CINEMA TICKET SALES FORECASTING

Time Series Analysis Project

ALENA MARIA THOMAS (MDS202303) alena.mds2023@cmi.ac.in
DIVYANSHI KUMARI (MDS202322) divyanshi.mds2023@cmi.ac.in

Abstract

This project aims to forecast cinema ticket sales to help cinema operators optimize staffing, inventory, and marketing strategies. Accurate sales forecasting is critical for improving operational efficiency and maximizing revenue. The project leverages time-series forecasting techniques, including ARIMA, SARIMA,exponential moving average and simple moving averages, to predict future ticket sales based on historical data. The dataset includes sales data across multiple cinemas, movies, and showtimes, spanning several months. Through data preprocessing, seasonal decomposition, and model evaluation, we compare the performance of different forecasting models using metrics like MSE, RMSE, and MAE.

Introduction

Cinema sales forecasting plays a crucial role in optimizing operations and maximizing revenue in the entertainment industry. Accurate sales forecasts allow cinema operators to plan staffing, inventory management, and marketing strategies more effectively, ensuring resources are allocated efficiently. Predicting ticket sales also helps in understanding customer demand patterns, particularly during peak seasons or for specific movie releases. However, the complexity of factors such as movie genres, showtimes, and seasonal variations makes this forecasting a challenging task.

Objective of the Project

The primary objective of this project is to improve the accuracy of cinema ticket sales forecasting. By employing time-series forecasting models such as ARIMA, SARIMA, moving averages, the project aims to develop models that can predict future sales based on historical data. The goal is to evaluate different models' effectiveness and identify the most accurate method for predicting sales, enabling cinema operators to make data-driven decisions.

Scope

The dataset used in this project spans from **February 21, 2018, to November 4, 2018**, containing sales data for multiple cinemas, movies, and showtimes. The dataset includes key variables such as total sales, occupancy percentage, and movie details.

Dataset Overview:

The data is collected from Kaggle.

- **film_code**: Unique identifier for the film being shown.
- **cinema_code**: Identifier for the cinema where the movie is being screened.
- **total_sales**: The total revenue generated from ticket sales.
- **tickets_sold**: The total number of tickets sold for a specific showtime.
- **tickets_out**: Number of tickets cancelled.
- **show_time**: The data is recorded at the **showtime level**, meaning for each specific showing of a movie, the sales (like total sales, tickets sold, occupancy percentage, etc.) are captured separately.
- **occu_perc**: occupation percent of cinema by means of available capacity
- **ticket_price**: The price of one ticket.
- **ticket_use**: A variable indicating whether a ticket has been used or not.
- **capacity**: The total seating capacity of the cinema.
- **date**: The date when the movie was shown.
- **month**: The month during which the show took place.
- **quarter**: The financial quarter during which the show took place.
- **day**: The day of the month when the movie was shown.

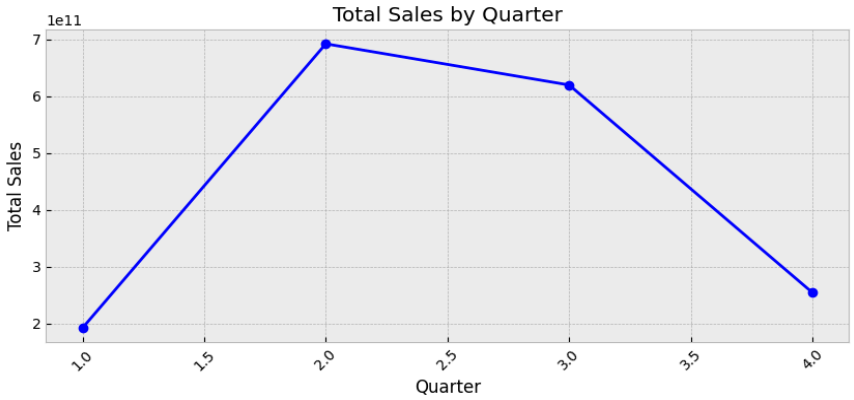
Data Cleaning:

We handled missing values (0.08% of data) by removing them, given their minimal impact. To identify rows uniquely, we created an ID column combining `film_code`, `cinema_code`, and `date`. Duplicate entries were removed due to their negligible presence.

