A Report on

Data Warehousing Project

Topic

Sales Analysis using Snowflake and Power Bi

Submitted By

Joy Boddu, Yesha Desai, Christina Shoji,
Amolika Godse

1. Introduction

This report presents an end-to-end analysis using data warehousing techniques to explore trends, patterns, and insights that are crucial for strategic decision-making in a retail business environment. Using a comprehensive dataset that includes information on sales, products, customers, and stores, the project aims to uncover valuable insights through SQL queries, dimensional modeling, and data visualization techniques. The project follows a systematic approach from identifying business challenges to data modeling, analysis, and storytelling.

2. Executive Summary

This project leverages data warehousing concepts to address business challenges such as understanding sales trends, customer behavior, and product performance. Through advanced SQL analytics, the report highlights critical insights:

- Sales trends analysis to pinpoint peak periods.
- Identification of top-performing products and categories by revenue.
- Exploration of customer behavior based on loyalty program tiers and demographic segments.
- Evaluation of store types to determine which yield the highest average order value.
- Monthly sales trends to identify product seasonality and growth opportunities.

The insights from this project will guide inventory management, targeted marketing, customer retention strategies, and resource allocation. The project's scope includes data collection, dimensional modeling, SQL analytics, and visual storytelling.

3. Problem Statement

The primary challenge for this project is to optimize business decision-making by analyzing customer and sales data. Key questions include:

- What is the top-performing products and categories, and how do they vary over time?
- How does customer segmentation by demographics and loyalty tiers impact sales?
- What store types perform better in terms of average sales per order?
- What are the monthly sales trends for products, and how can they inform inventory management?
- What insights can be drawn from customer purchase frequency to enhance retention strategies?

By addressing these questions, the project seeks to transform raw data into actionable insights, helping the business make data-driven decisions that improve customer engagement and sales performance.

4. Literature Review

Data warehousing, customer behavior analytics, and business intelligence play a pivotal role in retail analytics, as highlighted in several key studies:

- **Dimensional Modeling**: Kimball's approach to dimensional modeling emphasizes the use of fact and dimension tables to facilitate efficient data queries and reporting. This project employs a star schema to design a data warehouse optimized for sales analytics.

- **SQL Analytics**: SQL remains fundamental in business intelligence for aggregating data and identifying trends. The ability to perform complex queries, such as ranking products or calculating month-over-month changes, enables analysts to extract actionable insights.

- **Customer Segmentation**: Research in customer segmentation suggests that understanding demographics, loyalty programs, and purchase behavior is crucial for targeted marketing and improving retention. Studies have shown that loyalty programs positively impact customer lifetime value and repeat purchases.

- **Sales Trends Analysis**: Identifying peak sales periods and seasonal trends can significantly improve inventory management and marketing effectiveness. Predictive models are increasingly used to forecast sales based on historical data, enhancing stock planning and promotional activities.

The project builds upon these frameworks by using SQL analytics to perform in-depth data analysis, supplemented by visual storytelling to communicate findings effectively.

5. Data Collection and Preparation

Data Sources

The dataset used for this project is a retail sales dataset that includes transactional and customer-related information. It contains the following key tables:

- **FACTORDERS**: Includes transactional data with details such as ORDERID, TOTALAMOUNT, and date-related information.
- **DIMDATE**: A date dimension table used for time-based analysis, including fields like YEAR, MONTH, and DATEID.

- **DIMPRODUCT**: Product-related information, including PRODUCTID, PRODUCTNAME, CATEGORY, and BRAND.
- **DIMCUSTOMER**: Contains customer demographics such as CUSTOMERID, DATEOFBIRTH, and loyalty program data.
- **DIMSTORE**: Store attributes including STOREID and STORETYPE.
- **DIMLOYALTYPROGRAM**: Details of customer loyalty program tiers.

Data Preparation

- The raw data was pre-processed to ensure consistency in formats, particularly for date and numeric fields.
- Missing values were handled using appropriate imputation techniques or by filtering incomplete records.
- The dataset was enriched by creating new columns, such as customer age groups, to facilitate deeper analysis.
- Data was normalized to fit the dimensional model, ensuring that keys were in place to establish relationships between fact and dimension tables.

