**: A large disparity exists between the top-selling and lowest-selling products, indicating a very long tail in the sales distribution.

### 6.1.3 Products with Quantity Greater Than 5

- Several products have a quantity of 14, including "Avery Binder Covers, Economy", "Breville Microwave, Silver", and "Deflect-O Clock, Black".

**Interpretation**: A quantity of 14 may represent a common order quantity for certain products.

## 6.2 Profitability Analysis

### 6.2.1 Average Profit by Product

- "Canon imageCLASS 2200 Advanced Copier" has the highest average profit at $5040.0
- Other Canon copiers and office equipment show high average profits.

**Interpretation**: Canon copiers are highly profitable.

### 6.2.2 Products with Negative Profit

- "Cubify CubeX 3D Printer Double Head Print" has the lowest profit of -$6599.98

- "Apple Smart Phone, Full Size" is present in both the top-selling and negative profit lists.

**Interpretation**: Some products are generating significant losses. The presence of "Apple Smart Phone, Full Size" in both the high-sales and negative-profit lists is concerning.

### 6.3 Shipping Cost Analysis

#### 6.3.1 Total Shipping Cost by Product

- "Motorola Smart Phone, Full Size" has the highest total shipping cost at $8817.3.
- "Apple Smart Phone, Full Size" has the second-highest shipping cost at $8027.7.
- Smartphones, in general, have high shipping costs.
- Executive leather armchairs also have high shipping costs.

**Interpretation**: Smartphones and executive armchairs incur high shipping costs.

#### 6.3.2 Average Shipping Cost by Order Priority

- This image does not contain the actual average shipping costs for each order priority. It would be beneficial to have that data to complete the analysis.

#### 6.3.3 Average Shipping Cost by Ship Mode

- "Same Day" shipping has the highest average shipping cost at $42.9.
- "Standard Class" shipping has the lowest average shipping cost at $20.0.

**Interpretation**: Expedited shipping ("Same Day") is significantly more expensive than standard shipping.

#### 6.3.4 Total Count of Products for Each Ship Mode

- "Standard Class" is the most frequently used shipping mode with 30,775 products.
- "Same Day" is the least frequently used shipping mode with 2,701 products.

**Interpretation**: "Standard Class" is the most popular shipping mode, likely due to its lower cost.

## 7. Findings and Insights

I have found the following insights using this data

1. Retrieve all product names and their corresponding sales from Dataset .

```
SELECT
  "Product Name",
  SUM(Sales) AS Total_Sales  -- Added double quotes for consistency and to handle spaces
FROM
  product performance AS pp  -- Added table alias "pp"
WHERE
  Sales IS NOT NULL AND "Product Name" IS NOT NULL
GROUP BY
  "Product Name"
ORDER BY
  Total_Sales DESC;
```

2. Calculate the total sales for all products in Dataset .

```
SELECT SUM(Sales) AS Total_sales
FROM
  product_details
WHERE
  Sales IS NOT NULL
```

3. Find the product with the highest sales in Dataset .

```
SELECT
  Product_Name,
  Total_Sales
FROM (
  SELECT
    Product_Name,
    SUM(Sales) AS Total_Sales,
    RANK() OVER (ORDER BY SUM(Sales) DESC) AS Ranking
  FROM
    product_details
  GROUP BY
    Product_Name
) AS RankedSales
```

```
WHERE
   Ranking = 1
LIMIT 1;
```

**Alternative and Simpler Query:**

```
SELECT
   Product_Name,
   SUM(Sales) AS Total_Sales
FROM
   product_details
GROUP BY
   Product_Name
ORDER BY
   Total_Sales DESC
LIMIT 1;
```

4. Find the product with the lowest sales in Dataset .

```
SELECT
   Product_Name,
   SUM(Sales) AS Total_Sales
FROM
   product_details
WHERE
   Sales IS NOT NULL AND Product_Name IS NOT NULL
GROUP BY
   Product_Name
ORDER BY
   Total_Sales
LIMIT 1;
```

5. Retrieve all products with a quantity greater than 5 from Dataset .

```
SELECT
   Product_Name,
   Quantity
FROM
   product_details
WHERE
   Quantity > '5'
```

```
        ORDER BY
           Quantity DESC
```

6. Calculate the average profit for each product in Dataset.

```
SELECT
        Product_Name,
        CAST(AVG(Profit)AS numeric(10,1)) AS Avg_profit
FROM
        product_details
WHERE
        Product_Name IS NOT NULL
GROUP BY
        Product_Name
ORDER BY
        Avg_profit DESC
```

7. Find the products with negative profit in Dataset .

```
SELECT
        Product_Name,
        Profit
FROM
        product_details
WHERE
        Profit < 0
ORDER BY
        profit
```

8. Calculate the total shipping cost for each product in Dataset

```
SELECT
        Product_Name,
        CAST(AVG(Shipping_Cost)AS numeric(10,1)) AS Total_Shipping_Cost
FROM
        product_details
WHERE
        Product_Name IS NOT NULL
GROUP BY
```

```
        Product_Name
ORDER BY
         Total_Shipping_Cost DESC
```

