MAHMOUD
MOHAEMD

2025

EDA
REPORT

Table of Contents

BUSINESS &DATA UNDERSTANDING

BUSINESS CONTEXT:

In today's competitive retail landscape, understanding sales patterns and the factors influencing them is crucial for strategic decision-making. This project focuses on analyzing historical sales data from Walmart, one of the largest retail chains in the United States. The primary business objective is to develop accurate sales forecasts at the department level across multiple stores, accounting for various influencing factors such as holidays, markdowns, temperature, fuel prices, and consumer price index (CPI).

Effective forecasting allows Walmart to:

- Optimize inventory levels,
- Improve supply chain efficiency,
- Reduce operational costs,
- Align promotional activities with consumer demand.

DATA UNDERSTANDING:

The dataset offers a structured and detailed view of Walmart’s sales ecosystem, capturing the interplay between store operations, economic conditions, seasonality, and promotional strategies. By examining these dimensions, we can begin to uncover patterns and relationships that drive sales performance.

Key Observations:

- Temporal Granularity: Data is organized on a weekly basis, allowing for the detection of seasonal trends, holiday effects, and short-term fluctuations.
- Geographical Spread: With data from 45 different stores, the dataset enables regional comparisons and helps assess how store size and type influence sales outcomes.
- Department-Level Detail: Sales data is segmented by department, providing a fine-grained view of category-level performance within each store.
- External Influencers: Variables like temperature, fuel prices, CPI, and unemployment allow us to explore how macroeconomic and environmental factors impact consumer behavior.
- Promotions & Discounts: The inclusion of multiple markdown variables helps evaluate the effectiveness of various promotional strategies over time.
- Holiday Effects: Holiday indicators make it possible to isolate and assess the sales impact of major holidays — a key consideration in retail planning.

DATA DESCRIPTION:

The dataset includes several key columns that provide information about store operations, sales, and external factors. Below is a description of each column:

- Store: Unique ID for each store
- Dept: Department number within the store
- Date: Week of the sale (YYYY-MM-DD)
- IsHoliday: Whether the week includes a major holiday (True/False)
- Weekly_Sales: Cleaned weekly sales figures
- Holiday_Flag: Holiday indicator (0 or 1)
- Temperature: Average weekly temperature
- Fuel_Price: Regional fuel price
- CPI: Consumer Price Index
- Unemployment: Unemployment rate
- Markdown1–5: Promotional markdown values
- Type: Store type (A, B, or C)
- Size: Store size (in square feet)

Data Overview

This section provides a first look at the datasets used in this project, helping us understand their structure, coverage, and potential quality issues. The analysis covers:

- Dataset Dimensions: Understanding the number of rows and columns in the DataSet.

- Column Types & Nulls: Reviewing data types and identifying missing values that may require cleaning or imputation.

- Unique Values

- Date Range: Verifying the temporal coverage of the dataset, which spans weekly data from early 2010 to mid-2012.

DATA SET DIIMENSIONS

The dataset contains **421570** rows and **17** columns

COLUMN TYPES

The dataset includes a mix of numerical and categorical columns allowing for both quantitative analysis and time-based trend evaluation.

UNIQUE VALUES

We identified several unique values that provide key insights into the structure of the business operations:

- Stores: The dataset includes data from 45 distinct stores, each contributing to the overall performance metrics.

- Departments: There are 81 unique departments across all stores, showcasing the diverse range of products and services offered.

- Store Size Type: The stores are categorized into three types based on their size: Type A, Type B, and Type C. This classification helps differentiate between stores based on their operational scale and could influence sales performance and inventory management.

DATE RANGE

The dataset spans a period from February 5, 2010 to October 26, 2012, providing a comprehensive view of the data over nearly three years. This date range allows for an analysis of trends and patterns over time, including seasonal variations, promotions, and other time-dependent factors that may have impacted the data.

NULL VALUES

Upon inspecting the dataset, we found that the majority of key fields—such as Store, Dept, Date, IsHoliday, Weekly_Sales, Holiday_Flag, Temperature, Fuel_Price, CPI, Unemployment, Type, and Size all contain no missing values, with all having a 0% null percentage.

However, there are null values present in the MarkDown columns (MarkDown1, MarkDown2, MarkDown3, MarkDown4, and MarkDown5), with missing data percentages ranging from approximately 64% to 74%. This is a noteworthy observation, as it may reflect periods where promotional markdowns were not applied or where data collection for discounts was incomplete.

DATA QUALITY ISSUES & SOLUTIONS

Upon analyzing and reviewing the dataset, we identified several data quality issues that need to be addressed to improve the dataset's integrity and usefulness for analysis:

DataSet Issues

1-Data Type Inconsistencies:

The Date column is currently in object data type format, which limits its usefulness for time-based analysis, such as trend analysis or time series forecasting. For a more efficient and accurate analysis, it is crucial to convert this column to the appropriate DateTime format. Converting this column to the proper date and time data type will enable us to perform time-related operations, such as filtering, sorting, and aggregation by date, and facilitate any time series modeling we might need to conduct.

2-High Percentage of Missing Values in MarkDown Columns:

The MarkDown1 to MarkDown5 columns exhibit a significant number of missing values, with missing data percentages ranging from 64% to 74%. This level of missingness exceeds acceptable thresholds (typically 40-50%) and can have a considerable impact on analysis and modeling. The absence of this critical discount data could lead to incomplete insights into sales patterns, especially related to promotions.

