# Capstone Project – Walmart Sales Forecasting

-by Yoganshu Raikhere

Table of Contents

1. Problem Statement
2. Project Objective
3. Data Description
4. Data Pre-processing Steps and Inspiration
5. Choosing the Algorithm for the Project
6. Motivation and Reasons for Choosing the Algorithm
7. Assumptions
8. Model Evaluation and Techniques
9. Inferences from the Same
10. Future Possibilities of the Project
11. Conclusion
12. References

# Problem Statement

Our primary goal is to forecast the weekly sales for stores using the provided dataset, which includes features such as store number, weekly sales values, temperature, CPI, etc. This analysis will help us understand how various time-based factors (like day of the week, seasonal patterns, and holidays).

A critical aspect of this analysis is to investigate how holidays impact weekly sales. Holidays are expected to significantly boost store sales due to increased consumer spending during these periods. By incorporating holidays into our predictive models, we can better understand their effect and make more accurate sales forecasts.

This comprehensive approach will enable us to identify patterns and trends in the data, allowing for improved strategic decision-making and resource allocation to maximize sales and profitability.

# Project Objective

The objective of this project is to forecast weekly sales for stores using the provided dataset. With multiple combination of different time series models like AR, ARIMA, SARIMAX, we aim to achieve the best fit model that will forecast the weekly sales for 12 more weeks with minimum root mean squared error.

By incorporating these factors into predictive models, the study aims to provide more accurate sales forecasts and strategic insights to improve operational efficiency and profitability.

# Data Description

The dataset available is Walmart DataSet.csv

Data description, various insights from the data.

The walmart.csv contains 6435 rows and 8 columns.

|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| Store | Store number |
| Date | Week of Sales |
| Weekly\_Sales | Sales for the given store in that week |
| Holiday\_Flag | If it is a holiday week |
| Temperature | Temperature on the day of the sale |
| Fuel\_Price | Cost of the fuel in the region |
| CPI | Consumer Price Index |
| Unemployment | Unemployment Rate |

# Data Preprocessing Steps and Inspiration

The preprocessing of the data included the following steps:

1. **Checked for Null Values:** Identify any missing values in the dataset to ensure data completeness.
2. **Checked for Duplicates:** Detect and remove any duplicate entries to maintain data integrity.
3. **Checked for Data Types:** Verify the data types of each column to ensure they are appropriate for analysis.
4. **Converted Date Column Data Type:** Changed the data type of the date column from an object to a datetime format for accurate time-based analysis.
5. **Set Date Column as Index:** Set the date column as the index to facilitate time series analysis and improve the accessibility of time-related data.

These steps are crucial for preparing the dataset for further analysis and modelling.

The main inspiration for this study is to leverage the power of data analytics to enhance sales forecasting accuracy for Walmart stores.

# Choosing the Algorithm for the Project

Description for the Time Series algorithm for the project.

I have chosen the below algorithm for this project for the following reasons:

1. **Autoregressive (AR) Model**

**Reason**: The AR model is a fundamental time series model that uses the dependency between an observation and a number of lagged observations (previous values). It is useful for capturing the linear relationship between past and current values of the series, making it a good starting point for understanding time series data.

2**. Autoregressive Integrated Moving Average (ARIMA) Model**

**Reason**: The ARIMA model extends the AR model by incorporating differencing (to make the series stationary) and a moving average component (to account for lagged forecast errors). This combination makes ARIMA a powerful model for a wide range of time series forecasting problems, especially when data shows evidence of non-stationarity.

**3. Seasonal ARIMA (SARIMAX) Model**

**Reason**: The SARIMAX model further extends ARIMA by including seasonal components and external variables. This model is particularly valuable when the time series data exhibits seasonal patterns (e.g., monthly or quarterly cycles or in our case weekly) and when other external factors (like holidays, or economic indicators) impact the series.

**4. AUTO\_ARIMA Model**

**Reason**: AUTO\_ARIMA automates the process of finding the best parameters for the ARIMA model by performing a grid search over possible parameters. This automation simplifies the model selection process and ensures that the best ARIMA model configuration is chosen for the data, saving time and effort while maintaining accuracy.

