Walmart Sales Data Analysis and Forecasting

DA 460: Data Visualization

Dr. Zeyad Hailat

Team Members:

Mohammad Harb

Mohammad Abdullah Bani Younis

Mohammad Rami Bani Younis

Hashem Bani Younis

Mahmoud Aeideh

2. Introduction

This project aims to build an interactive visual data analysis solution using Microsoft Power BI to analyze historical sales data from Walmart. In the highly competitive retail sector, understanding sales trends and the impact of external factors—such as holidays, fuel prices, and economic indicators—is crucial for making informed business decisions. By combining exploratory analysis with explanatory storytelling, this project provides a comprehensive tool for both data analysts and stakeholders to gain deep insights into store performance and seasonal patterns.

To ensure a structured approach to the problem definition, we applied the "Five Ws" framework as follows:

- **What (The Data):** The analysis is based on a historical dataset covering weekly sales from 45 Walmart stores over a period of nearly three years (2010–2012). The data includes quantitative measures such as weekly sales figures and environmental factors (temperature, fuel price, CPI, unemployment), as well as categorical dimensions like store ID and holiday flags.
- **Why (The Goal):** The primary objective is to discover sales patterns, specifically the impact of major holiday events (e.g., Super Bowl, Christmas) on revenue. The solution aims to help management optimize inventory planning and understand how external economic factors correlate with consumer purchasing behavior.
- **Who (The Users):** The solution targets two distinct user groups:
    1. **Data Analysts:** Who require an exploratory dashboard to filter data, drill through details, and identify outliers.
    2. **Executive Stakeholders:** Who need a high-level explanatory dashboard (storytelling) to view Key Performance Indicators (KPIs) and actionable insights without getting lost in raw data.
- **Where (Context):** This dashboard is designed to be used in corporate headquarters during strategic planning sessions and quarterly performance reviews.

- **When (Timing):** The insights derived from this tool are most critical during pre-holid planning phases to ensure stores are prepared for demand surges, as well as for post-season analysis to evaluate performance.

3. Dataset Description

Source of the Dataset: The dataset selected for this project is the "Walmart Store Sales Forecasting" dataset. It was obtained from a reputable public repository (Kaggle) and originally sourced from Walmart's recruiting competition. The data covers a time span of approximately 33 months, from February 2010 to October 2012, satisfying the project requirement of at least 24 months of time-based data.

- **Dataset Link:** https://www.kaggle.com/datasets/asahu40/walmart-data-analysis-and-forcasting

Data Abstraction: To understand the underlying structure of the data and prepare it for analysis, we performed the following abstraction steps:

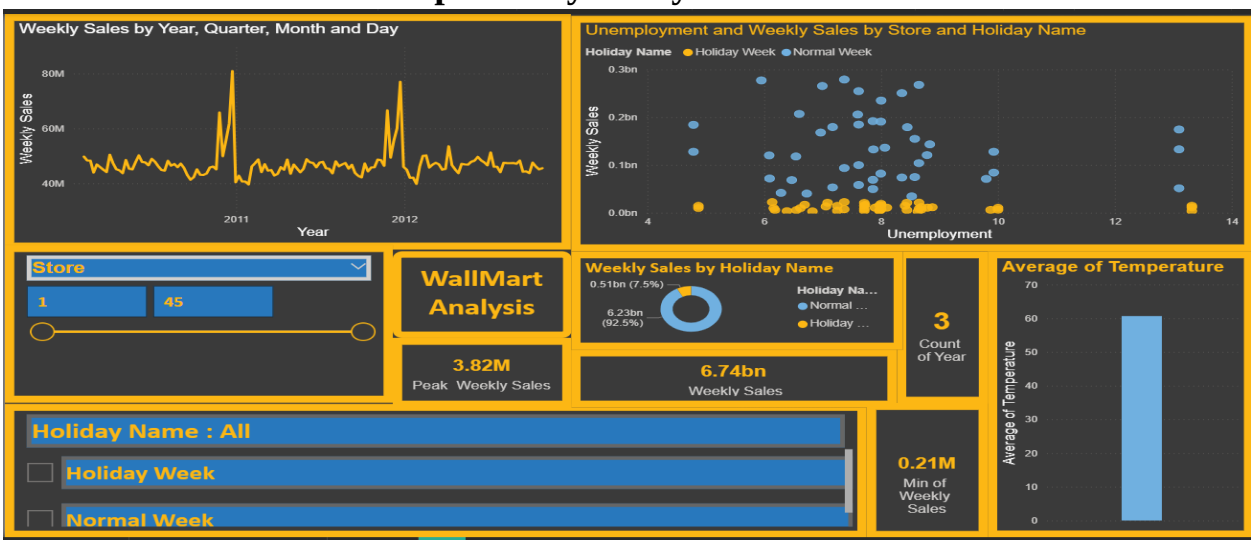
- **Dataset Type:** The original data is structured as a flat table (CSV format). In Power BI, we transformed this into a Relational Model (Star Schema) by separating fact tables (Sales) from dimension tables (Stores, Features) to ensure efficient filtering and accurate calculations.
- **Attribute Types:**
  - **Quantitative Attributes:** Weekly_Sales (the primary measure), Temperature, Fuel_Price, CPI, and Unemployment.
  - **Categorical Attributes:** Store (Nominal), Dept (Nominal), and Holiday_Flag (Binary/Nominal).
  - **Temporal Attributes:** Date (Continuous time-series with weekly granularity).
- **Cardinality:**
  - **Records:** The dataset contains approximately 421,570 transaction records.
  - **Dimensions:** It covers 45 distinct stores and 81 distinct departments.
  - **Time Points:** Data is recorded over 143 weeks.
- **Derived Attributes:** As required for deeper analysis, we computed the following derived attributes during the preprocessing phase:

