2024

Capstone Project- Walmart Sales Predictions

DATA SCIENECE & MACHINE LEARNING

BANANI KASHYAP

Table of Contents

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 provided with the weekly sales data for their various outlets. Use statistical analysis, EDA, outlier analysis, and handle the missing values to come up with various insights that can give them a clear perspective on the following:

1) If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?
2) If the weekly sales show a seasonal trend, when and what could be the reason?
3) Does temperature affect the weekly sales in any manner?
4) How is the Consumer Price index affecting the weekly sales of various stores?
5) Top performing stores according to the historical data.
6) The worst performing store, and how significant is the difference between the highest and lowest performing stores.
7) Use predictive modelling techniques to forecast the sales for each store for the next 12 weeks.

Problem Objective:

The primary objective of this analysis is to utilize data-driven insights to comprehensively understand Walmart's sales patterns. By delving into the diverse factors that impact sales performance and applying advanced modelling approaches, the goal is to uncover actionable insights that can enhance Walmart's sales strategies and promote long-term growth.

Data Description:

In this dataset, there are historical sales data of 45 Walmart stores based on store location and week. There are certain events and holidays which impact sales on each day. The business is facing a challenge due to unforeseen demands and runs out of stock some times. Walmart would like to predict the sales and demand accurately. The objective is to determine the factors affecting the sales and to analyse the impact of markdowns around holidays on the sales.

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

Data Analysis Overview:

The analysis of Walmart sales data encompasses several essential steps aimed at extracting insights to enhance business strategies and decision-making processes. It begins with loading the dataset from a CSV file using Pandas and preprocessing tasks to ensure data readiness, including data type conversion and handling missing values. Exploratory data analysis (EDA) forms a cornerstone, providing insights into the dataset's characteristics and relationships among variables through statistical analysis and visualization techniques such as histograms, box plots, and scatter plots. Time series analysis reveals temporal patterns and trends, employing decomposition techniques like seasonal decomposition and modelling with ARIMA and SARIMA to forecast future sales. Predictive modelling using the Random Forest Regressor forecasts sales, with evaluations based on metrics like MAE, MSE, and RMSE. The analysis culminates in actionable insights and recommendations, identifying top-performing stores, understanding sales influencers, and guiding future forecasting and strategy formulation. In essence, this comprehensive approach illuminates intricate patterns and insights within the Walmart sales dataset, empowering stakeholders with valuable information for informed decision-making and business growth.

Importing Important Libraries:

The analysis begins by importing essential libraries such as NumPy, Pandas, Matplotlib, and Seaborn. These libraries provide robust tools for data manipulation, visualization, and statistical analysis.

- **NumPy**: NumPy is used extensively for numerical computations and data manipulation tasks, such as handling arrays of data, performing mathematical operations, and statistical analysis.

- **Pandas**: Pandas is used for data manipulation tasks, such as loading, cleaning, transforming, and analysing data. It offers convenient methods for reading and writing data from various file formats and performing operations like filtering, grouping, and aggregation.

- **Matplotlib**: Matplotlib is used for data visualization tasks, such as creating line plots, scatter plots, histograms, bar charts, and more. It allows for customization of plot styles, axes, labels, and annotations to effectively communicate insights from the data.

- **Seaborn**: Seaborn is primarily used for statistical data visualization tasks, such as plotting complex relationships between variables, creating categorical plots, and visualizing distributions. It offers convenient functions for exploring data and generating insights through visualization.

- **SciPy:** SciPy is used for advanced scientific computing tasks, such as numerical integration, solving differential equations, and statistical hypothesis testing. It complements NumPy by offering a wide range of mathematical functions and algorithms for scientific analysis.

- **Scikit-learn (sklearn):** Scikit-learn is used for building and deploying machine learning models in Python. It offers a wide range of algorithms for classification, regression, clustering, dimensionality reduction, and model selection. Additionally, it provides utilities for data preprocessing, feature scaling, and model evaluation.

- **Prophet:** Prophet is used for time series forecasting tasks, such as predicting future values based on historical data. It automatically detects seasonal patterns, holidays, and other recurring events in the data and provides intuitive interfaces for model fitting, prediction, and visualization. Prophet is particularly useful for generating accurate forecasts with minimal manual intervention.

Data Preprocessing:

Data preprocessing plays a pivotal role in ensuring the integrity and quality of the dataset. In this phase, various preprocessing steps such as data type conversion, handling missing values, and setting the date column as the index are performed to prepare the data for further analysis.

1. Data Cleaning: Cleaning the data by removing missing values, outliers and other inconsistencies.
2. Data Exploration: Exploring the data to gain insights and understanding the data.
3. Data Visualization: Visualizing the data for better understanding.

Loading the Dataset:

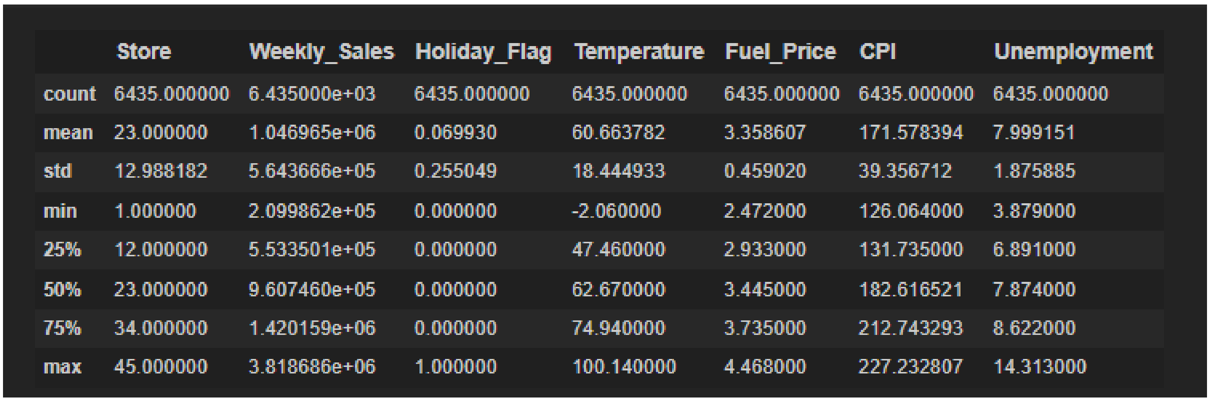
The Walmart sales dataset, a rich repository of historical sales data, is loaded into a Pandas Data Frame. This dataset serves as the cornerstone for all subsequent analyses and modelling endeavours.

```
Store  Date    Weekly_Sales        Holiday_Flag Temperature  Fuel_Price   CPI
Unemployment

0      1       05-02-2010   1643690.90  0       42.31  2.572  211.096358  8.106

1      1       12-02-2010   1641957.44  1       38.51  2.548  211.242170  8.106

2      1       19-02-2010   1611968.17  0       39.93  2.514  211.289143  8.106

3      1       26-02-2010   1409727.59  0       46.63  2.561  211.319643  8.106

4      1       05-03-2010   1554806.68  0       46.50  2.625  211.350143  8.106
```

Top 5 records of the dataset

Statistical Analysis:

A comprehensive statistical analysis is conducted to gain insights into the fundamental characteristics of the dataset. This phase includes examining the overall shape of the data, calculating summary statistics of numerical variables, and exploring key trends and patterns.



	Store	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
count	6435.000000	6.435000e+03	6435.000000	6435.000000	6435.000000	6435.000000	6435.000000
mean	23.000000	1.046965e+06	0.069930	60.663782	3.358607	171.578394	7.999151
std	12.988182	5.643666e+05	0.255049	18.444933	0.459020	39.356712	1.875885
min	1.000000	2.099862e+05	0.000000	-2.060000	2.472000	126.064000	3.879000
25%	12.000000	5.533501e+05	0.000000	47.460000	2.933000	131.735000	6.891000
50%	23.000000	9.607460e+05	0.000000	62.670000	3.445000	182.616521	7.874000
75%	34.000000	1.420159e+06	0.000000	74.940000	3.735000	212.743293	8.622000
max	45.000000	3.818686e+06	1.000000	100.140000	4.468000	227.232807	14.313000

Data Preprocessing:

Data preprocessing plays a pivotal role in ensuring the integrity and quality of the dataset. In this phase, various preprocessing steps such as data type conversion, handling missing values, and setting the date column as the index are performed to prepare the data for further analysis.

```
RangeIndex: 6435 entries, 0 to 6434
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Store         6435 non-null   category
 1   Date          6435 non-null   datetime64[ns]
 2   Weekly_Sales  6435 non-null   float64
 3   Holiday_Flag  6435 non-null   bool
 4   Temperature   6435 non-null   float64
 5   Fuel_Price    6435 non-null   float64
 6   CPI           6435 non-null   float64
 7   Unemployment  6435 non-null   float64
dtypes: bool(1), category(1), datetime64[ns](1), float64(5)
```

Exploratory Data Analysis (EDA):

The heart of the analysis lies in the exploratory data analysis (EDA), where a deep dive into the dataset unveils valuable insights. This phase involves several key steps:

- **Outlier Analysis**: Identifies anomalous data points in weekly sales, shedding light on exceptional sales events or irregularities that may warrant further investigation.