# Assumptions

The following assumptions were made in order to create the model for Walmart Sales project.

1. **Stationarity**: The dataset (or the transformed dataset) is assumed to be stationary, meaning its statistical properties like mean, variance, and autocorrelation are constant over time.

2. **Linearity**: The relationship between the current value and its past values (lags) is assumed to be linear.

3. **No Missing Values**: The dataset is assumed and checked to be complete with no missing values, or any missing values have been appropriately handled through imputation or other methods.

4. **Time-Ordered**: The data is time-ordered, meaning each observation is sequentially arranged according to time, which is crucial for time series analysis.

5. **Homogeneity**: The underlying process generating the data is assumed to be homogeneous over time, implying that the data-generating process remains the same throughout the period under analysis.

# Model Evaluation and Technique

The following techniques and steps were involved in the evaluation of the models like AR, ARIMA, SARIMAX, and AUTO\_ARIMA

**1. Train-Test Split**

**Purpose**: To assess how the model generalizes to new, unseen data. Split the dataset into a training set and a test set, usually in a chronological order to respect the time series nature.

**2. Dataset Split**

**Purpose: Split dataset for each store i.e., 45 to predict individual sales**

**3. Performance Metrics**

**Root Mean Squared Error (RMSE)**: The square root of MSE, providing error magnitude in the same units as the data.

RMSE=sqrt(MSE)

**4. Model Selection Criteria**

* **Akaike Information Criterion (AIC)**: A measure of the relative quality of statistical models for a given dataset. Lower values indicate a better fit. Used in pmdarima.

**5. Forecast Accuracy**

* **Purpose**: To ensure the model makes accurate future predictions.

Compare the model's forecasts with actual observations by plotting the predicted value over test values.

**6. Comparison of Models**

* **Purpose**: To select the best-performing model.

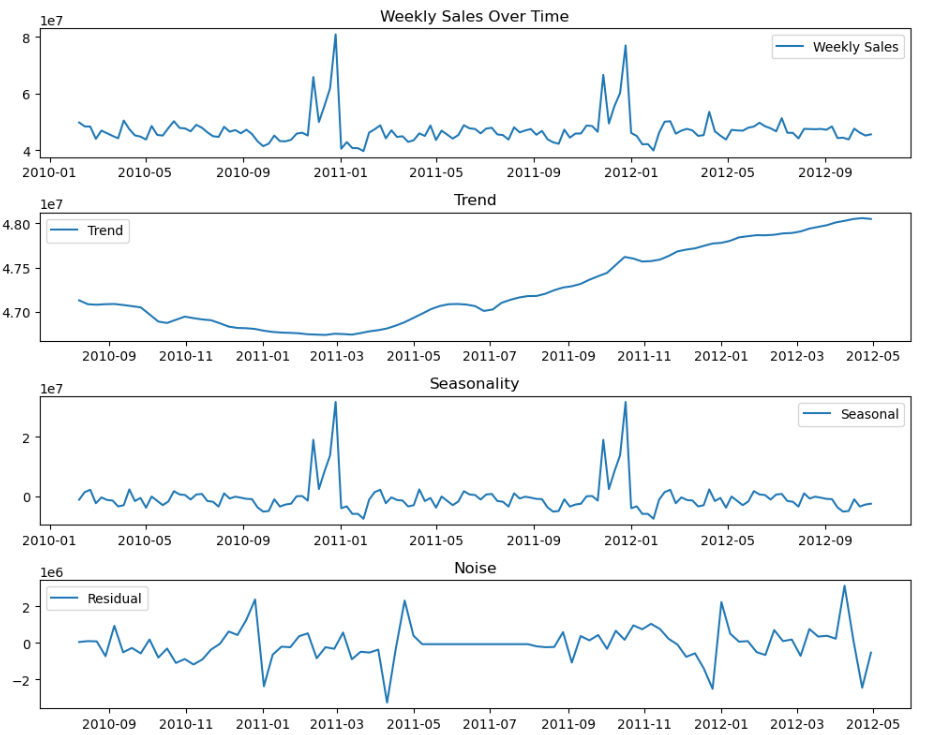
Evaluate each model (AR, ARIMA, SARIMAX, AUTO\_ARIMA) using the performance metrics and criteria, and choose the model that provides the best balance of accuracy and simplicity.

