Exploring Walmart Sales across 45 different stores

**Authors:** Anqa Javed, Humaira Enayetullah, Julio Carneiro, Shahab Eshghifard.

## Introduction

Retail businesses generate vast amounts of sales data and can provide valuable insights into customer behavior, seasonal trends, and business performance. Walmart, as one of the largest retail chains, offers a comprehensive dataset with sales information across multiple stores and departments. In this project, we will examine Walmart sales data from February 5, 2010, to November 1, 2012, to understand sales trends, external economic influences, and the impact of holidays on consumer spending.

## Data Analysis

We used **pandas** for cleaning our data and creating new columns for further analysis. We use **NumPy** and **SciPy** to do the data analysis and create linear regression models to identify if there are any external factors that influence sales. Finally, we utilized **pandas** and **matplotlib** for data visualization, generating bar graphs, pie charts, scatter plots, and line graphs to better understand sales patterns and trends.

## Summary of findings

* Weekly sales data showed a wide range, with most sales falling within expected thresholds.
* The top 10 stores had more than $2 million in weekly sales while the bottom stores had less than $0.5 million total weekly sales.
* More than 50% of sales are below $1 million, likely due to Walmart being used for daily needs.
* There appears to be a seasonal trend in weekly sales with the highest monthly sales happening in April and July (approximately $6.5 million) while the lowest sales happen in January when it is less than $3.5 million. This may be because people are saving after the Christmas and New Year holidays.
* Based on our scatter plots there is a weak correlation between weekly sales and average temperature, weekly sales and consumer price index, weekly sales and fuel price, and weekly sales and unemployment rate with R square values of 0.005.
* The weekly sales data is not normally distributed. It is positively skewed with the mean of weekly sales being 1,046,964.88, the median being 960,746.04, and the mode being 209,986.25.
* There were a few extreme values identified as potential outliers, which may indicate unusual sales events like holidays or promotions. There are 34 outliers in weekly sales when there is a high volume of sales and most of these happen prior to the holidays.

**Limitations**

* This dataset consists of 6435 rows and 8 columns, which may not be sufficient for robust correlation results.
* The dataset uses 45 different Walmart stores, but the data provides no information about the location of these stores. This limits our ability to do any regional analysis or consider any geographic factors that might influence sales.
* The Holiday flag column in the dataset only marks Super Bowl, Labor Day, Thanksgiving, Christmas as holidays. However, other holidays such as Easter, and other religious holidays were not included. Since Walmart often runs promotions during these events, incorporating them into the dataset would provide a more comprehensive understanding of holiday-related sales trends.
* There is no information on the units of certain columns such as temperature and fuel price.

**Future Suggestions**

* Develop a model to optimize stocking during peak sales periods, especially holidays.
* Perform sentiment analysis on customer reviews to understand consumer behavior.
* Expand the Dataset and look for additional sales datasets on Kaggle or other open data sources to increase the sample size.
* Include data on employee numbers for better management of stores.
* Combine multiple datasets to include more variables, for example customer demographics and promotions.

**References**

*M Yasser H (Owner), Walmart Dataset*. Kaggle. Retrieved from<https://www.kaggle.com/datasets/yasserh/walmart-dataset>

OpenAI. (2025). *ChatGPT –* Retrieved from<https://openai.com/chatgpt>

Real Python. (n.d.). *Correlation with NumPy, SciPy, and Pandas*. Retrieved from<https://realpython.com/numpy-scipy-pandas-correlation-python/>

GeeksforGeeks. Box Plot in Python using Matplotlib. Retrieved from <https://www.geeksforgeeks.org/box-plot-in-python-using-matplotlib/>

W3Schools. Retrieved from <https://www.w3schools.com/python/>