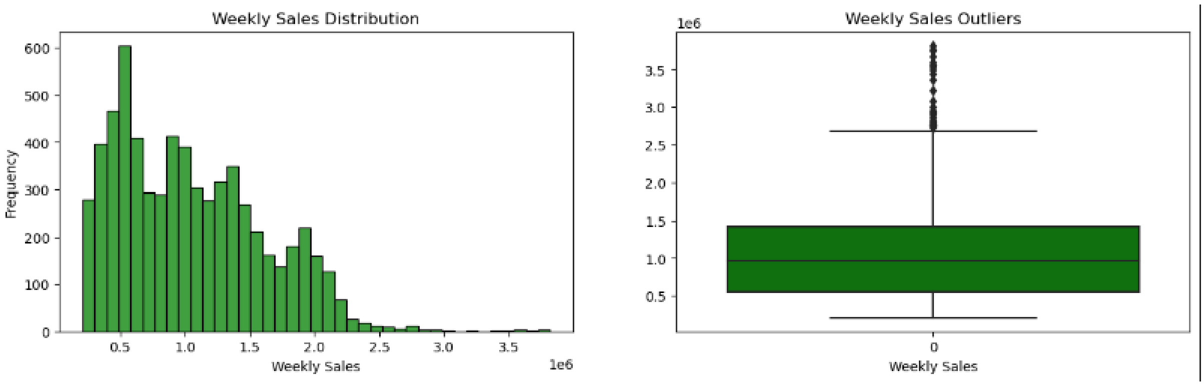


Fig: Weekly Sales Outliers

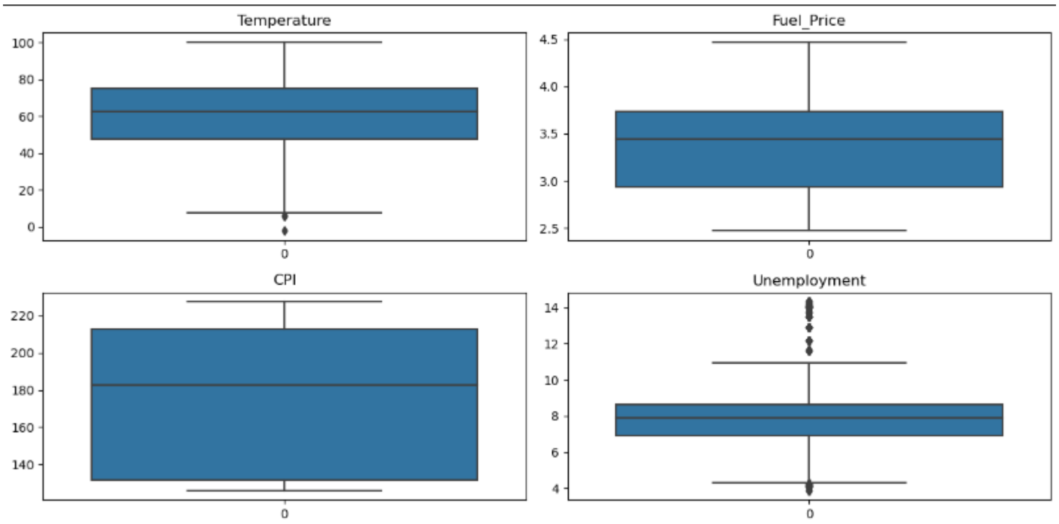
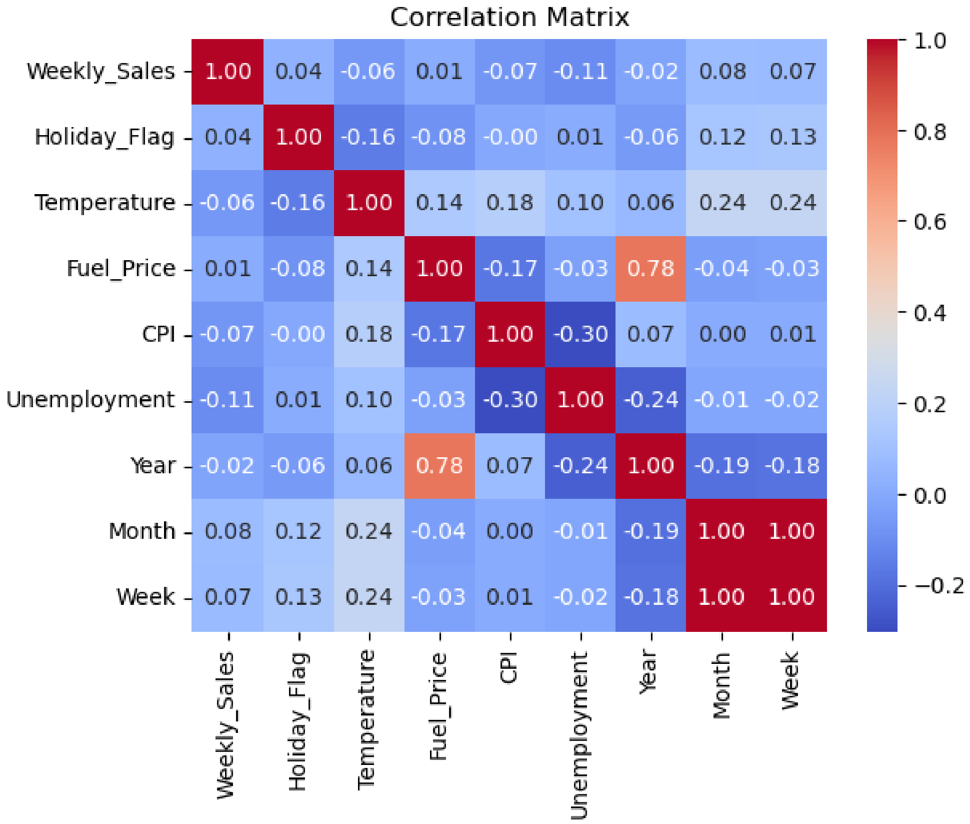