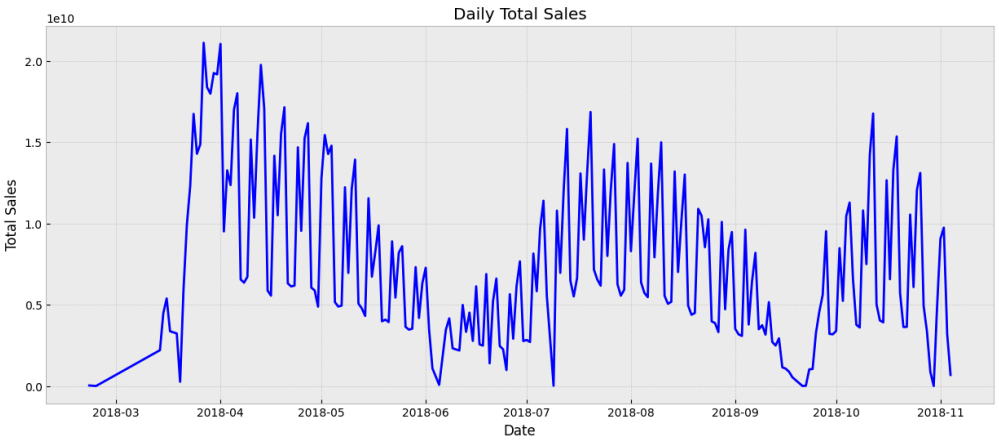
Exploratory Data Analysis

Given that our key variable is total_sales, we conducted EDA by resampling the data at quarterly, monthly, weekly, and daily levels to identify patterns and trends:

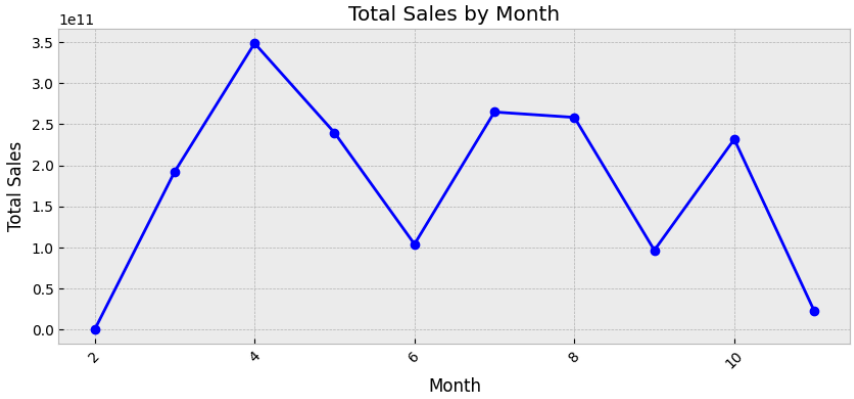
- Quarterly Trends: Sales are significantly higher in the 2nd and 3rd quarters, indicating peak cinema attendance during these periods.



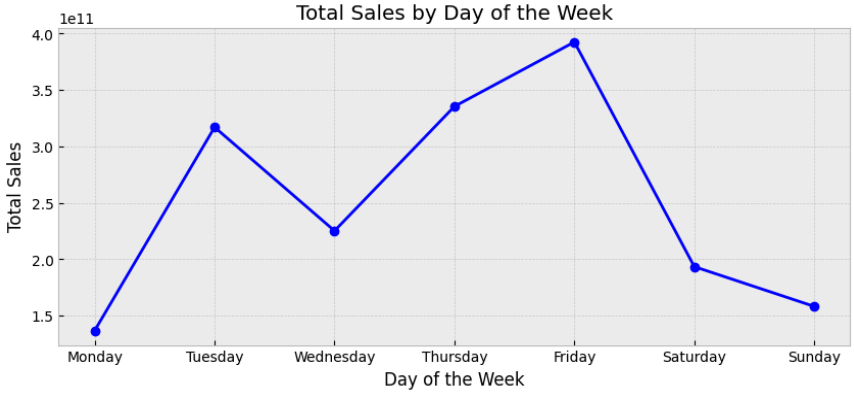
- Yearly Patterns: Sales decline notably at the start and end of the year, suggesting a seasonal dip in audience interest. However, due to lack of yearly data we cannot affirm our observation.



- Monthly Insights: September appears to be an unusually low month for cinema attendance, making it one of the least popular months for moviegoing.



- Day of the Week Analysis: Pre-weekend days tend to outperform weekends in ticket sales, indicating a preference for moviegoing just before the weekend.