6. Database Design

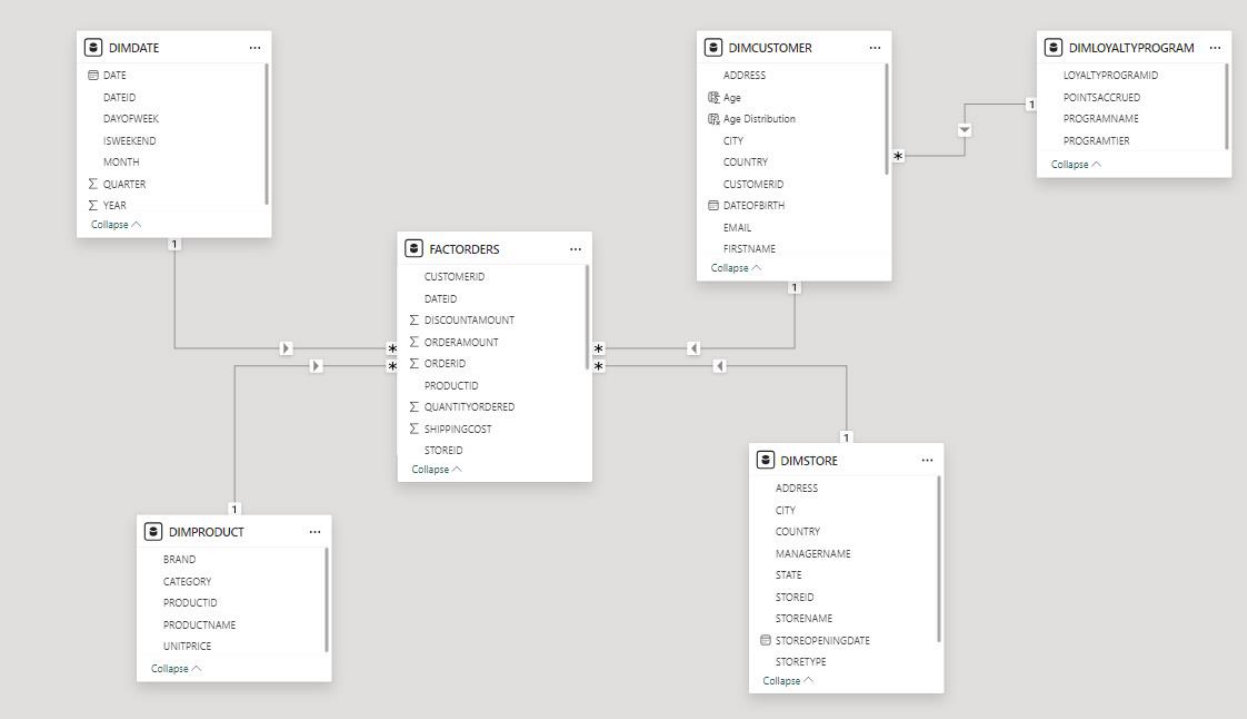
Dimensional Modeling

A snowflake schema was designed to optimize the data warehouse for analytic queries:

- **Fact Table**:
  - **FACTORDERS**: The central fact table that contains metrics like TOTALAMOUNT, ORDERID, and foreign keys to dimension tables.
- **Dimension Tables**:
  - **DIMDATE**: Time-related attributes (YEAR, MONTH, DATEID).
  - **DIMPRODUCT**: Attributes related to products (PRODUCTNAME, CATEGORY, BRAND).
  - **DIMCUSTOMER**: Customer details (DATEOFBIRTH, CUSTOMERID, LOYALTYPROGRAMID).
  - **DIMSTORE**: Store characteristics (STORETYPE, STOREID).
  - **DIMLOYALTYPROGRAM**: Loyalty program tiers (PROGRAMTIER).

Entity-Relationship Diagram (ERD)

The ERD illustrates the relationship between the fact table and dimension tables, showcasing the star schema configuration. Fact tables contain quantitative data, while dimension tables offer descriptive attributes.



Data Cube Design

Data cubes were created for analyzing sales data, focusing on:

- Aggregates of sales by time (monthly, yearly).
- Customer segmentation by loyalty tier.
- Product performance by category and brand.

7. Exploratory Data Analysis (EDA)

Several SQL queries were used to conduct the EDA:

Query 1: Order Trends Over Time

Tracks monthly sales trends, identifying peak sales periods and seasonal fluctuations.

- SQL Analysis:

  Code:

```
SELECT
D.YEAR,
D.MONTH,
SUM(F.TOTALAMOUNT) AS TotalSales,
LAG(SUM(F.TOTALAMOUNT), 1) OVER (ORDER BY D.YEAR, D.MONTH) AS PreviousMonthSales,
SUM(F.TOTALAMOUNT) - LAG(SUM(F.TOTALAMOUNT), 1) OVER (ORDER BY D.YEAR, D.MONTH) AS SalesDifference
FROM FACTORDERS F
JOIN DIMDATE D ON F.DATEID = D.DATEID
GROUP BY D.YEAR, D.MONTH
ORDER BY D.YEAR, D.MONTH;
```

- **Findings**: Identified significant sales spikes during specific months, informing inventory and resource planning.