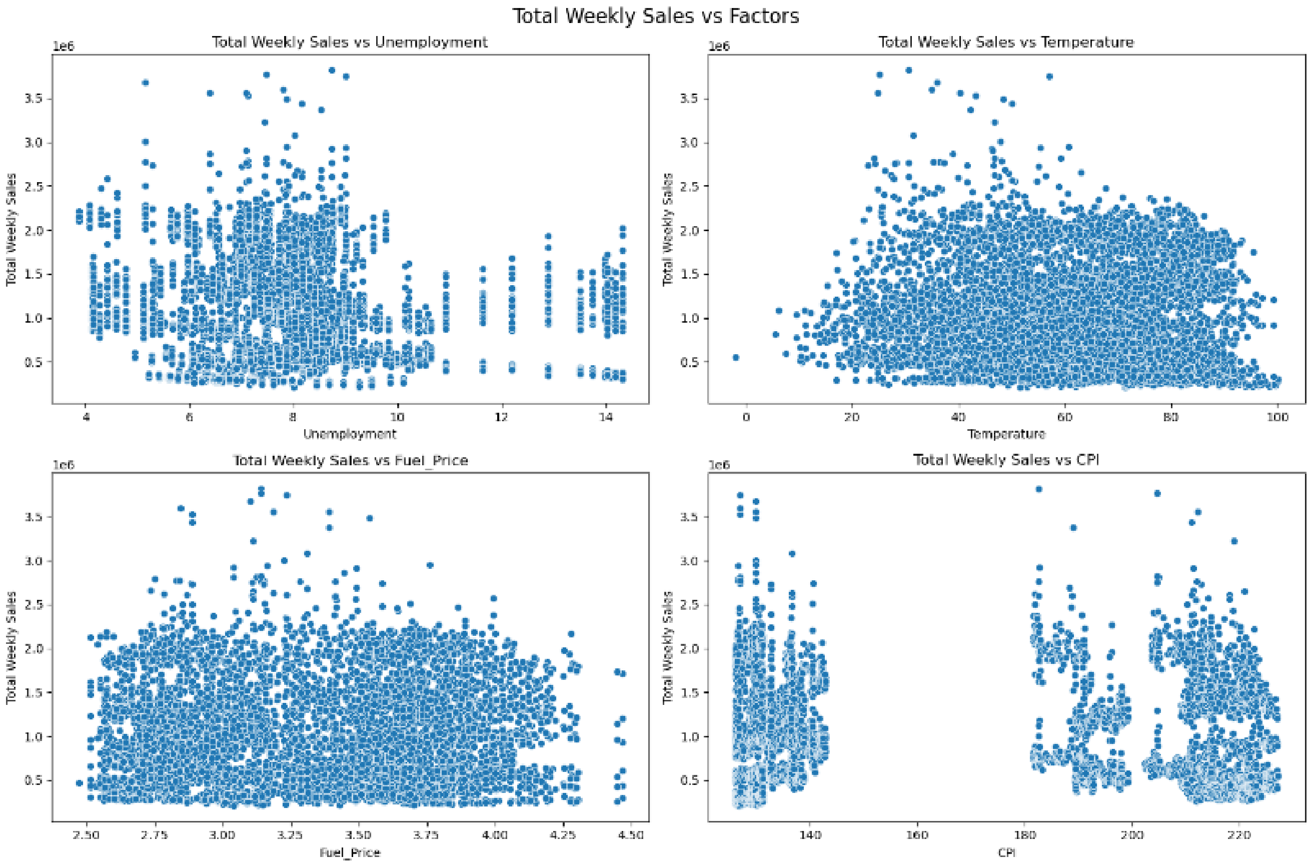
Fig: Outliers for features columns

- **Correlation Analysis**: correlation analysis to understand the relationships between different variables. A correlation plot, generated using the Seaborn library, visually represents the correlation matrix of the dataset. This plot provides insights into the strength and direction of correlations between variables, aiding in identifying potential factors that influence Walmart sales. By examining correlations, such as those between weekly sales and various factors like unemployment rate, temperature, fuel price, and CPI, valuable insights are gleaned to inform strategic decision-making processes. This analytical approach enhances the understanding of how different factors interact and impact sales performance, facilitating the formulation of targeted strategies to optimize sales growth and profitability.

  Reveals nuanced dependencies between total weekly sales and various factors such as holidays, temperature, fuel price, and CPI, providing insights into the underlying drivers of sales performance.



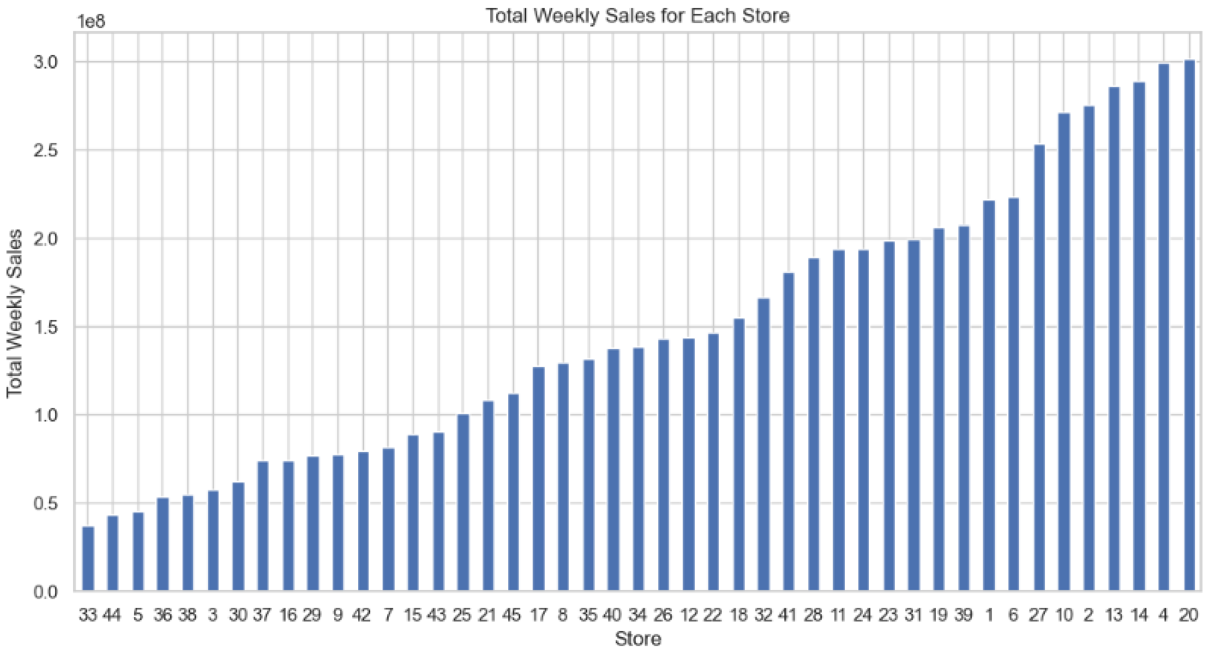
- **Scatter Plots**: Illustrating the absence of direct correlations between unemployment rate, temperature, CPI, and weekly sales, suggesting that these factors may have limited direct impact on sales dynamics



.

Upon examination of scatter plots depicting weekly sales against various features such as temperature, fuel price, CPI, and unemployment rate, it becomes evident that no discernible trend is observed. This absence of a clear trend suggests that there is no direct linear relationship between these features and weekly sales. In other words, the fluctuations in weekly sales cannot be solely attributed to any single factor among the features analyzed. Instead, the sales dynamics are likely influenced by a combination of factors or external variables not accounted for in this analysis. Consequently, further exploration and analysis may be warranted to uncover additional drivers of weekly sales and better understand the underlying factors contributing to sales fluctuations.

- **Bar Plots**: Bar plots revealed the top and worst performing sales based on which we can go and perform future predictions.