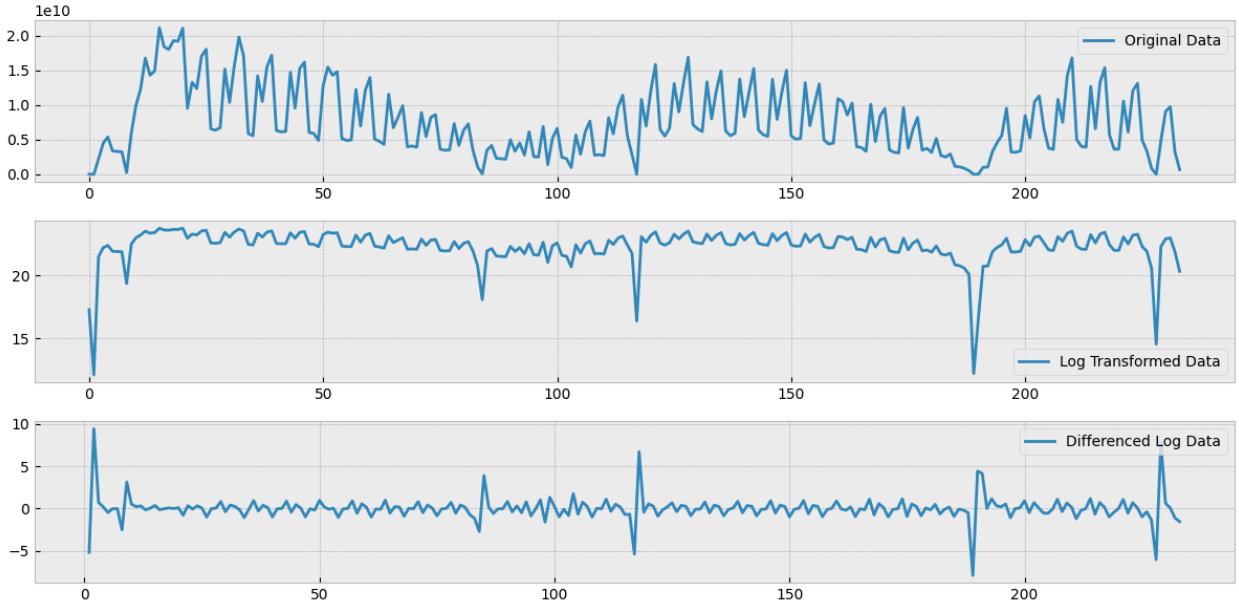


These insights highlight both seasonal and weekly patterns in cinema attendance, which can inform targeted forecasting and strategic planning.

- We also tried to analyse affects of film_code on total_sales weekly and monthly basis.Our analysis shows that **film_code** has an inconsistent impact on weekly and monthly sales. While some weeks see a dominant film, contributing over 99%, most weeks show much smaller contributions, typically ranging from 0.5% to 3%. Similarly, in most months, the top-performing film contributes less than 1% to total sales, with February being the only exception.Given these patterns, **film_code** appears to have a minimal and inconsistent influence on daily sales, making it an unreliable factor for forecasting. Instead, we will focus on other factors like trends, seasonality, and historical sales for more accurate predictions.

