Final Project DS5110: Customer Behavior and Sales Trend Analysis Dashboard

Karthik Nair, Filip Tomovski

Fall 2024

1 Introduction

This project aims to develop a real-time sales analytics dashboard integrating multiple data sources, including sales data, customer demographics, weather data, holidays, and search trends. The objective is to provide insights into the influence of external factors on customer behavior and sales, aiding strategic decisions in marketing, inventory management, and sales optimization.

2 Project Goals and Scope

2.1 Specific Goals

- Build a dashboard with real-time data that visualizes trends and behaviors.
- Provide actionable insights for marketing, inventory management, and sales optimization.
- Integrate external data sources (weather, holiday, search trends) to enhance analysis depth.

2.2 Project Scope

The scope is limited to the integration of specific data sources and creating a dashboard with interactive visualizations. This includes:

- Data integration, ETL (Extract, Transform, Load) processes.
- Dashboard UI design, data visualization, and interactivity.
- Excluding complex features beyond the primary goals, like unrelated data sources or advanced ML models.

3 Data and Data Processing

3.1 Data Sources

The project utilizes multiple data sources to provide a comprehensive view of customer behavior and sales trends. The key data sources include:

- **Sales Data**: Historical sales records including transactional details such as product type, quantity, price, and transaction timestamp.
- **Customer Demographics**: Customer data providing demographic insights such as age, gender, and location, aiding in segmentation analysis.
- **Weather Data (API)**: External weather information sourced via real-time API integration, offering daily and seasonal patterns that can affect consumer behavior.
- **Holiday Data**: A calendar of national and local holidays to assess variations in sales and customer engagement around specific dates.
- **Search Trends (API)**: Data from search trends that reflect consumer interest in certain products or categories, contributing to predictive insights.

3.2 Data Acquisition and Integration

The data acquisition process involves sourcing data from internal databases and external APIs. The primary sales and customer demographic data are sourced from an internal repository, while the weather and search trends data are gathered through API calls. These external data sources are integrated into the primary dataset to enable analysis of external factors' impact on sales.

3.3 ETL Pipeline

An ETL (Extract, Transform, Load) pipeline has been established to streamline the data ingestion process. This pipeline is responsible for:

- **Extracting** data from diverse sources, including internal databases and external APIs.
- **Transforming** data to ensure consistency across formats, handling missing values, and standardizing timestamps and categorical variables.
- **Loading** processed data into a structured database schema designed for efficient querying and data retrieval for dashboard visualization.

3.4 Data Cleaning and Preparation

Data cleaning is performed to address issues such as incomplete records, missing values, and format inconsistencies. For instance, missing demographic details are handled with imputation techniques, while outliers in sales data are reviewed to maintain data integrity. Additionally, categorical data are encoded, and continuous variables are normalized to facilitate smoother data processing and analysis.

3.5 Data Integration and Storage

The integrated dataset is stored in a relational database (MySQL or PostgreSQL), organized to support efficient querying for visualizations. Tables are designed to capture relationships between sales transactions, customer segments, weather conditions, and holiday data, enabling seamless access to combined insights. This database structure underpins the real-time sales dashboard, allowing for dynamic updates as new data is ingested.

3.6 Real-Time Data Processing and API Integration

For real-time insights, external APIs for weather and search trends are incorporated into the ETL pipeline, enabling automated, scheduled updates. This integration ensures that the dashboard reflects up-to-date information, providing insights on current trends that influence customer behavior. To handle real-time data streams, the system uses asynchronous API calls, buffering updates for consistent availability without overwhelming the database.

3.7 Exploratory Data Analysis (EDA)

Initial exploratory data analysis (EDA) was conducted to understand the distributions, correlations, and trends within the data. This step revealed key patterns, such as the impact of seasonal changes on product demand and the role of demographic factors in purchasing preferences. EDA insights guide the development of the dashboard's visualizations and help identify areas for potential predictive modeling.