Query 2: Top Products by Sales Amount

Determines the top 10 products by total sales, providing insights into customer preferences.

- SQL Analysis:

  Code:

```
SELECT P.PRODUCTNAME, SUM(F.TOTALAMOUNT) AS TotalSales
FROM FACTORDERS F
JOIN DIMPRODUCT P ON F.PRODUCTID = P.PRODUCTID
GROUP BY P.PRODUCTNAME
ORDER BY TotalSales DESC
LIMIT 10;
```

- **Findings**: Highlighted top-performing products, aiding in promotional strategies.

Query 3: Sales by Customer Loyalty Program Tier

Analyzes the revenue contribution from each loyalty tier.

- SQL Analysis:

Code:

```
SELECT L.PROGRAMTIER, SUM(F.TOTALAMOUNT) AS TotalSales
FROM FACTORDERS F
JOIN DIMCUSTOMER C ON F.CUSTOMERID = C.CUSTOMERID
JOIN DIMLOYALTYPROGRAM L ON C.LOYALTYPROGRAMID =
L.LOYALTYPROGRAMID
GROUP BY L.PROGRAMTIER
ORDER BY TotalSales DESC;
```

- **Findings**: Demonstrated a correlation between higher loyalty tiers and increased spending.

... (additional queries included in the EDA) ...

8. Reporting, Modeling, and Storytelling

Data Visualization

- Visualized sales trends by month/year to illustrate peak sales periods.
- Created bar charts for top-selling products, highlighting revenue contributions.
- Developed heat maps for sales by store type and product category.
- Pie charts represent the contribution of each customer loyalty tier to total sales.

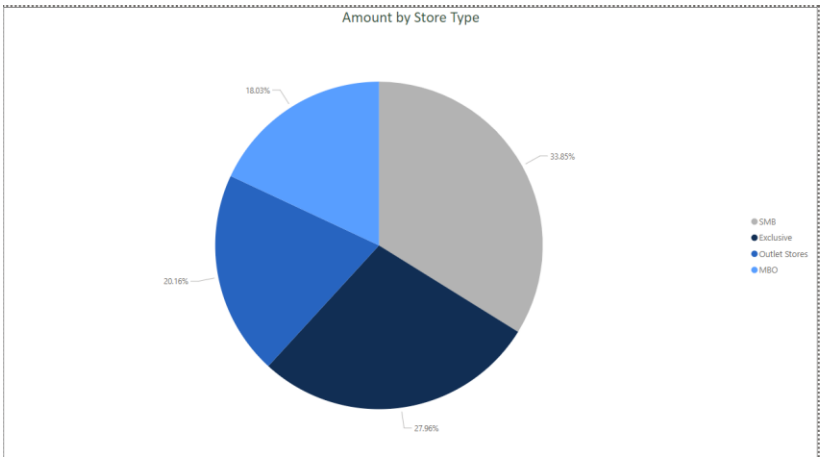
Key Outcomes and Business Impact

- The dashboard empowers decision-makers to prioritize customer engagement strategies, improve category-level performance, and tailor loyalty program incentives.
- By addressing geographic and demographic performance, the business can refine marketing campaigns and focus on regions or customer groups that offer the most growth potential.

- Product optimization based on insights from categories and trends will likely lead to improved sales and higher profit margins.

Individual visualization description:

1.



