Retail Store Chain Analysis

Dataset:

Superstore Sales Dataset (https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting)

Objective:

The primary goal of this project is to analyze retail sales data to gain insights into product performance, customer behaviors, and sales trends across different regions and categories. This will support strategic decision-making to enhance sales performance and improve customer satisfaction.

Core Requirements:

Data Pipeline:

- Clean retail data using Python (handling missing values and duplicates).
- Process sales history with PySpark.
- Design an efficient PostgreSQL schema for retail analytics.

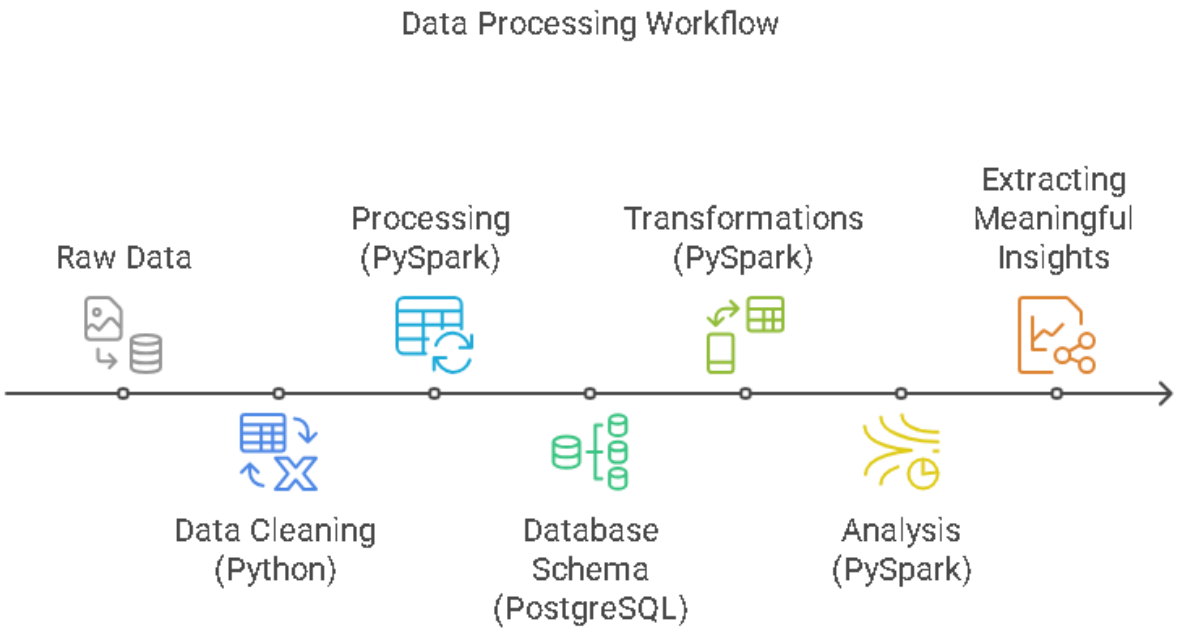
Data Transformations:

- Organize product hierarchy.
- Aggregate sales by region and category.
- Analyze customer purchase patterns.

Analysis Features:

- Product performance metrics.
- Regional sales comparison.
- Seasonal trend analysis.
- Customer segment profitability.
- Build KPI summary tables.

Data Flow Diagram:



Dataset Overview:

The dataset consists of 9800 records and 18 columns:

