WALMARTS' SALES FORECASTING

SAYANTANI SAHA, SUNEHA SEN
Roll No. MDS202345, MDS202351

Supervisor M.R. Srinivasan, Adjunct professor, Chennai Mathematical Institute

CHENNAI MATHEMATICAL INSTITUTE

1 Data and objective

Walmart Inc. is an American multinational retail corporation that operates a chain of hypermarkets (also called super centers), discount department stores, and grocery stores in the United States, headquartered in Bentonville, Arkansas. In India, Walmart operates under the name of Flipkart Wholesale.

As of July 31, 2022, Walmart has 10,585 stores and clubs in 24 countries, operating under 46 different names. Out of which, the dataset has 45 stores for basic analysis. The file has information about the Weekly Sales of 45 stores for the year 2010-2012 including the factors affecting Sales such as Holidays, Temperature, Fuel Price, CPI, and Unemployment.

Walmart is the world's largest company by revenue, with about USD$570 billion in annual revenue, according to the Fortune Global 500 list in May 2022.

About the Dataset:

1. Store: Identifier for the retail store.
2. Date: Date of sales record.
3. Holiday_Flag: Indicator for holiday week (1) or non-holiday week (0).
4. Temperature: Temperature in the region of the store.
5. Fuel_Price: Fuel price in the region.
6. CPI: Consumer Price Index.
7. Unemployment: Unemployment rate.

Objective:

- By having an estimate of the future sales, the stores can anticipate demand on certain weeks and arrange for staff accordingly.
- Manages the steps of supply chain.It uses the data to analyze transportation lanes and routes for the company's trucks. These data help Walmart keep transportation costs down and schedule an appropriate time for drivers.

