Information
Technology
Institute

Walmart Data Analytics Project: Final Documentation

Data Management: Team 5

Mohamed Hazem Mohamed Sakr
Rodina Ashraf Ahmed
Mahmoud Osama AbdelFatah
Yomna Safwat Afifi

1. Executive Summary

The **Walmart Data Analytics Project** was designed to replicate real-world decision-making scenarios in one of the largest retail chains globally. This project focuses on analyzing customer purchase behavior, optimizing store layouts, and enhancing inventory and promotion strategies to drive business growth. By creating a robust data warehouse, leveraging advanced SQL queries, and presenting actionable insights, this project highlights how data analytics can address business challenges in a highly competitive retail landscape.

The analysis uncovered key customer behavior patterns, identified frequently purchased product combinations, and provided strategic recommendations to improve cross-selling, promotional impact, and inventory management. Implementing these insights could significantly enhance Walmart's operational efficiency and profitability.

2. Data Warehouse Design

2.1 Business Problem

Walmart processes millions of daily transactions across various locations, making it challenging to extract meaningful insights from the raw data. A structured data warehouse is essential for storing and analyzing this data effectively.

Here is a **restructured and polished explanation** of how DBT (Data Build Tool) was used in your project, formatted for clarity and professionalism:

2.2 DBT Implementation in Our Data Warehouse Project

2.2.1 Overview

In this project, we utilized **DBT Cloud** and **BigQuery on GCP** to design, build, and model our data warehouse for analytics purposes. DBT allowed us to transform raw data into a structured, analytics-ready format through modular SQL and version-controlled workflows. The following steps detail the methodology used:

2.2.2 Data Connection and Schema Definition

- **Connecting to Datasets**:
    - We connected DBT Cloud to our **BigQuery datasets** to access and process the source data stored on GCP.
    - Configuration files and credentials ensured secure and seamless integration between DBT and BigQuery.
- **Defining the Schema**:
    - A schema.yml file was created to define the structure of our tables and schemas. This allowed DBT to understand:
        - The location of tables in the data source.
        - The metadata and dependencies for tables and models.

2.2.3. Data Validation and Testing

To ensure data quality and reliability, we implemented several **DBT tests and macros**:

- **Primary Key Validation**:
    - Ensured that all primary keys across the tables were **unique** and **not null** to maintain referential integrity.
- **String Validation**:
    - Verified that critical string fields (e.g., product_name, category_name) were **not null** to avoid incomplete records.
- **Relationship Testing**:
    - Tested and validated the relationships between **dimension tables** and the **fact table** to ensure that data integrity was preserved during joins.
- **Gender Validation**:
    - Ensured that the gender field contained only the allowed values: [Male, Female], using macros for strict formatting enforcement.

2.2.4. Staging Data

- **Source Selection**:
    - The raw data was selected from the source tables specified in schema.yml under the models/staging folder.
    - This staging layer acted as an intermediate step, where data was lightly cleaned and standardized for further processing.

2.2.5 Data Modeling

- **Core Data Models**:
    - Using the staged data, we built the core models for the data warehouse:
        - **Dimension Tables**:
            - Enriched and cleaned dimensions such as product_dim, customer_dim, and category_dim were created to provide context to transactional data.
        - **Fact Table**:
            - The fact_transactions table was modeled by aggregating and joining the staged data.
    - These models were defined in the models/core folder, demonstrating how the data was **cleaned**, **transformed**, and **modeled** to create analytics-ready tables.

2.2.6 Final Deliverables

- The DBT folder structure, including all models, configurations, and tests, is attached with the project to provide:
    - **Reproducibility**: All transformations and logic are stored in DBT files, enabling others to rebuild the data warehouse.
    - **Transparency**: Schema definitions, tests, and queries are documented within the DBT project.