  1. **Year and Month:** Extracted from the original Date column to enable hierarchical drill-down (Year > Month) and seasonal analysis.
  2. **Holiday_Type:** A categorical column derived from the numeric `Holiday_Flag`. Values were mapped from "1/0" to descriptive labels ("Holiday" / "Normal Day") to improve chart readability and storytelling.

4. Dashboards

This section presents the interactive visualization solution developed in Power BI. The report is divided into two distinct dashboards, each serving a specific user intent as defined in the project requirements.

- ### 4.1. Dashboard 1: Exploratory Analysis



Exploratory Analysis The first dashboard is designed for Exploratory Data Analysis (EDA). Its main goal is to allow the user to navigate through the data, discover seasonal trends, and find correlations between variables.

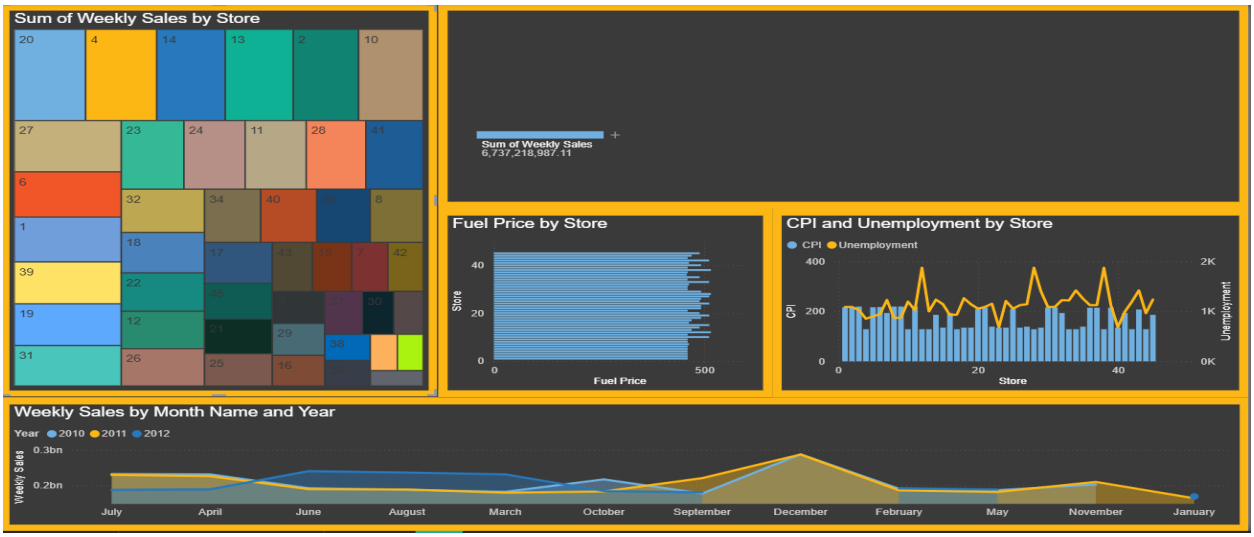
Key Features & Rationale:

Interactive Exploration: We implemented slicers for "Date" and "Holiday Name" to allow users to filter the results dynamically.

