Name

- Afrose Ahamed Fathima

Course

- Advanced Certification in Data Science and Artificial Intelligence

Capstone Project

Walmart

Date: 27-09-2024

Table of Contents

Acknowledgment

I would like to express my sincere gratitude to the Intellipaat Team and IIT Madras for providing me with the opportunity to enhance my skills and knowledge through their comprehensive course. The curriculum, coupled with expert guidance, has significantly contributed to my understanding of data science, machine learning, and artificial intelligence. I appreciate the dedication and support of the instructors and the valuable resources that were made available throughout the program. I am eager to apply what I have learned in my future endeavours.

Problem Statement

A retail store that has multiple outlets across the country are facing issues in managing the inventory - to match the demand with respect to supply. You are a data scientist, who has to come up with useful insights using the data and make prediction models to forecast the sales for X number of months/years.

1. Using the above data, come up with useful insights that can be used by each of the stores to improve in various areas.
2. Forecast the sales for each store for the next 12 weeks.

Project Objective:

The primary objective of this project is to develop a robust data-driven approach for optimizing sales strategies across multiple Walmart retail outlets. The project aims to achieve the following:

1. **Sales Forecasting**:

- Build predictive models to accurately forecast future sales for each Walmart store over the next 12 weeks, enabling better demand planning and inventory management.

2. **Sales Trend and Performance Analysis**:

- Analyze historical sales data to identify key trends, seasonality, and patterns (e.g., monthly, and weekly sales trends).
- Compare sales performance across different stores and determine factors driving variations in sales.

3. **Impact of Holidays/Promotions**:

- Analyze the impact of holidays and promotions (if available) on sales, enabling stores to plan future promotional events effectively to boost revenue.

Data Description

Feature Name	Description	Data Type
Store	Store number	Integer
Date	Week of Sales	DateTime
Weekly_Sales	Sales for the given store in that week	Float
Holiday_Flag	If it is a holiday week	Integer
Temperature	Temperature on the day of the sale	Float
Fuel_Price	Cost of the fuel in the region	Float
CPI	Consumer Price Index	Float
Unemployment	Unemployment Rate	Float

The dataset contains 6435 rows and 8 columns.

Data Pre-processing Steps and Inspiration

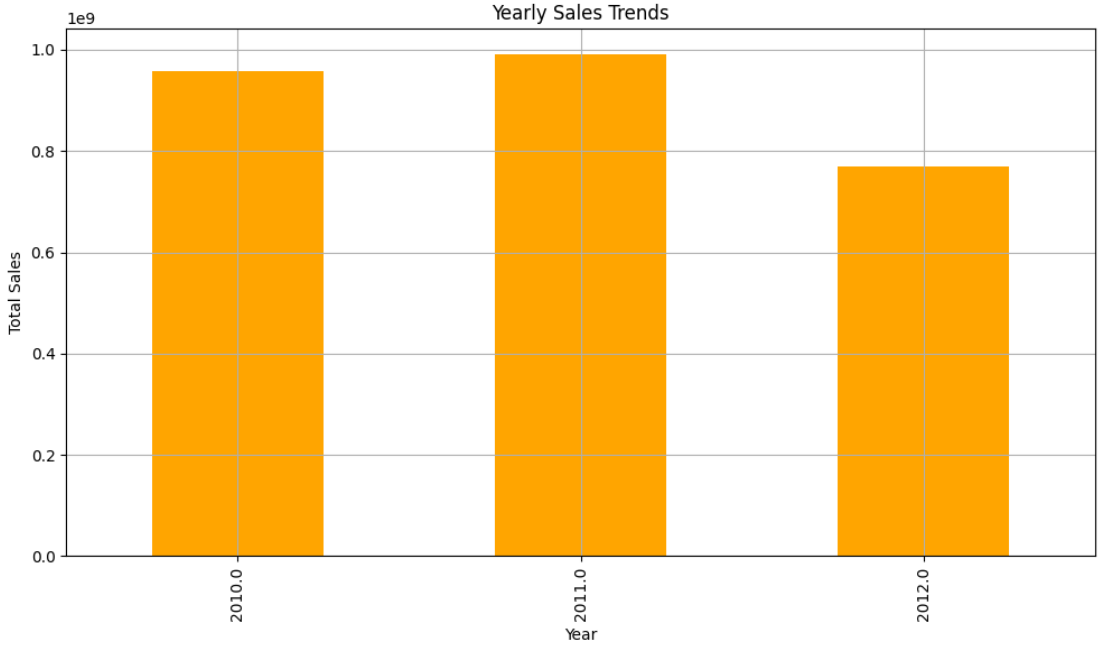
The main goal of data preprocessing is to handle missing values, deal with any inconsistencies, format data correctly, and prepare the dataset for further analysis and forecasting.