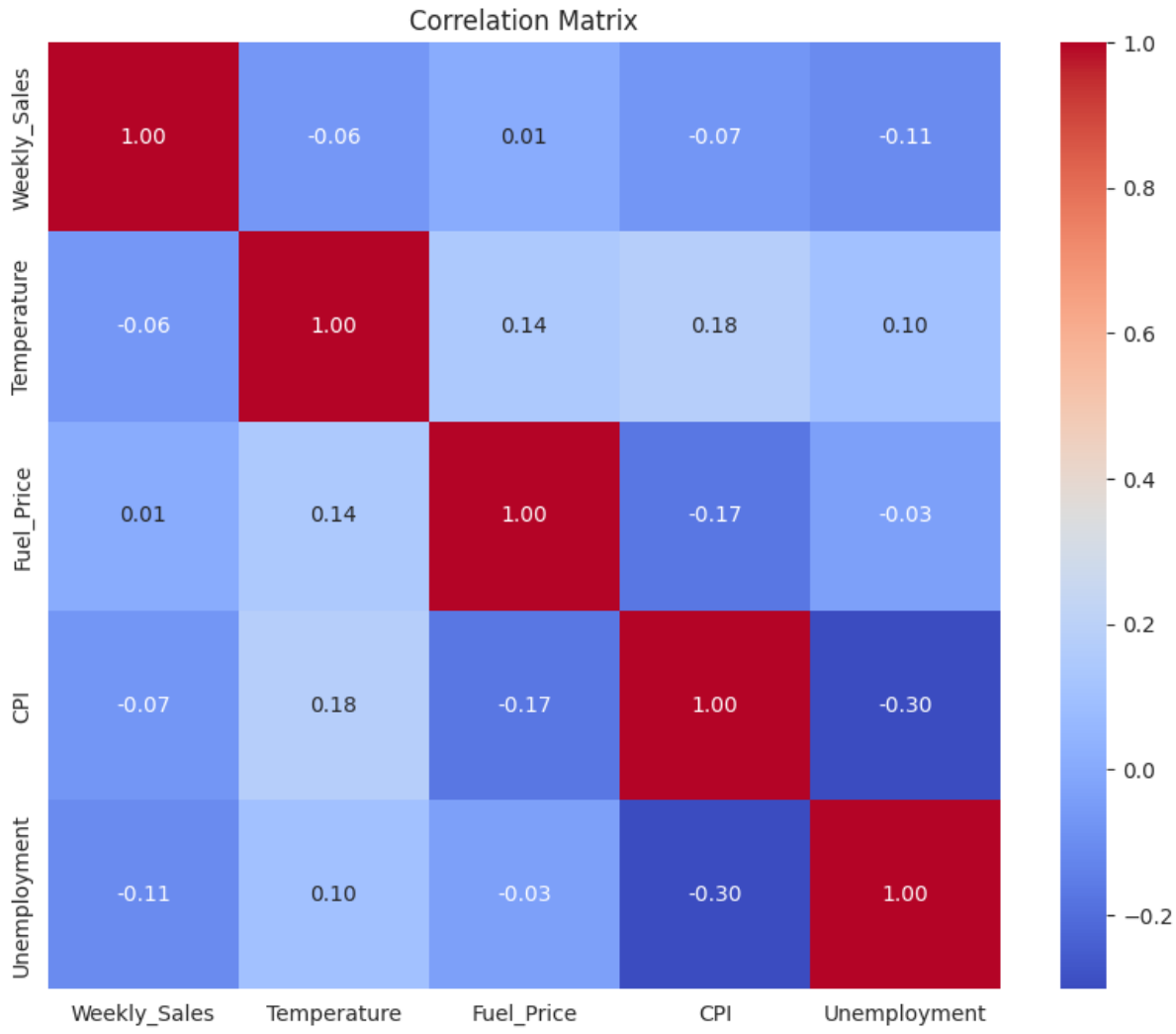
2 Methodology

2.1 Data Preparation

- The dataset was loaded, and it was checked for missing values. For numerical columns the missing values if found, were replaced by the mean and for categorical columns, by the mode. For this particular dataset, we did not have any missing values. Duplicate rows, if any were removed to ensure data integrity.
- The **Date** column was converted from an object to a date-time format for easier time based analysis. New features such as **days of the week**, **month** and **year** were extracted from the **Date** column. Additionally, a **season** feature was created to categorize months into their respective seasons (such as Winter, spring).

2.2 Exploratory Data Analysis

- The distribution of the features were explored through count plots, histograms and correlation heat maps. For **Store** and **Holiday_Flag**, count plots are drawn. From correlation matrix, we observed no columns were highly correlated.

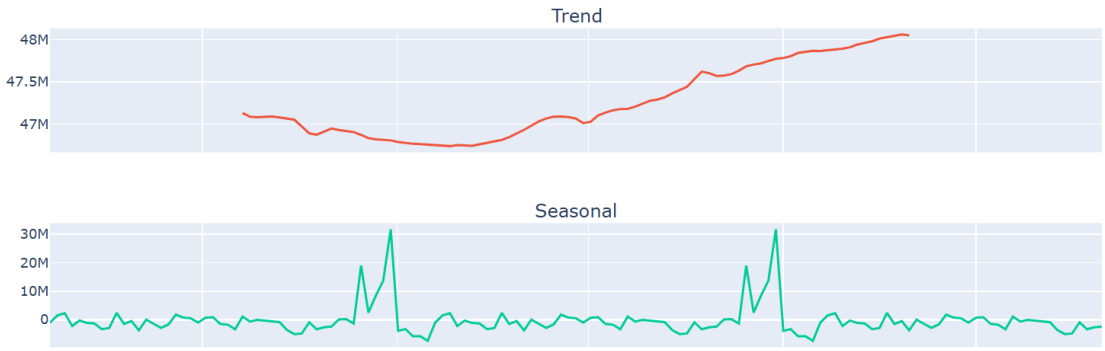


- Top 10 stores' contributing to the total sales were identified and their percentage contribution calculated. And we observe that 53% of total sales coming from top 10 stores.
- Weekly sales trends were analyzed across different seasons, years and months using point plots to understand how sales varied over time.

2.3 Time Series Analysis

Resampling and aggregation:

- The data was resampled to a weekly frequency to obtain the total **Weekly Sales** for each week, and to also ensure that the week ends on a Sunday.
- Time series decomposition was performed to separate the **Weekly Sales** into trend, seasonal and residual components using an additive model.