Trend Discovery: The Line Chart provides a clear view of how sales peak during holiday seasons over the 3-year period.

Relationship Analysis: The Scatter Plot is used to explore if there is a direct correlation between the Unemployment rate and Weekly Sales, with color-coding to distinguish between holiday and normal weeks.

- ## 4.2 Dashboard 2: Explanatory Storytelling



The second dashboard focuses on Explanatory Analysis, where we communicate specific, confirmed insights to the stakeholders. It follows a narrative flow to explain "why" certain sales figures occurred.

Key Features & Rationale:

Coherent Narrative: We used a Decomposition Tree to break down the total sales of $6.74bn into categories, showing the exact contribution of each holiday type.

Comparative Insights: The Treemap provides an immediate visual comparison of store performance, highlighting the top-performing branches.

Economic Context: By combining CPI and Unemployment in one chart, we tell a story about the regional economic environment and its impact on purchasing power.

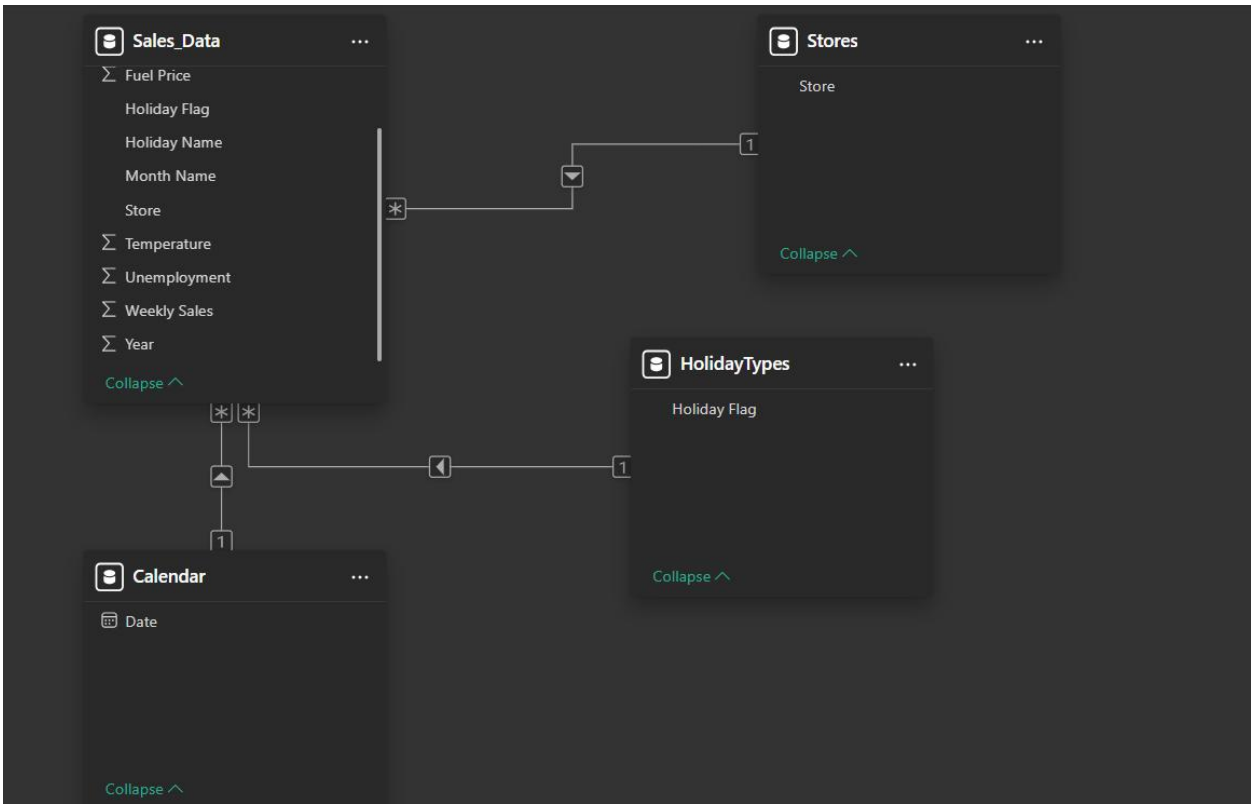
5. Preprocessing Steps

Data preprocessing is a critical phase to ensure the accuracy and reliability of the visual insights. We followed a rigorous **ETL (Extract, Transform, Load)** process using Power Query to prepare the Walmart dataset.

- **5.1 Data Cleaning and Transformation * Handling Missing Values:** We scanned the dataset for null values, specifically in economic indicators like CPI and Unemployment,

ensuring that every record used for analysis was complete and accurate. * **Data Type Standardization:** We ensured that Date fields were recognized as temporal objects and Weekly_Sales as decimal numbers to allow for correct mathematical aggregations.

- **Holiday Flag Transformation:** The raw data used a binary (0/1) for holidays; we transformed this into descriptive text labels ("Holiday Week" and "Normal Week") to enhance user readability in slicers and legends.
- **5.2 Data Augmentation (Derived Attributes)** As required by the project guidelines, we derived several key attributes to enrich the analysis:
- **Calendar Table Creation:** We generated a dedicated **Calendar Dimension Table** using DAX to extract Year, Month Name, from the raw date column.
- **5.3 Data Modeling (Star Schema)** To optimize performance and interactivity, we moved away from a flat table structure to a **Star Schema**.



- **Fact Table:** The main Sales_Data table containing measures like Weekly_Sales

