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 that integrates multiple data sources, including sales data, customer demographics, holidays, and search trends. The goal is to provide actionable insights into customer behavior and sales trends, aiding strategic decision-making in areas like marketing, inventory management, and sales optimization.

The dashboard emphasizes ease of use, interactivity, and comprehensive visualizations, enabling businesses to make informed decisions based on data.

2 Project Goals and Scope

2.1 Specific Goals

- Develop a real-time dashboard that visualizes customer behavior, sales trends, and geographic distribution.

- Provide actionable insights for business operations, including marketing campaigns, inventory allocation, and seasonal sales planning.

- Incorporate external data sources (holidays, search trends) to enhance analytical depth and accuracy.

2.2 Project Scope

- Data collection and processing, including sales records, customer demographics, and holiday trends.

- Design of a robust data pipeline and database architecture for efficient storage and retrieval.

- Development of an intuitive dashboard with advanced interactivity and data exploration capabilities.

- Creation of visualizations that highlight trends in sales, customer behavior, and geographic performance.

3 System Design and Architecture

3.1 Entity-Relationship Diagram

The Entity-Relationship (ER) diagram, shown in Figure 1, forms the foundation of the data schema for this project. It captures the relationships between sales transactions, customer demographics, store locations, and holidays.

- **Key Entities:**

- **Sales:** Contains transactional data such as sale ID, date, product category, sales amount, and links to customer and location.
- **Customer Demographics:** Stores attributes like age, gender, income level, and location, enabling segmentation analysis.
- **Locations:** Includes geographical data such as city, state, latitude, and longitude for spatial analysis.
- **Holidays:** Tracks holiday dates, names, and associated regions for analyzing seasonal trends.

- **Relationships:**

  - Sales are linked to customers through 'customer$_i$d'andtolocationsthrough'location$_i$d'.Holidaysarelinkedtosalesviathe'd $relatedsalesspikes.$
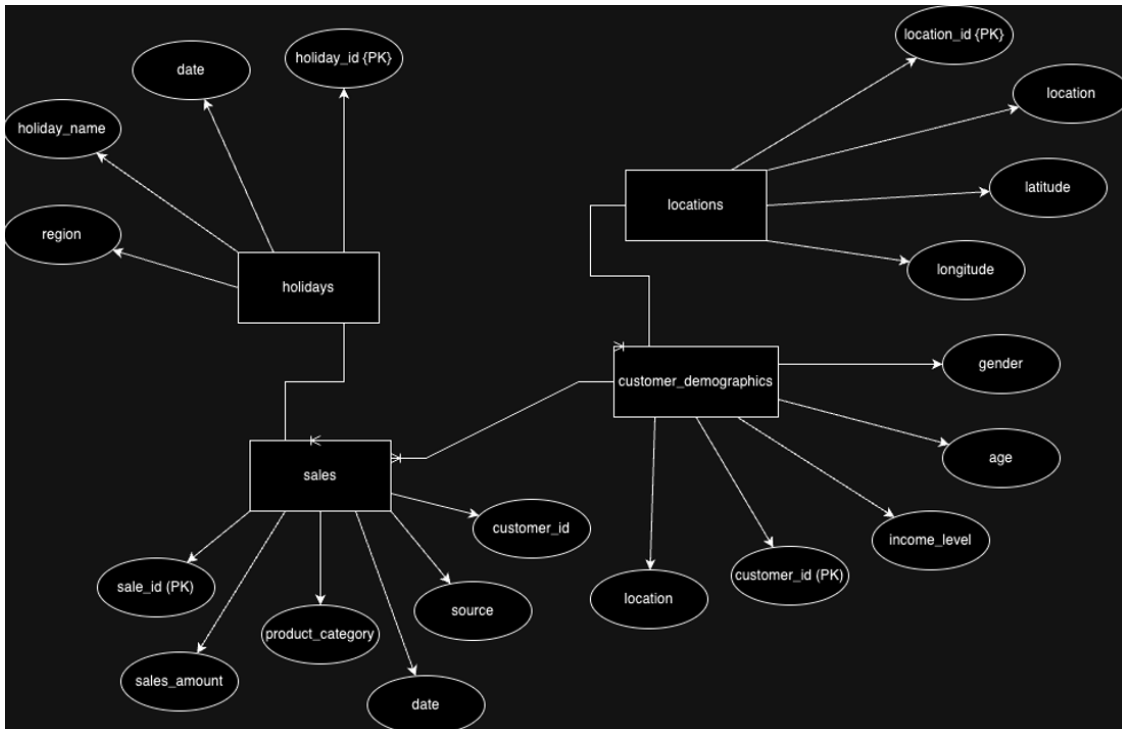

Figure 1: Entity-Relationship Diagram of the Data Schema

3.2 System Architecture

The system architecture, depicted in Figure 2, demonstrates the flow of data from raw inputs to the dashboard. This architecture ensures seamless integration, processing, and visualization of data.