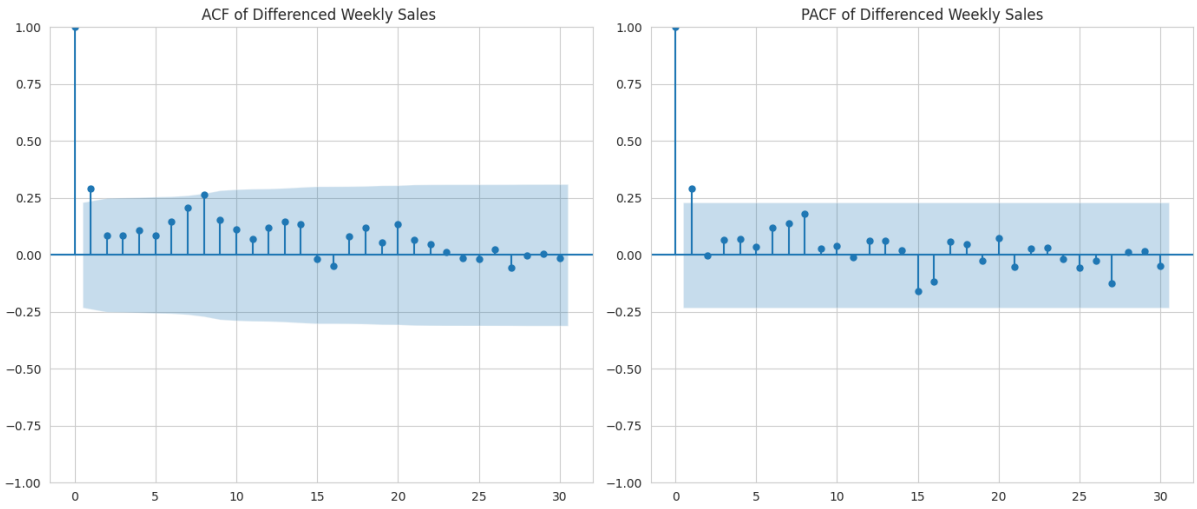


We can see a slight trend, decreasing first and then going upwards and a clear seasonality.

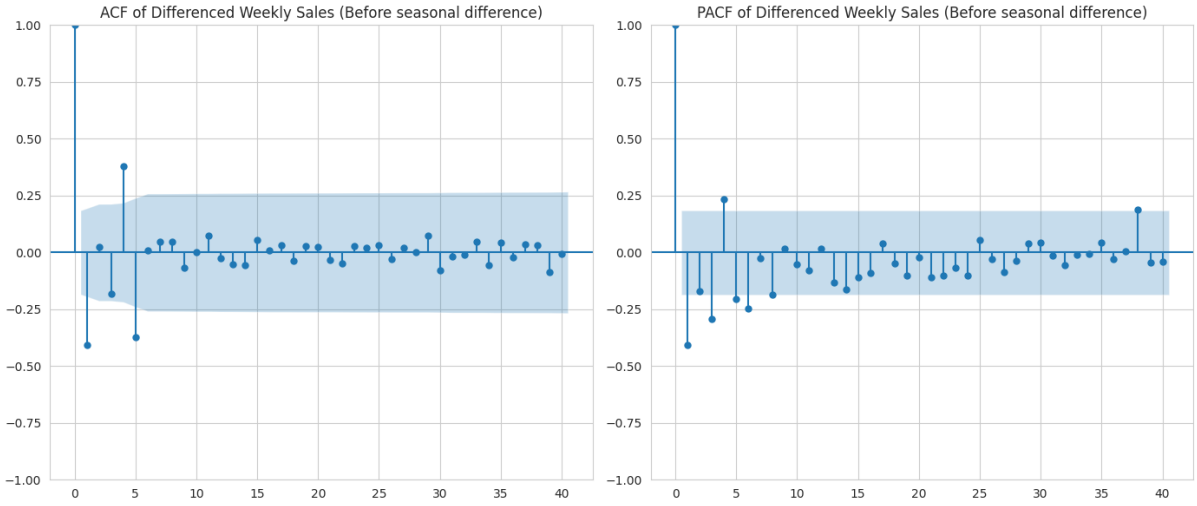
- The ADF test was conducted to check for stationarity. Differencing was applied to make the data stationary, and the ADF test was repeated on the differenced data, confirming stationarity.