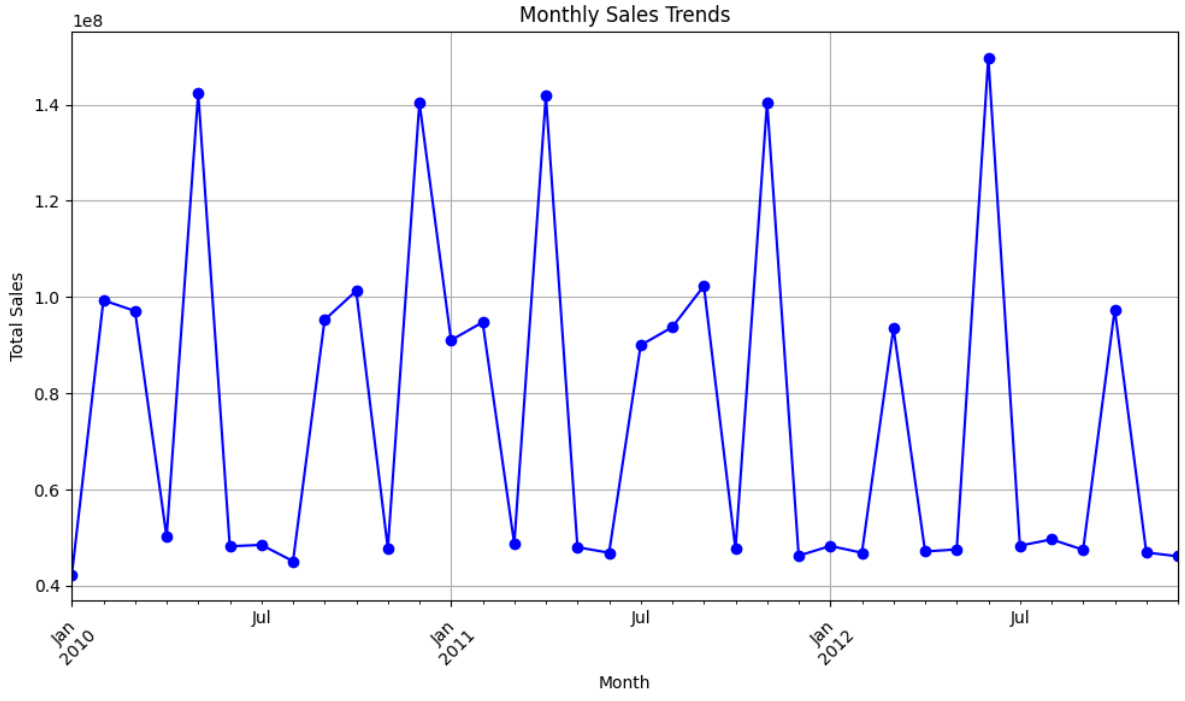
Data Pre-processing Steps:

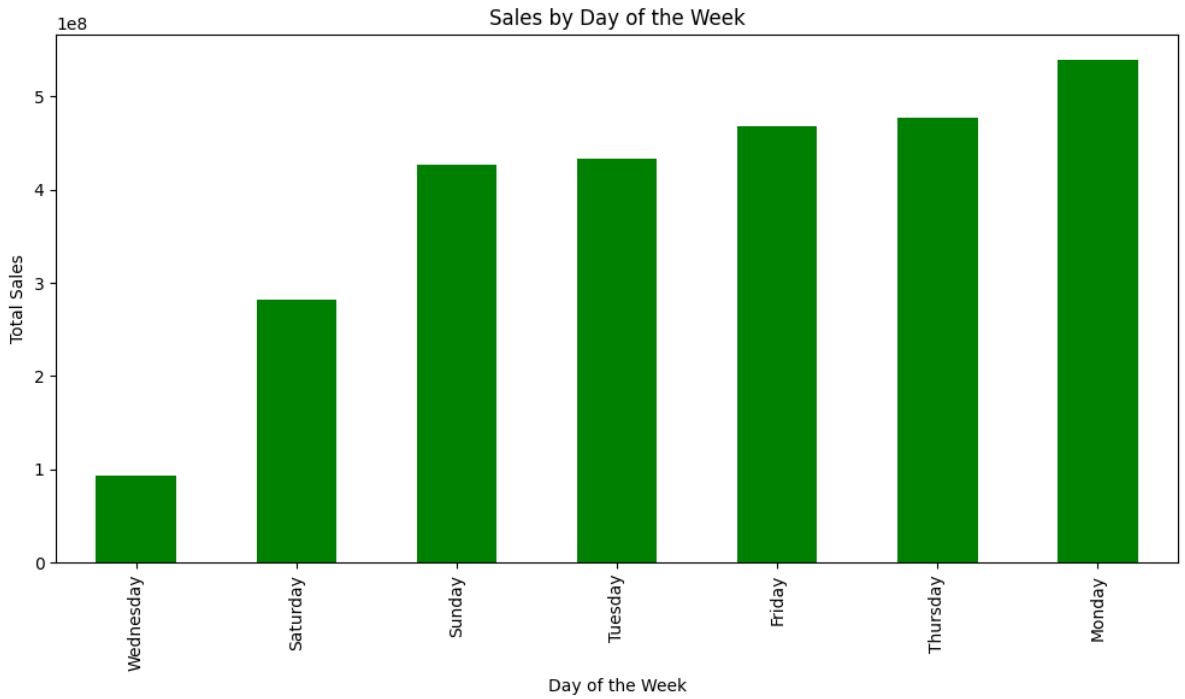
1. **Loading the Data**:
    - **Objective**: Loading the dataset into a panda DataFrame for further analysis and manipulation.
2. **Handling Missing Values**:
    - **Objective**: Identify and address missing values that could affect the analysis or model performance.
    - **Actions**:
        - Drop rows or columns with too many missing values.
        - Imputing missing values using appropriate methods (mean, median, or forward-fill).