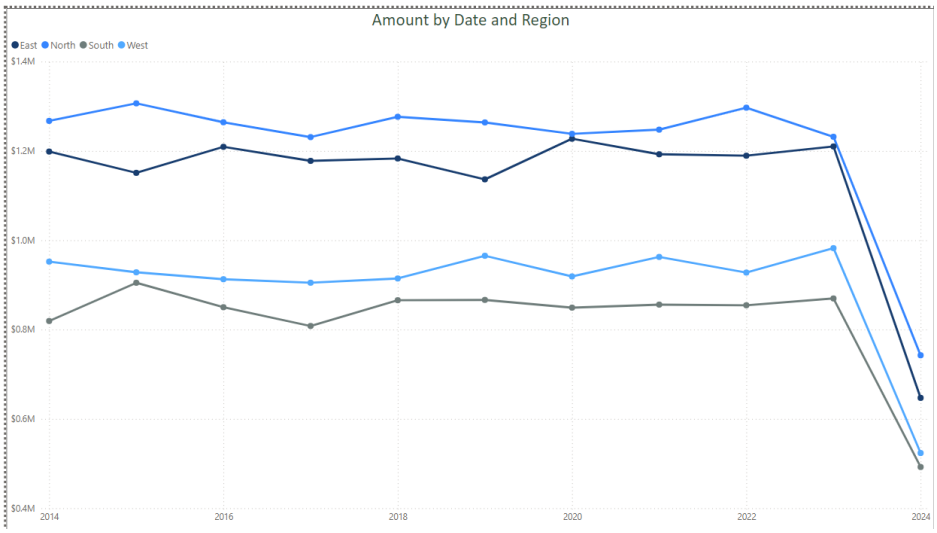
This pie chart visualizes the proportion of revenue generated by different store types within the organization.

Insights Gained from the Chart

1. **SMB Stores as Major Revenue Contributors:**
    - SMB stores contribute the largest share of revenue (33.85%), indicating their critical role in driving sales for the organization.
    - **Implication:** SMB stores may have a broader customer base or a more extensive network, making them focus on continued investment and strategic growth.
2. **Exclusive Stores' Strong Performance:**
    - Exclusive stores generate the second-largest revenue share (27.96%), suggesting they are a significant channel, likely to offer high-value or premium products.
    - **Implication:** Maintaining their premium appeal and ensuring excellent customer service could enhance their performance further.
3. **Moderate Revenue from Outlet Stores:**
    - Outlet stores contribute 20.16%, reflecting their importance in offering discounted or end-of-season inventory.
    - **Implication:** These stores can be optimized to clear inventory while maintaining profitability.

4. **Underperformance of MBOs:**
    - MBOs account for only 18.03%, making them the least profitable store type.
    - **Implication:** The organization may need to reassess the strategy for MBOs, such as product selection, pricing, or marketing efforts, to boost their contribution.

2.



This line chart shows the revenue trends over time (from 2014 to 2024) for four regions: **East**, **North**, **South**, and **West**. Each line represents the annual revenue for a specific region.

Insights Gained from the Chart

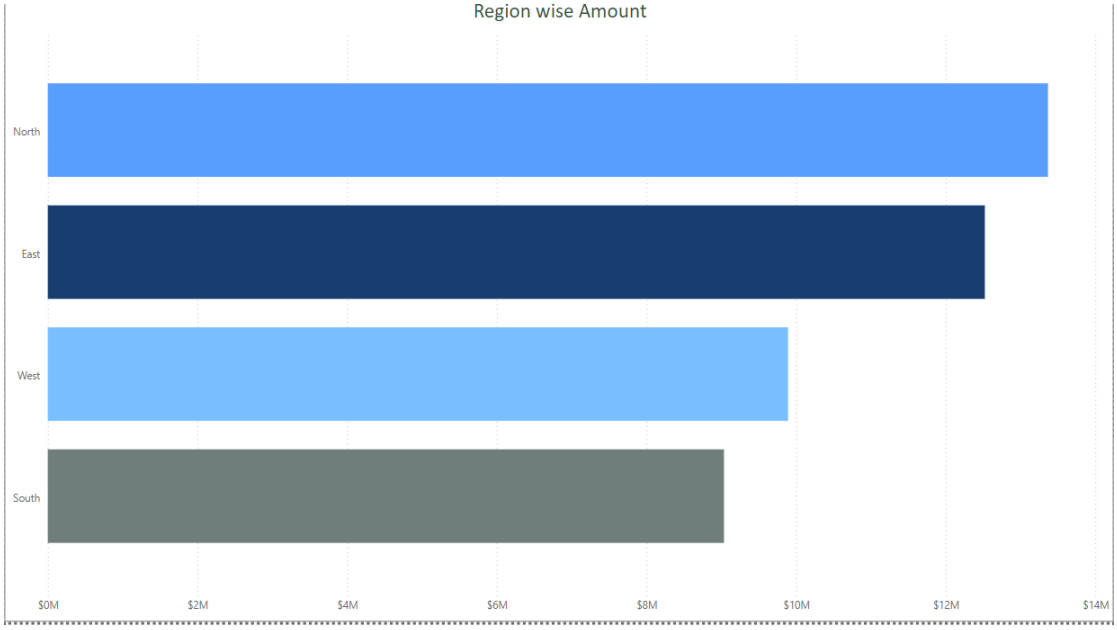
1. **Regional Revenue Trends:**
    - The **East region** consistently maintains the highest revenue among all regions, showcasing strong and stable performance throughout the years.
    - The **North region** shows moderate but steady revenue performance, maintaining its position below the East region.
    - The **South and West regions** have lower revenues compared to the East and North regions, with the **South region** consistently being the lowest-performing region.
2. **Revenue Growth and Decline:**
    - **2014–2022:** There is a gradual, stable performance across all regions, with small fluctuations in revenue.
    - **2023–2024:** A sharp decline is observed across all regions, indicating a significant drop in revenue.