Model selection:

- The ARIMA model was applied to the differenced data. The model's parameters for ARIMA(1,2,1) were chosen based on the ACF and PACF plots. From ACF plot, we get q=1, from PACF plot, we get p=1 and after differencing, we get d=2.



- The model was trained on 80% and tested on the remaining 20%.
- To account for seasonality, a Seasonal ARIMA (SARIMA) model was used, with parameters (1,1,1) for ARIMA and (1,1,1,52) for the seasonal component.
  We get the value of the parameters from ACF and PACF plots(before seasonal difference) and since our seasonal component exists, we take D=1.



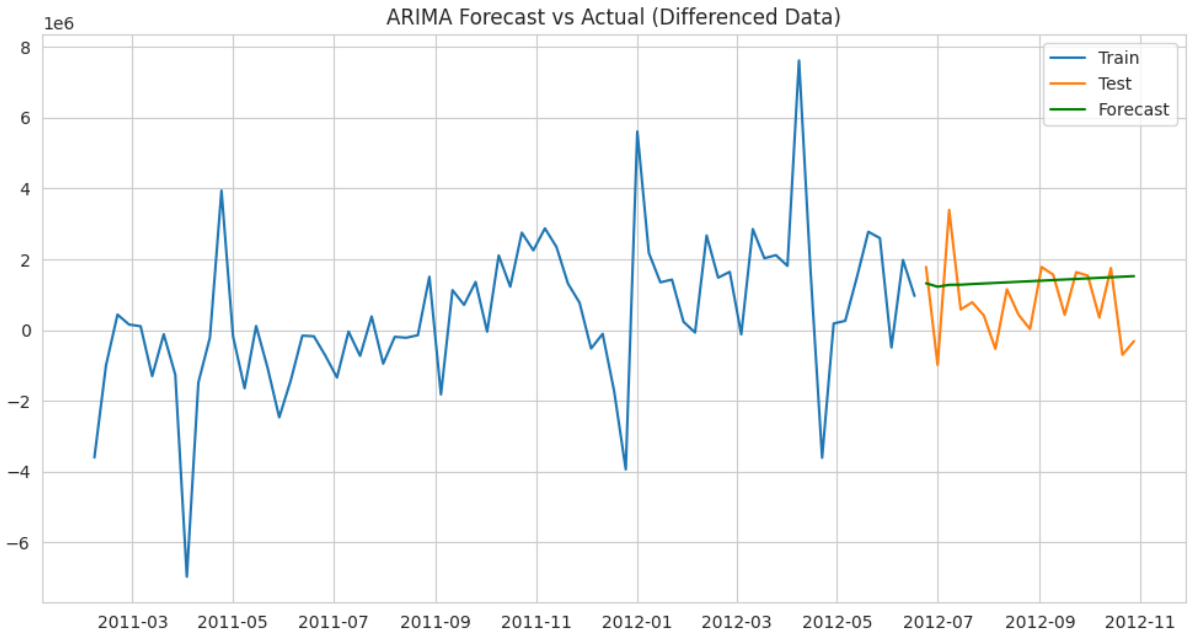
- The SARIMA model was evaluated on both differenced data and the original scale.
- The **LSTM** (Long Short-Term Memory) model was implemented to capture complex temporal dependencies in the data. An LSTM is particularly well-suited for handling sequential data, making it a good choice for time series forecasting.
- Forecast accuracy was evaluated using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).

2.4 Forecasting

- We do a **Test period forecast**. Predictions were made for the test set using the SARIMA model, with results compared to the actual weekly sales values.
- A 52-week future forecast was generated using the SARIMA model, providing a prediction for the next year. Forecast values were brought back to the original scale for interpretability.