The analysis highlights significant disparities in weekly sales performance across various Walmart stores, with stores 20, 4, 14, 13, and 2 consistently outperforming others, while store 33 lags behind with notably lower sales figures. This discrepancy prompts a deeper dive into understanding the underlying factors contributing to the success of these top-performing stores. By examining factors such as location, customer demographics, store layout, promotional strategies, and operational efficiencies, we aim to uncover the unique attributes and strategies driving their exceptional sales performance. Additionally, exploring the reasons behind the poor sales performance of store 33 offers valuable insights into potential challenges or shortcomings that may need addressing. By dissecting the success stories of the top-performing stores and identifying areas for improvement in underperforming ones, we can develop targeted strategies and best practices to enhance overall sales performance across the Walmart retail chain. Through this approach, we strive to leverage insights from both success and failure to inform actionable recommendations that optimize sales strategies, drive revenue growth, and foster sustained success for all Walmart stores.

Statistical Analysis:

Augmented Dickey-Fuller (ADF) Test:

As part of the statistical analysis, we conducted an Augmented Dickey-Fuller (ADF) test to determine whether the Walmart sales data exhibits stationarity. Stationarity is a crucial assumption for time series analysis, indicating that the statistical properties of the data, such as mean and variance, remain constant over time. The ADF test evaluates the presence of a unit root in the time series data, with the null hypothesis being that the data is non-stationary. By applying the ADF test to our dataset, we obtained the test statistic along with the associated p-value. If the p-value is less than a predetermined significance level (commonly set at 0.05), we reject the null hypothesis, concluding that the data is stationary. Conversely, if the p-value exceeds the significance level, we fail to reject the null hypothesis, indicating non-stationarity. This analysis is essential for ensuring the validity of time series models and forecasting techniques, as stationarity facilitates the application of predictive models and enhances the reliability of future projections.

```
ADF test result for Store 20:
ADF Statistic: -5.3937386928548285
p-value: 3.4912952838128635e-06
Critical Values:
1%: -3.47864788917503
5%: -2.882721765644168
10%: -2.578065326612056
The data is stationary
```

```
ADF test result for Store 4:
ADF Statistic: -2.8793819840147084
p-value: 0.047798662236698805
Critical Values:
1%: -3.4793722137854926
5%: -2.8830370378332995
10%: -2.578233635380623
The data is stationary
```

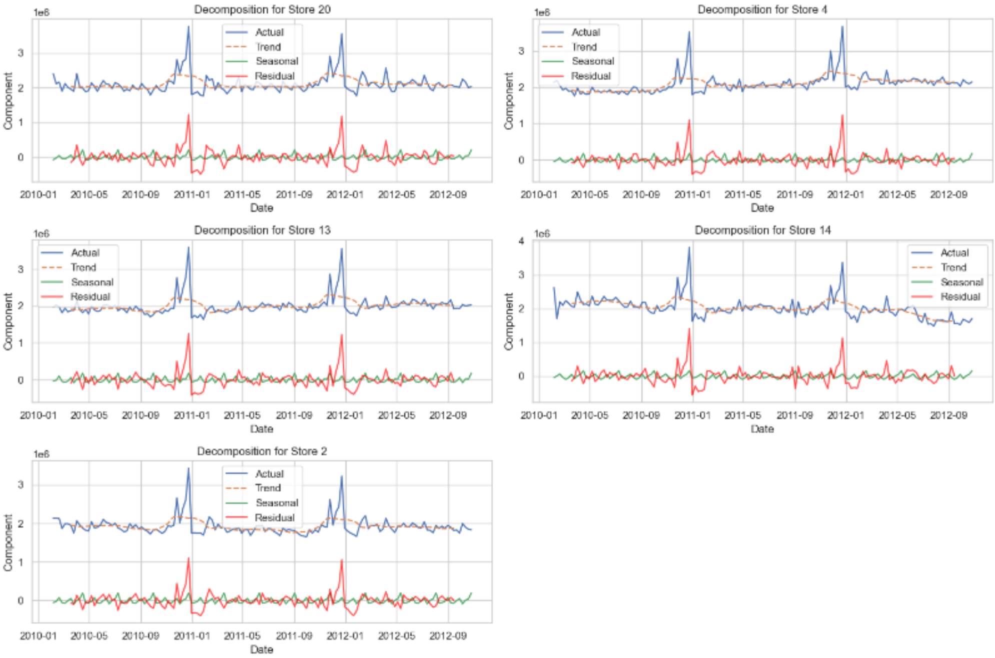
```
ADF test result for Store 13:
ADF Statistic: -5.502481711233357
p-value: 2.05644619369346e-06
Critical Values:
1%: -3.47864788917503
5%: -2.882721765644168
10%: -2.578065326612056
The data is stationary
```

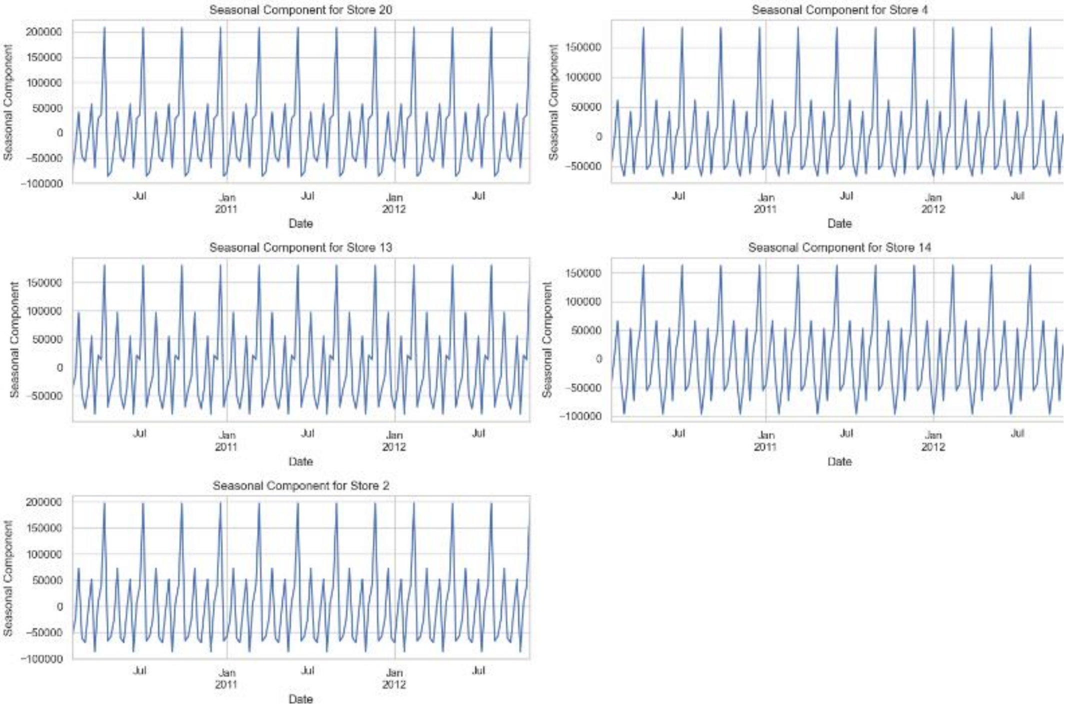
```
ADF test result for Store 14:
ADF Statistic: -2.7368866106752017
p-value: 0.06786986708375066
Critical Values:
1%: -3.4793722137854926
5%: -2.8830370378332995
10%: -2.578233635380623
The data is not stationary
```

```
ADF test result for Store 2:
ADF Statistic: -3.708862572618915
p-value: 0.0039902070890662795
Critical Values:
1%: -3.4793722137854926
5%: -2.8830370378332995
10%: -2.578233635380623
The data is stationary
```