9. Calculate the average discount for each product in Dataset .

```
SELECT
        Product_Name,
        CAST(AVG(Discount)AS numeric(10,1)) AS Avg_discount
FROM
        product_details
WHERE
        Product_Name IS NOT NULL
GROUP BY
        Product_Name
ORDER BY
        Avg_discount DESC
```

10. Calculate the total quantity for each product in Dataset .

```
 SELECT
        Product_Name,
        CAST(SUM(Quantity)AS numeric(10,1)) AS Total_quantity
FROM
        product_details
WHERE
        Product_Name IS NOT NULL
GROUP BY
        Product_Name
ORDER BY
        Total_quantity DESC
```

11. Find the average shipping cost for each order priority in Dataset .

```
SELECT
        Order_Priority,
        CAST(AVG(Shipping_Cost)AS numeric(10,1)) AS Avg_shipping_cost
FROM
        product_details
```

```
WHERE
        Order_Priority IS NOT NULL
GROUP BY
        Order_Priority
ORDER BY
        Avg_shipping_cost DESC
```

12. Find the sum of the profit for each order priority in Dataset .

```
SELECT
        Order_Priority,
        CAST(SUM(Profit)AS numeric(10,1)) AS TOTAL_PROFIT
FROM
        product_details
WHERE
        Order_Priority IS NOT NULL
GROUP BY
        Order_Priority
ORDER BY
        TOTAL_PROFIT DESC
```

13. Using Dataset 1, find the total sales for each product category

```
 SELECT
        Category,
        CAST(SUM(Sales)AS numeric(10,1)) AS Totalsales
FROM
        product_details
WHERE
        Sales IS NOT NULL
GROUP BY
        Category
ORDER BY
        Totalsales DESC
```

14. Using Dataset 1, find the average shipping cost for each ship mode.

```
SELECT
        Ship_Mode,
        CAST(AVG(Shipping_Cost)AS numeric(10,1)) AS Avg_shipping_cost
FROM
        product_details
WHERE
        Sales IS NOT NULL
GROUP BY
        Ship_Mode
ORDER BY
        Avg_shipping_cost DESC
```

15. Find the total count of products for each ship mode.

```
SELECT
        Ship_Mode,
        COUNT(*) AS total_countofproducts
FROM
        product_details
WHERE
        Ship_Mode IS NOT NULL
GROUP BY
        Ship_Mode
ORDER BY
        total_countofproducts DESC
```

16. Find the correlation between discount and profit in Dataset

```
SELECT
        CORR(Discount, Profit) AS correlation
FROM
        product_details;
```

17. Rank products by profit in descending order in Dataset

```
SELECT
  Product_Name,
  Profit,
  RANK() OVER (ORDER BY Profit DESC) as profit_rank
```

```
FROM product_details
ORDER BY profit_rank;
```

18. Create a view that shows the product name, sales, and profit for products with negative profit in Dataset.

```
CREATE VIEW negative_profit_products AS
SELECT
   Product_Name,
   Sales,
   Profit
FROM product_details
WHERE Profit < 0;
```

19. Calculate the cumulative sales for each product in Dataset

```
SELECT
   Product_Name,
   Sales,
   SUM(Sales) OVER (PARTITION BY Product_Name ORDER BY Sales ASC) as cumulative_sales
FROM product_details
ORDER BY Product_Name, Sales ASC;
```

20. Identify the products with the highest shipping cost relative to their sales in Dataset

```
SELECT
   Product_Name,
   Shipping_Cost,
   Sales,
   CAST((Shipping_Cost / Sales) AS DECIMAL(10, 2)) AS shipping_cost_to_sales_ratio
FROM product_details
WHERE Sales > 0  -- To avoid division by zero
ORDER BY shipping_cost_to_sales_ratio DESC;
```

21. Create a report showing the product name, average sales, average profit, and average discount for each product in Dataset

```
SELECT
  Product_Name,
  AVG(Sales) AS average_sales,
  AVG(Profit) AS average_profit,
  AVG(Discount) AS average_discount
FROM product_details
GROUP BY Product_Name;
```

22. Using Dataset 1, identify the customer segment with the highest total profit.

```
SELECT
  Segment,
  SUM(Profit) AS Total_Profit
FROM
  orders_table
GROUP BY
  Segment
ORDER BY
  Total_Profit DESC
LIMIT 1;
```

23. Using Dataset 1, calculate the percentage of total sales for each product category.

```
SELECT
  Category,
  SUM(Sales) AS CategorySales,
  (SUM(Sales) / (SELECT SUM(Sales) FROM orders_table)) * 100 AS
PercentageOfTotalSales
FROM
  orders_table
GROUP BY
  Category;
```

24. find the products with sales greater than the average sales in their respective category.

```
SELECT
  Product_Name,
  Sales,
  Category,
  AverageCategorySales
FROM (
  SELECT
    Product_Name,
    Sales,
    Category,
    AVG(Sales) OVER (PARTITION BY Category) AS AverageCategorySales
  FROM
    orders_table
) AS SalesByCategory
WHERE Sales > AverageCategorySales
ORDER BY Category, Sales;
```

## 8. Recommendations

Based on the analysis:

- **Product Strategy**:
  - Focus on high-selling products like smartphones but address the negative profitability of some smartphone models.
  - Analyse the profitability of Canon copiers and consider strategies to maximize profits from this product line.
  - Evaluate the low sales of "Eureka Disposable Bags" and consider discontinuing or restructuring its pricing.
- **Pricing and Profitability**:
  - Review the pricing and cost structure of products with negative profits, such as the "Cubify CubeX 3D Printer" and "Apple Smart Phone, Full Size".
- **Shipping and Logistics**:
  - Investigate the high shipping costs associated with smartphones and executive armchairs.