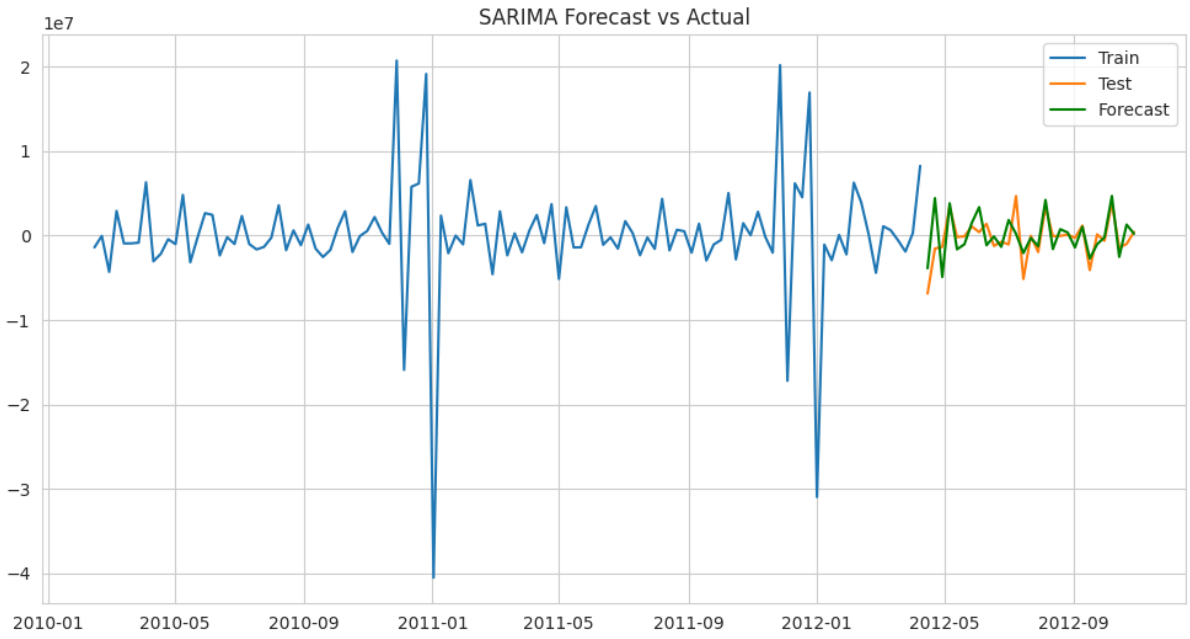
3 Result Analysis

- The **ARIMA** model did not perform well as expected, capturing short-term trends and patterns. Error metrics for the ARIMA model were calculated to check the accuracy of the model.



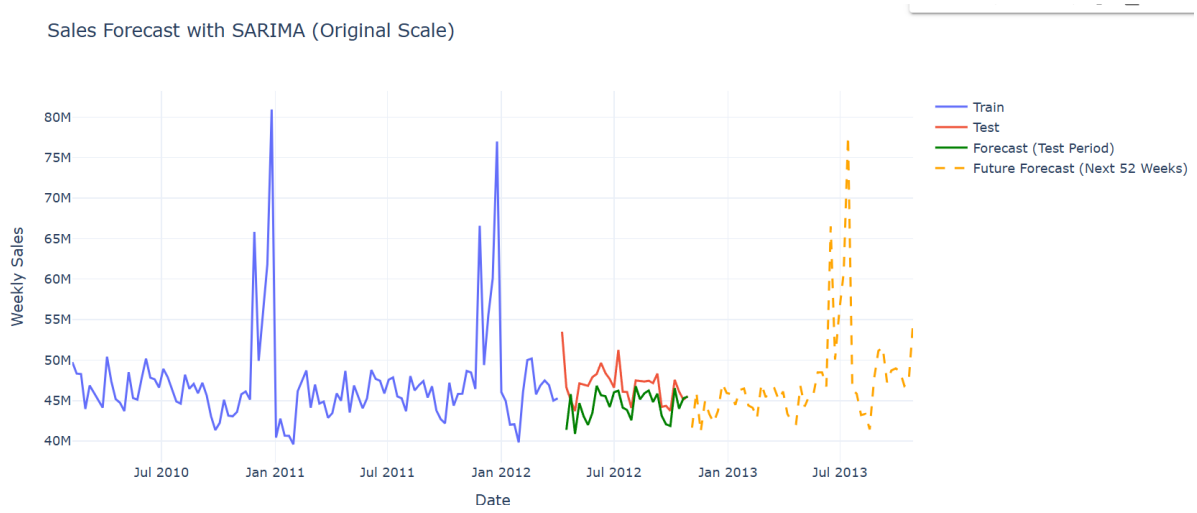
From the plot we can say that ARIMA model did not forecast the test data well. The forecast appeared smoother and less volatile, indicating that the ARIMA model might be underestimating the variability in the test data, suggesting that the model is good at capturing the general direction but not the short-term changes or seasonality.