# Inferences from the Project

The model performance, inferences-

**Impact of Time-Based Factors:**

Seasonal trends indicate that certain times of the year, particularly during holidays and special events, lead to spikes in sales, confirming the importance of seasonality in forecasting models.

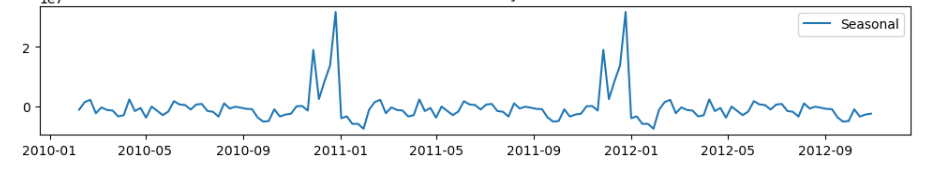


As we can clearly see from seasonality graph (graph 3) there is a spike in December – January month due to holidays over state.

**Effect of Holidays:**

The inclusion of holidays in the dataset highlights a noticeable increase in sales during these periods, validating the hypothesis that holidays drive higher consumer spending.

Promotional activities and strategic marketing during holidays appear to effectively boost sales, emphasizing the need for targeted campaigns during these times.



**Model Performance:**

The ARIMA, SARIMAX, and AUTO\_ARIMA models demonstrate varying degrees of accuracy in forecasting weekly sales, with SARIMAX and AUTO\_ARIMA often outperforming simpler models due to their ability to account for seasonality and exogenous variables

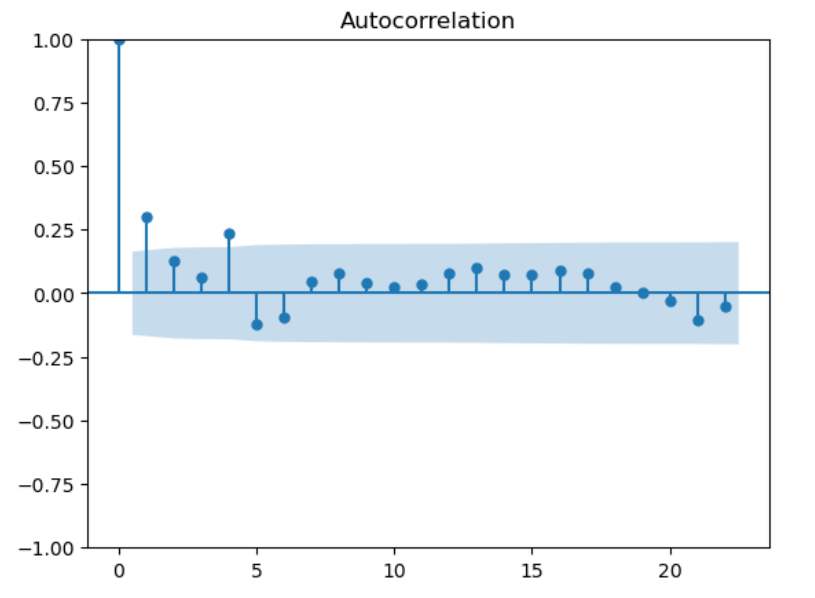
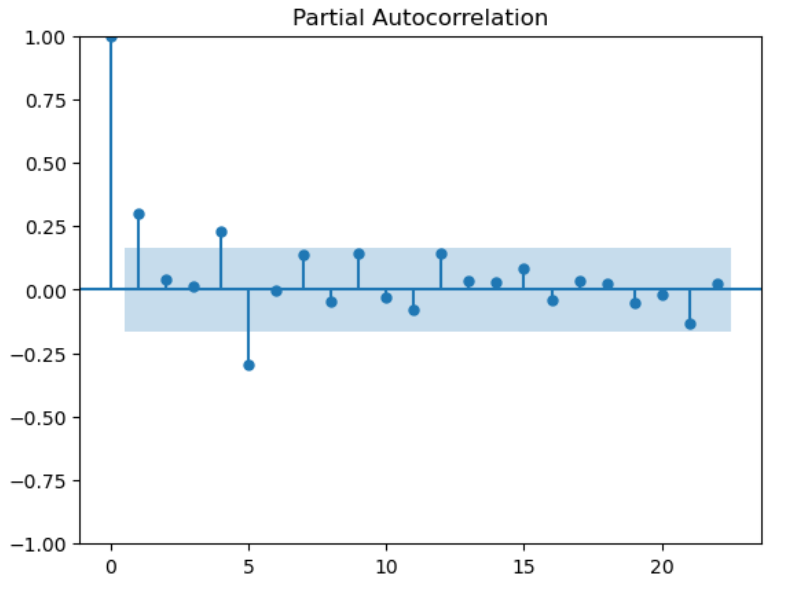
ADFULLER Test: - Dataset has p value much lesser than 0.05 which depicts that data is stationary.