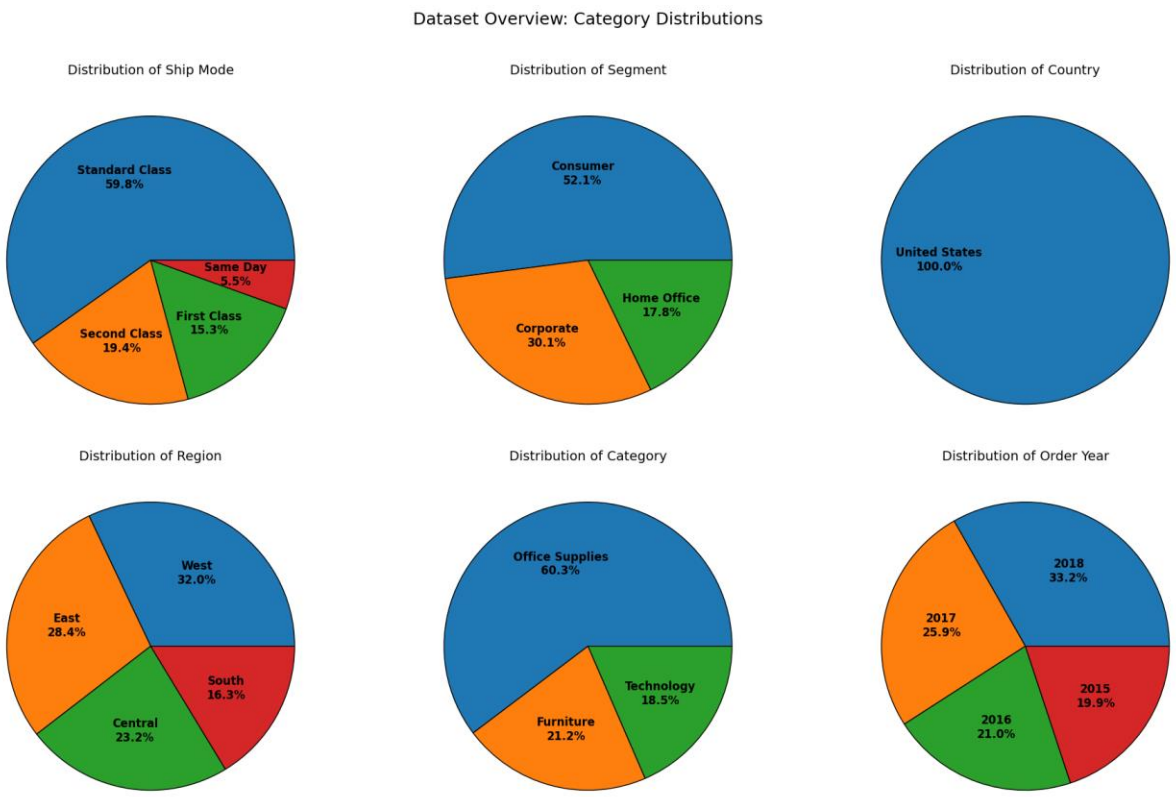
- Row ID: Unique identifier for each record.
- Order ID: Unique identifier for each order.
- Order Date: Date when the order was placed.
- Ship Date: Date when the order was shipped.
- Ship Mode: Mode of shipment used.
- Customer ID: Unique identifier for each customer.
- Customer Name: Name of the customer.
- Segment: Business segment of the customer.
- Country: Country where the order was placed.
- City: City where the order was placed.
- State: State where the order was placed.

- Postal Code: Postal code of the delivery location.
- Region: Geographic region of the order.
- Product ID: Unique identifier for each product.
- Category: Product category.
- Sub-Category: Sub-category under the product category.
- Product Name: Name of the product.
- Sales: Sales amount for the product.

Dataset Column Information

Column	Non-Null Count	Null Value Count	Dtype
Row ID	9800	0	int64
Order ID	9800	0	object
Order Date	9800	0	object
Ship Date	9800	0	object
Ship Mode	9800	0	object
Customer ID	9800	0	object
Customer Name	9800	0	object
Segment	9800	0	object
Country	9800	0	object
City	9800	0	object
State	9800	0	object
Postal Code	9800	11	float64
Region	9800	0	object
Product ID	9800	0	object
Category	9800	0	object
Sub-Category	9800	0	object
Product Name	9800	0	object
Sales	9800	0	float64

Missing Values: Postal Code has 11 missing values.
Data Types: Mostly object (strings) and float64 (numeric).