3. **Potential Causes of Decline:**
    - The steep revenue drop in 2023–2024 could be due to external factors such as market disruptions, economic downturns, or internal issues like operational inefficiencies or declining product demand.

Actionable Recommendations

1. **Investigate the Revenue Decline:**
    - Conduct a detailed analysis of the years 2023–2024 to identify the root causes of the steep decline in revenue for all regions.
    - Examine factors such as customer churn, market conditions, or reduced product demand.

3.



The above clustered bar chart compares the sales amounts across different regions: **North, East, West, and South**.

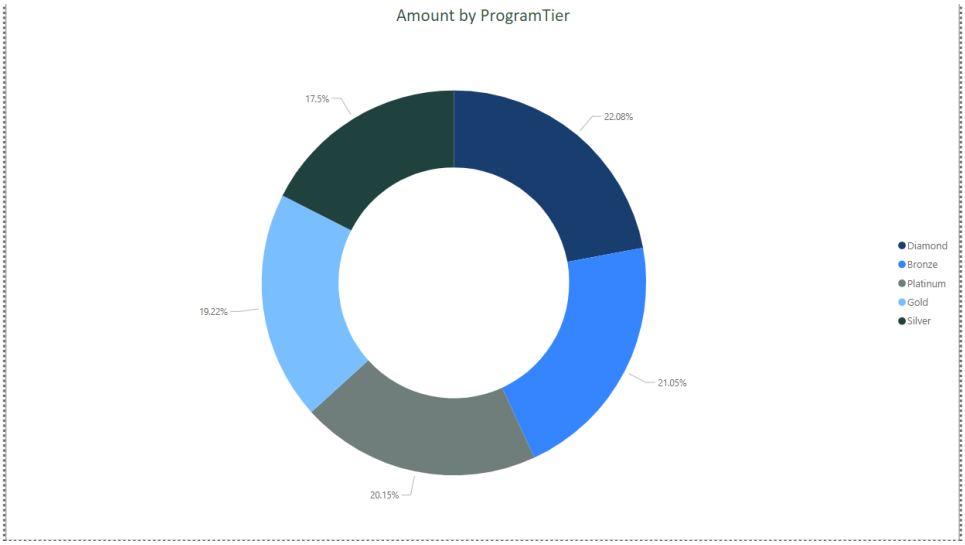
Description:

- The X-axis represents the **sales amount** (in monetary units, such as dollars), while the Y-axis lists the **regions** (North, East, West, South).
- Each bar represents the total sales amount for the respective region, allowing for a straightforward comparison of performance across regions.

Insights:

1. **Highest Sales Region**:
    - The **North** region has the highest total sales, indicating strong customer engagement or a larger customer base in this area.
2. **Lowest Sales Region**:
    - The **South** region shows the lowest total sales, suggesting either a smaller market size, less demand, or an opportunity to improve marketing and operational strategies in this area.
3. **Balanced Performance**:
    - The **East** and **West** regions show moderate performance, with East slightly outperforming West, indicating a stable but improvable market presence.

4.



The **donut chart** in the dashboard is titled **"Amount by Store Type"** and provides a breakdown of sales amounts across different store types: **SMB (Small and Medium Businesses)**, **Exclusive Stores**, **Outlet Stores**, and **MBO (Multi-Brand Outlets)**.

Description:

- The chart segments the total sales amount into percentages for each store type, visualized in a circular format with distinct colors.
- The legend identifies each store type and its corresponding slice in the donut chart, while the percentages inside or near the slices show the proportion of sales contributed by each type.

