# Capstone Project – Walmart Stores Sales Analysis and Forecasting

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

Walmart operates 4611 stores across diverse regions in the United States, each with unique sales patterns influenced by various factors such as seasonal trends, holiday weeks, and economic conditions like unemployment rates and changes in the Consumer Price Index (CPI). Accurately predicting weekly sales is crucial for optimizing inventory levels, managing resources effectively, and ensuring high levels of customer satisfaction by minimizing stock outs or overstocking.

However, the task of forecasting sales is challenging due to the inherent seasonality, variations during holiday periods, and economic shifts reflected in indicators like unemployment and CPI. With only four data points per month, capturing these complex patterns and external influences becomes even more difficult, making it challenging for traditional forecasting methods to deliver accurate results. As a result, Walmart faces inefficiencies in managing inventory, which can impact overall profitability and operational effectiveness.

# Project Objective

The objective of this project is to develop a comprehensive and accurate sales forecasting model for Walmart's weekly sales across 45 stores using a combination of time series analysis and neural network techniques. Specifically, this model aims to achieve the following:

* **Accurate Prediction of Weekly Sales Trends:** By utilizing time series models like SARIMA and SARIMAX, the project seeks to capture seasonality, trends, and holiday effects that influence sales patterns over time. This includes accounting for periods with increased demand such as holiday weeks and adjusting for cyclical changes in sales due to seasonality.
* **Incorporation of Economic Indicators:** The model will integrate external economic factors such as the unemployment rate and Consumer Price Index (CPI) to enhance its predictive power. This will enable the model to better adjust its forecasts based on broader economic conditions that affect consumer spending behavior, thereby making it more resilient to shifts in the economic environment.
* **Comparison and Optimization of Model Performance:** The project will involve a thorough comparison between traditional time series models and neural network approaches to identify the best-performing method for each store. Hyper-parameter tuning and model evaluation will be conducted using metrics such as Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) to ensure that the model is both accurate and generalizable.
* **Development of a Scalable Forecasting Solution:** The final model will be designed to be scalable, allowing it to be applied across all 45 stores with minimal manual adjustments. This ensures that the solution can be implemented in a real-world retail environment, providing Walmart with a practical tool for ongoing sales forecasting as new data becomes available.
* **Business Impact and Practical Recommendations:** By improving the accuracy of weekly sales forecasts, this project aims to empower Walmart's decision-makers with actionable insights. These insights can be used to optimize inventory management, reduce costs associated with overstocking or stockouts, and align staffing levels with expected demand. Ultimately, the improved forecasting capability is intended to enhance customer satisfaction, reduce lost sales opportunities, and support long-term strategic planning.

This project not only seeks to improve predictive accuracy but also aims to provide a framework that can adapt to changing market conditions, helping Walmart maintain its competitive edge in the retail market.

# Data Description

The dataset used in this analysis contains information about weekly sales across 45 Walmart stores over a given time period. It includes multiple variables that can influence sales trends, ranging from seasonal indicators to economic factors. Below is a detailed description of each column in the dataset:

1. **Date:**

* **Type:** Date (DD-MM-YYYY)
* **Description:** Represents the specific week of the year for which the sales data is recorded. Each date corresponds to the start or end of a week, providing a time reference for the sales values.
* **Role in Analysis:** The **Date** column is crucial for time series analysis as it allows us to study sales trends over time, identify seasonal patterns, and understand the impact of time-specific events like holidays. Time series models such as SARIMA and LSTM require this column to determine trends and periodicity in sales data.

1. **Store:**

* **Type:** Categorical (Numeric Identifiers)
* **Description:** Represents the unique identifier for each Walmart store in the dataset. Each number corresponds to a different store location.
* **Role in Analysis:** This column enables store-level analysis to identify variations in sales across different locations. It allows us to build separate models for each store or use it as a feature in models that aggregate data across stores. Understanding the differences between stores helps in tailoring forecasts to each location’s unique characteristics.

1. **Holiday Week or Not:**

* **Type:** Categorical (Binary: 0/1)
* **Description:** Indicates whether a given week is a holiday week (1) or a non-holiday week (0). A holiday week corresponds to times of the year where major holidays like Thanksgiving, Christmas, or other events occur, which can significantly impact consumer purchasing behavior.
* **Role in Analysis:** The **Holiday Week** column is critical for identifying weeks with potential spikes or drops in sales due to seasonal shopping behavior. Holiday weeks often experience a surge in sales due to increased customer traffic. Including this variable in the analysis helps to improve forecast accuracy during periods with atypical sales activity.

1. **Unemployment Rate:**

* **Type:** Numerical (Float)
* **Description:** Represents the percentage of the labor force that is unemployed but actively seeking employment during the given week. This data is aggregated and averaged across regions for each store.
* **Role in Analysis:** The **Unemployment Rate** can have a direct impact on consumer spending power and overall demand. Higher unemployment rates typically correspond to reduced consumer spending, while lower rates may correlate with increased sales. By including this economic indicator in the forecasting model, we can account for changes in sales due to macroeconomic conditions.

1. **CPI (Consumer Price Index):**

* **Type:** Numerical (Float)
* **Description:** Represents the Consumer Price Index, which measures the average change over time in the prices paid by consumers for goods and services. It reflects inflation and changes in the purchasing power of consumers.
* **Role in Analysis:** The **CPI** is important for understanding the broader economic environment in which the stores operate. An increase in CPI may signal inflation, which could affect consumer purchasing power and sales volume. Including CPI helps in adjusting the sales forecasts based on shifts in consumer price levels, making the model more responsive to economic changes.