- **Steps in the Pipeline:**

  1. **Raw Data Sources:** Data files such as CSVs containing sales and customer demographics are uploaded manually or through automated processes.
  2. **Airflow:** Orchestrates ETL (Extract, Transform, Load) tasks, ensuring scheduled data ingestion and preprocessing.
  3. **Kafka:** Facilitates real-time streaming of external data, including holiday and trend information.
  4. **Postgres Database:** Serves as the central repository for integrated datasets, optimized for querying and aggregation.

5. **Dashboard:** Fetches processed data and presents it in an interactive format for analysis.
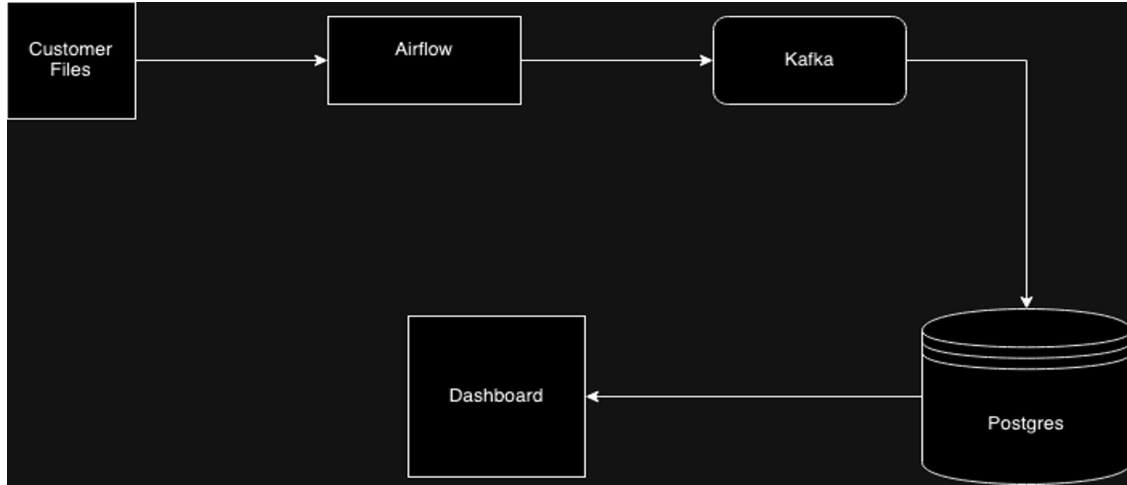

Figure 2: System Architecture of the Data Pipeline and Dashboard

4 Analysis and Results

4.1 Key Features

The dashboard incorporates several features to deliver actionable insights:

- **Key Metrics:** Displays high-level statistics, including total customers, total sales, and the most purchased product.

- **Sales Analysis:** Explores trends over time and by geographic location.

- **Heat Map of Sales:** Highlights sales density across regions using a geographic heatmap.

- **Holiday Sales Comparison:** Examines sales performance during holidays compared to daily averages.

4.2 Key Metrics

The dashboard highlights key metrics such as:

- **Total Customers:** 49,934 unique customers.

- **Total Sales:** $163,946,338.48.