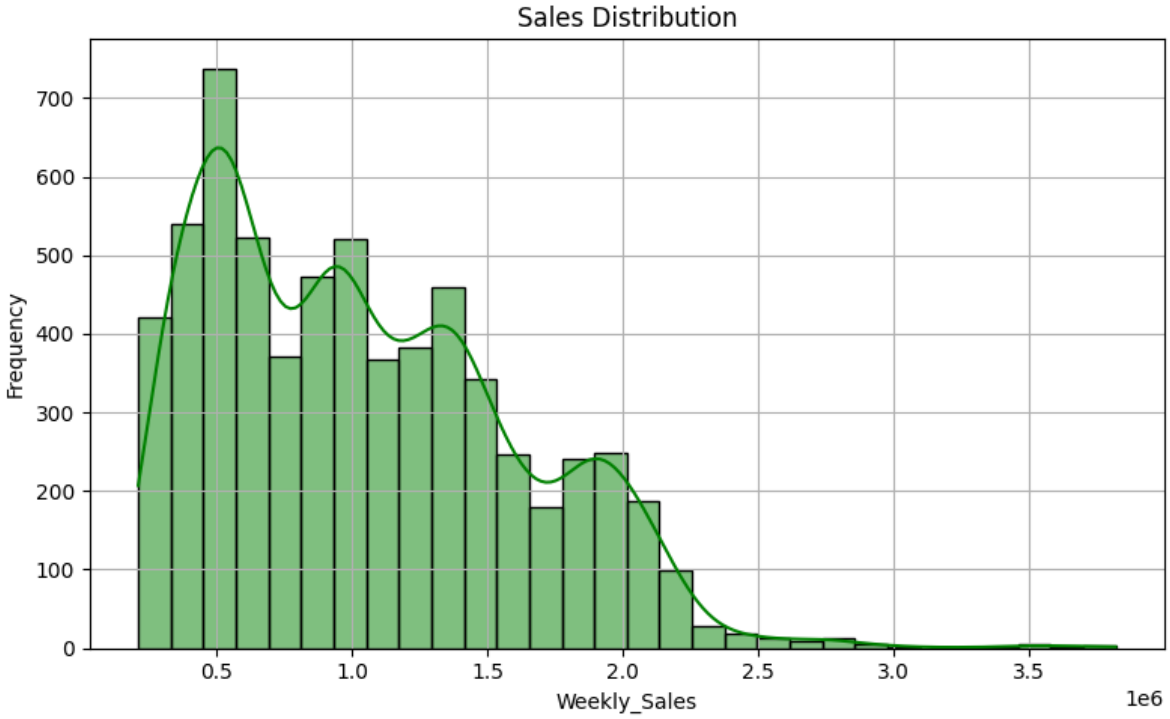
    There are no missing values in this dataset.
3. **Converting Data Types**:
    - **Objective**: Ensuring all columns have the correct data types (e.g., converting the 'Date' column to a datetime type).
4. **Feature Engineering**:
    - **Objective**: Create useful new features, such as Year, Month, or Week, from the Date column.
5. **Handling Outliers**:
    - **Objective**: Detect and handle outliers in the data to avoid skewing the forecasting models.
    - **Actions**:
        - Identify outliers using statistical techniques (e.g., Z-scores or IQR).
        - Handle outliers by capping, transforming, or removing them.
6. **Splitting the Data**:
    - **Objective**: Divide the dataset into training and testing sets
    - **Action**: Split the dataset into a train-test split to evaluate forecasting models.