- To capture the seasonality we fit the **SARIMA** model on differenced data and noticed that it performed better than ARIMA model and forecast the test data very well.

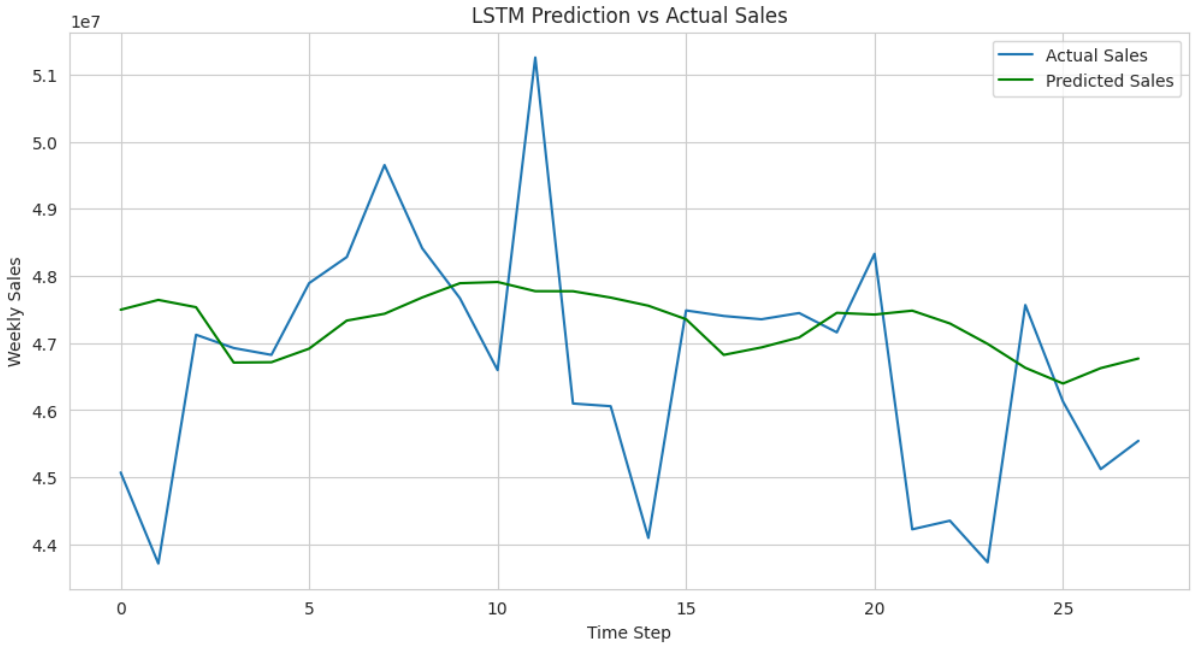


- The **SARIMA** model outperformed the ARIMA model due to its ability to handle seasonality explicitly. The forecast for the test period aligned well with the actual values, showing the model's capability to capture both short-term and seasonal trends.

The **forecast** for the next 52 weeks suggested a continuation of seasonal patterns seen in historical data, with peaks during specific seasons.



- The LSTM model's predictions were visualized alongside the actual sales values, demonstrating its effectiveness in capturing the temporal patterns present in the dataset.



From the plot, we can observe that the LSTM model captures the overall trend of the weekly sales but struggles with sharp fluctuations. While the model captures general seasonality and trend, it may not be as responsive to sudden changes in the data, possibly due to its inherent smoothing behavior on trainning data.