Key Benefits of Using DBT

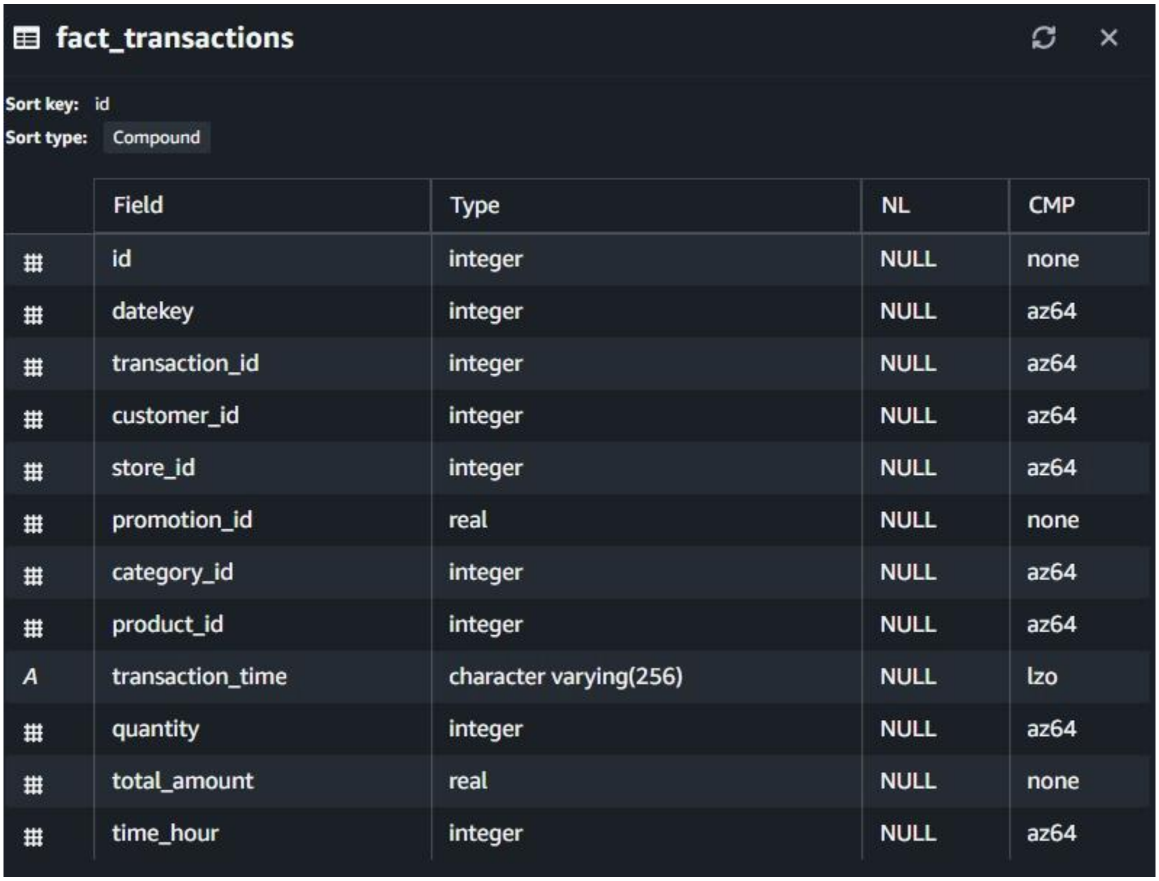
1. **Modular Data Transformation**:
    - DBT's model structure allowed us to break down transformations into easily manageable and reusable steps.
2. **Data Quality Assurance**:
    - Built-in tests and macros ensured high-quality, validated data throughout the pipeline.
3. **Cloud Scalability**:
    - Leveraging **BigQuery on GCP** ensured scalable storage and processing power for our data warehouse.

2.3 DataWarehouse Schema

To address the business problem, the data warehouse was designed using a **Star Schema**. The schema includes the following:

Fact Table: Transactions Fact Table

- **Columns**: ID, Date Key, Transaction ID, Customer ID, Store ID, Promotion ID, Category ID, Product ID, Transaction Time, Current Time, Quantity, Total Amount, Time Hour
- **Purpose**: To store transactional data capturing all purchases made by customers.
- **Granularity:** one row per product purchased in a transaction.



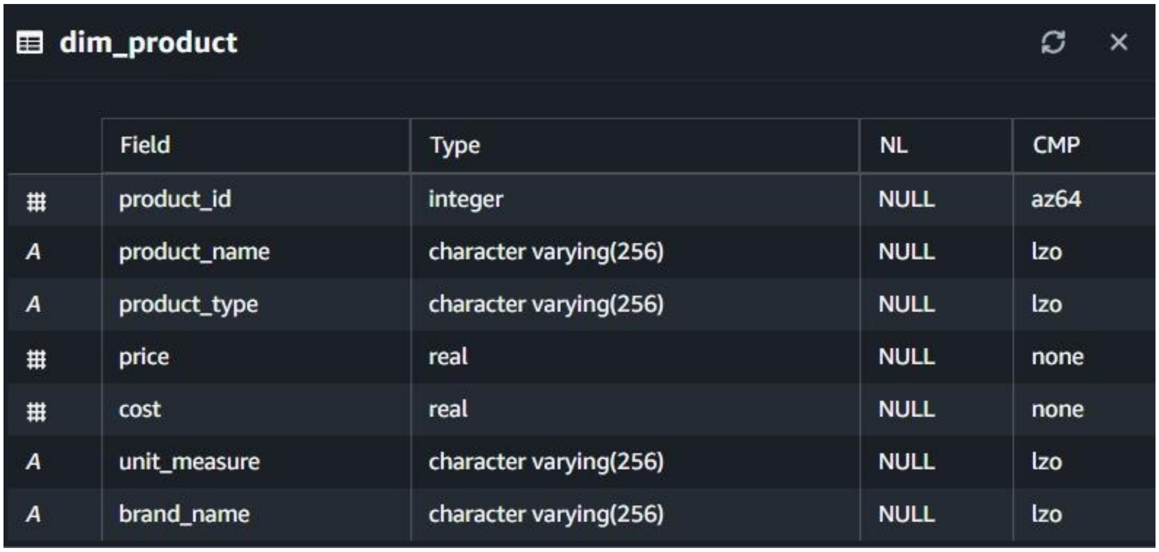
fact_transactions

Sort key: id
Sort type: Compound

	Field	Type	NL	CMP
#	id	integer	NULL	none
#	datekey	integer	NULL	az64
#	transaction_id	integer	NULL	az64
#	customer_id	integer	NULL	az64
#	store_id	integer	NULL	az64
#	promotion_id	real	NULL	none
#	category_id	integer	NULL	az64
#	product_id	integer	NULL	az64
A	transaction_time	character varying(256)	NULL	lzo
#	quantity	integer	NULL	az64
#	total_amount	real	NULL	none
#	time_hour	integer	NULL	az64