Stationarity Test

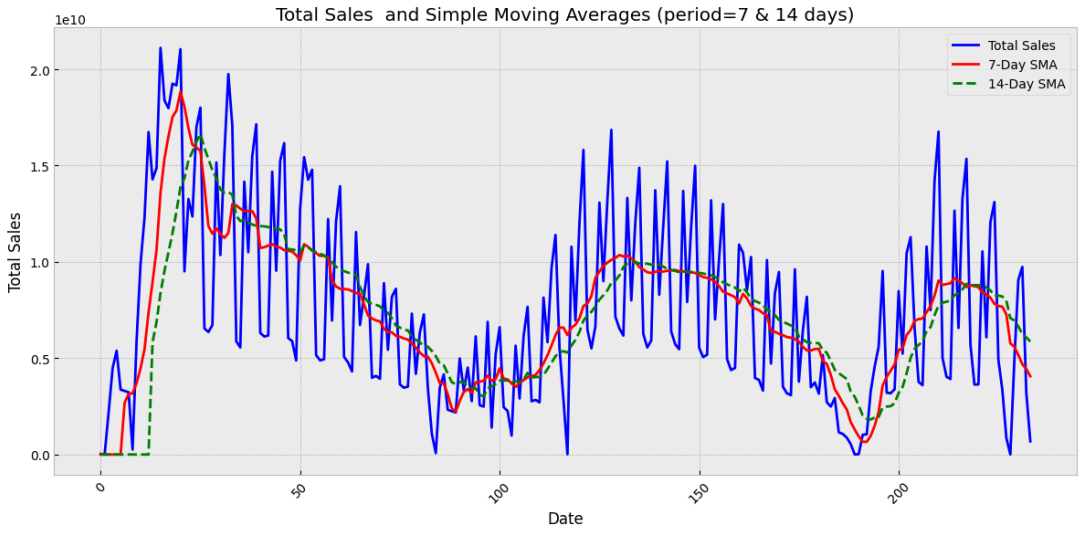
To assess stationarity in our cinema sales data, we performed the Augmented Dickey-Fuller (ADF) test. The ADF test results indicated that the data was non-stationary.To address this, we applied a log transformation to reduce the impact of large scale values and stabilize variance. Following this transformation, we implemented first-order differencing to remove trends and make the data closer to stationary, preparing it for more reliable time series modeling.

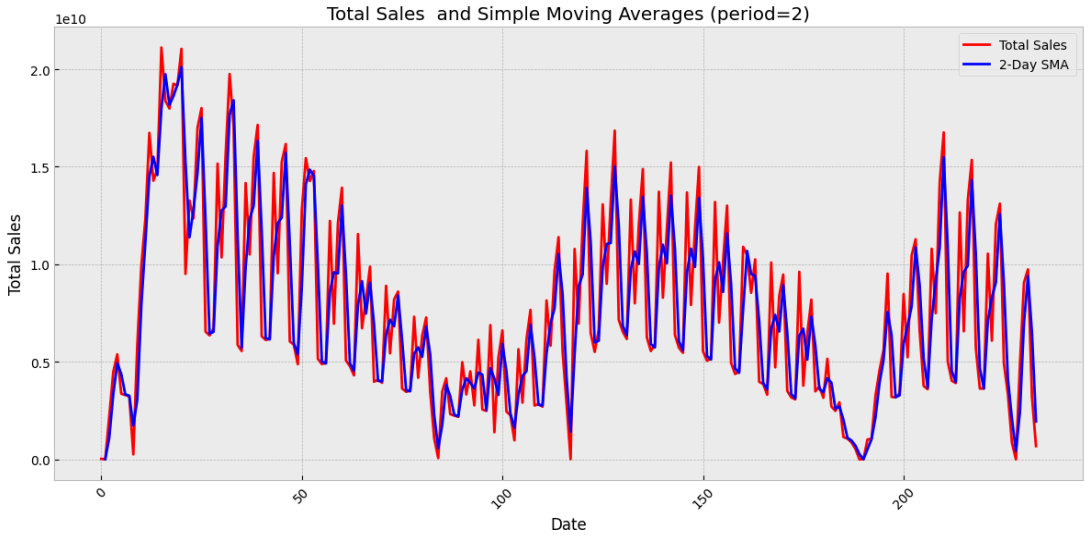


Simple Moving Average & Exponential Moving Average

In our cinema sales forecasting project, we applied Simple Moving Average (SMA) and Exponential Moving Average (EMA) to smooth data and observe trends.

- **SMA**:
  - **7-day MA and 14-day MA**:
    - Both models act more like a **mean model**, smoothing the data excessively and losing critical patterns, especially spikes. This behavior is expected since a larger window size prioritizes long-term trends over short-term variations, effectively flattening short-term fluctuations.
  - By averaging sales over a **2-day period**, SMA provided a smooth overview of cinema sales trends, reducing day-to-day variability and revealing general seasonal patterns. However, it lacked precision in capturing sudden spikes in sales.

