

**Amrita Vishwa Vidyapeetham**

**Centre for Excellence in Computational Engineering and Networking**

**Amrita School of Artificial Intelligence, Coimbatore**

**Topic: Superstore sale analysis**

**21AIE304**

**Submitted by:**

**Batch A Group – 6**

**Team members:**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **NAME** | **ROLL.NO** |
| **1.** | **MARREDDY MOHIT SASANK REDDY** | **CB.EN.U4AIE21031** |
| **2.** | **YARRAM SRI SATHWIK REDDY** | **CB.EN.U4AIE21077** |
| **3.** | **P.SAI VIJAY KUMAR** | **CB.EN.U4AIE21040** |
| **4.** | **B. HARISH BALAJI** | **CB.EN.U4AIE21007** |

**Supervised by:**

**Dr. Sanjanasri**

An End Semester Project submitted to the CEN department as a part of course evaluations of **Big Data and Database Management**  for B. Tech in **Artificial Intelligence**

**ACKNOWLEDGMENT**

We are deeply thankful to Centre for Excellence in Computational Engineering and Networking (CEN) at Amrita Vishwa Vidyapeetham, Coimbatore for providing us such a wonderful environment to pursue our research. We would like to express our sincere gratitude to Dr. Sanjanasri, Asst. Professor, Department of Centre for Excellence in CEN, Amrita Vishwa Vidyapeetham. We have completed our research under her guidance. We found the research area, topic, and problem with her suggestions. We would also like to acknowledge our team members for supporting each other and be grateful to our university for providing this opportunity for us. Lastly special thanks to Centre for Excellence in CEN for providing this opportunity to research in this field.

**Abstract**

The orders Spark application is a robust data analytics tool designed to analyze and derive insights from sales data stored in a MySQL database. Leveraging Apache Spark, the application integrates seamlessly with the database to perform a wide array of analyses.The dataset captures sales and order information for a retail business operating in the United States. The data spans multiple years and includes details such as order ID, order date, shipping information, customer details, product information, and sales figures. The dataset encompasses a variety of product categories, including furniture, office supplies, and technology.

The project analysis involves examining patterns and trends within the dataset to gain insights into the company's performance. Key areas of focus may include sales distribution across different regions, customer segments, and product categories. Analyzing the shipping modes and order processing times could provide valuable information on logistics efficiency.

**MOTIVATION OF PROJECT**

In today's highly competitive business landscape, data-driven decision-making is crucial for companies aiming to thrive and adapt to changing market dynamics. This project seeks to leverage comprehensive sales and order data to extract meaningful insights that can drive strategic initiatives for the retail business under consideration.

Understanding customer preferences, optimizing logistics and supply chain operations, and identifying high-performing product categories are integral to staying ahead in the market. By delving into the dataset, we aim to uncover patterns and trends that can inform key business decisions. Additionally, the project will explore the effectiveness of different shipping modes, assess regional variations in sales, and identify opportunities for targeted marketing efforts.

**INTRODUCTION**

The dataset, comprised of diverse details such as order ID, order date, shipping information, customer profiles, and product specifics, offers a panoramic view of the company's interactions with its market. This project endeavors to uncover hidden patterns and trends within the data, providing a deeper understanding of customer behavior, product preferences, and the efficiency of supply chain logistics.

Through a combination of descriptive statistics, data visualization techniques, and targeted analyses, this project seeks to address key questions surrounding sales distribution, customer segmentation, shipping efficiency, and the performance of distinct product categories. The ultimate goal is to equip the retail business with the knowledge required to make informed decisions, capitalize on emerging opportunities, and navigate challenges adeptly in the pursuit of sustained success.

**Database Setup and Data Loading Report:**

The script initializes a database named "super\_store" and creates a table "orders" with various columns to store retail data.

Data is loaded into the "orders" table from a CSV file using the LOAD DATA INFILE statement, specifying field and line terminators while ignoring the CSV header.

The script displays all rows from the "orders" table to inspect the imported data.

Safe updates are temporarily disabled, and the "Ship\_Date" column is updated to ensure consistent date formatting using MySQL functions.

The script facilitates a comprehensive setup, data loading, inspection, and necessary data manipulation for subsequent analysis of the retail dataset.

**Integration with Database**

The application integrates with a MySQL database named "vegetable\_db" using the JDBC connector. The database connection details, such as URL, username (root), and password, are specified in the code. The data is loaded into Spark DataFrames using the get\_table method, which leverages Spark's JDBC data source.