Dimension Tables:

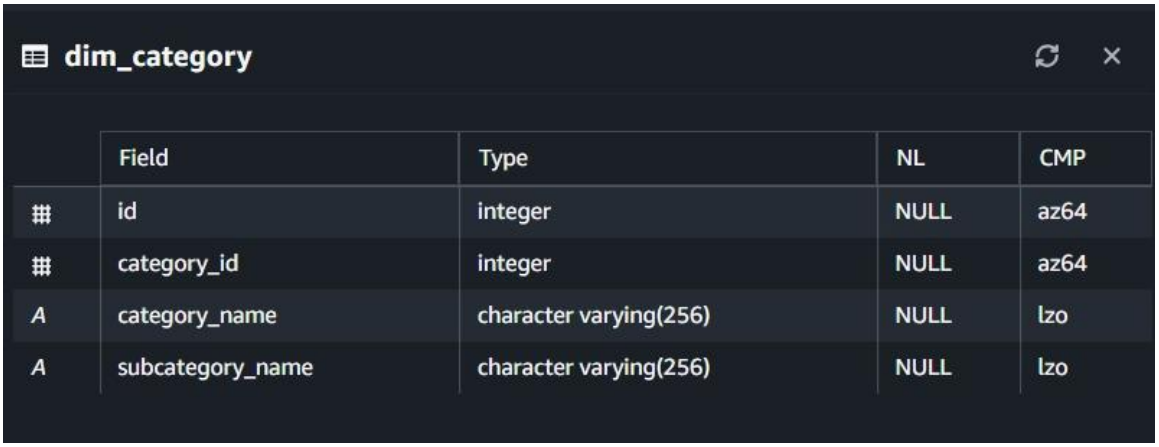
1. **Products Dimension**:
   - Attributes: Product ID, Product Name, Product Type, Price, Cost, Unit Measure, Brand Name.
   - Purpose: To provide detailed information about products.



dim_product

	Field	Type	NL	CMP
#	product_id	integer	NULL	az64
A	product_name	character varying(256)	NULL	lzo
A	product_type	character varying(256)	NULL	lzo
#	price	real	NULL	none
#	cost	real	NULL	none
A	unit_measure	character varying(256)	NULL	lzo
A	brand_name	character varying(256)	NULL	lzo

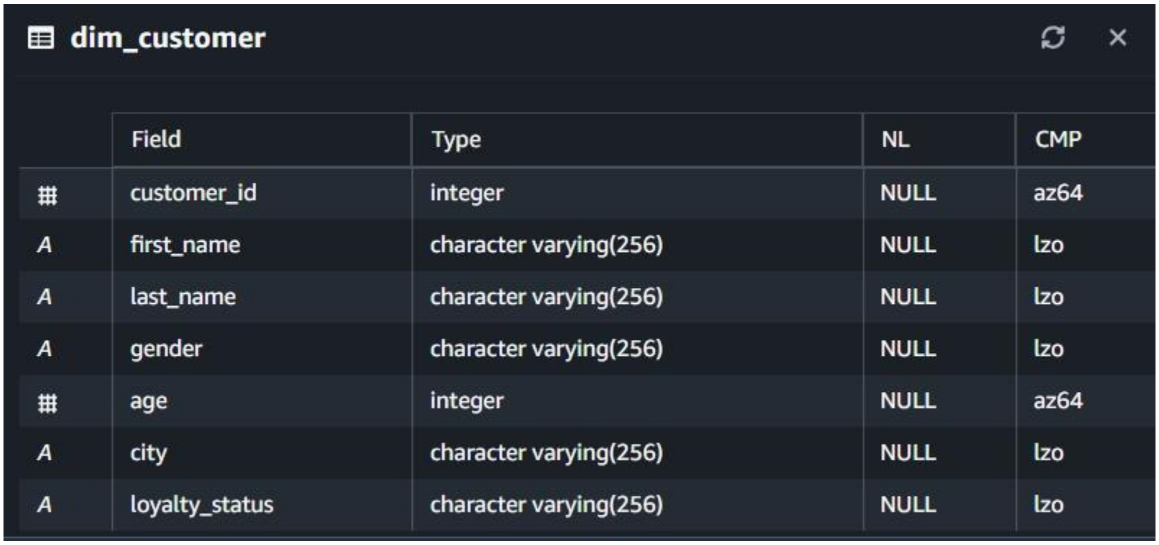
2. **Category Dimension**:
   - Attributes: Customer ID, Category Name, Sub Category Name, Gender.
   - Purpose: To group products into categories for aggregated analysis and to identify category-level trends and performance.



dim_category

	Field	Type	NL	CMP
#	id	integer	NULL	az64
#	category_id	integer	NULL	az64
A	category_name	character varying(256)	NULL	lzo
A	subcategory_name	character varying(256)	NULL	lzo