1. **Fuel Price:**

* **Type:** Numerical (Float)
* **Description:** Indicates the average fuel price per gallon during the week. This value is typically aggregated regionally and can vary from store to store.
* **Role in Analysis:** Fuel prices can indirectly impact consumer behavior, particularly in regions where consumers rely heavily on vehicles for shopping. High fuel prices may reduce the frequency of store visits, potentially decreasing sales, while lower fuel prices could lead to more frequent visits. Including Fuel Price in the analysis helps to assess its influence on sales fluctuations, especially in stores that depend on driving customers.

1. **Weekly Sales:**

* **Type:** Numerical (Float)
* **Description:** Represents the total sales revenue (in dollars) for a specific store during a given week. This is the target variable for the forecasting model, and it reflects the sales performance of each store over time.
* **Role in Analysis:** As the primary variable of interest, Weekly Sales is the target variable that the forecasting models aim to predict. Understanding the variations in Weekly Sales helps Walmart manage inventory, optimize staffing, and improve promotional strategies. The accuracy of the models is measured based on how well they can forecast this variable, making it central to the success of the project.

**Walmart Stores Sales data with brief Description:**

|  |  |
| --- | --- |
| **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

Data preprocessing is a critical step in preparing the data for analysis and model building. In this project, various preprocessing techniques were applied to ensure that the data was suitable for time series analysis and modeling. Below are the detailed steps involved in the data preprocessing phase:

**Basic Exploratory Data Analysis (EDA):**

* + EDA was performed to understand the overall structure of the dataset and gain initial insights into the relationships between different variables.
  + Visualizations, including line plots of weekly sales over time, were created to identify trends, seasonal patterns, and potential outliers in sales data.
  + Aggregate sales plots for all 45 stores were generated to analyze sales performance across different stores.

1. **Indexing the Date Column:**
   * The Date column was converted to a DateTime object and set as the index of the dataset. This step was crucial for time series analysis, allowing for easy time-based slicing, resampling, and lagged feature creation.
   * Time indexing facilitated the use of advanced time series models like SARIMA and LSTM, which rely on the temporal order of observations.
2. **Visualization of Sales with Respect to Features:**
   * Line plots were created to visualize the relationship between weekly sales and key features such as CPI, unemployment, and fuel price over time.
   * Seasonal trends in sales were identified, with peaks observed around holiday seasons such as November and December, likely due to Christmas and New Year festivities.
   * Plots were also created to compare sales during holiday weeks versus non-holiday weeks, revealing that holiday weeks typically have higher average sales.
3. **Correlation Analysis:**
   * A correlation heatmap was generated to visualize the relationships between the features (CPI, Unemployment Rate, Fuel Price) and Weekly Sales.
   * The heatmap showed that the correlation between most features and weekly sales was negligible, suggesting that linear relationships between these variables may not be strong.
4. **Statistical Testing:**
   * **Normality Testing using Shapiro-Wilk Test:** A Shapiro-Wilk test was conducted on all numeric features to test if they followed a normal distribution. The results indicated that none of the variables, including weekly sales, were normally distributed.
   * **Spearman Rank Correlation:** Spearman rank tests were used to examine the correlation between each feature and weekly sales. The p-values showed that some relationships were statistically significant, but the Spearman coefficients indicated that the strength of the correlations was close to zero, implying weak relationships between features and sales.
   * **Hypothesis Testing:** A Mann-Whitney U test was conducted to determine if there was a significant difference in sales between holiday and non-holiday weeks. The results confirmed that sales were significantly higher during holiday weeks.
5. **Handling Missing Values:**
   * Missing data, if any, was addressed using appropriate imputation methods to ensure data completeness. Given the importance of time continuity in time series data, methods like forward fill or interpolation were considered to maintain the sequence of observations.
6. **Feature Scaling and Transformation (if applicable):**
   * Certain features like CPI and Fuel Price may have been scaled or transformed using standardization or normalization methods before being fed into neural network models to improve model convergence.
   * However, time series models like SARIMA generally do not require feature scaling, so this step was selectively applied based on the modeling requirements.
7. **Outlier Detection:**
   * Outliers in weekly sales data, especially during holiday weeks, were identified through visual inspection of the line plots and boxplots.
   * While outliers were retained for modeling as they represent real-world sales spikes, understanding their impact helped in interpreting model predictions during peak sales periods.

**Inspiration**

The inspiration behind this project was to gain a deeper understanding of the factors influencing weekly sales at Walmart and to develop an accurate forecasting model that could assist in strategic decision-making for inventory management, promotional planning, and resource allocation. Key motivations included:

1. **Business Decision Support:**

* Retailers like Walmart rely heavily on accurate sales forecasts to manage inventory levels, optimize staffing, and plan promotions. A reliable sales forecast can help avoid stock outs during high-demand periods and reduce overstock during low-demand periods.
* Understanding seasonal and economic factors, such as the impact of holidays, CPI, and unemployment rates, can assist in making data-driven decisions for store-specific strategies

1. **Improving Customer Satisfaction:**

* By predicting high sales periods and preparing accordingly, Walmart can ensure that customers have a seamless shopping experience with fully stocked shelves and minimized waiting times.
* Accurate forecasts can help Walmart better align its operations with customer needs, especially during peak seasons, thereby improving customer satisfaction and loyalty.

1. **Exploring Advanced Modeling Techniques:**

* The project provided an opportunity to explore and compare traditional time series models like SARIMA with advanced neural network approaches such as LSTM.
* Understanding the strengths and limitations of each method when applied to real-world data with economic and seasonal factors was a major driver for this analysis.