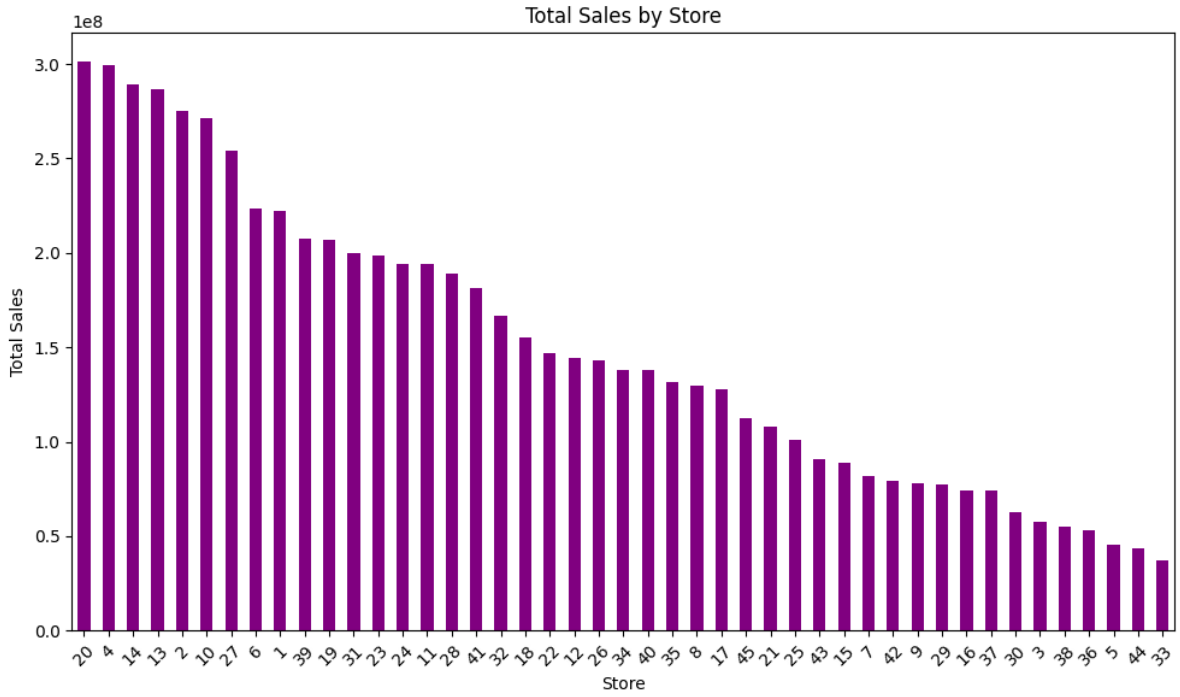
Data Visualization

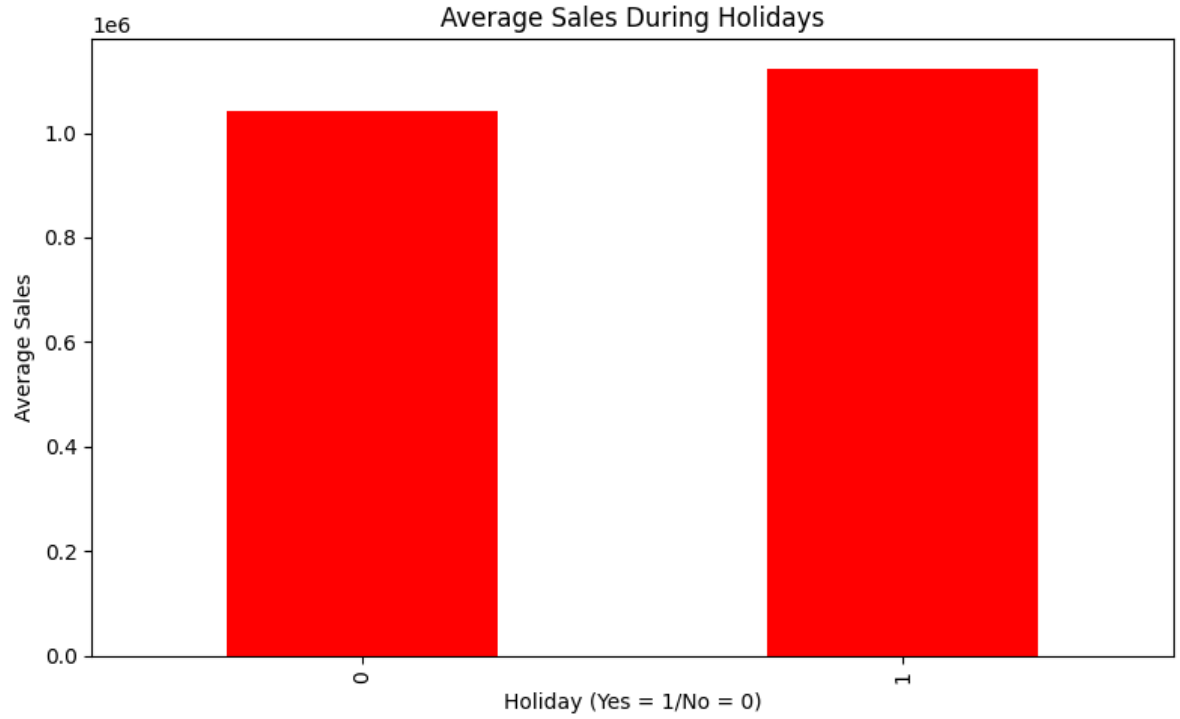


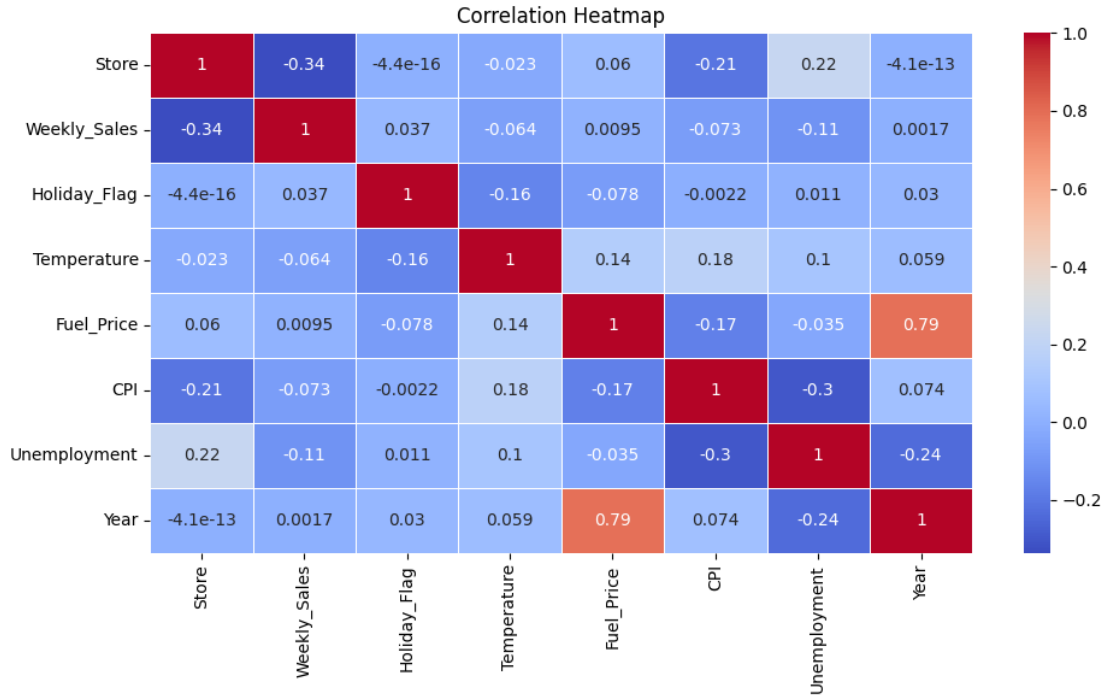



Sales by Day of the Week
1e8
Total Sales
0
1
2
3
4
5
Wednesday
Saturday
Sunday
Tuesday
Friday
Thursday
Monday
Day of the Week


Sales Distribution
Frequency
0
100
200
300
400
500
600
700
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Weekly_Sales
1e6


Total Sales by Store
1e8
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Total Sales
20 4 14 13 2 10 27 6 1 39 19 31 23 24 11 28 41 32 18 22 12 26 34 40 35 8 17 45 21 25 43 15 7 42 9 29 16 37 30 3 38 36 5 44 33
Store


Average Sales During Holidays
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Average Sales
0
1
Holiday (Yes = 1/No = 0)



Choosing the Algorithm for the Project

Selecting the appropriate algorithm for forecasting sales in a time series dataset is crucial for obtaining accurate predictions. The right model depends on the complexity of the data, including seasonality, trends, holidays, and external factors like weather or fuel prices.

1. ARIMA (Auto-Regressive Integrated Moving Average)

When to Use:

- Use ARIMA if the sales data follows a linear trend with no strong seasonality.
- Good for simple datasets where sales are relatively stable, without drastic fluctuations or holiday effects.

Key Features:

- ARIMA models the linear relationship between the current and past values of the series.
- The model is defined by three parameters: p (lag order), d (degree of differencing), and q (order of moving average).

Pros:

- Works well for univariate time series (e.g., sales over time).
- Simple and interpretable.

Cons:

- Does not handle seasonality well.
- Assumes a linear relationship in the data, which might not be suitable for complex sales patterns.

2. SARIMA (Seasonal ARIMA)

When to Use:

- Use SARIMA if the sales data exhibits clear seasonal patterns (e.g., spikes during holidays, weekends, or certain months).
- Suitable for datasets with both trend and seasonality.

Key Features:

- An extension of ARIMA that adds seasonal components.
- Includes seasonal parameters (P, D, Q, s), where "s" refers to the seasonal length (e.g., weekly, monthly).