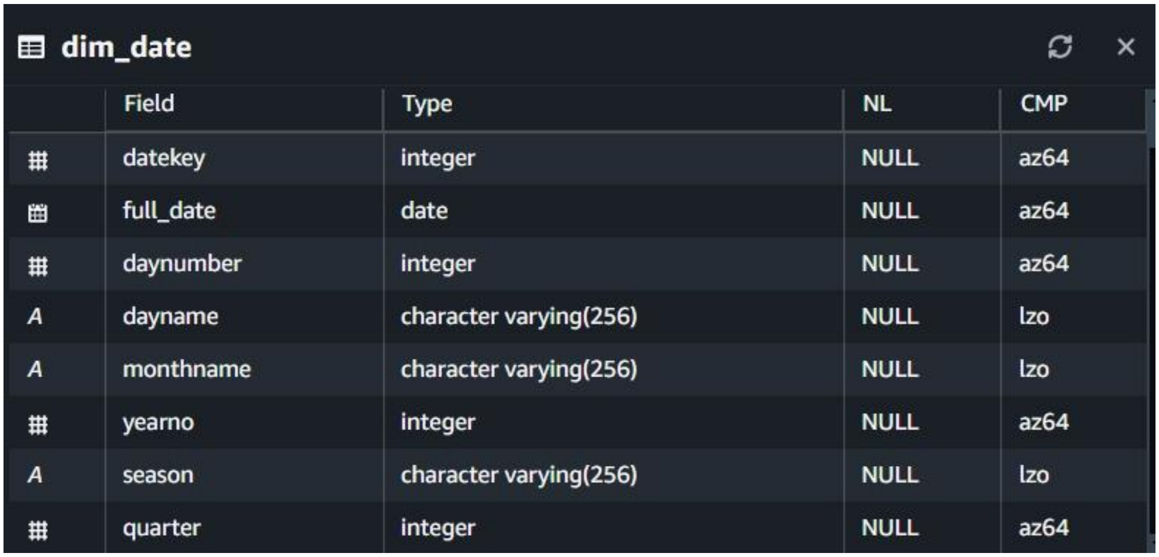
3. **Customers Dimension**:
   - Attributes: Customer ID, First Name, Last Name, Gender, Age, City, Loyalty Status.
   - Purpose: To segment customers based on demographic and behavioral data.



dim_customer

	Field	Type	NL	CMP
#	customer_id	integer	NULL	az64
A	first_name	character varying(256)	NULL	lzo
A	last_name	character varying(256)	NULL	lzo
A	gender	character varying(256)	NULL	lzo
#	age	integer	NULL	az64
A	city	character varying(256)	NULL	lzo
A	loyalty_status	character varying(256)	NULL	lzo

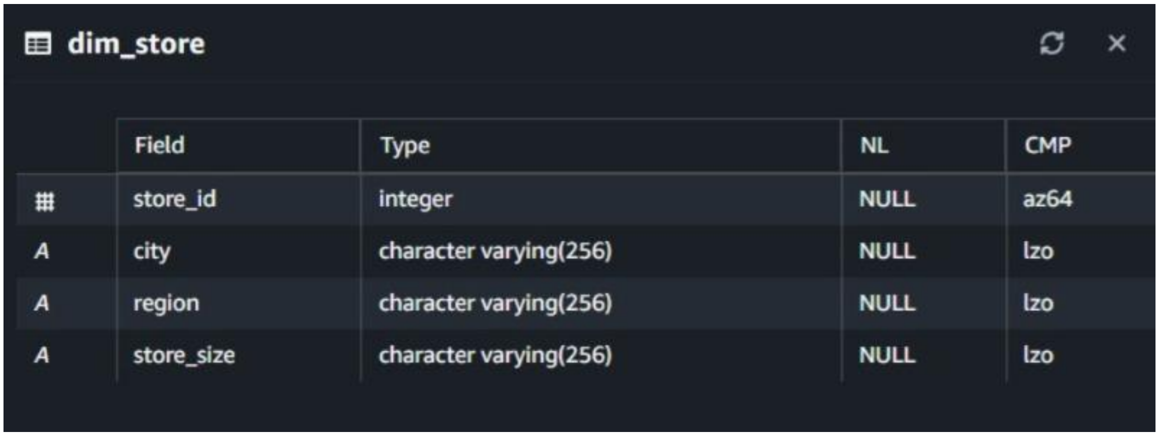
4. **Date Dimension**:
   - Attributes: DateKey, Full Date, Day Number, Day Name, Month Name, Year No, Season Quarter.
   - Purpose: To analyze temporal trends in sales.



dim_date

	Field	Type	NL	CMP
#	datekey	integer	NULL	az64
📅	full_date	date	NULL	az64
#	daynumber	integer	NULL	az64
A	dayname	character varying(256)	NULL	lzo
A	monthname	character varying(256)	NULL	lzo
#	yearno	integer	NULL	az64
A	season	character varying(256)	NULL	lzo
#	quarter	integer	NULL	az64