- **Dimension Tables:** We created separate tables for Calendar, Stores, and HolidayTypes. * **Relationships:** We established **One-to-Many (1:*)** relationships between dimensions and the fact table, ensuring that filters applied to a store or a date propagate correctly across all visuals.

6. Discussion

6.1 Technical Challenges & Solutions:

Challenge: Initial difficulty in finding significant "Key Influencers" for sales increases using AI visuals.

Solution: Switched to a **Decomposition Tree** to manually deconstruct sales by Year and Holiday type.

6.2 Analytical Insights:

Cross-Highlighting: Enabled interaction between the **Treemap** and economic charts to reveal store-specific correlations.

7. Conclusion and Lessons Learned

- ## 7.1 Key Outcomes

- **Holiday Sales Dominance:** The analysis confirmed that "Holiday Weeks" trigger massive revenue spikes compared to normal weeks, regardless of the fiscal year.
- **Economic Resilience:** Data trends showed that Walmart’s sales remain remarkably stable despite fluctuations in **Unemployment rates** and the **Consumer Price Index (CPI)**.
- **Store Performance Variance:** We successfully identified top-performing branches (e.g., Store 20 and Store 4) using the **Treemap** visual, enabling data-driven management decisions.

- ## 7.2 Lessons Learned

- **Power BI Mastery:** Gained hands-on experience with advanced AI visuals, specifically the **Decomposition Tree**, to provide deep-dive analytical insights.
- **Data Modeling:** Learned the critical importance of the **Star Schema** and data preprocessing to ensure high dashboard performance and interactivity.
- **Storytelling vs. Exploration:** Developed the skill to distinguish between exploring data for patterns (**Exploratory**) and presenting confirmed insights to stakeholders (**Explanatory**).
- **Visual Hierarchy:** Mastered the use of formatting (Borders, Color schemes, and Layout) to guide the user's attention toward the most significant data points.

8. References

- **Data source :**
- Walmart Recruiting - Store Sales Forecasting. Retrieved from Kaggle.
- **URL : https://www.kaggle.com/datasets/asahu40/walmart-data-analysis-and-forcasting**
- **Software and Tools :**
- Microsoft Power BI Desktop (Version 2025) for data modeling and visualization
- **Course Materials:** DA 460 - Final Project, Dr. Zeyad Hailat, Yarmouk University
- **Methodology References :**
- "The Five Ws" framework for data context and visualization design principles
- **Libraries and Functions :**
- Data Analysis Expressions (DAX) for derived attributes and measures