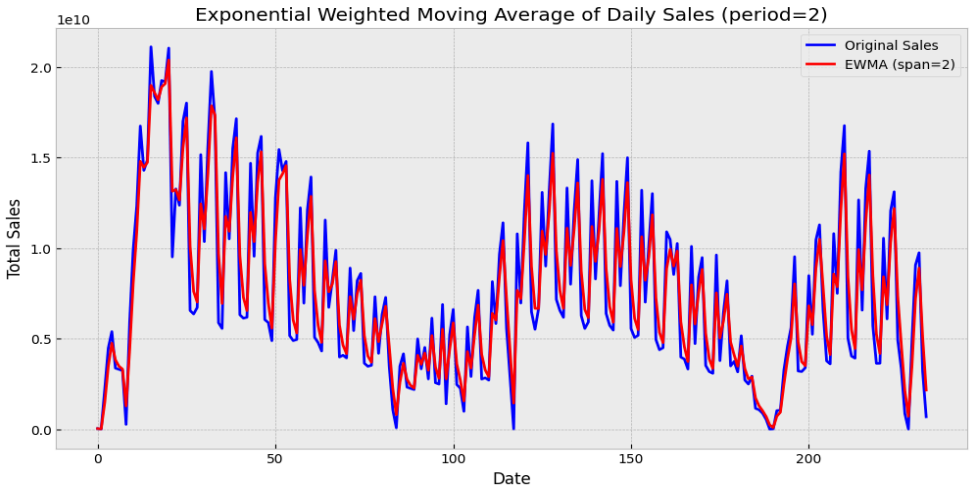


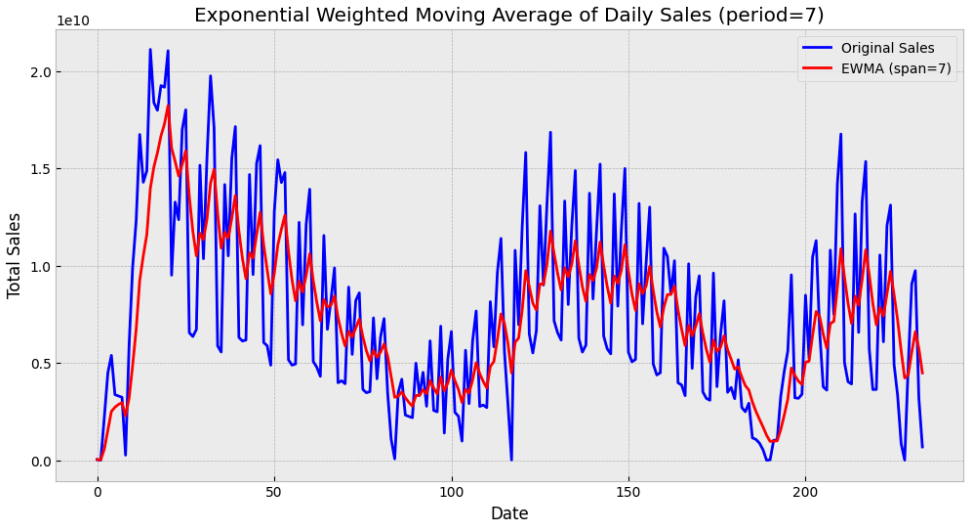


Moving averages work well for general trend analysis but struggle with datasets containing frequent, varying spikes.

A **simple moving average** is inherently limited because it gives equal weight to all observations within the window, which may not be appropriate for data with abrupt changes or uneven spikes.

- **EMA**: With a 2-day span, EMA assigned more weight to recent data, making it more responsive to recent changes. This responsiveness allowed EMA to better detect shifts and sudden increases in sales, offering valuable insights for timely decisions. However, for 7 day span, it performed close enough to a mean model.





TRAIN-TEST split

The data has been divided into training and testing sets, with the last 15 days reserved as the test set. This approach allows the model to learn from historical data and be evaluated on its ability to predict recent, unseen cases, ensuring a reliable assessment of forecasting performance.

ARIMA

ARIMA Parameters: We used the parameters (3,1,2) for the ARIMA model, selected via AutoARIMA to minimize the AIC value.

Model Performance on Training Set:

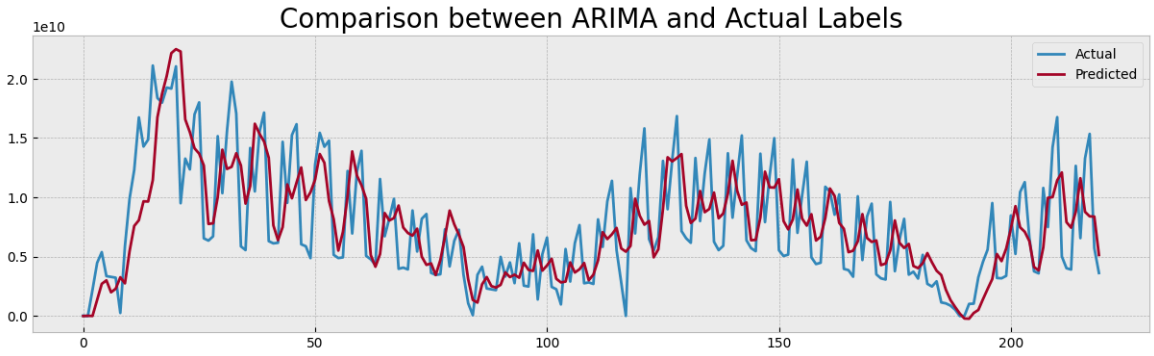
- The model successfully captured the **broader trend**s in the data.
- However, it **failed to capture the spike patterns**, particularly the sudden increases and decreases in the data.

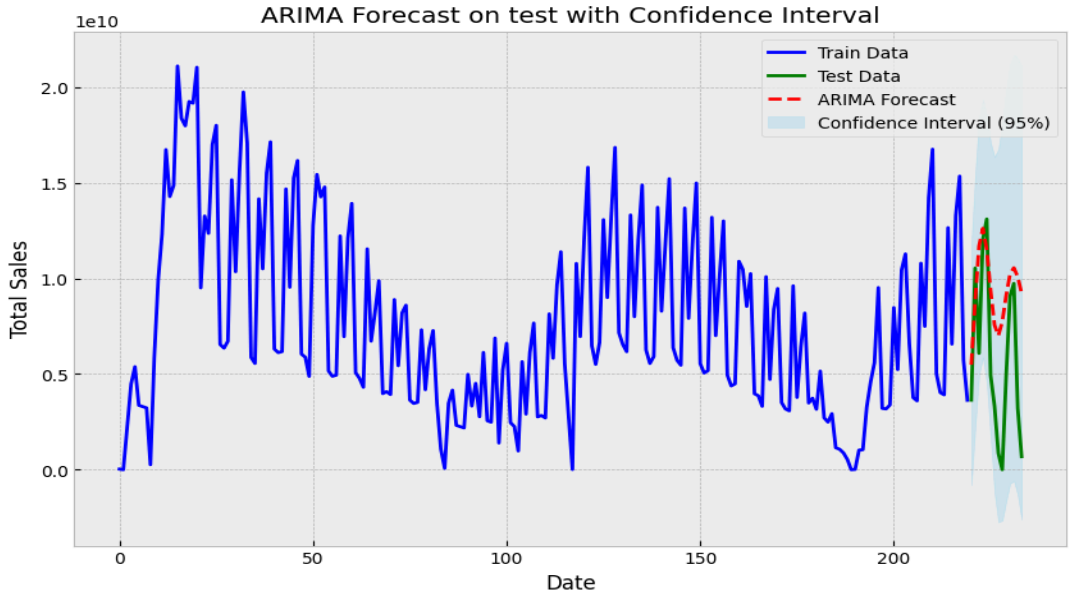
Model Performance on Test Set:

- The test set performance reflected the same limitations, with the model unable to identify or predict the **downward spikes** in the data.

Implication:

- Despite optimizing the ARIMA model for trend prediction, it struggled to account for irregular fluctuations or spikes, suggesting that ARIMA may not be the best fit for datasets with sharp, short-term variations.





SARIMA

SARIMA Parameters: Using AutoARIMA, the SARIMA model was configured with parameters (0,1,2) for the non-seasonal part and (1,0,2) for the seasonal part, with a seasonal period of 7.

Model Performance on Training Set:

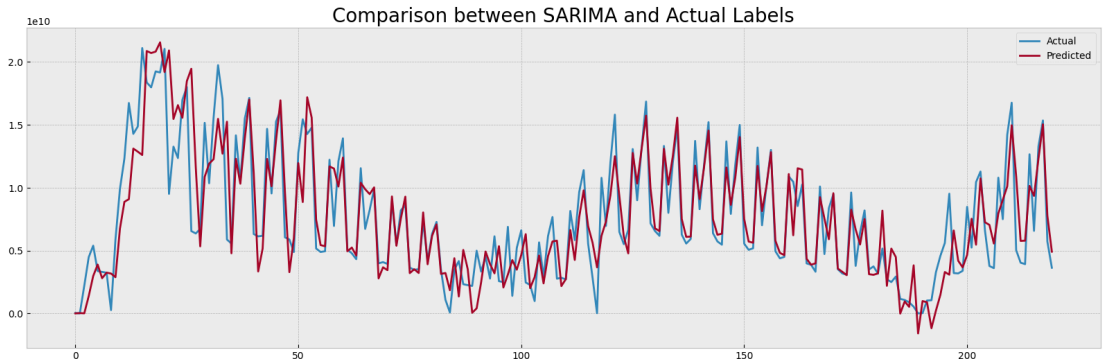
- The SARIMA model performed significantly better than ARIMA in capturing the trend and spike patterns in the data.
- It was able to capture most of the fluctuations, including upward and downward spikes, more accurately.

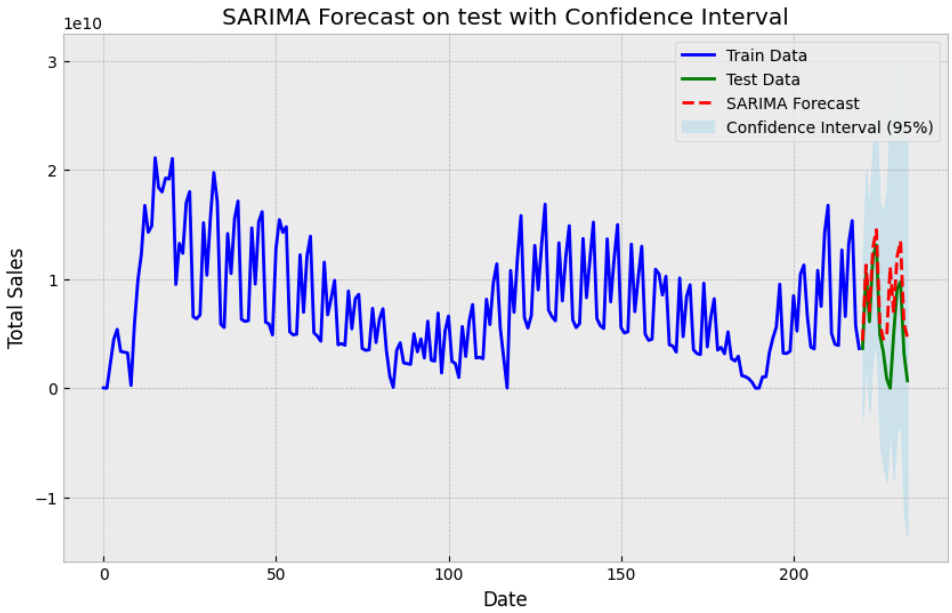
Model Performance on Test Set:

- The forecast on the test set was also superior to the ARIMA model, as it captured the broader patterns and trends effectively.
- However, while it did attempt to capture the downward spike in the test set, it failed to do so accurately, indicating that SARIMA still struggles with capturing certain irregular fluctuations in the data.

Implication:

- SARIMA improved upon ARIMA by capturing both trends and spikes more effectively, but it still struggled with precise prediction of sharp downward spikes, suggesting that further refinements or different models might be needed for more accurate spike forecasting.





Conclusion

Model	MSE	MAE	RMSE
ARIMA	2.258526e+19	3.967438e+09	4.752395e+09
SARIMA	1.388197e+19	2.652119e+09	3.725851e+09
2-day SMA	4.846730e+18	1.796510e+09	2.201529e+09
7-day SMA	1.667416e+19	3.729618e+09	4.083401e+09
14-day SMA	1.895032e+19	4.022291e+09	4.353196e+09
EMA 2 day	2.204286e+18	1.378041e+09	1.484684e+09

EMA 7 day	1.110475e+19	3.153407e+09	3.332378e+09

Metrics Evaluation: For the test set, we evaluated the performance of various models, including:

- 2-day, 7-day, and 14-day Simple Moving Averages (SMA)
- 2-day and 7-day Exponentially Weighted Moving Averages (EWMA)
- ARIMA and SARIMA

Best Performing Model: Among all the models, the 2-day EWMA performed the best, as it exhibited the lowest RMSE, MSE, and MAE compared to the other models.

The 2-day EWMA effectively captured short-term fluctuations and spikes, making it the most suitable model for forecasting in this case, outperforming both ARIMA and SARIMA, as well as the moving averages.