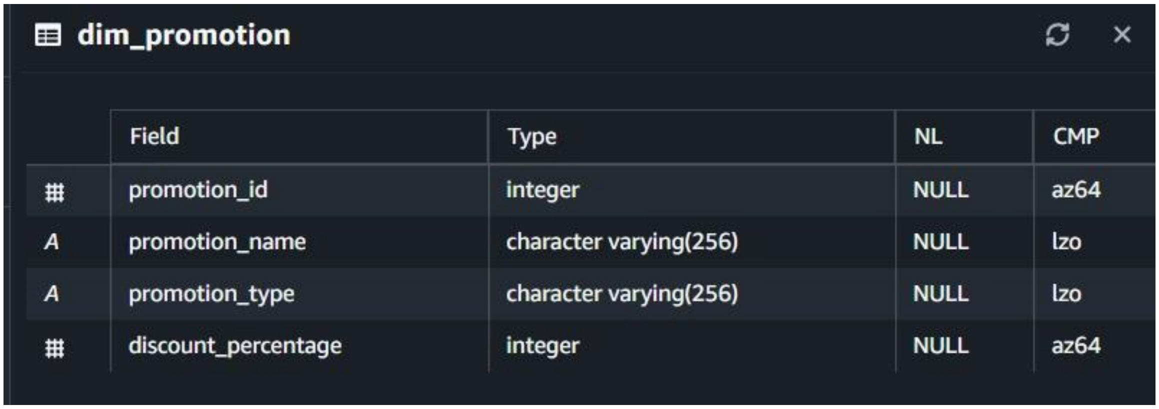
5. **Store Dimension**:
   - Attributes: Store ID, City, Region, Store Size.
   - Purpose: To analyze sales across different store locations.



dim_store

	Field	Type	NL	CMP
#	store_id	integer	NULL	az64
A	city	character varying(256)	NULL	lzo
A	region	character varying(256)	NULL	lzo
A	store_size	character varying(256)	NULL	lzo

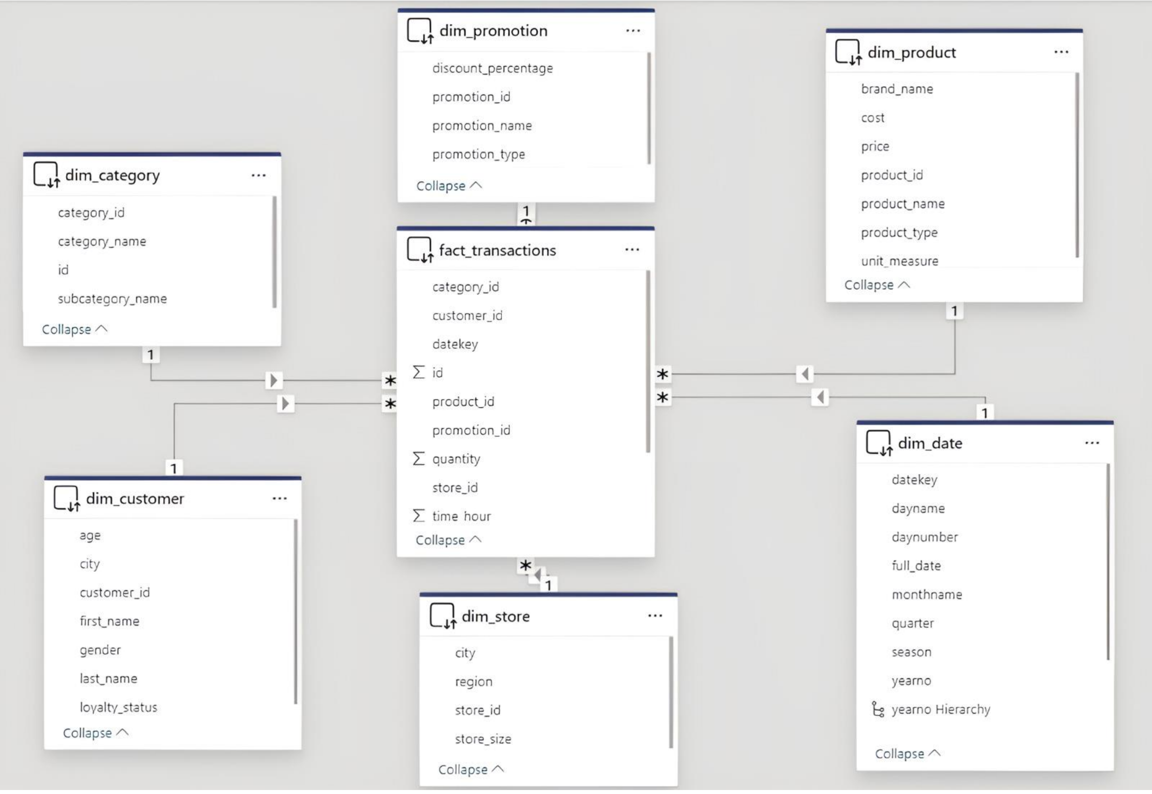
6. **Promotions Dimension**:
   - Attributes: Promotion ID, Promotion Name, Promotion Type, Discount Percentage.
   - Purpose: To measure the effectiveness of promotional campaigns.



dim_promotion

	Field	Type	NL	CMP
#	promotion_id	integer	NULL	az64
A	promotion_name	character varying(256)	NULL	lzo
A	promotion_type	character varying(256)	NULL	lzo
#	discount_percentage	integer	NULL	az64

Schema Diagram:



Design Justification

- The **star schema** was chosen for its simplicity and performance in analytical queries.
- Dimension tables provide context to transactional data, enabling granular analysis across multiple perspectives.
- The design supports scalability for future data growth and integration with business intelligence tools.