Therefore, no data transformation is performed including boxcox transformation and differencing

**ARIMA: -**

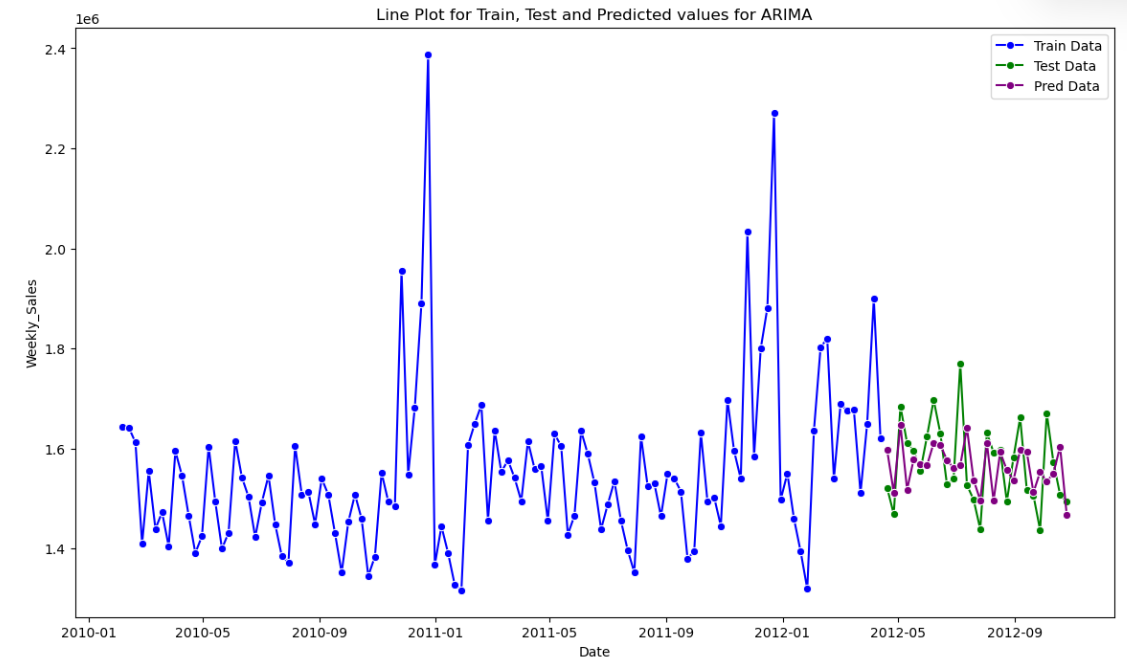
Since dataset has seasonal component, AR model would have given higher RMSE value leading to incorrect sales forecast.

