**Walmart Store Sales Forecasting and Analysis**

Insights into Trends, Performance, and External Factors

Pavan Kalyan Imadabathini

01/27/2025

Table of Contents

[**1.** **Project Overview** 3](#_Toc189066392)

[**2.** **Problem Definition** 4](#_Toc189066393)

[**3.** **Objectives** 5](#_Toc189066394)

[**4.** **Datasets Used** 6](#_Toc189066395)

[**5.** **Methodology** 7](#_Toc189066396)

[**6.** **Insights and Key Findings** 11](#_Toc189066397)

[**7.** **Challenges and Assumptions** 13](#_Toc189066398)

[**8.** **Recommendations** 14](#_Toc189066399)

[**9.** **Conclusion** 15](#_Toc189066400)

[**Bonus Task** 16](#_Toc189066401)

# **Project Overview**

This project leverages Walmart store sales data to uncover trends, identify performance patterns, and analyze the impact of external factors. The objective is to provide actionable insights for store management and strategy optimization by performing Python-based analysis and creating Tableau visualizations.

# **Problem Definition**

Walmart experiences variability in sales influenced by multiple factors such as holidays, temperature fluctuations, and unemployment rates. Understanding these patterns is critical for optimizing business strategies.

**Key Goals:**

* Analyze historical sales data to identify trends and patterns.
* Evaluate the influence of external factors (temperature, CPI, unemployment).
* Create an interactive dashboard to present findings and inform decision-making.

# **Objectives**

1. **Data Preparation:**
   * Clean and preprocess sales data for reliable analysis.
2. **Exploratory Data Analysis (EDA):**
   * Identify insights such as monthly trends, holiday impacts, and temperature-driven patterns.
3. **Visualizations:**
   * Build interactive dashboards to:
     + Highlight sales trends.
     + Examine external factors affecting performance.
     + Identify top-performing stores and departments.

# **Datasets Used**

The datasets were sourced from the [Kaggle 5.almart Store Sales Forecasting Competition](https://www.kaggle.com/competitions/walmart-recruiting-store-sales-forecasting/overview).

**Dataset Overview:**

* **Train Data:** Weekly sales data by store and department.
* **Features Data:** External factors (temperature, fuel price, CPI, etc.).
* **Stores Data:** Metadata for each store, including size and type.

# **Methodology**

**1. Data Preprocessing**

**Steps Taken:**

* **Merging Datasets:**  
  Combined Train, Features, and Stores datasets using Store and Date as primary keys.
* **Handling Missing Values:**  
  Filled markdown columns with 0 and dropped records with critical missing data.
* **Feature Engineering:**
  + **Created new features:**
    - Temperature Category: Binned temperatures into categories like "Cold," "Moderate," etc.
    - Holiday Boost: Multiplication of IsHoliday and Weekly\_Sales.
    - Sales Per Square Foot: Calculated as Weekly\_Sales / Store Size.
* **Duplicate Removal & Normalization:**
  + Removed duplicate rows.
  + Standardized text formatting for categorical fields.

**Key Outputs:**

* Clean, merged dataset ready for analysis and visualization.
* Enhanced granularity through newly engineered features.

2**. Data Analysis**

**Techniques and Insights:**

* **Correlation Analysis:**
  + Heatmap revealed positive correlations between holiday weeks and sales.
  + External factors like unemployment and CPI showed a negative correlation with sales.
* **Clustering:**
  + Used K-Means clustering to group stores by size and sales, helping to segment stores for targeted strategies.
* **Top Performers**:
  + Identified top stores and departments contributing to revenue.