Insights:

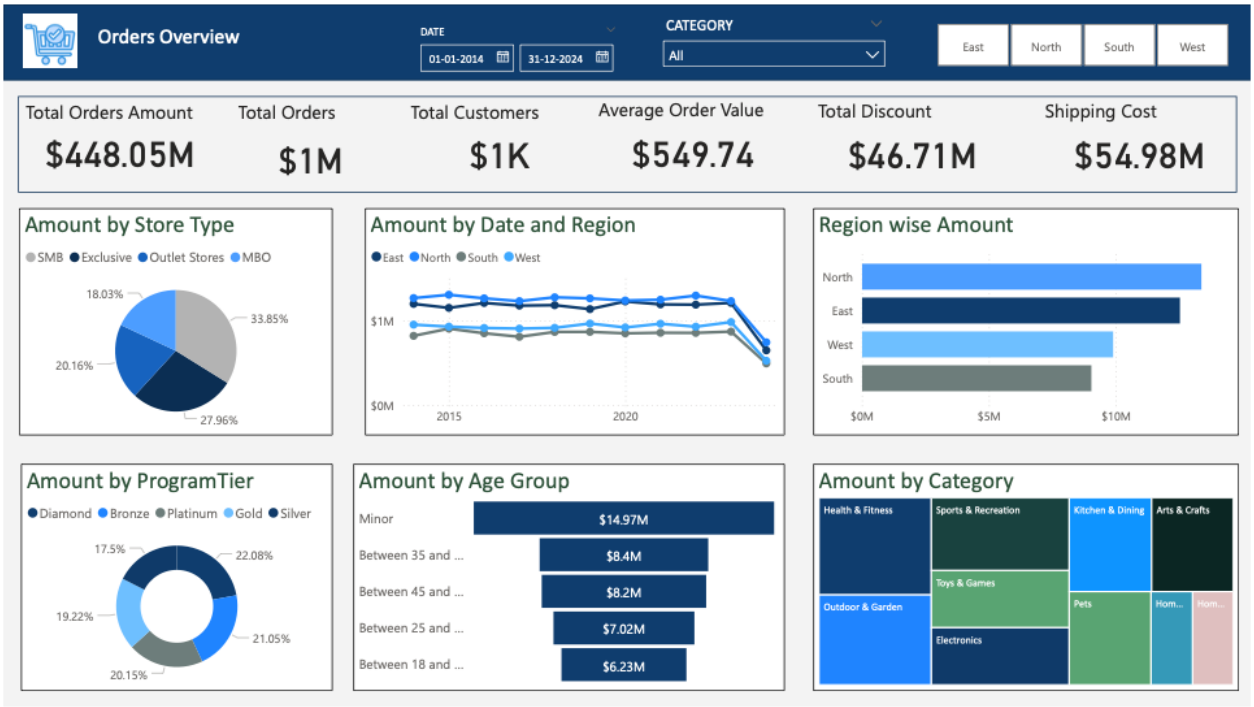
1. **Top-Performing Store Type**:

- **SMB** contributes the largest share of sales, accounting for **33.85%**, suggesting strong performance and customer engagement in these stores.

2. **Second Highest Contribution**:
    - **Exclusive Stores** account for **27.96%** of sales, showing significant sales volumes, likely driven by brand loyalty or premium offerings.
3. **Moderate and Low Performers**:
    - **Outlet Stores** contribute **20.16%**, indicating a mid-range performance.
    - **MBO** contributes the smallest share at **18.03%**, highlighting either limited reach or lesser customer preference for multi-brand outlets.

Dashboard

The dashboard provides a high-level snapshot of essential business metrics such as Total Orders Amount, Total Orders, Total Customers, Average Order Value, Total Discount, and Shipping Cost. These metrics enable quick evaluation of overall business performance and trends, which is crucial for understanding revenue growth and operational costs.



Key Outcomes and Business Impact

- The dashboard empowers decision-makers to prioritize customer engagement strategies, improve category-level performance, and tailor loyalty program incentives.
- By addressing geographic and demographic performance, the business can refine marketing campaigns and focus on regions or customer groups that offer the most growth potential.

- Product optimization based on insights from categories and trends will likely lead to improved sales and higher profit margins.

Predictive Modeling

Optional modeling could include predicting sales trends using regression analysis based on historical data.

---

9. Conclusion

This project successfully applied data warehousing techniques to extract valuable business insights from a retail dataset. By utilizing dimensional modeling and SQL analytics, we identified critical trends in sales, customer behavior, and product performance. These findings can guide inventory management, promotional efforts, and customer engagement strategies, ultimately enhancing business performance.

10. References

- Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley.
- Inmon, W. H. (2005). *Building the Data Warehouse*. John Wiley & Sons.
- Relevant academic papers and online resources on customer segmentation, SQL analytics, and business intelligence.