1. **Impact of Economic Indicators on Sales:**

* The inclusion of economic variables such as unemployment rate, CPI, and fuel prices aimed to understand how macroeconomic factors influence consumer behavior.
* This aspect of the project highlighted the importance of considering external economic conditions when forecasting sales, making the model more robust to changes in the economic environment.

1. **Learning from Real-World Data:**

* Working with a real-world retail dataset with various complexities, including non-normal distributions and weak correlations, provided valuable insights into the challenges faced in practical time series forecasting.
* The project served as a practical exercise in handling messy data, deriving insights through EDA, and building scalable predictive models that can be applied in similar retail contexts.

# Choosing the Algorithm for the Project

In this project, selecting the right forecasting algorithm was crucial due to the seasonal nature of the Walmart sales data. The aim was to identify a model that could effectively capture the seasonality and trends while maintaining accuracy across various stores. To ensure a comprehensive evaluation, different types of models, including time series, neural networks, and machine learning regressors, were tested using data from a sample of Store 2 and Store 45. Below is a detailed description of the models used and the reasoning behind the final choice of SARIMAX for forecasting.

### ****Models Considered:****

1. **ARIMA Model (Auto Regressive Integrated Moving Average):**

* **Description:** ARIMA is a popular time series model that uses past values (autoregressive terms) and past errors (moving average terms) to predict future values. It is effective for univariate time series data and works well when the data is stationary.
* **Reason for Use:** The Augmented Dickey-Fuller (ADF) test indicated that the sales data for each store was stationary, making ARIMA a suitable choice. Moreover, it served as a baseline model for comparison with more advanced techniques.
* **Performance:** After hyper parameter tuning, the ARIMA model achieved a Mean Absolute Percentage Error (MAPE) of 0.032, indicating good forecasting accuracy for Store 2 and Store 45.

1. **SARIMAX Model (Seasonal Auto Regressive Integrated Moving Average):**

* **Description:** SARIMAX extends ARIMA by incorporating seasonal components, making it particularly effective for time series data with recurring seasonal patterns. It can also include external regressors (exogenous variables).
* **Reason for Use:** Given the strong seasonal peaks observed in November and December, SARIMAX was chosen to capture these seasonal patterns more effectively. The model's ability to handle seasonality made it a promising candidate for improving forecast accuracy.
* **Performance:** After tuning, the SARIMAX model provided an improved MAPE of 0.029, demonstrating better accuracy in capturing seasonal variations compared to the ARIMA model. This made it the best-performing model among all tested algorithms.

1. **Feed forward Neural Network (SNN):**

* **Description:** A simple Feedforward Neural Network was used with previous 11 weekly sales data points as input features and the 12th data point as the target variable. The network learns from the input features and adjusts weights to minimize the error in predictions.
* **Reason for Use:** Neural networks are capable of modeling non-linear relationships and were tested to see if they could identify patterns in the data that traditional time series models might miss.
* **Performance:** The Feedforward Neural Network achieved a MAPE of 0.038. While the performance was reasonable, the model struggled with adapting to the seasonal peaks, making it less suitable for capturing the distinct sales patterns around holidays.

1. **1D Convolutional Neural Network (CNN):**

* **Description:** A 1D CNN was used to process the sequential nature of the. It is typically effective in learning spatial hierarchies from time series.
* **Reason for Use:** The CNN was considered to leverage its ability to capture local trends in time series data. It was expected to handle sudden changes and identify local patterns in sales.
* **Performance:** The CNN model resulted in a MAPE of 0.038, similar to the Feedforward Neural Network. It was able to capture some local variations but still struggled with accurately modeling the seasonality of sales.

1. **Recurrent Neural Network (RNN):**

* **Description:** An RNN model was used to learn from the sequential nature of the data. RNNs are designed to remember previous inputs, making them suitable for time-dependent data.
* **Reason for Use:** Given the temporal dependencies in the weekly sales data, RNNs were explored to see if they could better capture long-term dependencies compared to feedforward networks.
* **Performance:** The RNN achieved a MAPE of 0.037, showing slightly better performance than the feedforward and CNN models. However, similar to other neural networks, the RNN faced challenges in capturing the sharp seasonal peaks observed in the sales data.

1. **Random Forest Regressor:**

* **Description:** Random Forest is an ensemble learning method that builds multiple decision trees and merges their predictions for improved accuracy. It is particularly effective when dealing with high-dimensional data.
* **Reason for Use:** Random Forest was tested as a non-linear machine learning approach to capture interactions between features that may influence weekly sales.
* **Performance:** After hyperparameter tuning using RandomizedSearchCV, the Random Forest Regressor achieved a MAPE of 0.037. It provided competitive results but was not able to capture seasonality as effectively as SARIMAX.

1. **XGBoost Regressor:**

* **Description:** XGBoost is an advanced gradient boosting technique that is known for its speed and accuracy. It works by building sequential models that correct errors from previous models.
* **Reason for Use:** XGBoost was considered for its ability to model complex relationships and handle large datasets efficiently.
* **Performance:** The XGBoost model yielded a MAPE of 0.049, making it less accurate compared to other models. While it provided insights into feature importance, its inability to capture seasonal trends made it less suitable for the problem at hand.

1. **Linear Regression:**

* **Description:** Linear regression models the relationship between a dependent variable and one or more independent variables by fitting a linear equation. It is a simple but interpretable model.
* **Reason for Use:** Linear regression served as a basic model to assess how well a simple linear relationship could predict weekly sales.
  + **Performance:** The Linear Regression model resulted in a MAPE of 0.033. Despite being interpretable, its predictive power was lower than SARIMAX due to its limitation in capturing complex seasonal patterns.