3. **Visualizations (Tableau Dashboards)**

**Key Dashboards and Insights:**

* **Sales Trends Over Time:**
  + **Visualization:** A bar chart showing monthly sales patterns.
  + **Purpose:** Highlights the impact of seasonality and holidays on sales performance.
  + **Key Insight:** Strong seasonal peaks during colder months and holiday seasons.
* **Top-Performing Stores:**
  + **Visualization:** Ranked bar chart showing top stores based on revenue.
  + **Purpose:** Identifies high-performing stores for resource allocation and strategy refinement.
  + **Key Insight:** Stores 20, 5, and 10 consistently outperform others.
* **Average Sales by Temperature:**
  + **Visualization:** A bar chart showing average weekly sales by temperature categories.
  + **Purpose:** Evaluates how temperature influences customer behavior and spending.
  + **Key Insight:** Moderate and cold temperatures are associated with higher sales.
* **Weekly Sales vs. Temperature (Scatter Plot):**
  + **Visualization:** Scatter plot showing the relationship between weekly sales and temperature.
  + **Purpose:** Provides granular insights into temperature's impact on sales patterns.
  + **Key Insight:** Extreme weather conditions (hot or cold) lead to reduced sales.
* **Sales by Holiday Status and Temperature**:
  + **Objective:** Combine holiday and temperature effects.
  + **Insight:** Holidays boosted sales across all temperature categories.
* **Impact of Unemployment and Fuel Prices:**
  + **Visualization:** Scatter plot showing the correlation between sales, unemployment, and fuel prices.
  + **Purpose**: Demonstrates how macroeconomic conditions affect consumer spending.
  + **Key Insight:** Rising unemployment and fuel prices negatively affect sales.
* **Top 10 Departments Weekly Sales:**
  + **Objective:** Highlight the most profitable departments.
  + **Insight:** Departments 79, 40, and 13 contributed the most to total revenue.
* **Comprehensive Dashboard:**
  + **Visualization:** An interactive Tableau dashboard integrating all key insights.
  + **Features:**
    - Filters for Year, Store, Department, and Temperature Category.
    - Dynamic visualizations for deeper analysis.
  + **Purpose:** Enables decision-makers to explore data interactively and uncover actionable insights.

# **Insights and Key Findings**

**1. Monthly Sales Trends:**

* **Seasonal Impact:**Sales show a clear peak during colder months, such as November and December, likely driven by holiday shopping.
* **March Spike:**March demonstrates an unexpected sales spike, possibly due to specific promotions, seasonal changes, or fiscal calendar events.
* **Consistency Across Other Months:**While colder months dominate, summer months display relatively steady performance, albeit lower than winter.

**2. Impact of Temperature:**

* **Higher Sales in Moderate Temperatures:**Moderate and cold temperatures correspond to higher average weekly sales.
  + **Reason:** Moderate weather might increase foot traffic and shopper engagement.
* **Lower Sales in Extreme Temperatures:**Very hot and very cold temperatures negatively impact sales.
  + **Reason:** Adverse weather conditions reduce customer willingness to visit stores.

**3. Top Performers:**

* **Top-Performing Stores:**Stores 20, 5, and 10 consistently lead in total sales across all time periods.
  + These stores may benefit from better location, larger customer bases, or stronger management strategies**.**
* **Top-Performing Departments:**Departments 79 and 40 generate the highest revenue.
  + Strategic focus on these departments could help maintain and grow overall store profitability.

4. Economic Factors:

* **Unemployment:**Higher unemployment correlates with reduced sales.
  + Actionable Insight: Track unemployment trends to adjust pricing and promotions accordingly**.**
* **Fuel Prices:**Increases in fuel prices negatively impact sales, as consumers may prioritize essential goods and reduce discretionary spending.

**5. Holiday Impact:**

* **Boost Across Categories:**Holidays consistently drive higher sales, regardless of temperature conditions.
  + This insight emphasizes the importance of preparing holiday-specific strategies for promotions and inventory management.

# **Challenges and Assumptions**

**Challenges Faced:**

1. **Data Cleaning:**
   * Addressing missing values in critical fields (e.g., markdowns).
   * Ensuring datasets were correctly merged despite missing unique identifiers.
2. **Feature Engineering:**
   * Developing new features like Holiday Boost and Temperature Category required careful validation.
3. **Handling Large Datasets:**
   * Iterative validation was required to maintain data integrity.

**Assumptions:**

* Sales are influenced by external factors such as temperature, unemployment, and fuel prices.
* Holiday weeks universally boost sales, regardless of external factors.
* Clustering stores by sales and size provides meaningful segmentation for strategic planning.

# **Recommendations**

Based on the analysis and visualizations, the following recommendations are proposed:

1. **Promotions During Cold Months:**
   * Increase marketing and promotions during colder months (November to March) and holiday periods to capitalize on higher sales potential.
2. **Strategic Focus on Top Performers:**
   * Allocate more resources (inventory, staffing, promotions) to top-performing stores (20, 5, 10) and departments (79, 40).
3. **Economic Monitoring:**
   * Closely monitor trends in unemployment and fuel prices.
   * Consider adjusting pricing and promotions during periods of economic uncertainty to maintain customer engagement.
4. **Tailored Strategies for Clusters:**
   * Use clustering insights to develop targeted strategies for underperforming stores.
   * For example, focus on improving product mix, customer engagement, or localized promotions.
5. **Inventory Optimization:**
   * Use temperature and holiday analysis to align inventory levels with predicted demand.