Solutions

1-Convert Date Column to DateTime Format:

The Date column should be converted from its current object data type to the DateTime data type.

2-Handle Missing Values in MarkDown Columns:

Given the high percentage of missing data in the MarkDown1 to MarkDown5 columns, and considering the data's importance in promotional analysis, the best solution is to drop these columns entirely. This is because the missing data exceeds the acceptable threshold and would likely lead to unreliable analysis if imputed or filled with arbitrary values.

3-Dropping Redundant &unnecessary Columns [Date,IsHoliday]

FEATURE EXTRACTION

To enrich the dataset and enable more meaningful analysis, we engineered a series of time-related features derived from the original Date column. These new features provide additional context that can help capture seasonal trends, promotional effects, and time-based patterns in the data:

- Month, Year, Quarter, WeekOfYear: Extracted individual time components to facilitate temporal grouping and trend analysis across months, years, and quarters.

- Season: Categorized each date into one of four seasons (Winter, Spring, Summer, Fall) based on the month. This allows for seasonal performance analysis.

- IsPromoWeek: A binary flag indicating whether a date aligns with major promotional events such as the Super Bowl, Labor Day, Thanksgiving, or Christmas. This helps identify sales spikes and promotional effects.

These features provide the temporal granularity and context needed to perform detailed time series analysis, detect seasonality, and understand promotional impacts on sales.

UNDERSTANDING COORELATION

A correlation heatmap was generated to explore the linear relationships between the remaining numerical variables in the dataset. This analysis helps identify which features may be more influential or redundant when building predictive models.

Key observations:

- Weekly_Sales shows a moderate negative correlation with Store (-0.31), which could suggest that certain stores consistently underperform compared to others.

- CPI (-0.08), Unemployment (-0.10), and Temperature (-0.04) also show weak negative correlations with Weekly_Sales. These weak associations imply that while these external factors may influence sales slightly, they do not have a strong linear relationship.

- Size and Dept (0.79) demonstrate a strong positive correlation, indicating that larger stores tend to have more departments—a logical and expected relationship.

- Temporal variables, including Month, WeekOfYear, Year, and Quarter, are strongly correlated with each other:
    - Month and WeekOfYear (0.97) and WeekOfYear and Quarter (0.96) show particularly high correlations, reflecting the natural structure of calendar data.
    - This high multicollinearity among time features suggests that not all should be included in certain models simultaneously, to avoid redundancy and distortion of variable importance.

Since the MarkDown columns were dropped due to excessive missing values, they were excluded from consideration in both the correlation analysis and future modeling. This helps ensure a cleaner dataset and reduces noise from incomplete promotional data.

OUTLIERS DETECTION & HANDLING

To ensure the reliability and accuracy of our analysis, we conducted a thorough outlier detection process across key numerical features. The goal was to identify extreme values that could distort statistical measures or model performance, and to make informed decisions based on domain knowledge and distribution characteristics.

Features Analyzed for Outliers:

We focused on the following columns where outliers could significantly affect results:

- Weekly_Sales
- Temperature
- Unemployment
- CPI
- Size
- Fuel_Price

Visual Techniques Employed:

To understand the distribution and detect potential outliers, we utilized several visual tools:

- Histograms to observe the overall distribution and skewness
- Boxplots based on the IQR method to flag statistical outliers
- Violin plots to illustrate distribution density and variation

Findings & Decisions:

Each feature was evaluated individually, considering both statistical thresholds and contextual realism:

- Unemployment had ~32,000 outliers, with a maximum value (14.31) only slightly exceeding the upper bound. Given its plausibility, the values were retained without modification.

- CPI, Fuel_Price, and Size showed no statistical outliers and were left unchanged.
- Temperature had 69 flagged values ranging from -2.06°F to 100.14°F. These are extreme but realistic, so they were kept.

- Weekly_Sales had a significant number of outliers (2,179), with maximum values far exceeding the upper IQR limit. Given the nature of sales data, these spikes could reflect actual promotional events or seasonal trends. However, to prevent these from disproportionately influencing models, a capping strategy is recommended—particularly around the 95th to 99th percentile.

Final Strategy Summary:

Retained without changes:

- Unemployment, Temperature, CPI, Fuel_Price, Size

Capped at upper threshold:

- Weekly_Sales → Suggested approach: clip values at 95th or 99th percentile

By selectively applying outlier handling, we maintain the dataset's integrity while minimizing the distortion caused by extreme values. This balance ensures that our models and insights remain both robust and reflective of real-world patterns.

Conclusion

This report presented a comprehensive analysis of the Walmart sales dataset, focusing on data quality assessment, exploratory insights, and preparation for modeling. Key findings included a diverse store network across 45 locations and 81 departments, with three store types (A, B, C), and a sales timeframe ranging from February 2010 to October 2012.

Data cleaning efforts addressed issues such as improper data types and high missing values. The Date column was successfully converted to datetime format to enable time-based analysis, and the MarkDown features were dropped due to excessive null values. Outlier detection guided the capping of extreme Weekly_Sales values at the 95th–99th percentile to reduce skewness while preserving important trends.

Output:

- Cleaned dataset with consistent types and structure
- Null values eliminated or columns dropped based on severity
- Time-based features enabled via Date conversion
- Outliers capped selectively to enhance model robustness
- Ready-to-use dataset for Extracting Insights

We thank you for your continued support

MAHMOUD
MOHAEMD

2025

EDA
REPORT