Pros:

- Handles seasonality, making it suitable for sales data with periodic fluctuations.
- Provides good performance on univariate datasets.

Cons:

- Requires manual tuning of both non-seasonal (p, d, q) and seasonal (P, D, Q) parameters.
- Doesn’t incorporate external variables like holidays or promotions easily.

This dataset involves seasonality, holiday effects, and possibly external factors (like fuel prices, CPI, and unemployment). Therefore, *SARIMA* is chosen for forecasting.

Motivation and Reasons For Choosing the Algorithm

In the context of forecasting sales for Walmart stores, several factors make SARIMA an ideal choice over other algorithms.

1. **Seasonality and Trends**: Walmart's sales data likely exhibits strong seasonal patterns, making SARIMA ideal for capturing these.
2. **Univariate Focus**: Since we're forecasting sales based on historical data without requiring external regressors, SARIMA is well-suited for univariate time series forecasting.
3. **Flexibility**: SARIMA offers the flexibility to adjust seasonal (P, D, Q, s) and non-seasonal (p, d, q) components to optimize performance.
4. **Interpretability**: SARIMA provides an interpretable framework that balances complexity and ease of explanation.
5. **Proven Track Record**: The effectiveness of SARIMA in time series forecasting, especially in retail, makes it a reliable choice.

Assumptions

When using the **SARIMA** (Seasonal ARIMA) model for time series forecasting, several underlying assumptions need to be satisfied to ensure the model's validity and accuracy.

1. **Stationarity**: The data is stationary or can be made stationary through differencing.
2. **Linear Relationship**: Future sales are linearly dependent on past sales and errors.
3. **Constant Seasonality**: The seasonality patterns do not change significantly over time.
4. **No Significant Outliers**: The data does not contain many outliers or structural breaks.
5. **Uncorrelated Residuals**: The residuals are uncorrelated and normally distributed.
6. **Additive Components**: The seasonal and trend components are additive.
7. **Constant Parameters**: The model parameters remain constant over time.

Model Evaluation and Techniques

1. Train-Test Split

Before evaluating the model, it is important to split the data into **training** and **test sets**:

- The training set is used to fit the model.
- The test set is used to evaluate the model's predictive power on unseen data.

In the case of time series forecasting:

- **The most recent data points** (e.g., the last few weeks/months of sales) should be used as the test set.
- **Earlier periods** should be used for training the model.

All the data except the last 12 weeks are used for the training set.

The last 12 weeks are used for the test set.

2. Metrics for Model Evaluation

a) Mean Absolute Error (MAE)

- Measures the average magnitude of the errors between predicted and actual sales values.
- **Interpretation**: The lower the MAE, the better the model's performance in terms of absolute errors.

b) Mean Squared Error (MSE)

- Measures the average of the squared differences between predicted and actual sales.
- **Interpretation**: A lower MSE indicates a better fit. However, MSE can amplify large errors due to squaring.

c) Root Mean Squared Error (RMSE)

- The square root of MSE, which brings the metric back to the same unit as the data (sales).

- **Interpretation**: RMSE is commonly used for interpretability since it provides the error in the same unit as the sales data. Lower RMSE values indicate better performance.

d) Mean Absolute Percentage Error (MAPE)

- Measures the percentage error between predicted and actual sales values.
- **Interpretation**: MAPE expresses error as a percentage of actual values, making it easier to understand model performance in a percentage-based context. Lower MAPE values mean the model predictions are closer to the true values.

Below is the evaluation metrics table after building the model and evaluating it.

Evaluation Metrics	Value
Mean Absolute Error (MAE)	195935.2543263406
Mean Squared Error (MSE)	90882082742.26907
Root Mean Squared Error (RMSE)	301466.55327294447
Mean Absolute Percentage Error (MAPE)	20.323792942531234

Residual Analysis

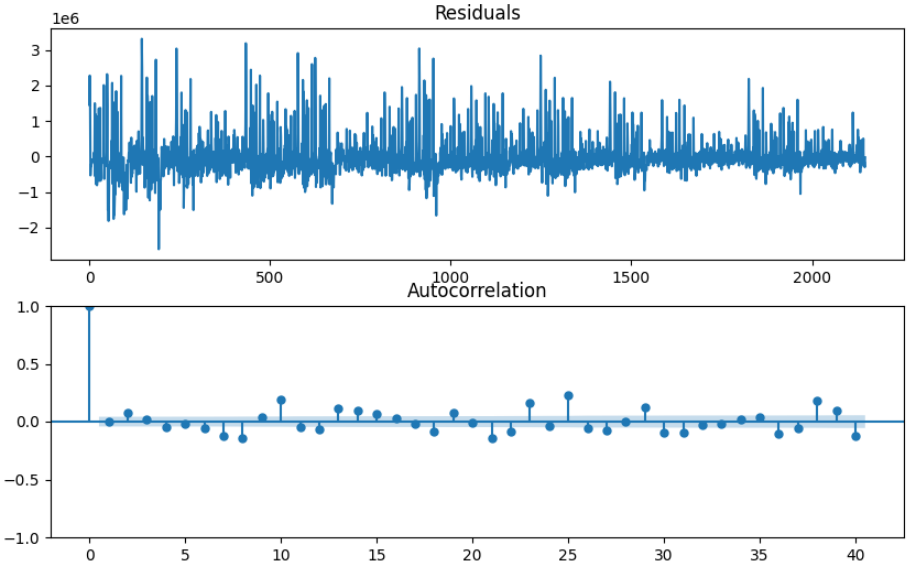
a) Check for Autocorrelation in Residuals