Manual calculation of ACF (q) and PACF (p) shows multiple combination with P = 1,4,5 and Q = 1,4 and D = 0.

The best combination came out to be (1,0,4) having least Root Mean Squared Error of **75547.29** which is still very high but much lesser when compared to RMSE of different combinations.

The Predicted values overlapped for the most part as shown below:

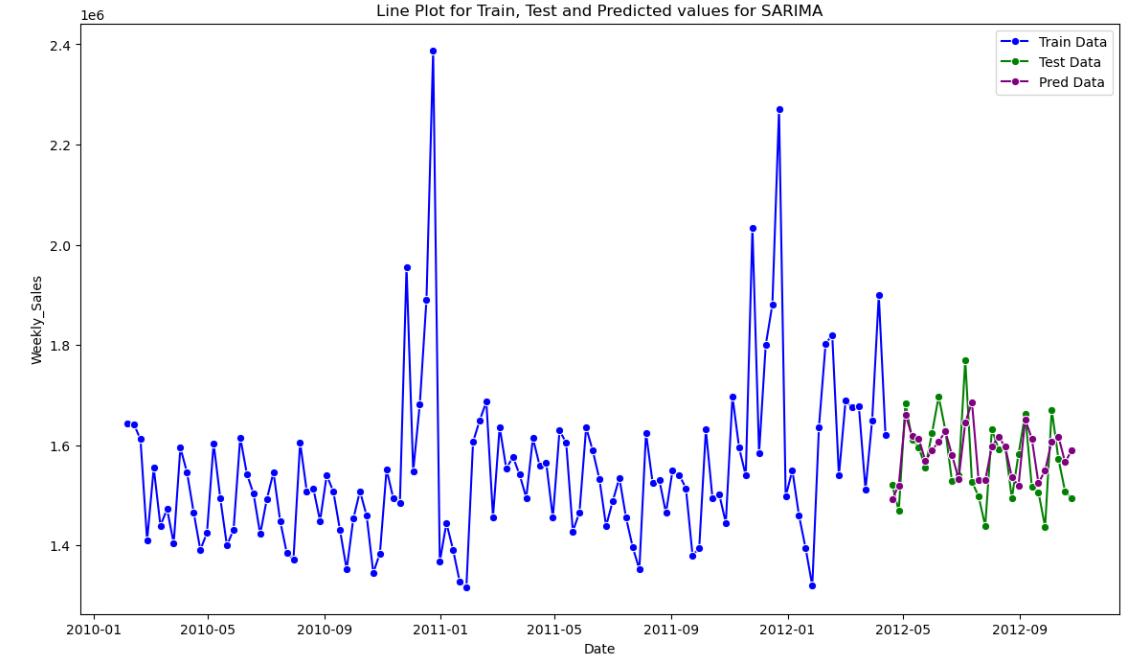


A separate Store\_1 Data Frame has been created just for store 1 data to identify which model has better performance.

Since dataset has seasonality so will try SARIMA for better performance.