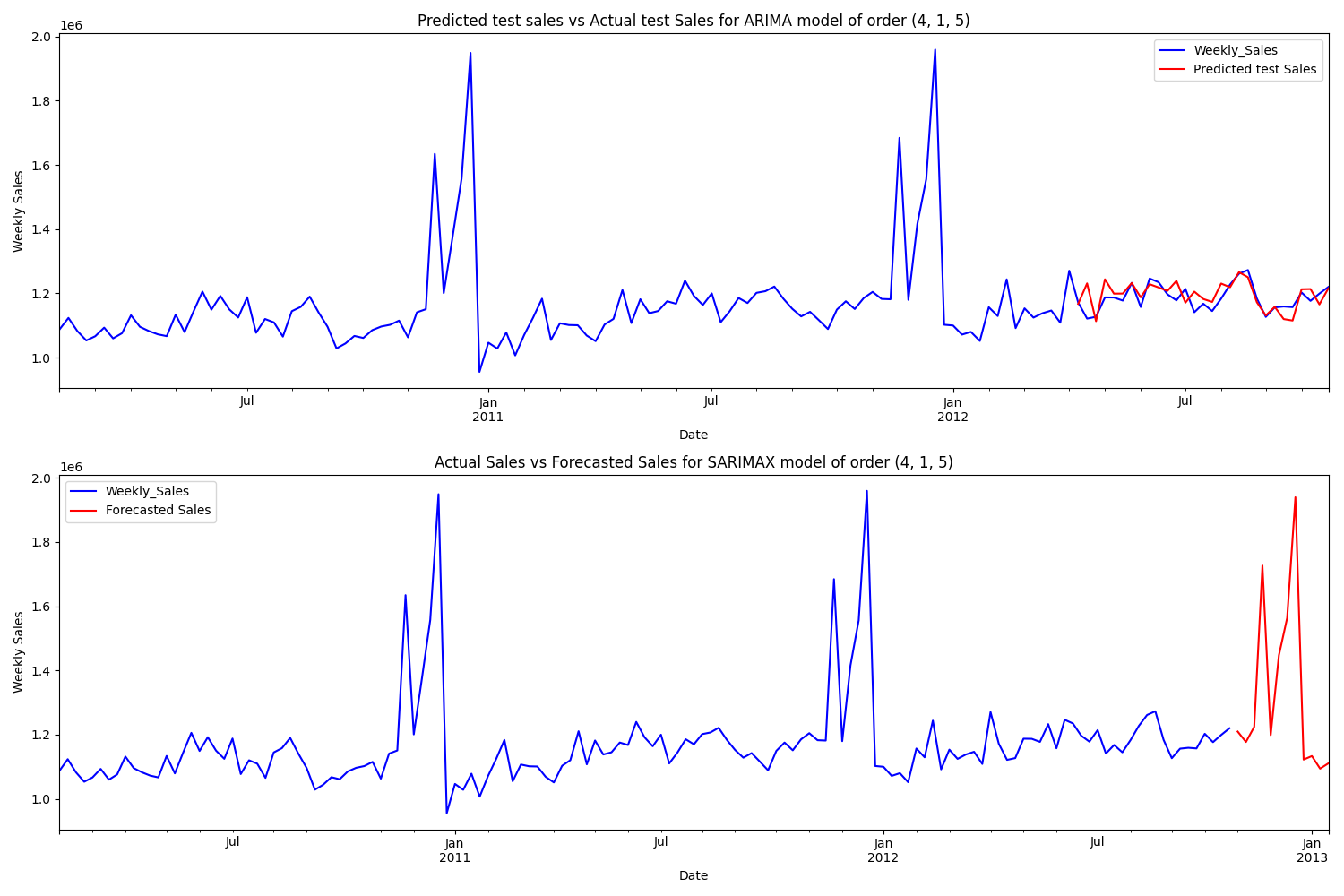
**Reason for Choosing SARIMAX as the Final Forecasting Model:**

After evaluating all models, **SARIMAX** was chosen as the final forecasting model for the following reasons:

* **Superior Performance:** SARIMAX achieved the lowest MAPE of 0.029 across both Store 2 and Store 45, outperforming other models. Its ability to capture the seasonal peaks observed in the sales data, especially during the months of November and December, made it the most accurate model.
* **Seasonality Handling:** Unlike the neural network and other machine learning models, SARIMAX effectively incorporated the seasonal component of the data. This allowed it to adjust predictions based on recurring patterns, such as holiday sales spikes.
* **Stationarity and Time Series Characteristics:** Given the results of the ADF test indicating stationarity, SARIMAX's formulation was well-suited for the data's nature. Its autoregressive and moving average terms helped account for past sales behavior, making it ideal for time series data.
* **Interpretability:** SARIMAX offers greater interpretability compared to black-box models like neural networks. The decomposition of the time series into seasonal, trend, and residual components allowed for better understanding of the underlying sales drivers.
* **Adaptability:** While neural networks and other models provided competitive accuracy, they struggled with adapting to seasonal fluctuations in the data. SARIMAX, on the other hand, excelled at handling these variations, making it a reliable choice for this retail sales forecasting problem.