Ship Mode: Standard Class, Second Class, First Class, Same Day.
Segment: Consumer, Corporate, Home Office.
Country: United States.
Region: West, East, Central, South.
Category: Office Supplies, Furniture, Technology.
Order Year: 2018, 2017, 2016, 2015.

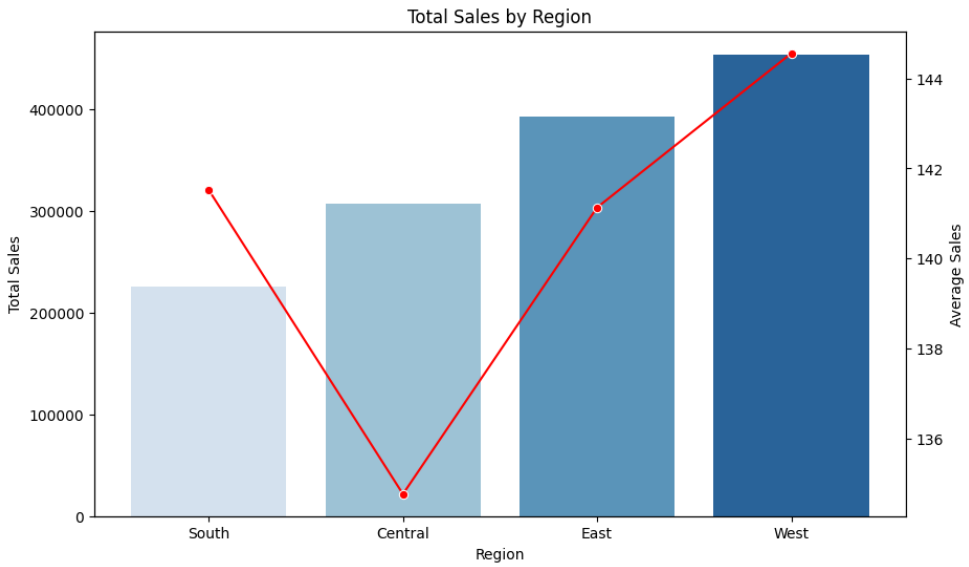
Steps of Operations Performed:

1. The dataset is loaded and inspected to understand the structure, data types, and missing values.
2. Missing values in the postal code column are identified and handled appropriately.
3. Duplicate records are checked and removed if necessary to ensure data integrity.
4. Data types are adjusted where needed, converting date and category fields into appropriate formats for analysis.
5. During sales history processing with PySpark, additional columns are created to extract useful information such as order year and month.
6. The cleaned data is processed and structured for efficient analysis.

7. The dataset is stored in a relational database for further processing and querying.
8. Data transformations are applied to prepare the dataset for deeper analysis.
9. Customer and sales-related insights are derived to understand purchasing patterns and trends.
10. The final dataset is analyzed to evaluate overall sales performance, customer segments, and regional distribution.
11. The results are visualized using charts and graphs to present key findings in an understandable format.
12. Performed sales forecasting for the next year using time-series features and machine learning models, incorporating additional features such as categorical variables, quarters, and one-hot encoding to enhance prediction accuracy.