# **Conclusion**

This project delivers an in-depth analysis of Walmart's sales data, highlighting significant insights into sales trends, external influences, and the performance of stores and departments. The Tableau dashboard enables interactive exploration for decision-makers, providing a comprehensive platform to assess and act on data-driven findings effectively.

**Key Outcomes:**

* **Seasonal Impact:** Sales peak during colder months and holidays, with March and December showing notable spikes, emphasizing the importance of seasonal promotions.
* **External Factors:** Moderate temperatures are associated with higher sales, while extreme conditions and economic factors such as unemployment and fuel prices negatively impact performance.
* **Performance Insights:** Stores 20, 5, and 10 consistently outperform others, while departments 79 and 40 generate the highest revenue, highlighting strategic focus areas. Meanwhile, certain departments and clusters show potential for growth with targeted interventions.

**Business Value:**

* These insights provide Walmart's management with actionable recommendations to refine sales strategies, optimize resource distribution, and improve profitability. By leveraging seasonal trends, addressing external challenges, and focusing on underperforming areas, Walmart can enhance operational efficiency and maintain a competitive edge in the retail sector.

# **Bonus Task**

The work accomplished in the bonus tasks, detailing the methodologies employed and the insights gained from each.

**1. Automating the Data Pipeline:**

**Objective:** To streamline the process of data extraction, transformation, and loading (ETL) by automating routine tasks, ensuring efficiency and consistency.

**Approach:**

* ETL Functions: Developed modular functions in Python:
  + extract\_data(): Reads raw data from a specified CSV file.
  + transform\_data(): Processes the data, such as converting date columns to datetime objects.
  + load\_data(): Saves the transformed data to a new CSV file for subsequent analysis.
* **Automation:** Encapsulated the ETL process within an etl() function, which sequentially calls the extraction, transformation, and loading functions. This script can be scheduled to run at predetermined intervals using task schedulers like cron (on Unix-based systems) or Task Scheduler (on Windows).

**Outcome:** The automated pipeline ensures that data is consistently updated and pre-processed without manual intervention, enhancing efficiency and reducing the potential for errors.

**2. Predicting Future Sales:**

**Objective:** To forecast upcoming sales trends using established time series analysis techniques, aiding in inventory management and strategic planning.

**Approach:**

* **Data Preparation:** Aggregated sales data every week to capture temporal patterns.
* **ARIMA Model:** Implemented the Autoregressive Integrated Moving Average model:
  + Model Fitting: Configured the model with parameters (p=5, d=1, q=0) based on data characteristics.
  + Forecasting: Predicted sales for the subsequent 12 weeks.
  + Visualization: Plotted historical sales alongside ARIMA forecasts to visualize trends.
* **Exponential Smoothing:** Applied the Holt-Winters Exponential Smoothing method:
  + Model Configuration: Set parameters to account for additive trends and seasonality with a periodicity of 52 weeks.
  + Forecasting: Generated 12-week sales forecasts.
  + Visualization: Compared historical sales data with Exponential Smoothing forecasts.

**Outcome:** Both models provided projections of future sales, offering valuable insights into decision-making. The visualizations facilitated a clear comparison between actual sales and predicted trends.

**3. Analyzing Marketing Campaign Effectiveness:**

**Objective:** To assess the impact of a specific marketing campaign on sales by comparing sales figures before, during, and after the campaign period.

**Approach:**

* Campaign Period Definition: Specify the campaign's start and end dates (e.g., '2011-10-08' to '2011-10-13').
* Data Segmentation: Divided the sales data into:
  + Pre-Campaign Sales: Sales data prior to the campaign.
  + During-Campaign Sales: Sales data during the campaign period.
  + Post-Campaign Sales: Sales data following the campaign.
* Statistical Analysis: Conducted an independent t-test to compare pre-campaign and during-campaign sales, evaluating the null hypothesis that there is no significant difference between the two periods.

**Outcome:** The t-test yielded a t-statistic and p-value, which were interpreted to determine the campaign's effectiveness. A non-significant p-value (e.g., p > 0.05) suggests that the campaign did not lead to a statistically significant change in sales during the campaign period compared to the pre-campaign period.

**Conclusion:**

The marketing campaign executed between October 8 and October 13, 2011, did not result in a measurable change in sales during the campaign period compared to the pre-campaign period. This finding implies that the implemented campaign was ineffective in influencing sales. It is recommended that the campaign's strategy, including its content, target audience, timing, and delivery channels, be reviewed to identify potential areas for improvement in future marketing efforts.