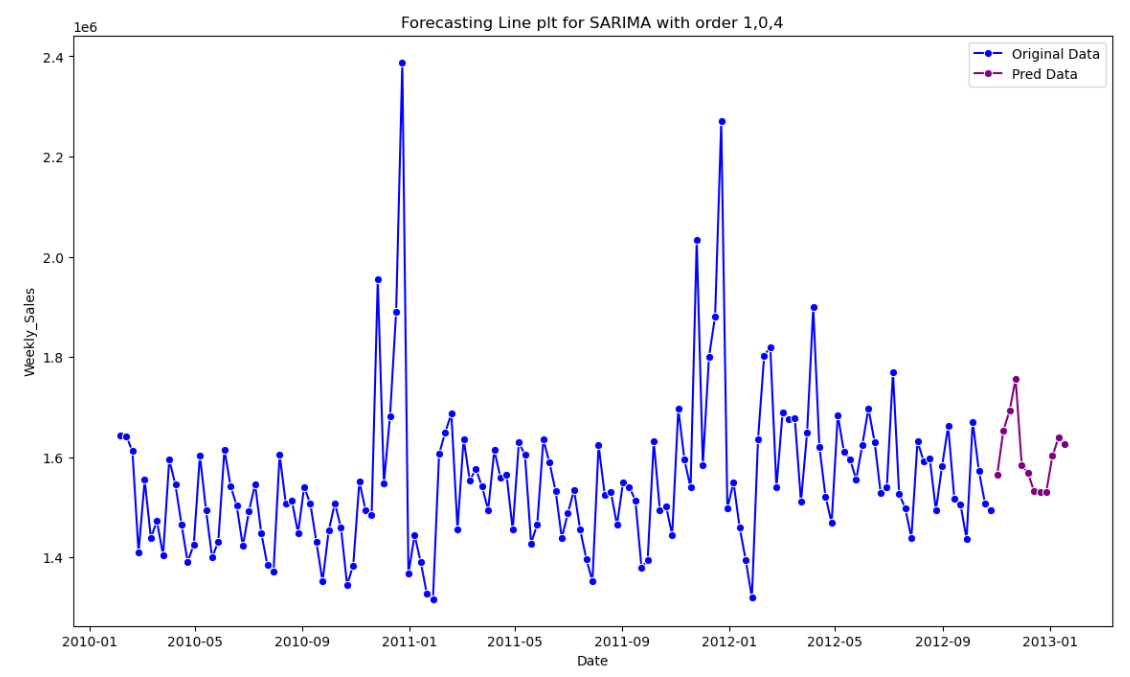
**SARIMA:**

With the same combination of P, D and Q values SARIMA model was able to provide lesser RMSE value accounting to **64200.52** which showed improvement from ARIMA model.



**Forecasting**:

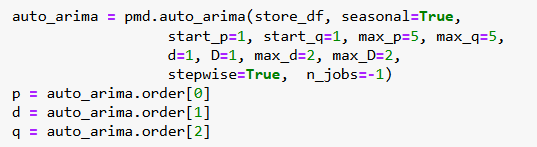
We will forecast the weekly sales for next 12 weeks using the model which we created for SARIMA using forecast function.



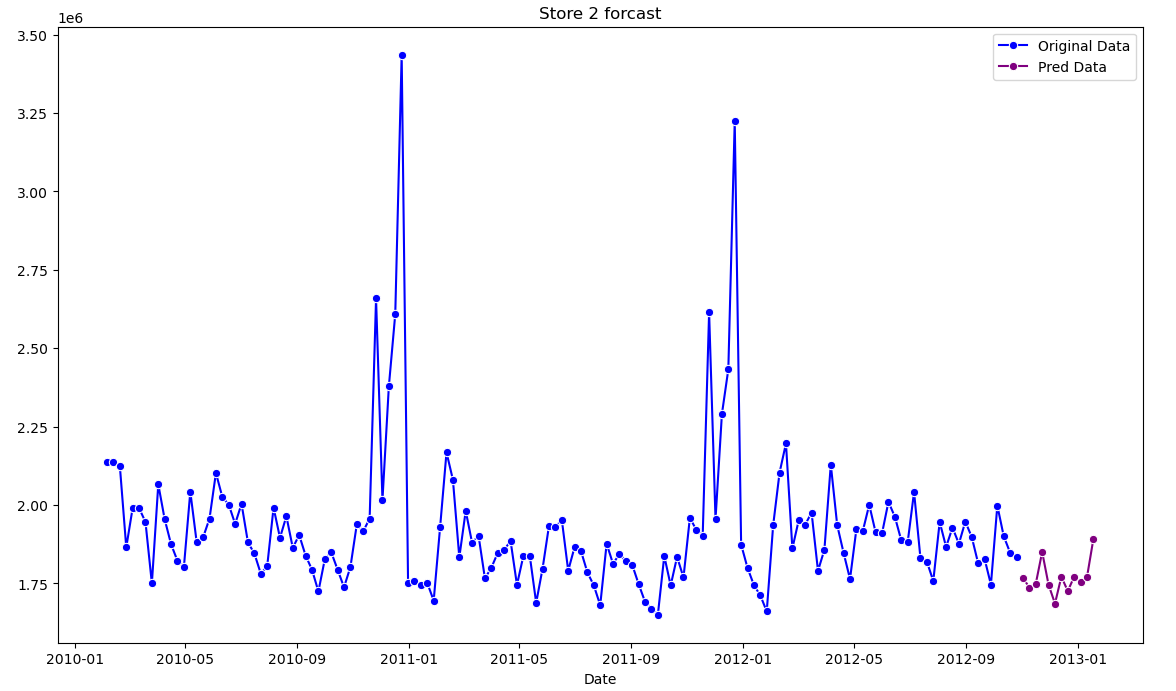
Prediction is shown in purple line and marker and the forecasted value will be shown in generalized function.