- The residuals (errors between actual and predicted values) should be **uncorrelated**.
- If the residuals are autocorrelated, it indicates that the model has not captured all patterns in the data and may require further tuning.

b) Plotting Residuals

- The residuals should be normally distributed with a mean of zero.
- A histogram or **Q-Q plot** can help assess whether the residuals follow a normal distribution.
- The **Autocorrelation Function (ACF)** plot of the residuals should not show significant lags.

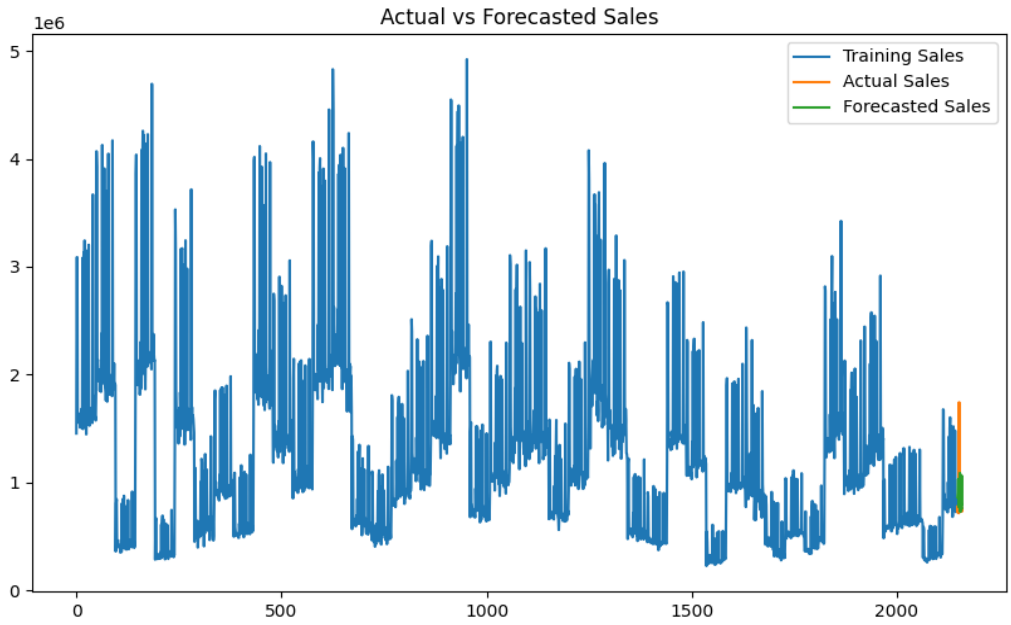
Below is the residual plot from our model.



Forecast Accuracy on the Test Set

After fitting the model, it's important to evaluate its forecasting ability on unseen data (test set) by comparing predicted values with actual sales data.

Below is the graph of forecast of sales



Inferences from the Same

After evaluating the SARIMA model using various techniques, we can derive several inferences about sales trends, patterns, and the model's effectiveness in predicting future sales for Walmart. Below are key insights based on the model's performance:

1. Sales Seasonality is Strong and Predictable

Inference:

- The SARIMA model successfully captured the seasonal patterns in sales data. This means that the sales at Walmart stores are influenced by specific time-based factors, likely driven by recurring events such as holidays, seasonal promotions, and economic cycles.

Example:

- During the holiday season (e.g., Black Friday, Christmas), there is a notable spike in sales, which repeats annually. The SARIMA model identifies this pattern and uses it to forecast sales during similar periods in future years.

Actionable Insight:

- Walmart can leverage this information by ensuring adequate inventory, staff, and marketing resources during these high-demand periods. Additionally, promotional strategies can be optimized to further boost sales during these key times.

2. Trend and Growth Patterns are Captured

Inference:

- The SARIMA model indicates that the sales data shows not only seasonality but also a **long-term trend**. This trend can either be positive (growing sales over time) or negative (declining sales). If the model captures a steady upward trend, it indicates growing demand for Walmart's products.