Time Series Decomposition:

Time series decomposition is employed to decompose the weekly sales data into its constituent components, including trend, seasonality, and residual. This technique helps identify underlying patterns and trends in the data, facilitating better understanding and forecasting.





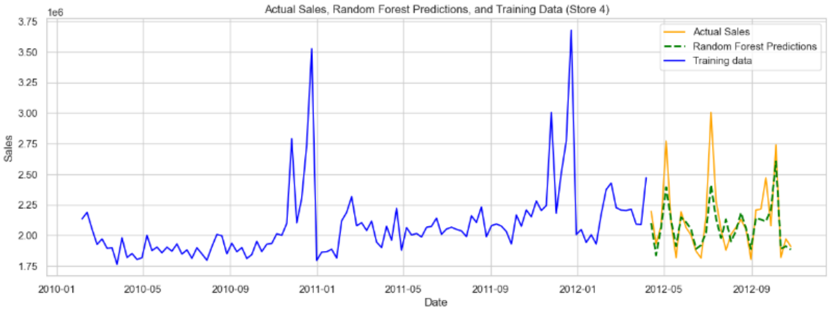
Model Evaluation:

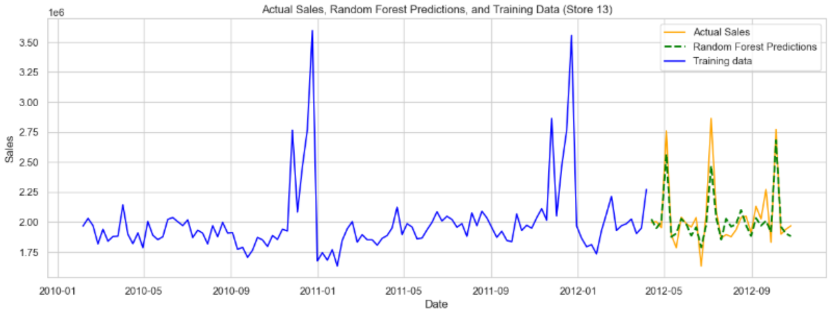
Harnessing the power of machine learning, predictive models such as Random Forest Regressor, ARIMA, and SARIMA are trained and evaluated to forecast weekly sales for Walmart's top-performing stores. Model evaluation metrics offer insights into the predictive performance and accuracy of these models, guiding strategic decision-making.

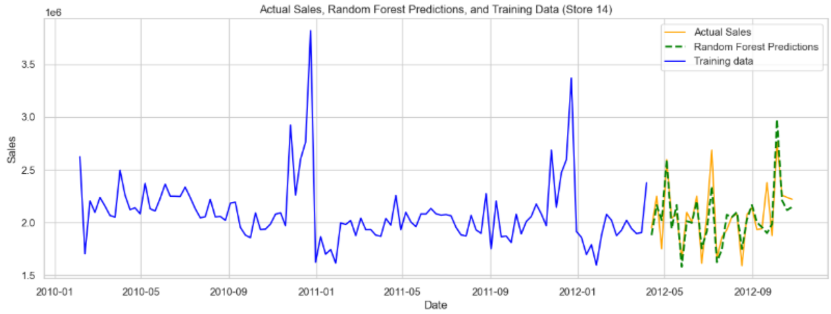
Random Forest Regressor:

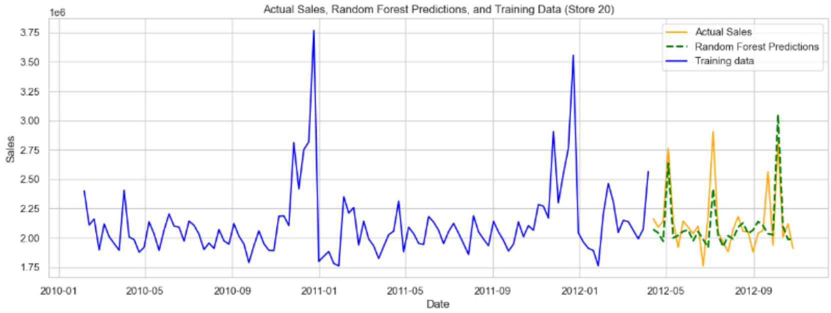
Random Forest Regressor is a powerful machine learning algorithm that leverages the strength of ensemble learning to make accurate predictions. It works by constructing multiple decision trees during training and outputting the average prediction of the individual trees. In this analysis, Random Forest Regressor is applied to forecast weekly sales, taking into account various features and factors influencing sales performance.