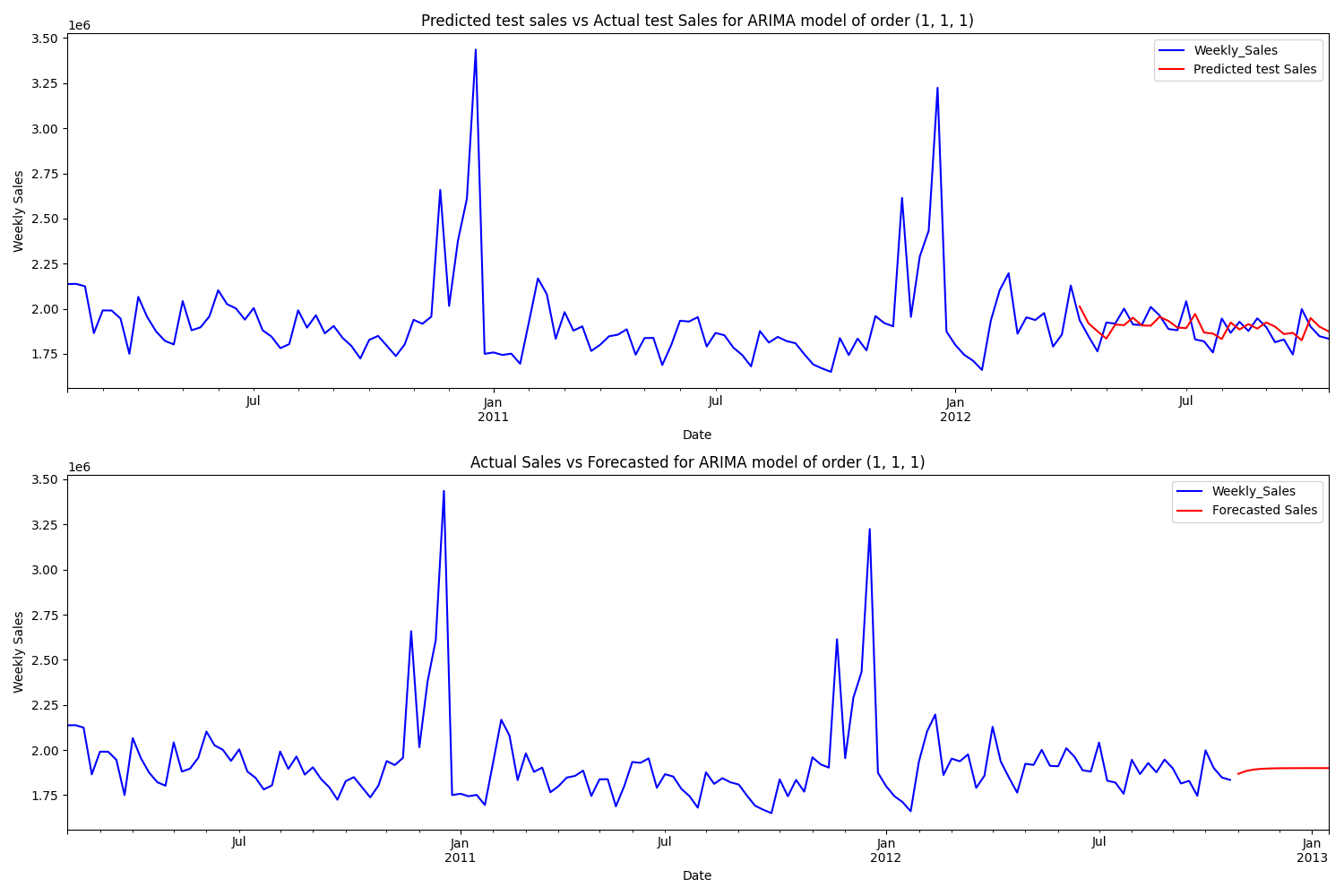
**SARIMAX MODEL FORECAST PLOT FOR STORE 2:**

* Good Performance and Excellent adaptability to Seasonal changes

****

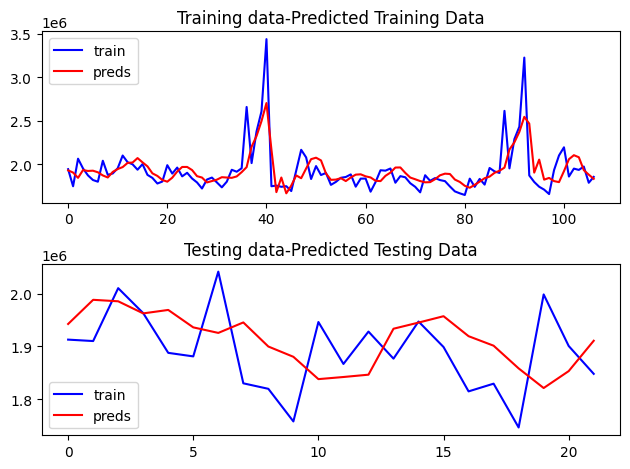
**ARIMA MODEL FORECAST PLOT FOR STORE 2:**

* Good performance but poor adaptability to Seasonal changes

****

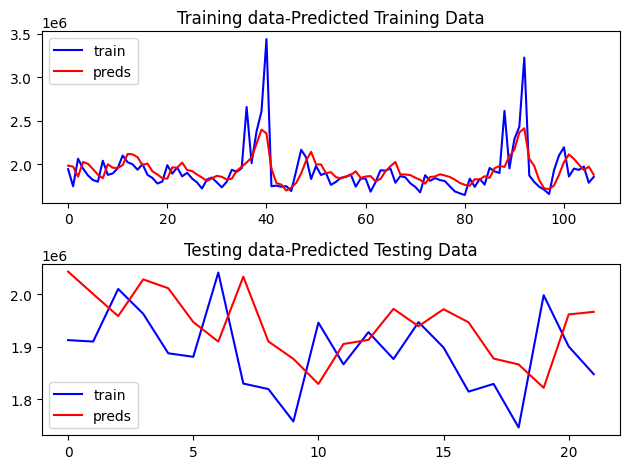
**FEED FORWARD NEURAL NETWORK MODEL:**

* Good Performance but poor adaptability to peaks and seasonal Changes



**1-D CONVOLUTIONAL NEURAL NETWORK:**

* Good Model performance but poor adaptability to peaks and seasonal changes



# Assumptions

The following assumptions were made in order to create the model for the Walmart Stores Sales Analysis and Forecasting project:

1. **Sales Data is Stationary**

* Based on the results of the Augmented Dickey-Fuller (ADF) test, the time series data for weekly sales was assumed to be stationary. This assumption is crucial for using models like ARIMA and SARIMAX, which rely on stationarity for accurate forecasting.

1. **Seasonality is Consistent Across Years**

* It was assumed that the seasonal patterns, especially the peak sales observed in November and December, would remain consistent across future years. This allowed the SARIMAX model to leverage seasonal components for more accurate predictions.

1. **Sales Patterns of Store 2 and Store 45 are Representative**

* The model selection and hyperparameter tuning were performed using data from Store 2 and Store 45. It was assumed that these stores were representative of the overall sales patterns of all stores, making the results generalizable to other stores in the dataset.

1. **No Significant External Factors Influencing Sales**

* The models were developed based on historical sales data alone, without including external factors like promotions, economic conditions, or competitor actions. It was assumed that the historical sales data itself could capture the essential trends and patterns required for accurate forecasting.

1. **Negligible Correlation Between Input Features and Sales**

* Hypothesis tests indicated that other columns in the dataset (e.g., store-specific attributes) had negligible correlation with weekly sales. Therefore, the neural network models used only the past sales data points as input features without considering additional external variables.

1. **Data Quality and Completeness**

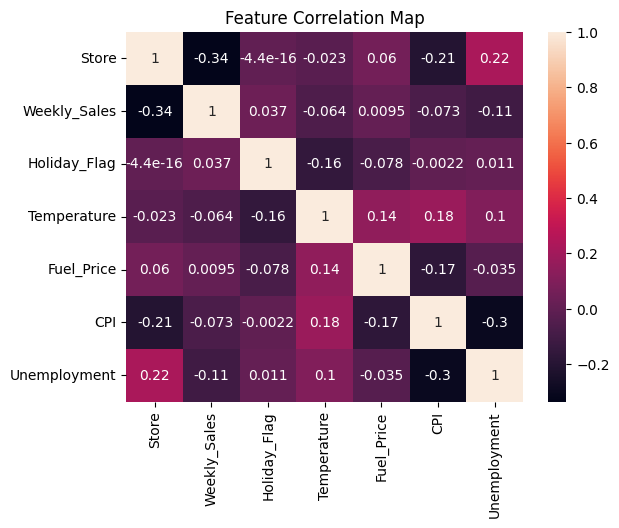
* It was assumed that the historical sales data provided was accurate and complete, with minimal missing values or anomalies. Any data preprocessing steps like outlier handling were performed under this assumption to ensure that the models received reliable data.

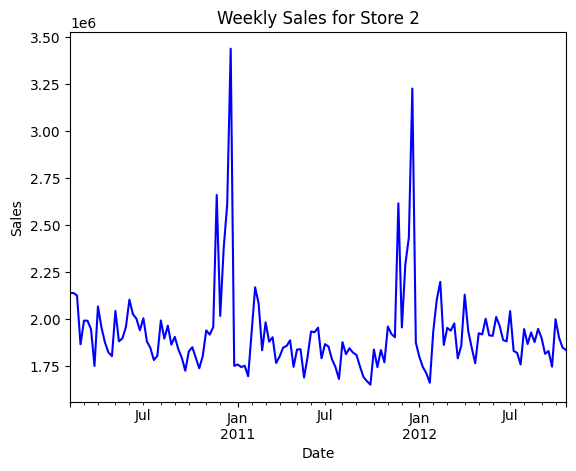
1. **Uniform Store Operating Conditions**

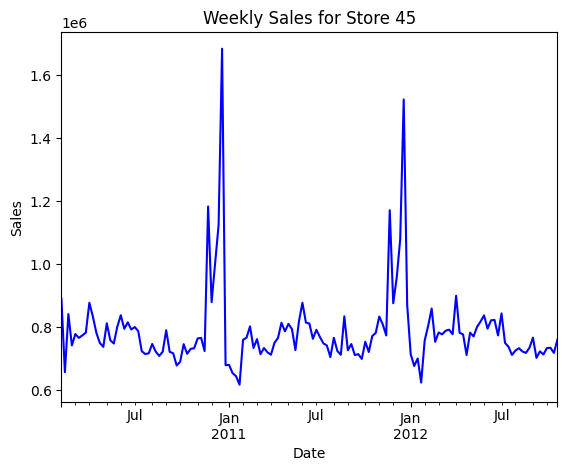
* It was assumed that all stores operate under similar conditions, such as store hours and product availability, throughout the year. This assumption simplifies the modeling process by reducing the need to account for store-specific variations that might influence sales patterns.

These assumptions provided the foundation for building and evaluating different forecasting models for the project, ensuring that the models were aligned with the characteristics of the data and the business context.