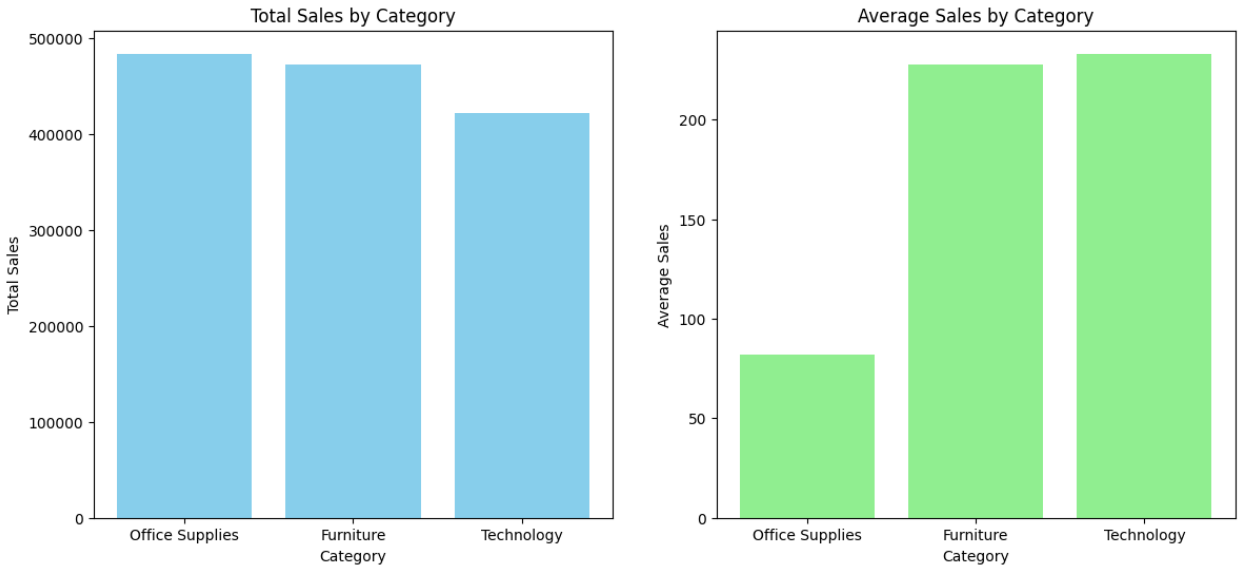
Analysis Findings:

Regional Performance:

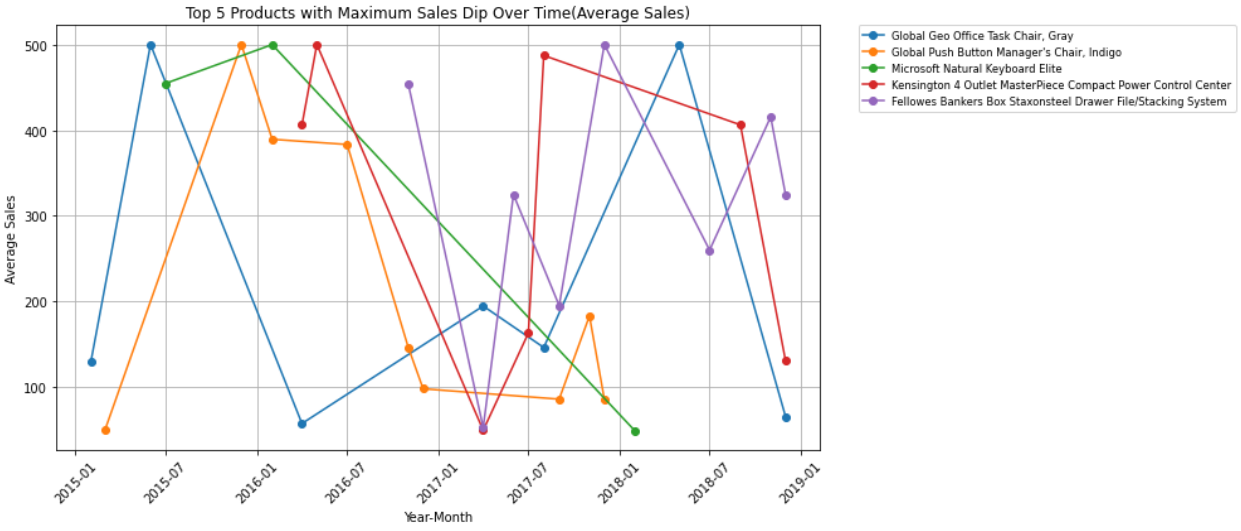


- The West region consistently outperforms others in total sales, contributing significantly to overall revenue.
- The East and Central regions show moderate performance, while the South region lags behind in total sales.