3. Business Insights

1. **Top-Selling Products:**
   - Superhero Action Figures (451 units) and Lego Sets (452 units) dominate the Toys category, reflecting strong demand for themed or gift-oriented products.
   - Tents (349 units) and Dumbbells (340 units) are leaders in the Sports category, driven by seasonal trends and fitness demand.
   - Toasters (255 units) in the Home Essentials category maintain steady year-round demand, suggesting these products are essential purchases rather than seasonal.
   - MacBook Bro: 425 K of our Sales.
2. **Customer Demographics:**
   - Age-Based Patterns:
     - The 30–40 age group accounts for most purchases in Toys and Home Essentials, likely representing families and homeowners.
     - The 20–30 age group drives fitness-related purchases (e.g., Dumbbells) and outdoor equipment (e.g., Tents), reflecting lifestyle preferences.
   - Loyalty Status:
     - Gold-tier customers generate higher revenue per transaction. These customers are likely long-term, high-value shoppers who should be prioritized for retention efforts.
3. **Seasonal Trends:**
   - Camping Gear (Tents, Sleeping Bags) peaks in summer and during holiday seasons, highlighting the importance of seasonal inventory planning.
   - Breakfast Staples (Milk, Cornflakes) exhibit consistent demand, making them reliable revenue drivers.

Key Insights:

1. **Top Product Pairs:**
   - Lego Sets and Superhero Action Figures (375 times): Highlight the popularity of related toys. This pairing suggests parents and gift buyers prefer themed items.
   - Milk and Cornflakes (1510 times): Reflect a strong breakfast staple pairing. This cross-category pair drives sales across Dairy and Grocery.
   - Tents and Sleeping Bags (278 times): Suggest a seasonal surge in camping gear, particularly during summer and holiday seasons.
2. **Anchor Products:**
   - Products like Milk and Cornflakes are “anchor items,” frequently driving the purchase of other products.
   - Tents act as a gateway item, encouraging purchases of additional camping gear like sleeping bags and lanterns.

3. **Cross-Category Patterns:**
    - Items like Bread and Butter show clear interdependencies, encouraging cross-aisle placements.
    - High-frequency product pairs span categories like Home Essentials, Grocery, and Toys, highlighting cross-category purchase behaviors.

Strategic Recommendations:

- **Store Layout Optimization:**
    - Co-locate frequently bought-together items to improve shopping convenience (e.g., place Cornflakes near the Dairy aisle).
    - Use endcap displays to feature pairs like Tents and Sleeping Bags during peak seasons.
- **Bundling Promotions:**
    - Offer value packs (e.g., "Camping Starter Kit" including a tent, sleeping bag, and flashlight) to encourage higher basket sizes.
    - Create meal bundles, such as "Breakfast Combos" (Milk + Cornflakes + Bread), with slight discounts to drive sales.

- **Personalized Marketing:**
    - Leverage purchase history to design personalized email or in-app promotions. For example:
        - Recommend Milk to customers frequently purchasing Cornflakes.
        - Offer discounts on Action Figures to customers buying Lego Sets.

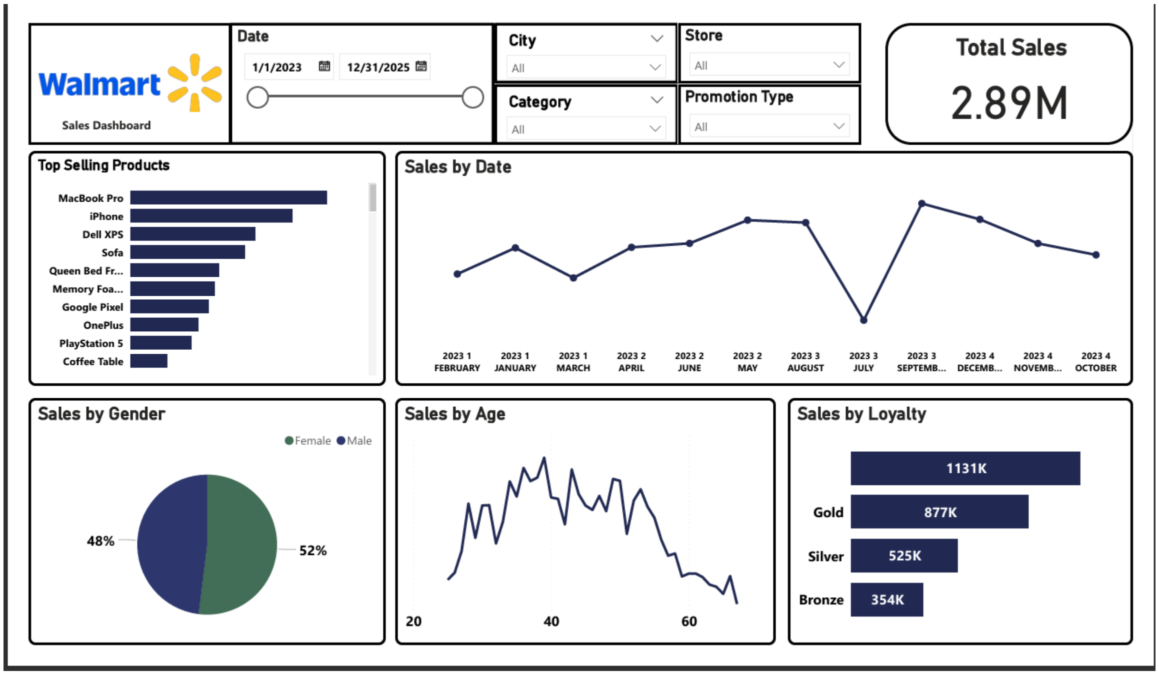
4. **Revenue Drivers:**
    - High-ticket items like Tents ($85,000 total revenue) contribute significantly to overall sales in the Sports category.
    - Bulk sales (e.g., Bread + Milk) drive Grocery revenue, indicating the importance of high-turnover staples.