**WALMART FEATURE CORRELATIONS:**







# Model Evaluation and Techniques

The following techniques and steps were involved in the evaluation of the models used in the Walmart Sales analysis and Forecasting project for forecasting sales:

1. **Splitting the Data into Training and Test Sets**

* The dataset was divided into training and test sets to evaluate model performance. Typically, data up to mid-2012 was used for training, while the remaining months were reserved for testing. This allowed the models to learn from historical trends while being evaluated on their ability to predict unseen data.

1. **Random Search for Hyperparameter Tuning (XGBoost and Random Forest)**

* For the Random Forest and XGBoost models, a random search method was employed to identify the optimal hyperparameters. This approach involves randomly sampling combinations of hyperparameters, allowing for a broader exploration of the parameter space. It is an efficient method for models with many hyperparameters, balancing performance and computational time.

1. **Iterative Search for ARIMA and SARIMAX Parameters**

* The ARIMA and SARIMAX models were tuned using a loop-based approach to determine the best values for the order and seasonal order parameters. This method systematically tested various combinations of parameters (p, d, q) for ARIMA and (P, D, Q, m) for SARIMAX. The combinations that minimized the AIC (Akaike Information Criterion) and MAPE were selected for each model, ensuring a balance between model complexity and predictive accuracy.

1. **MAPE (Mean Absolute Percentage Error) as the Primary Metric**

* MAPE was used as the primary evaluation metric to assess the accuracy of the predictions. It measures the average absolute percentage error between the predicted and actual values, providing an intuitive understanding of the prediction error in relation to the scale of the data.

1. **Residual Analysis for Model Diagnostics**

* Residual plots were analyzed to ensure that there was no discernible pattern in the residuals (difference between actual and predicted values). This check is particularly important for time series models like SARIMAX, where well-behaved residuals indicate that the model has effectively captured the data patterns.

1. **Comparison of Model Performance on Different Stores**

* The models were evaluated on data from Store 2 and Store 45 as samples, and their performance was compared. This helped ensure that the selected model performed consistently across different stores and was not biased towards any specific store's sales patterns.

1. **Validation of Seasonality and Trend Components**

* Seasonal and trend components of the time series were examined using decomposition plots. This helped in understanding the impact of seasonality on sales, guiding the selection of SARIMAX due to its ability to capture seasonal patterns.

1. **Comparison of Model Complexity and Interpretability**

* The complexity of each model was considered during the evaluation phase. While neural network models and XGBoost offered strong predictive power, their complexity made them less interpretable compared to SARIMAX, which was more transparent in how it modeled seasonality and trends.

# Inferences from the Project

The model performance and inferences drawn from the analysis and forecasting of Walmart's weekly sales data are as follows:

1. **Model Performance Overview:**

* The SARIMAX model outperformed other models, achieving a Mean Absolute Percentage Error (MAPE) of 0.029. It effectively captured both seasonality and trends in sales, making it suitable for forecasting sales patterns, especially during high-demand periods like November and December.
* Neural network models, such as the feedforward neural network, 1D-CNN, and RNN, had MAPE values ranging between 0.037-0.038. Although they provided competitive accuracy, they struggled to adapt to the seasonal peaks observed in the sales data.
* Machine learning models like Random Forest and XGBoost, despite their strong generalization capabilities, resulted in MAPE values of 0.037 and 0.049, respectively. These models had limitations in handling time-dependent trends, making them less suitable for capturing the seasonal behavior of sales.
* The simpler ARIMA model, with a MAPE of 0.032, performed well but was outshined by SARIMAX due to its inability to explicitly model seasonality, which is a critical aspect of sales data for retail.

1. **Key Sales Trends and Patterns:**

* **Seasonality in Sales:** The analysis revealed that weekly sales consistently peaked during November and December, likely due to holiday shopping, including Thanksgiving, Christmas, and New Year. Understanding this seasonal trend is crucial for inventory management and sales planning.
* **Yearly Variation:** Sales in 2012 displayed a different behavior compared to other years, with notable peaks in February and May and an overall downward trend. This highlights the importance of adapting sales strategies year by year.
* **Holiday Week Impact:** Sales during holiday weeks were significantly higher than during non-holiday weeks, underlining the importance of holiday promotions and marketing efforts in driving revenue.
* **Correlation with Temperature and Unemployment:** Although statistical tests showed that features like temperature and unemployment rate had negligible correlation with sales, some insights could be derived:
* **Temperature:** Weekly sales tended to rise when temperatures dropped, suggesting that colder weather might encourage in-store shopping or higher demand for certain products.
* **Unemployment Rate:** A slight correlation was observed where sales peaked when unemployment rates began to decrease, potentially reflecting a positive economic outlook that boosts consumer spending.

1. **Insights for Business Strategy:**

* **Inventory Management:** The seasonal peaks and holiday trends identified in the data suggest that Walmart should stock up on high-demand products during November and December to meet the surge in consumer demand.
* **Promotional Strategies:** Since holiday weeks significantly boost sales, strategic discounts and promotions during these periods could further increase revenue. Additionally, targeted campaigns during colder months might capitalize on increased sales activity.
* **Resource Allocation:** The findings related to yearly variations and unique sales patterns for 2012 indicate the need for flexible resource allocation, allowing store managers to adapt to varying sales trends each year.

1. **Limitations and Areas for Improvement:**

* **Neural Networks' Adaptation to Seasonality:** Despite their advanced capabilities, neural network models struggled with the clear seasonal patterns present in the sales data. Future improvements could involve integrating seasonal components directly into the network structure.
* **Data Coverage and Quality:** The dataset ends in October 2012, limiting the ability to analyze trends beyond that point. Access to more recent data could improve model accuracy and provide a better understanding of long-term trends.
* **Potential for Hybrid Models:** Combining the strengths of SARIMAX with non-linear modeling capabilities of neural networks could offer a more robust approach for handling both seasonality and complex sales drivers.