Now using **PMDARIMA** we will generalize the p, d, q values for each store and forecast next 12 weeks sales value for each store.

Using AIC, we have determined all the hyperparameters and have used to forecast the weekly values.



Now let’s plot the store 2 forecast in generalized manner.



Forecast for store 2

2012-11-02 1.768679e+06

2012-11-09 1.736760e+06

2012-11-16 1.748760e+06

2012-11-23 1.850558e+06

2012-11-30 1.745936e+06

2012-12-07 1.686048e+06

2012-12-14 1.771647e+06

2012-12-21 1.725878e+06

2012-12-28 1.771756e+06

2013-01-04 1.755400e+06

2013-01-11 1.770553e+06

2013-01-18 1.892122e+06

Like this for all the store forecasts has been done along with the values.

# Future Possibilities

**Incorporation of Additional Data Sources:**

* **Weather Data**: Include weather data to analyse its impact on sales, as certain weather conditions can significantly influence shopping patterns.
* **Economic Indicators**: Use macroeconomic indicators such as GDP growth, inflation rates, and consumer confidence indices to enhance the forecasting models.

**Advanced Modelling Techniques:**

* **Deep Learning Models**: Implement advanced deep learning techniques like Long Short-Term Memory (LSTM) networks or Convolutional Neural Networks (CNNs) for better capturing complex patterns and long-term dependencies in sales data.

**Real-Time Forecasting and Monitoring:**

* **Customer Segmentation**: Perform detailed customer segmentation to understand diverse consumer groups and their buying behaviours. This can help tailor marketing and promotional strategies more effectively.
* **Promotional Impact Analysis**: Analyse the impact of various promotional activities and campaigns on sales to optimize future promotional strategies.

**User-Friendly Visualizations and Dashboards:**

* **Interactive Dashboards**: Develop more interactive and user-friendly dashboards that allow stakeholders to explore data, trends, and forecasts easily.
* **Scenario Analysis Tools**: Incorporate tools for scenario analysis to simulate various business conditions and their potential impact on sales.

# Conclusion

The primary goal of this project was to accurately forecast weekly sales for Walmart stores using time series models such as AR, ARIMA, SARIMAX, and AUTO\_ARIMA. Through a meticulous process of data pre-processing, including checking for null values, duplicates, and ensuring proper data types, we laid a strong foundation for building robust models.

Our analysis revealed several key insights:

* **Time-Based Factors**: Holidays significantly impact sales, with higher consumer turnout observed during these periods. Seasonal trends also play a crucial role in sales patterns.
* **Effect of Holidays**: Incorporating holidays into the models demonstrated a clear spike in sales, emphasizing the importance of strategic promotional activities during these times.

The models, particularly SARIMAX and AUTO\_ARIMA, showcased their ability to capture complex patterns in the data, providing reliable forecasts. These forecasts are invaluable for optimizing inventory, staffing, and marketing strategies, ultimately enhancing operational efficiency and profitability.

References

* <https://otexts.com/fpp2/AR.html>
* <https://www.statsmodels.org/dev/generated/statsmodels.tsa.seasonal.seasonal_decompose.html>
* <https://www.geeksforgeeks.org/sarima-seasonal-autoregressive-integrated-moving-average/>