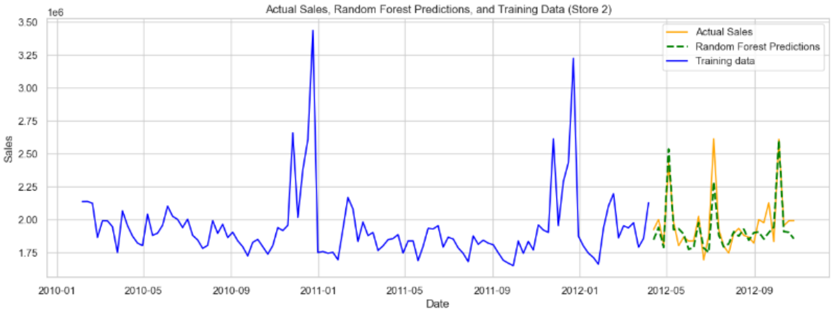
```
           Model Data_Type            MAE           MSE        RMSE  \
0  Random Forest  Store 20  131802.868076  3.196937e+10  363.046647
1  Random Forest   Store 4  114309.262472  2.911291e+10  338.096528
2  Random Forest  Store 13   88233.810783  1.418694e+10  297.041766
3  Random Forest  Store 14  107344.851734  2.242270e+10  327.635242
4  Random Forest   Store 2   81003.872338  1.093658e+10  284.611792

      Max_Diff      Mean_Diff    Min_Diff
0  529659.7122  131802.868076  16431.2420
1  592365.9355  114309.262472   6992.8766
2  396369.4810   88233.810783   2351.3071
3  476626.4874  107344.851734    395.5649
4  328171.3695   81003.872338   6774.3737
```


Actual Sales, Random Forest Predictions, and Training Data (Store 4)


Actual Sales, Random Forest Predictions, and Training Data (Store 13)


Actual Sales, Random Forest Predictions, and Training Data (Store 14)

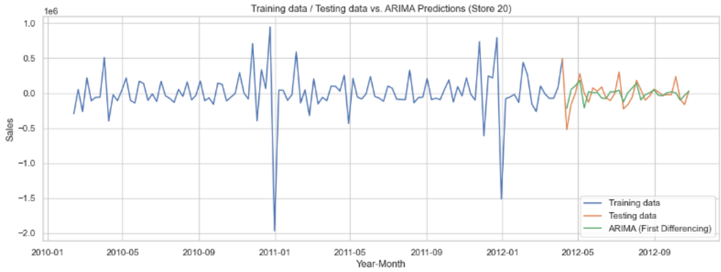

Actual Sales, Random Forest Predictions, and Training Data (Store 20)


Actual Sales, Random Forest Predictions, and Training Data (Store 2)

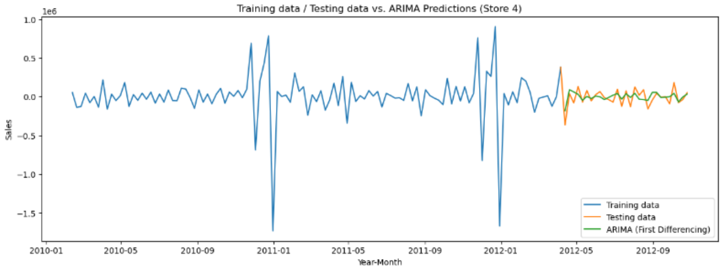
ARIMA (Autoregressive Integrated Moving Average):

ARIMA is a popular time series forecasting model that captures the autocorrelation and seasonality present in the data. It comprises three components: auto-regression (AR), differencing (I), and moving average (MA). By analysing past data points and identifying patterns, ARIMA can make predictions for future values. This model is particularly useful for capturing the underlying patterns and trends in time series data.

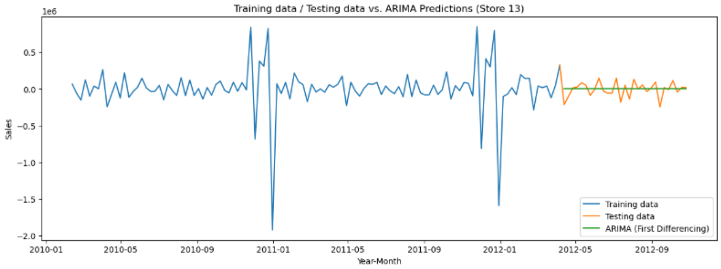
```
   Model Data_Type           MAE          MSE       RMSE  Min_Diff (%)  \
0  ARIMA  Store 20  102329.746400  1.735231e+10  319.890210     7.33244
1
1  ARIMA   Store 4   66425.661414  6.648409e+09  257.731763    13.202427

2  ARIMA  Store 13   75757.147040  9.897236e+09  275.240162    96.67534
8
3  ARIMA  Store 14  120431.397030  2.863904e+10  347.032271    22.17066
6
4  ARIMA   Store 2   76724.236583  8.624386e+09  276.991402     0.309464


   Mean_Diff (%)  Max_Diff (%)
0    445.842369   5164.340939
1    108.725871    478.564836
2    100.408870    115.300780
3    100.138815    173.320813
4    193.278235    922.919130
```



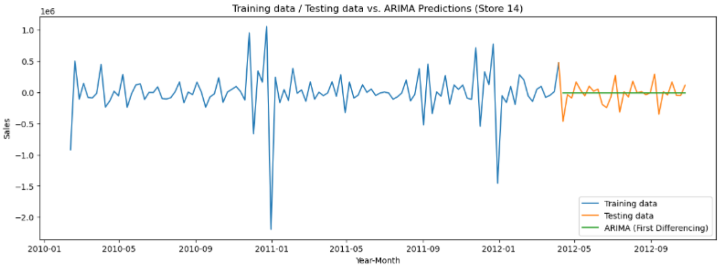
Best model: ARIMA(2,0,3)(0,0,0)[0]
Total fit time: 0.847 seconds