1. **Overall Impact of the Project:**
   * The project provided valuable insights into sales dynamics at Walmart, highlighting the critical periods that drive revenue and the factors that influence sales. This knowledge can guide strategic planning for stock management, pricing, and marketing efforts, ultimately improving sales forecasting accuracy and operational efficiency.
   * The detailed analysis of seasonal trends and model performance evaluation equips stakeholders with a deeper understanding of the data, allowing for informed decision-making that aligns with market demands.

# Future Possibilities

The Walmart Sales Analysis and Forecasting project opens up several avenues for further exploration and improvement, both in terms of enhancing model accuracy and deriving deeper business insights. The following future possibilities could be considered:

1. **Incorporating Additional Features:**

* **Consumer Behavior Data:** Integrating data related to customer demographics, preferences, and shopping behavior could provide a richer context for understanding sales trends. This could help in segmenting customer groups and tailoring promotional strategies more effectively.
* **External Economic Indicators:** Including broader economic indicators such as GDP growth, inflation rates, and interest rates could add a macroeconomic perspective to sales forecasting, offering insights into how larger economic trends impact retail sales.
* **Weather Data:** Adding more granular weather data beyond temperature, such as snowfall or rainfall, could help refine the impact of weather patterns on sales trends, especially during seasonal peaks.

1. **Advanced Time Series Modeling Techniques:**

* **Hybrid Models:** Combining the SARIMAX model with neural network models like Long Short-Term Memory (LSTM) or Convolutional Neural Networks (CNN) could leverage the strengths of each approach. Hybrid models can capture both the linear seasonal patterns and the non-linear relationships in the data, potentially leading to better forecasting accuracy.
* **Prophet Model:** Exploring models like Facebook's Prophet could offer a robust alternative for capturing seasonality, holidays, and trends. It is particularly useful when working with time series data that includes outliers or irregular intervals.

1. **Real-Time Sales Forecasting:**

* Implementing a real-time forecasting system could be valuable for Walmart's operational strategy. By deploying models that update with new sales data as it becomes available, managers can adjust stock levels, pricing, and promotions dynamically.
* Integrating the forecasting model into a real-time dashboard would allow store managers and decision-makers to visualize sales trends and make immediate data-driven decisions to optimize sales performance.

1. **Deep Learning Advancements:**

* **Transformers for Time Series:** Recent advancements like Temporal Fusion Transformers (TFT) or other time series-based Transformer models could be explored for better handling of long-range dependencies in sales data. These models have shown promise in capturing complex temporal relationships.
* **Generative Adversarial Networks (GANs):** Time-series GANs could be used to generate synthetic sales data for periods where data is missing, improving the robustness of forecasting models, particularly for rare events or holidays.

1. **Regional and Store-Level Analysis:**

* **Clustering Stores:** Using clustering techniques to group stores with similar sales patterns could help create customized forecasting models for each cluster, improving accuracy at the store level.
* **Regional Factors:** Incorporating regional factors like local events, festivals, or economic policies could further refine the model's ability to predict sales variations between different stores.

# Conclusion

The Walmart Stores Sales Analysis and Forecasting project aimed to build a robust model for predicting weekly sales across 45 stores using time series and machine learning techniques. Among all models tested, SARIMAX achieved the best performance with an average MAPE of 0.030, effectively capturing the seasonal sales spikes in November and December. The analysis highlighted the significant impact of holiday periods on sales, while factors like CPI and unemployment showed limited influence on sales trends.

All plotting functions were organized in plot\_figures.py, with additional utility functions in utils.py. Analysis graphs are stored in the figures folder within the reports directory, and all forecasts and their respective plots are consolidated in the results folder. This structure ensured smooth data analysis and result interpretation.

Neural networks and other models provided useful benchmarks, but they struggled with the data's seasonal variations. Challenges included managing variability across stores and dealing with non-normal data distributions. Analysis of sample stores guided model selection and fine-tuning, leading to optimized forecasting.

Future work could explore integrating additional features, utilizing hybrid models, and incorporating more recent data for improved accuracy. This would further enhance forecasting, aiding Walmart's inventory management and strategic planning efforts. The project underscored the value of data-driven insights in supporting retail operations.

# References

* **Github link for the Project: https://github.com/Yeshasvin/WalmartStoresSales\_Analysis\_and\_Forecasting**
* **Time Series Analysis**: Box, G.E.P., Jenkins, G.M., & Reinsel, G.C. (2008). Time Series Analysis: Forecasting and Control. This book provided the theoretical background for ARIMA and SARIMAX models used in the project.
* **Machine Learning Models**: Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. The book helped in understanding and implementing various machine learning models like Random Forest and XGBoost.
* **Statistical Tests**: Shapiro-Wilk and Spearman Rank tests were performed for normality and correlation analysis, guided by resources from Python Data Science Handbook by Jake VanderPlas.
* **Neural Networks**: Chollet, F. (2018). Deep Learning with Python. This resource guided the implementation of neural networks, including the recurrent and convolutional models used for time series forecasting.
* **Seasonality and Stationarity Testing**: Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. The ADF test was applied to confirm the stationarity of the sales data.
* **SARIMAX Model**: Documentation from the statsmodels library was referenced for implementing and tuning the SARIMAX model in Python. Available at: <https://www.statsmodels.org/>.
* **Hyperparameter Tuning**: Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. This paper provided insights into using Random Search for optimizing hyperparameters in Random Forest and XGBoost models.