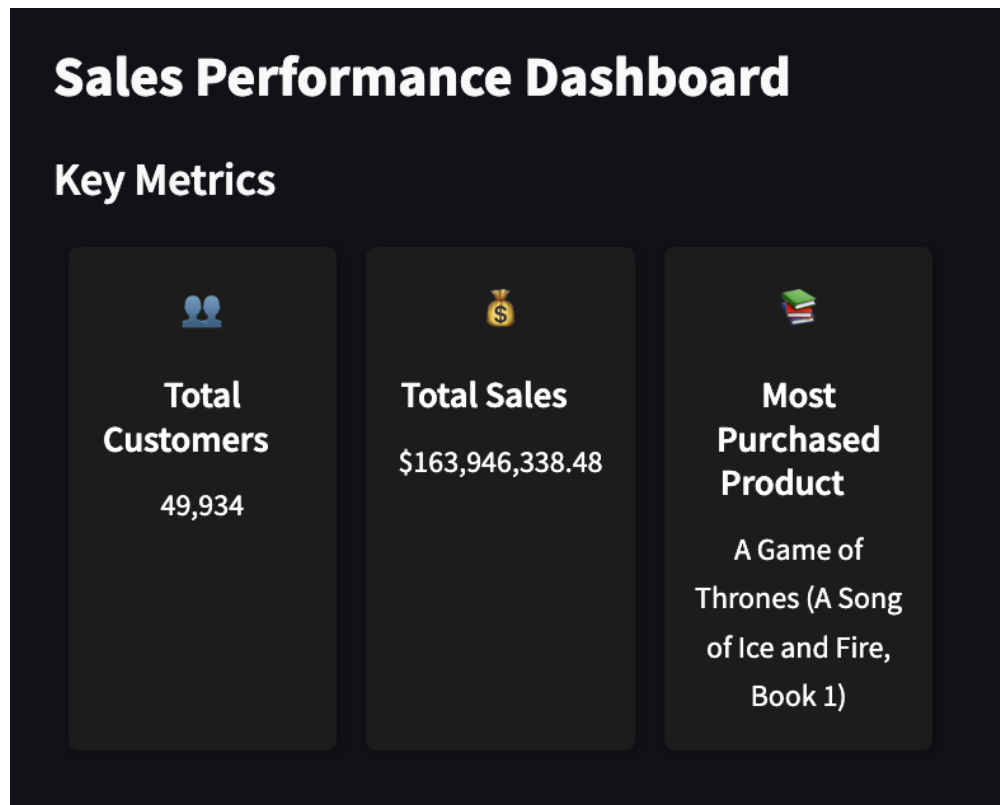- **Most Purchased Product:** "A Game of Thrones (A Song of Ice and Fire, Book 1)".

Figure 3: Key Metrics Overview on the Dashboard

4.3 Sales Analysis

The sales analysis section includes:

- **Sales Over Time:** A line chart visualizing seasonal trends and periodic spikes, particularly during holidays.
- **Sales by Location:** A bar chart identifying top-performing cities.
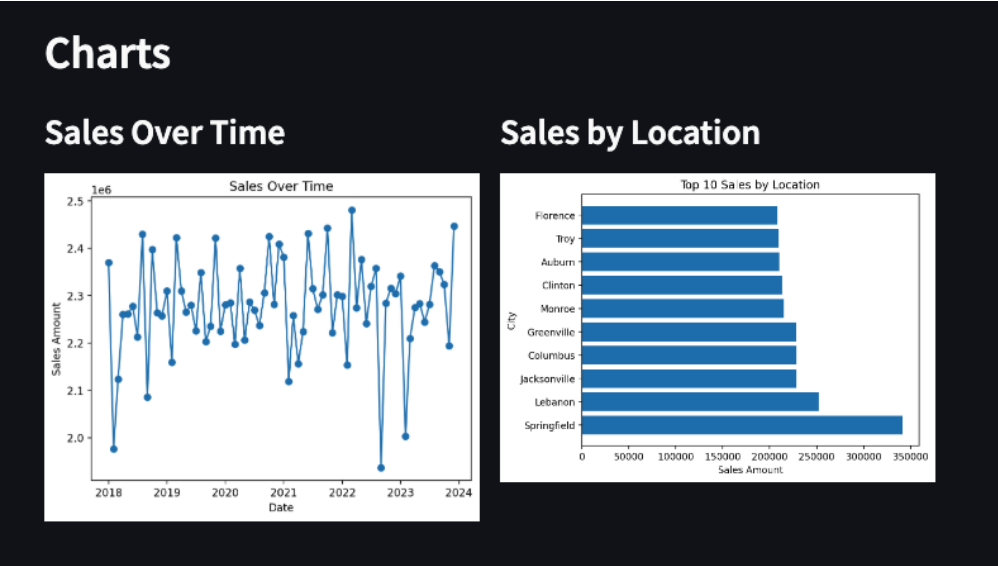
Figure 4: Sales Analysis Visualizations: Over Time and by Location

4.4 Heat Map of Sales by Location

The heatmap in Figure 5 visualizes sales density across the United States, revealing key trends:

- Urban areas exhibit higher sales volumes due to greater customer density.