Product Performance:

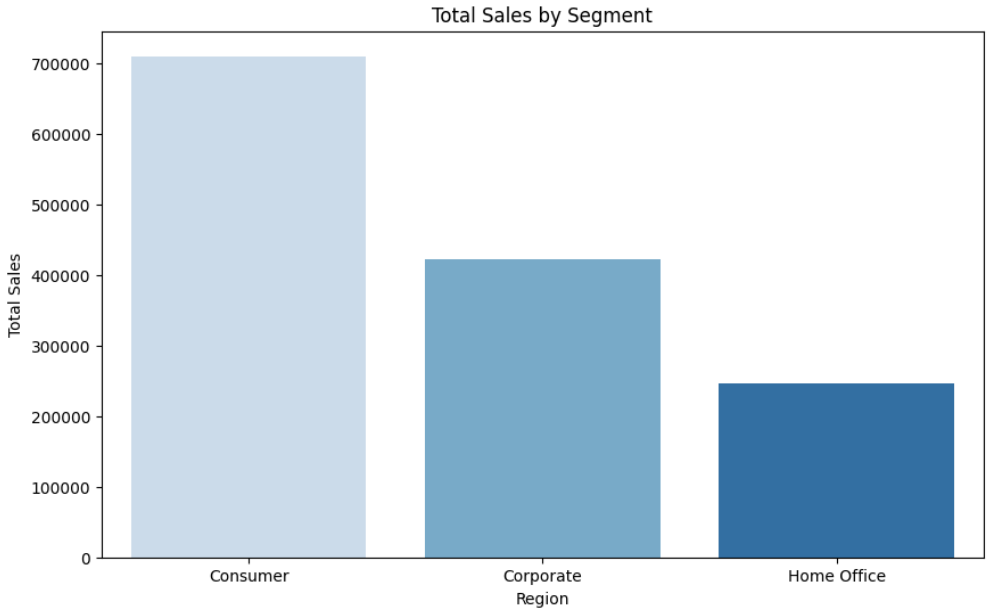


- Technology products generate the highest sales, with Phones being the top-selling sub-category. Despite having fewer products, Technology leads in profitability due to higher average sales per product.
- Furniture and Office Supplies show steady performance but lag behind Technology in terms of average revenue.



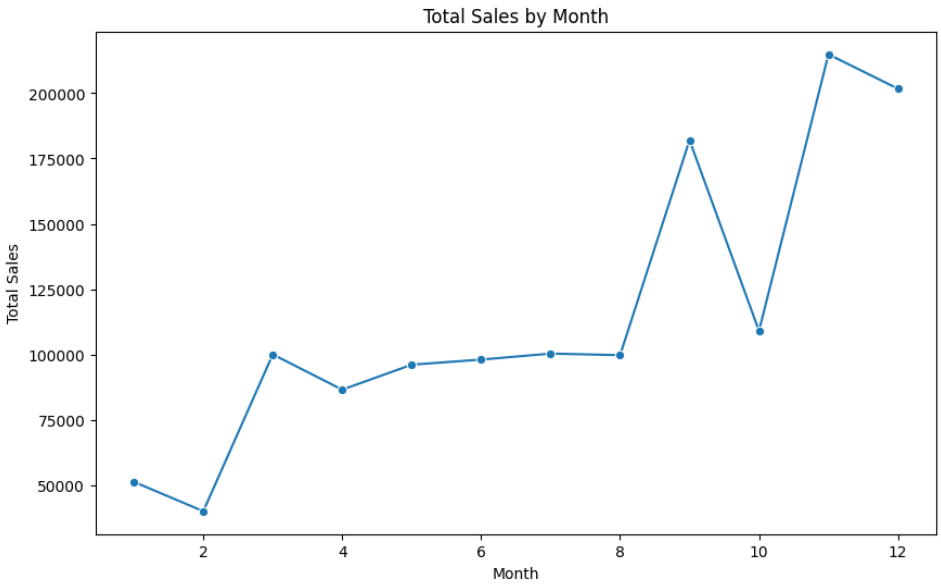
- Certain products experience significant dips in sales during specific periods.
- Seasonal trends or market competition might be influencing these drops.