Strategic Recommendations:

- **Inventory Management:**
    - Increase inventory for high-demand seasonal items (e.g., Tents) before summer, and ensure year-round availability of staples like Milk.
    - Use predictive analytics to forecast inventory needs based on historical trends.

- **Targeted Promotions:**
  - Create family-focused promotions for Home Essentials and Toys to align with the spending patterns of the 30–40 age group.
  - Offer fitness-focused promotions for Sports items like Dumbbells to cater to younger, active shoppers.
- **Personalized Loyalty Campaigns:**
  - Send personalized offers to Gold-tier customers, focusing on their preferred categories. For example, offer exclusive discounts on high-margin items like Lego Sets.
- **Store Layout Enhancements:**
  - Dedicate high-visibility sections for frequently purchased pairs, such as an aisle for "Quick Breakfast Items" (Milk, Cornflakes, Bread).
  - Enhance in-store signage to direct shoppers to bundled promotions like “Camping Essentials” or “Toy Combos.”

Dashboard Reflecting Insights:



Walmart Sales Dashboard

Date: 1/1/2023 – 12/31/2025

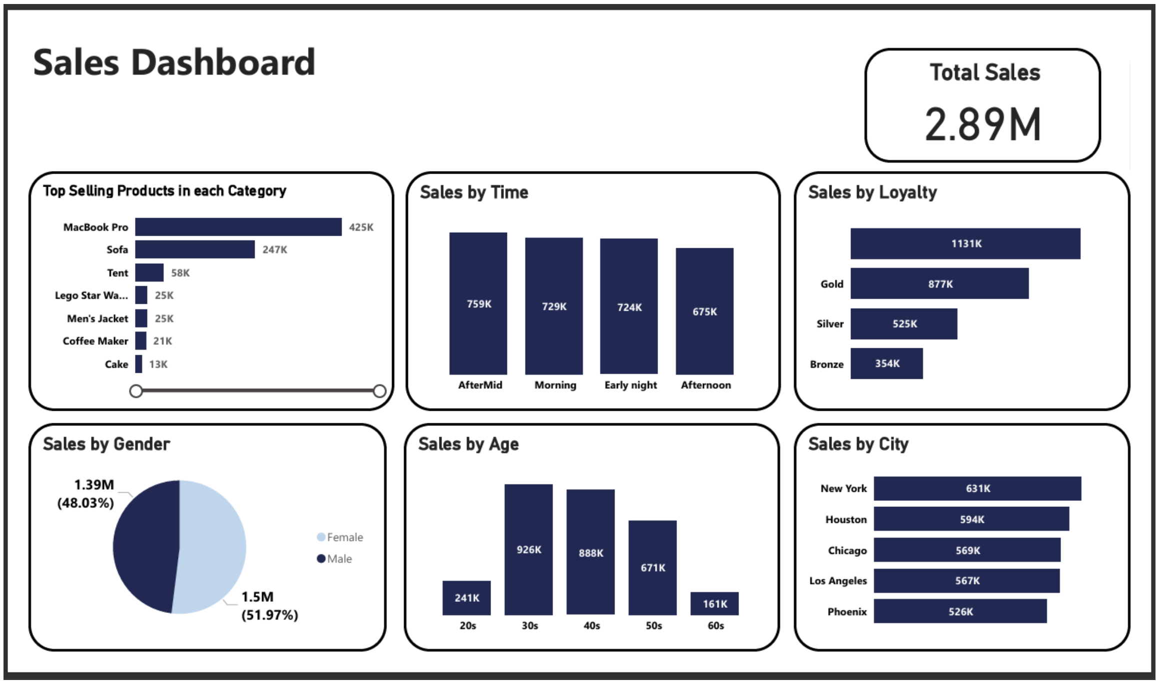
City: All

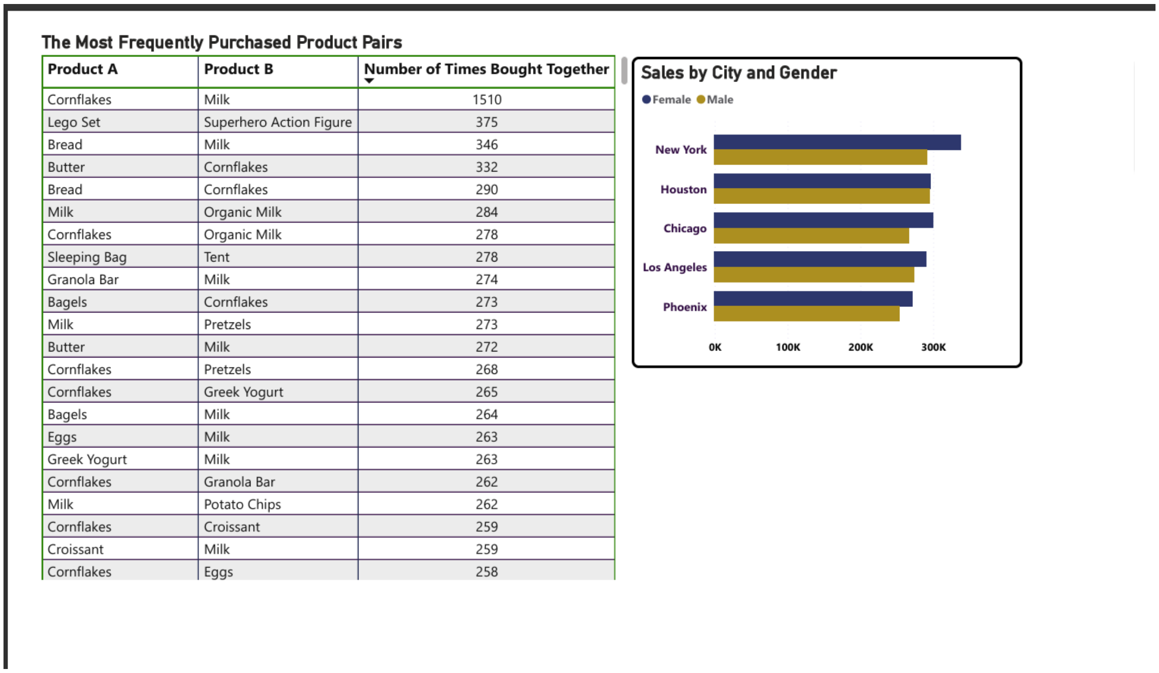
Store: All

Category: All

Promotion Type: All

Total Sales: 2.89M





The Most Frequently Purchased Product Pairs

Product A	Product B	Number of Times Bought Together
Cornflakes	Milk	1510
Lego Set	Superhero Action Figure	375
Bread	Milk	346
Butter	Cornflakes	332
Bread	Cornflakes	290
Milk	Organic Milk	284
Cornflakes	Organic Milk	278
Sleeping Bag	Tent	278
Granola Bar	Milk	274
Bagels	Cornflakes	273
Milk	Pretzels	273
Butter	Milk	272
Cornflakes	Pretzels	268
Cornflakes	Greek Yogurt	265
Bagels	Milk	264
Eggs	Milk	263
Greek Yogurt	Milk	263
Cornflakes	Granola Bar	262
Milk	Potato Chips	262
Cornflakes	Croissant	259
Croissant	Milk	259
Cornflakes	Eggs	258

4. Recommendations

