**The Walmart Sales Exploratory**

**Data Analysis And Clustering**

# CONTENTS

# **Setting**

The Walmart Sales Forecast dataset available on Kaggle furnishes crucial sales data from 45 Walmart stores situated in the United States, which grants us valuable insights into several factors that have an impact on weekly sales, such as store type, size, promotional events, date, and unemployment. The main purpose of scrutinizing this dataset is to pinpoint the significant performance indicators that impact retail sales and establish a clustering model that can assist in concentrating additional focus on stores belonging to clusters exhibiting lower sales performance.

By identifying and analysing essential performance metrics such as store type, size, and promotional events, Walmart can optimize their operations and enhance their revenue streams. Additionally, understanding the influence of external factors such as unemployment rates in the surrounding areas, historical sales trends, and seasonality can aid in developing effective strategies for targeted improvements in the retail industry.

By utilizing clustering models to group stores based on their sales performance, Walmart can identify the underperforming stores and focus additional resources on those stores to help them achieve better sales outcomes. This approach can enable Walmart to streamline their operations, improve efficiency, and increase profitability, ultimately leading to better overall sales performance.

# Data Description

**Dataset Name:** Walmart Sales - Source: Kaggle (<https://www.kaggle.com/code/kishorekumar1398/analysis-prediction-of-walmart-sales/edit>)

**Description:**

The Walmart Sales dataset contains historical sales data for 45 Walmart stores located in different regions. The dataset consists of the following columns:

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Store | A unique identifier for each store (integer). |
| Date | The date of the recorded sales (date). |
| Weekly\_Sales | The total sales for the given week in a particular store (float). |
| Holiday\_Flag | A binary indicator (0 or 1) that denotes whether the week contains a holiday or not (integer). |
| Temperature | The average temperature in the region of the store for the given week (float). |
| Fuel\_Price | The cost of fuel in the region of the store for the given week (float). |
| MarkDown1 to MarkDown5 | Additional promotional markdowns (numeric). |
| CPI (Consumer Price Index) | The Consumer Price Index for the region of the store for the given week (float). |
| Unemployment | The unemployment rate in the region of the store for the given week (float). |

# **Analysis and Discussion:**

**Approach:**

1. *Data Cleansing*: The initial step in our analysis entailed meticulous data cleansing, wherein we eliminated duplicate values and addressed any missing data points. Additionally, we converted the "Store" column into a factor to ensure proper order for subsequent visualization.

2. *Exploratory Data Analysis (EDA)*: We conducted comprehensive EDA to glean meaningful insights from the data. This included calculating the total sales for each store, identifying the store with the highest sales, and determining the standard deviation of weekly sales for each store to pinpoint the one with the most significant sales fluctuations.

3. *Data Visualization*: To effectively communicate our findings, we utilized various data visualization techniques. For instance, we employed a bar chart to visually represent the total sales for each store and a density plot to showcase the sales fluctuation pattern for store 14.

4. *Time Series Analysis*: We further conducted time series analysis by creating a new column "m\_Year" that extracted the month and year from the "Date" column. We then filtered the data based on the quartiles of the "m\_Year" column to gain insights into the sales trend over time.

**Findings:**  
1. Store 20 emerged as the top performer with the highest total sales of $301,397,792.46, while store 33 had the lowest total sales of $37,160,222.

2. Store 14 exhibited the highest standard deviation of weekly sales, signifying considerable sales fluctuations compared to other stores

3. The density plot analysis revealed that store 14 experienced the highest sales fluctuation in the earlier weeks, as evident from the plot's distribution pattern.

4. The time series analysis unveiled a seasonal trend in sales, with the first quartile (January to March) witnessing the highest sales, possibly due to holiday sales, a slight dip in the second quartile (April to June), and a marginal increase in the third quartile (July to September).

**Discussion:**  
The analysis of the Walmart sales data has yielded significant insights. Store 20's robust sales performance, as evidenced by the highest total sales, highlights its strong market performance, while store 33's comparatively lower sales necessitate further investigation into potential underlying reasons.

Store 14's notable sales fluctuations, as indicated by the highest standard deviation of weekly sales, may point to issues in demand forecasting, inventory management, or market dynamics, warranting further investigation to optimize sales performance.

The time series analysis revealing a seasonal trend in sales can aid Walmart in identifying peak sales periods, such as the first quartile, and planning inventory and marketing strategies accordingly. However, the dip in the second quartile may necessitate a deeper investigation to identify the underlying reasons and take corrective actions.

In conclusion, the findings from this analysis can provide valuable inputs for strategic decision-making at Walmart, including identifying top-performing stores, addressing sales fluctuation issues, and optimizing inventory and demand forecasting processes to enhance overall sales performance and profitability. Further in-depth analysis and investigation can be conducted to identify and mitigate the underlying factors impacting sales performance, and to develop data-driven strategies for sustained growth and success.

**Visualization:**

Chart, histogram

Description automatically generated

**A picture containing plot, diagram, screenshot

Description automatically generatedA picture containing diagram, line, plot, screenshot

Description automatically generated**

**A picture containing line, plot, text, font

Description automatically generated**

**A picture containing line, plot, text, diagram

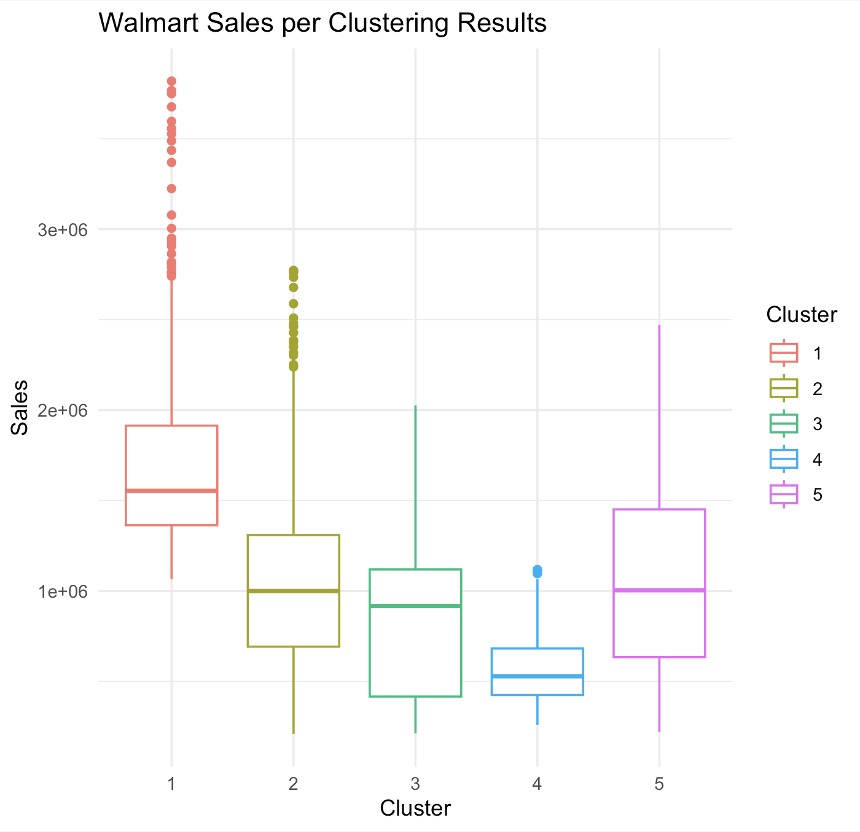
Description automatically generated**

**A picture containing line, plot, diagram, screenshot

Description automatically generated**

**A picture containing line, plot, diagram, screenshot

Description automatically generated**

****