Customer Insights:



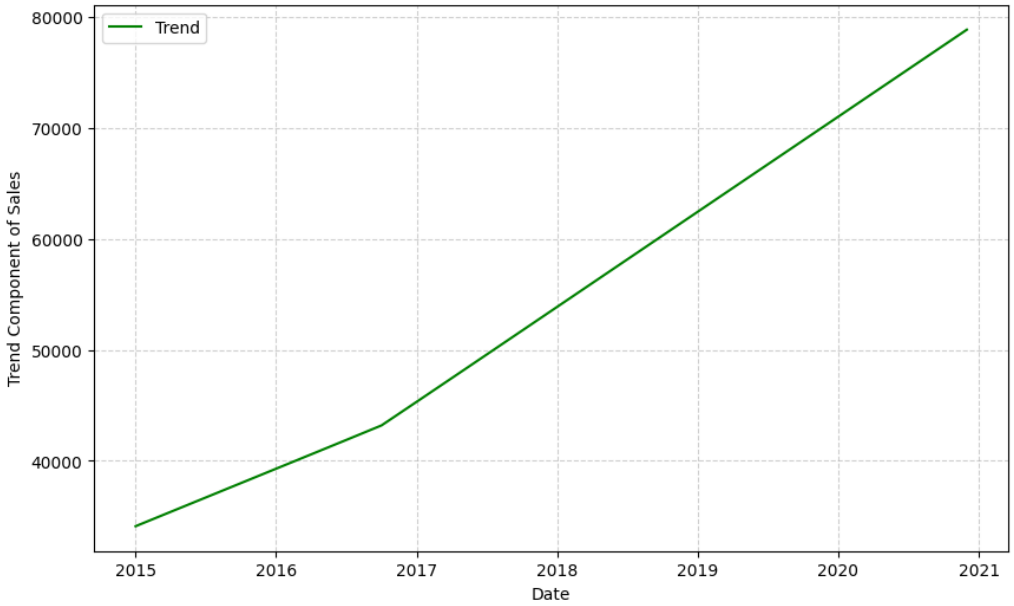
- Consumers are the largest customer segment, driving the majority of sales.
- Corporate customers have the highest average sales per transaction, indicating a preference for high-value purchases.
- Home Office customers contribute the least to total sales but show consistent purchasing patterns.

Seasonal Trends:



- Sales peak during the holiday season (November and December), driven by increased consumer spending.
- A noticeable dip in sales occurs in January and February, likely due to post-holiday spending fatigue.

Sales Forecasting :



- The trend analysis indicates a steady increase in sales, suggesting continued growth based on past patterns and seasonality.

Challenges:

- Assigning valid postal codes required domain expertise (e.g., "05401" for Burlington, Vermont).
- Ensuring unique records while merging data across tables was challenging due to potential conflicts in primary keys.
- Inconsistent data types, such as strings being interpreted as numbers, required careful handling during data transformation.

Learnings:

- Data cleaning and preprocessing are critical for ensuring accurate and reliable analysis.
- PySpark's distributed processing capabilities are invaluable for handling large datasets efficiently.
- Visualizations (e.g., bar plots, line charts) help communicate insights effectively to stakeholders.

Conclusion:

The project effectively demonstrated how to analyze retail data by cleaning, processing, and performing in-depth analytics using Python, PySpark, and PostgreSQL. The results provided valuable insights into customer behavior and product performance, guiding strategic decisions for the retail chain’s future.

Team Members:

Varsha U

Muneer A