- Differences in shopping behaviors across the country provide valuable insights for targeted marketing and resource allocation.
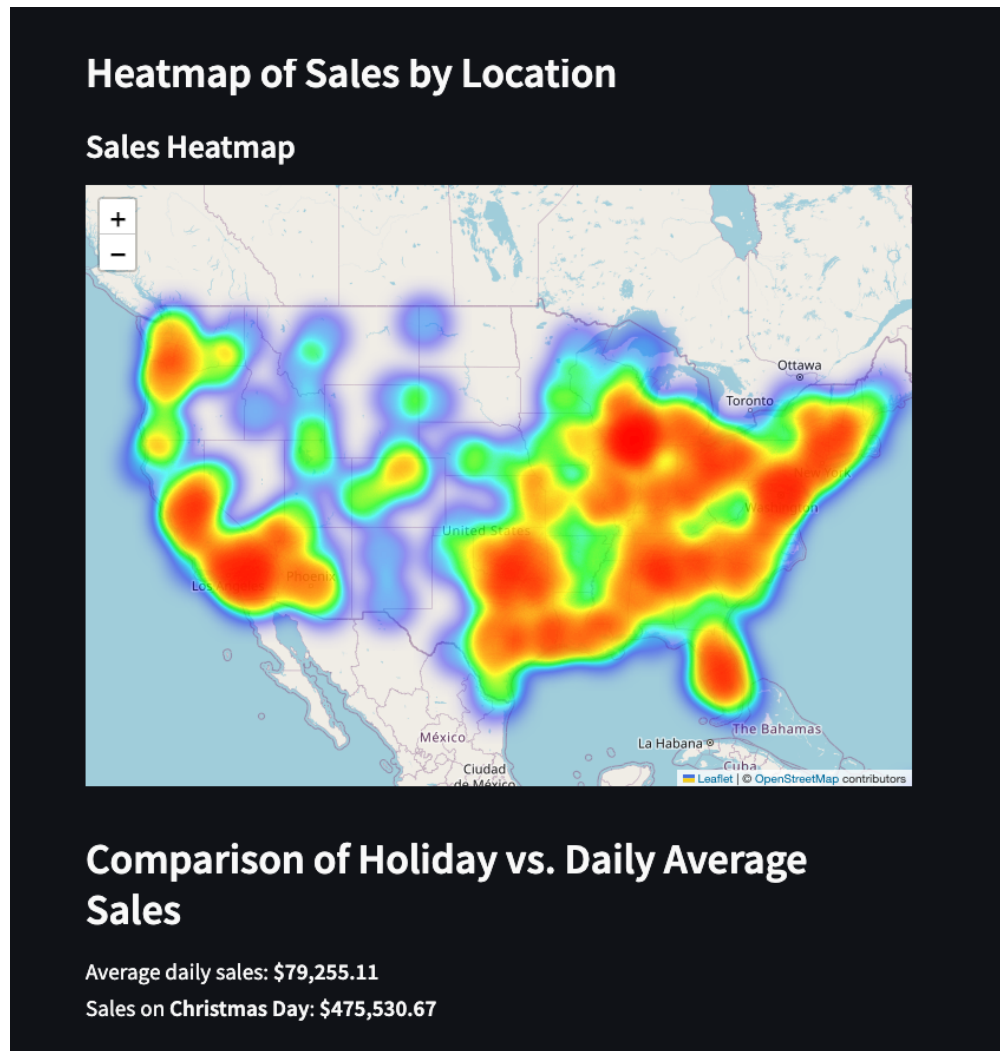
Figure 5: Heat Map of Sales by Location Across the United States

4.5 Holiday Sales Comparison

Holiday sales are compared to daily averages, highlighting the impact of holidays on consumer behavior:

- **Daily Average Sales:** $79,255.11.

- **Christmas Day Sales:** $475,530.67.

Listing 1: Code for Holiday Sales Comparison

```
holiday_sales = filtered_data[filtered_data["holiday_name"] == selected_holiday]["sales_amount"].sum()
average_daily_sales = filtered_data["sales_amount"].mean()
```

5   Future Improvements

- **Enhanced Predictive Analytics:** Integrate machine learning models to forecast sales based on historical and external factors.

- **Improved Geographic Analysis:** Incorporate customer density maps for regional marketing strategies.

- **Additional Data Sources:** Include social media sentiment analysis to understand customer opinions on products or services.

- **Real-Time Updates:** Implement faster real-time data pipelines using tools like Spark Streaming.

6 References

- Streamlit Documentation: `https://docs.streamlit.io/`

- PostgreSQL Documentation: `https://www.postgresql.org/docs/`

- Airflow Documentation: `https://airflow.apache.org/`