**Analysis and Methodology**

1. Analysis 1 - Customer Order Analysis:

* Method:
  + main (includes the entire analysis)
* Purpose:
  + Calculates the total orders and average sales per customer.
  + Determines customer categories (Gold, Silver, Bronze) based on average order values.

1. Analysis 2 - Weekly Sales Analysis:

* Method:
  + df\_weekly
* Purpose:
  + Analyzes and displays total sales on each day of the week.

1. Analysis 3 - Shipment Analysis:

* Method:
  + avgDaysByCategoryAndShipMode
* Purpose:
  + Calculates the average days between order and shipment for specific shipping modes and categories.

1. Analysis 4 - Seasonal Analysis:

* Method:
  + mostOrderedCategoryBySeason
* Purpose:
  + Analyzes and displays the total orders by season and category.

1. Analysis 5 – Sales and Revenue Percentage Analysis

* Method:
  + totalSalesByCategoryAndSubcategory
  + totalRevenueByCategory
  + revenuePercentageByCategoryAndSubcategory
* Purpose:
  + Calculates the percentage of total revenue for each subcategory within each category.
  + Orders and displays the resulting DataFrame showing the revenue percentage for each subcategory within each category.

1. Analysis 6 – Revenue Analysis

* Method:
  + revenueByRegionAndCategory
* Purpose:
  + Computes total revenue by region and category.

1. Analysis 7 - Sales Analysis:

* Method:
  + yearlyMonthlySales
* Purpose:
  + Displays yearly and monthly sales for the year 2017.

1. Analysis 8 - Delay Analysis

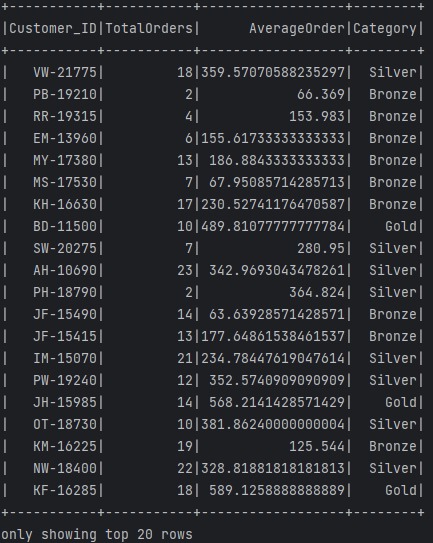
* Method:
  + citiesWithMostDelay
  + citiesWithLeastDelay
* Purpose:
  + Calculates shipment delays by category and city.
  + Identifies cities with the most and least delay for each category.

1. Analysis 9 - Yearly and State Profitability Analysis

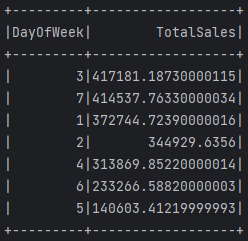
* Method:
  + mostProfitableStateYearPerRegionDF
  + leastProfitableStateYearPerRegionDF
* Purpose:
  + Analyzes and displays total sales for each year and month.
  + Identifies the most and least profitable states per region based on total sales.

**Results**

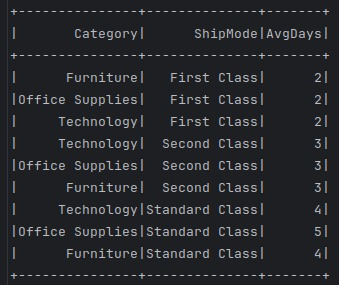
1. Analysis 1 - Customer Order Analysis:



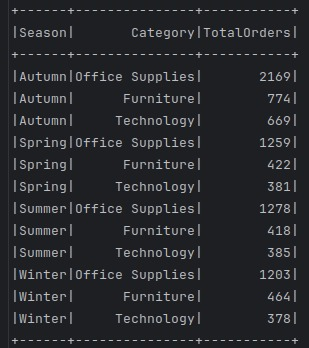
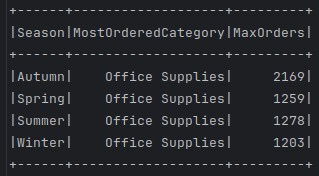
1. Analysis 2 - Weekly Sales Analysis:



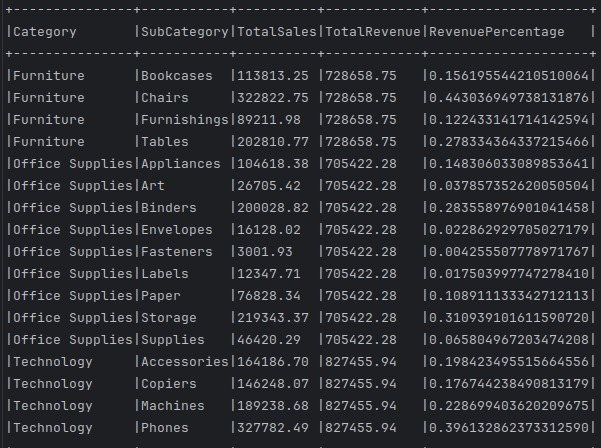
1. Analysis 3 - Shipment Analysis:



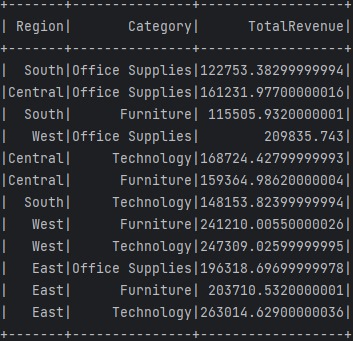
1. Analysis 4 - Seasonal Analysis:

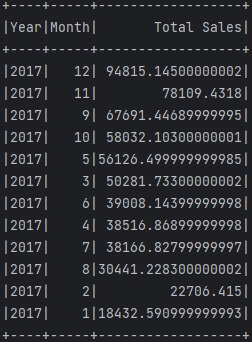
1. Analysis 5 – Sales and Revenue Percentage Analysis



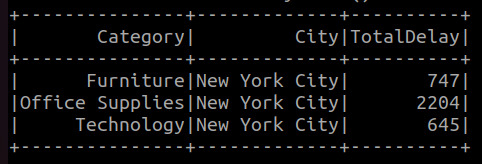
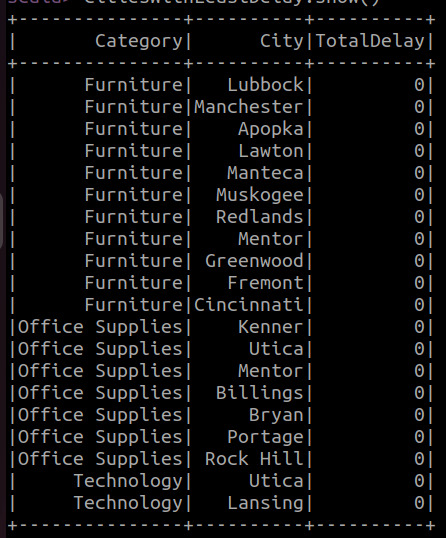
1. Analysis 6 – Revenue Analysis



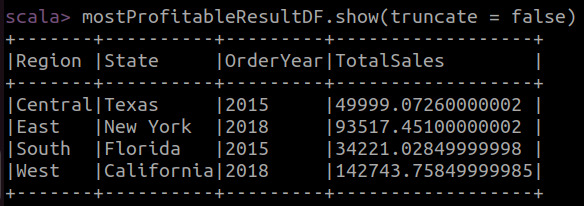
1. Analysis 7 - Sales Analysis:

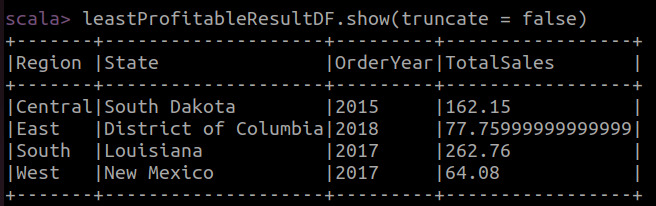


1. Analysis 8 - Delay Analysis

1. Analysis 9 - Yearly and State Profitability Analysis





**Conclusion**

The examination of the Superstore Sales Dataset, covering a four-year period, has yielded valuable insights across various dimensions of the retail domain. Customer segmentation has facilitated the categorization of clients into Gold, Silver, and Bronze, guided by their average order values, enabling tailored strategies for diverse customer segments.

Operational efficiency insights, particularly regarding order-to-shipment times, highlight opportunities for process optimization and improved customer satisfaction. Uncovering seasonal trends offers a nuanced understanding of customer preferences during different times of the year, supporting strategic planning.

The analysis of total revenue by region and category contributes to a clearer picture of the geographical distribution of sales, aiding in resource allocation and targeted marketing efforts. Identification of the most and least profitable states per region provides actionable intelligence for refining sales strategies.

This data-centric approach underscores the pivotal role of data in decision-making, offering a comprehensive view of product performance, profitability, and regional dynamics, crucial for sustained growth in the dynamic global retail landscape.