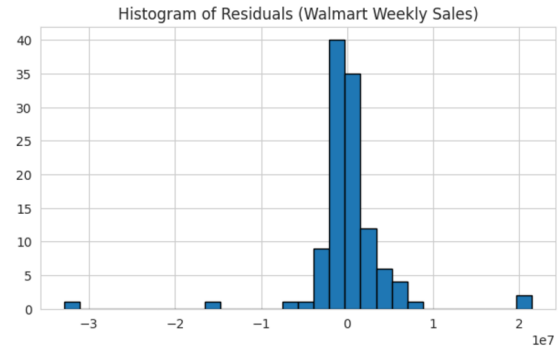
- Since **SARIMA** model captured the seasonality very well on our dataset, we choose the **SARIMA** model best among all models.

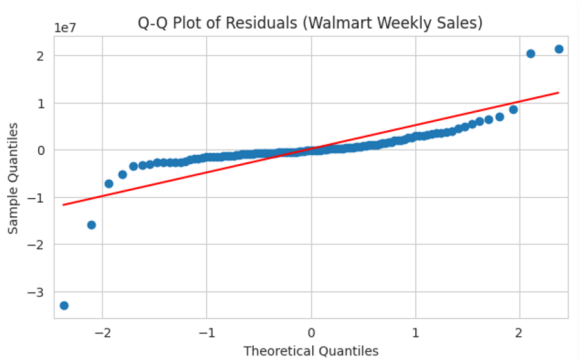
Table 1: Comparison of Error Metrics for ARIMA, SARIMA, and LSTM Models

Error Metric	**ARIMA**	**SARIMA**	**LSTM**
MSE	1478588517169.99	7270995838097.99	3852988216522.62
RMSE	1215972.25	2696478.41	1962903.01
MAE	972102.82	2292301.60	1471322.32
MAPE	691.90	4.87	3.22

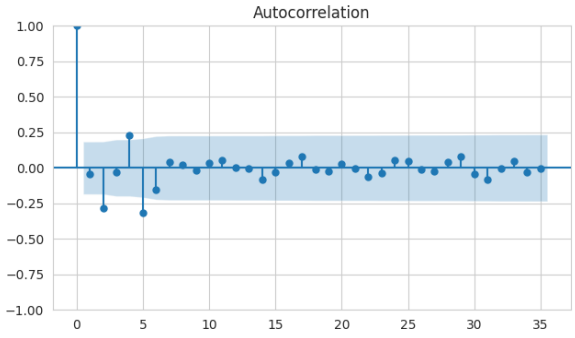
4 Residual Analysis

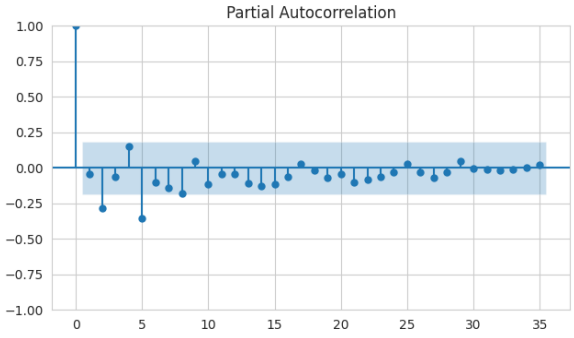
- Histogram and Q-Q plots were drawn for residual analysis. From the plots the residual analysis suggested that residuals were not perfectly normally distributed, but randomness was present, indicating no systematic bias.





- Despite some deviations from normality, the residual patterns showed no significant autocorrelation, suggesting the SARIMA model captured most of the underlying dynamics in the data.





5 Conclusion

This analysis provided a detailed exploration of Walmart's weekly sales data and demonstrated the effectiveness of SARIMA for time series forecasting. The SARIMA model proved to be a suitable choice for handling seasonal fluctuations in Walmart's sales data. The generated forecasts offer valuable insights for decision-making, assisting Walmart in optimizing stock levels, planning marketing campaigns, and preparing for seasonal demand changes.