4.1 Store Layout Optimization

- Place frequently bought-together items in proximity to encourage cross-selling.
    - Example: Position Milk and Cornflakes in the same aisle to reduce customer effort and increase basket size.

4.2 Bundle Promotions

- Introduce discounts on frequently purchased product pairs to boost sales.
    - Example: Offer a 10% discount for purchasing Diapers and Baby Wipes together.

4.3 Inventory Adjustments

- Increase stock levels for high-demand products during peak seasons.
    - Example: Increase stock of cold beverages during summer months.
- Use demographic insights to stock products favored by specific customer segments in relevant regions.

4.4 Promotional Strategies

- Tailor promotions to customer demographics and preferences.
    - Example: Target young adults with discounts on electronics during holiday seasons.

---

5. Tools and Technologies

5.1 Data Management

- **Database System**: ORACLE DBMS.
- **Amazon S3:** Storing dimensional data before loading it to Amazon Redshift.
- **Amazon RedShift**: data warehousing and query execution.
- **ETL Tools**: dbt cloud for data extraction, transformation, and loading.

5.2 Data Analysis

- **Query Language**: SQL for data analysis and identifying patterns.
- **Visualization Tools**: Power BI for presenting insights.

6. Conclusion

This project provided a detailed analysis of Walmart's transactional data, uncovering patterns in customer behavior and identifying opportunities for growth. The proposed recommendations—when implemented—are expected to enhance operational efficiency, increase customer satisfaction, and drive revenue growth.

---

7. Appendices

- **Appendix A**: SQL Queries
  - Example queries for finding frequently bought-together items and analyzing demographic trends.
- **Appendix B**: Visualizations
  - dashboards and charts created in Power BI.
- **Appendix C**: DBT Files