Example:

- Over the last few months/years, Walmart might see a general increase in sales, indicating growing customer demand, potentially due to factors such as improved customer experience, product assortment, or broader market trends.

Actionable Insight:

- If the sales trend is positive, Walmart can invest in store expansion or increasing product lines to meet the rising demand. If the trend is negative, this may call for investigating factors causing the decline, such as competition, customer satisfaction, or external economic factors.

3. Residuals Show Random Noise, Model is a Good Fit

Inference:

The analysis of residuals shows that they are mostly uncorrelated and follow a normal distribution with no significant autocorrelation. This indicates that the SARIMA model captures the core patterns (seasonality, trend) in the sales data effectively, leaving behind only random noise in the residuals.

Example:

If the residuals do not show any significant patterns or lags in the autocorrelation plot, it means that the SARIMA model is performing well and is not missing any major temporal structure in the data.

Actionable Insight:

Since the model fits the data well, Walmart can use it to make reliable sales forecasts. The team can now extend this model to new stores, regions, or product lines to make predictions.

4. Long-Term Planning for New Stores and Product Lines

Inference:

- The sales forecasts generated by SARIMA over the next 12 weeks can provide insights into the **long-term demand**. By comparing predictions for different stores, Walmart can identify regions with stronger demand growth, which may indicate a need for **new store openings** or **expansion of product lines**.

Example:

- If Store 7 shows consistently higher forecasted sales than other stores, it may indicate an opportunity for growth, warranting the introduction of more product categories or expanding store capacity.

Actionable Insight:

- **Store-specific insights** enable Walmart to prioritize certain stores for new product launches or expansions, while underperforming stores might benefit from promotions or changes in

product mix. This data-driven approach can help Walmart grow in regions with strong demand while optimizing resources in slower markets.

Summary of Key Inferences:

1. **Sales seasonality is strong and predictable**, especially during holidays and promotional periods.
2. The model successfully captures **long-term trends** in sales, indicating growing or declining demand.
3. The model fits well, as **residuals are random** and normally distributed, meaning core patterns in sales have been captured.
4. The **model's forecasts are accurate** within acceptable error margins and useful for inventory and staffing decisions.
5. Forecasts provide insights for **long-term planning**, including store expansions and product line introductions.

Future Possibilities of the Project

1. Real-time Forecasting and Dynamic Updates

Possibility:

Moving the model towards **real-time forecasting** by continuously feeding the latest sales data into the system would provide Walmart with **dynamic, up-to-date forecasts**. This would help Walmart respond swiftly to unexpected changes in demand.

Example:

If sales unexpectedly spike at a particular store, the real-time model can detect this change and adjust predictions for the next few weeks, alerting store managers to restock products more quickly.

Benefits:

Increased agility in managing inventory and supply chain.

Ability to **react in real-time** to sales fluctuations, minimizing stockouts or overstocking.

Implementation:

Develop a **data pipeline** that ingests new data continuously, updates the SARIMA model dynamically, and generates forecasts in real-time for all stores.

2. Demand Forecasting for Individual Products or Categories

Possibility:

Currently, the focus is on overall sales forecasts for each store. Expanding the project to **predict sales for individual products or product categories** (e.g., electronics, groceries) can provide more granular insights for inventory management and product planning.

Example:

Walmart can predict demand for high-margin products or products with long supply chains (e.g., electronics, apparel), enabling them to adjust their inventory to maximize profits and minimize costs.

Benefits:

Optimized inventory for specific product lines, reducing waste and increasing profitability.

Better visibility into which products are driving revenue growth or decline at different times of the year.

Implementation:

Train separate SARIMA or ML models for individual product categories or even specific products, using relevant features such as promotions, seasonality, and historical trends.

3. Profitability and Cost Forecasting

Possibility:

Besides sales forecasting, Walmart can extend the project to **forecast profits and costs** by incorporating data on **cost of goods sold (COGS)**, **operational expenses**, and **gross margins** into the model.

Example:

If sales of a certain product line are expected to increase, Walmart can forecast how this will affect overall profit margins by including costs and anticipated expenses for handling the increased demand.

Benefits:

Helps Walmart not just forecast sales but also **project profitability**, ensuring they maintain healthy margins as they scale operations.

Provides a clear understanding of the financial implications of sales trends, enabling better **budgeting and financial planning**.

Implementation:

Use sales forecasts alongside cost data to develop financial forecasting models that predict **profit and costs**. This can be done using a combination of SARIMA and financial analysis models.

Conclusion

The Walmart sales forecasting project has the potential to evolve into a sophisticated, real-time forecasting system that uses a combination of time series models, machine learning techniques, and external factors. These future possibilities open up opportunities to optimize inventory management, improve profitability, enhance customer engagement, and support long-term strategic planning. By continuously refining the forecasting models and integrating additional data sources, Walmart can further streamline its operations and remain competitive in an increasingly data-driven retail environment.