- Evaluate the pricing of "Same Day" shipping and explore cost-reduction strategies.
- Promote the use of "Standard Class" shipping to cost-sensitive customers

## 9. Conclusion

**Summary of Key Takeaways and Internship Reflections**

This internship, focused on the analysis of the Global Superstore dataset, provided a valuable opportunity to apply data analysis techniques to a real-world business scenario. Through the comprehensive exploration of sales, profitability, customer behaviour, and shipping operations, several key takeaways emerged.

Firstly, the analysis highlighted significant disparities in product performance. While certain products, such as "Apple Smart Phone, Full Size" and "Canon image CLASS 2200 Advanced Copier," drove substantial sales, they also presented challenges in terms of profitability. The "Apple Smart Phone, Full Size" case, in particular, revealed the potential for high-volume sales to coexist with negative profit margins, emphasizing the need for meticulous cost management and pricing strategies.

Secondly, customer segmentation and purchase pattern analysis revealed valuable insights into customer behaviour. The "Consumer" segment emerged as the largest sales driver, while the "Corporate" segment exhibited higher average order values and profit margins. Understanding these nuances allows for targeted marketing and customer retention strategies.

Thirdly, the shipping analysis underscored the importance of optimizing shipping modes and costs. The "Standard Class" shipping mode, while cost-effective and popular, resulted in longer delivery times. Conversely, "Same Day" shipping, though expensive, offered faster delivery. Balancing these factors is crucial for customer satisfaction and operational efficiency. The need to analyse the shipping priority average costs was also noted, and this is a point that needs further analysis.

Geographical analysis, facilitated by Power BI visualizations, revealed regional variations in sales and profitability. The "US" market, particularly the "West" region, demonstrated strong sales performance, while other regions presented unique challenges and opportunities.

From this internship, I gained practical experience in the entire data analysis pipeline, from data cleaning and preprocessing to exploratory data analysis and visualization. I learned to leverage tools like PostgreSQL, MySQL, Pandas, and Power BI to extract meaningful insights from large datasets. The experience enhanced my ability to identify patterns, trends, and anomalies, and to communicate findings effectively through visualizations and reports.

**Limitations and Future Research**

While this analysis provided valuable insights, it is essential to acknowledge its limitations and potential areas for future research.

Firstly, the dataset, while comprehensive, had some limitations. The "Customer Lifetime Value (CLV)" analysis, for instance, was limited by the available data. A more detailed dataset with customer interaction history and retention metrics would enable a more accurate CLV calculation.

Secondly, the analysis focused primarily on descriptive statistics and visualizations. Advanced analytical techniques, such as predictive modelling and machine learning, could provide deeper insights into customer behaviour and sales forecasting. For example, regression analysis could be used to predict sales based on factors like discounts, shipping costs, and customer demographics.

Thirdly, the analysis of shipping costs could be further refined. A more granular analysis of shipping costs by product dimensions and weight could provide more accurate cost estimations. Additionally, incorporating external data, such as economic indicators and competitor pricing, could enhance the analysis and provide a broader context.

Future research could explore the following areas:

- **Predictive Modelling:** Developing predictive models to forecast sales, customer churn, and shipping costs.
- **Customer Segmentation:** Employing advanced clustering techniques to identify more nuanced customer segments and tailor marketing strategies.
- **Supply Chain Optimization:** Analysing inventory management and supply chain logistics to minimize costs and improve efficiency.
- **Product Recommendation Systems:** Developing recommendation systems to suggest products based on customer purchase history and preferences.
- **Dynamic Pricing Strategies:** Implementing dynamic pricing strategies based on demand, competition, and customer behaviour.
- **Shipping priority average costs analysis:** Investigating the average shipping costs for each priority level.
- **Discount optimization:** Analysing the optimal discount levels for different product categories and customer segments.

In conclusion, this internship provided a valuable learning experience in data analysis. By addressing the limitations and exploring the potential areas for future research, the insights gained from this analysis can be further leveraged to improve the Global Superstore's performance and drive strategic decision-making.

## 10. Bibliography

a. **Books and Articles:**

- "Data Science for Business" by Foster Provost and Tom Fawcett
- "Marketing Analytics: Data-Driven Techniques with Microsoft Excel" by Wayne L. Winston
- "Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die" by Eric Siegel

b. **Online Resources:**

- Kaggle: http://www.kaggle.com/
- Towards Data Science: http://www.towardsdatascience.com/