Walmart Sales Data Processing & Analysis

Author: ISTIAK ALAM

Affiliation: DataCamp Real World Project

Executive Summary

This report documents a data engineering and analytics project focused on Walmart's grocery sales data, complemented by external factors such as holiday indicators, temperature, fuel prices, CPI, unemployment rates, promotions, and store characteristics. The primary task was to integrate and clean these two disparate datasets — one from a PostgreSQL database (`grocery_sales`) and another from a Parquet file (`extra_data.parquet`) — into a single `clean_data` dataset suitable for analysis.

Key objectives included:

- Merging and harmonizing data sources.
- Cleaning and transforming data to retain essential fields.
- Aggregating sales information to analyze monthly trends.
- Generating the `agg_data` dataset for managerial insights.

The final outputs include two CSV files: `clean_data.csv` and `agg_data.csv`, which are ready for further business intelligence or forecasting tasks.

Table of Contents

Introduction

Background

Walmart, America’s largest retail corporation, generated over $80 billion in e-commerce sales by 2022, accounting for 13% of total revenue. Seasonal and public holidays like the Super Bowl, Labour Day, Thanksgiving, and Christmas significantly influence shopping behaviors. Understanding these temporal effects on sales is crucial for aligning supply chain and promotional strategies.

Objectives

- Merge transactional sales data with complementary contextual data.
- Transform weekly sales into monthly aggregates for high-level trend analysis.
- Identify the sales patterns during holiday and non-holiday periods.
- Store cleaned datasets for use in forecasting, inventory, and marketing analysis.

Data Sources

1. **PostgreSQL Database: grocery_sales Table**
    - `Store_ID` – Identifier of the store
    - `Date` – Week of sales
    - `Dept` – Department number
    - `Weekly_Sales` – Weekly sales for the store-department combination
2. **Parquet File: extra_data.parquet**
    - `IsHoliday` – Boolean indicator for holiday weeks
    - `Temperature` – Average temperature (°F)
    - `Fuel_Price` – Average fuel price ($)
    - `CPI` – Consumer Price Index

    - `Unemployment` – Unemployment rate (%)
    - `MarkDown1` to `MarkDown4` – Promotional markdowns
    - `Size` – Store size (sq.ft)
    - `Type` – Store type (A, B, C)

Methodology

Step 1: Data Extraction

- Queried the PostgreSQL database to retrieve the `grocery_sales` table using SQL.
- Loaded the `extra_data.parquet` file into a pandas DataFrame.

Step 2: Data Merging

- Merged datasets on common keys (`Store_ID`, `Date`, `Dept`) ensuring alignment of sales with corresponding contextual variables.

Step 3: Data Transformation

- Extracted **month** from the `Date` column and created a new `Month` feature.
- Filtered and kept only key columns:
  `Store_ID, Month, Dept, IsHoliday, Weekly_Sales, CPI, Unemployment`
- Ensured correct data types for numerical and date columns.

Step 4: Aggregation

- Grouped by `Month` to create an aggregated dataset (`agg_data`) showing average `Weekly_Sales` per month.

Step 5: Export

- Saved the transformed `clean_data` and aggregated `agg_data` as CSV files for further analysis.

Results

Sample of Cleaned Data

Store_ID	Month	Dept	IsHoliday	Weekly_Sales	CPI	Unemployment
1	2	1	0	24924.50	211.096	8.106
1	2	1	1	46039.49	211.242	8.106

Aggregated Monthly Sales

Month	Weekly_Sales
1	33174.18
2	34333.33
3	...

These results allow quick identification of seasonal peaks and troughs.

Discussion

Key findings:

- **Holiday Periods** show noticeable spikes in sales, confirming the business hypothesis about seasonal demand surges.
- **Aggregated Monthly Sales** facilitate high-level strategic planning and inventory allocation.
- Data merging across files ensures a holistic view combining both transactional and contextual factors.

Limitations:

- Sales patterns by department were not deeply analyzed in this phase.
- Inflation-adjusted sales and store-type segmentation could yield more actionable insights.

Conclusion & Recommendations

- The data pipeline successfully integrates sales and contextual datasets into analysis-ready formats.

- Monthly average sales trends can support inventory optimization, workforce scheduling, and promotional planning.
- Future enhancements:
    - Integrate machine learning models for sales forecasting.
    - Incorporate promotional markdown effectiveness analysis.
    - Segment analysis by store type and size for operational decision-making.

Appendix

Project Deliverables:

- `clean_data.csv` – Merged, cleaned dataset with key metrics.
- `agg_data.csv` – Monthly aggregated sales summary.

Visualizations (Optional for PDF Report):

- Monthly sales trend line chart.
- Holiday vs Non-Holiday weekly sales boxplot.
- Correlation heatmap for CPI, unemployment, and sales.