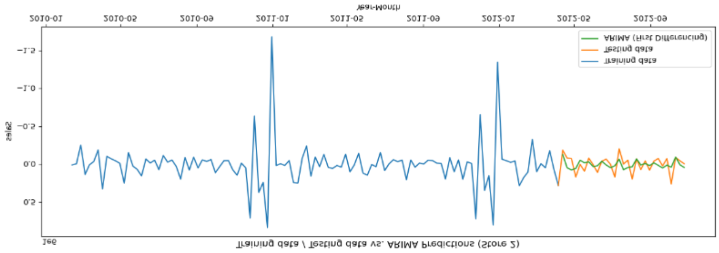
Best model: ARIMA(0,0,5)(0,0,0)[0]
Total fit time: 0.372 seconds



Best model: ARIMA(0,0,5)(0,0,0)[0]
Total fit time: 0.375 seconds



Best model: ARIMA(2,0,3)(0,0,0)[0]
Total fit time: 0.442 seconds

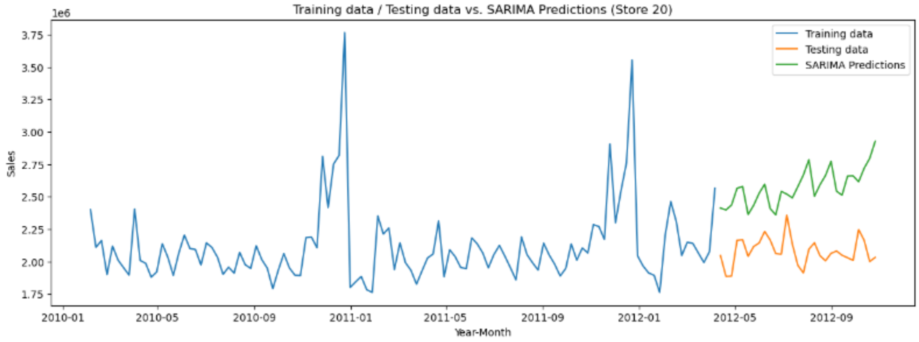


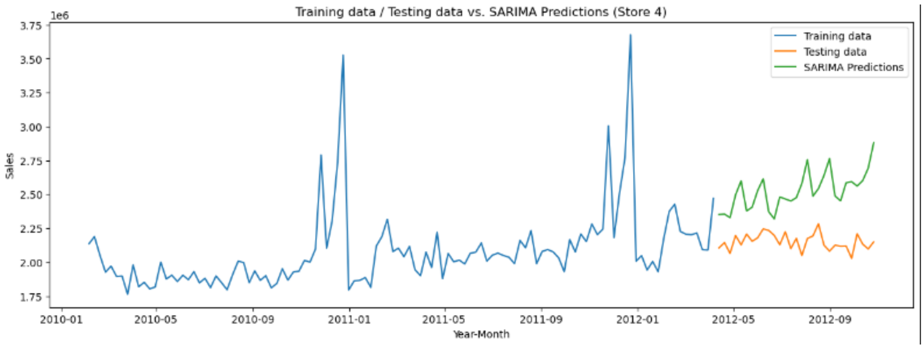
Best model: ARIMA(5,0,0)(0,0,0)[0]
Total fit time: 4.097 seconds

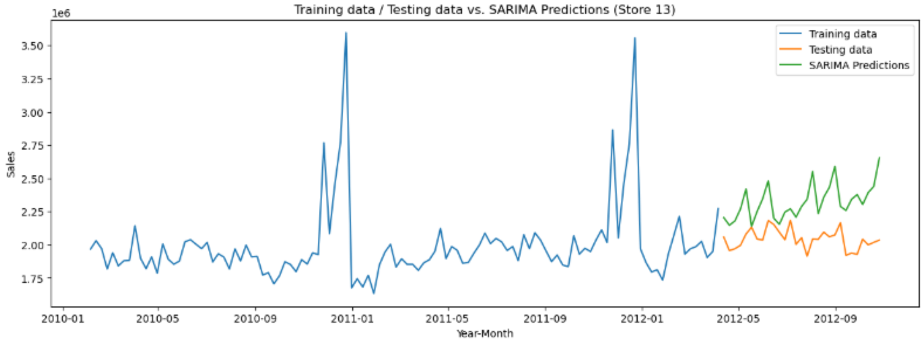
SARIMA (Seasonal Autoregressive Integrated Moving Average):

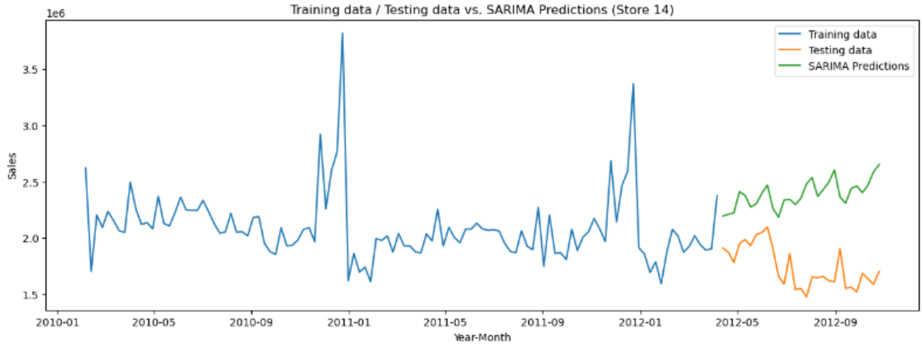
SARIMA is an extension of the ARIMA model that incorporates seasonal components into the forecasting process. It takes into account seasonal variations in the data and adjusts the model parameters accordingly. SARIMA is well-suited for time series data with distinct seasonal patterns, making it an effective tool for forecasting sales data that exhibit seasonality.

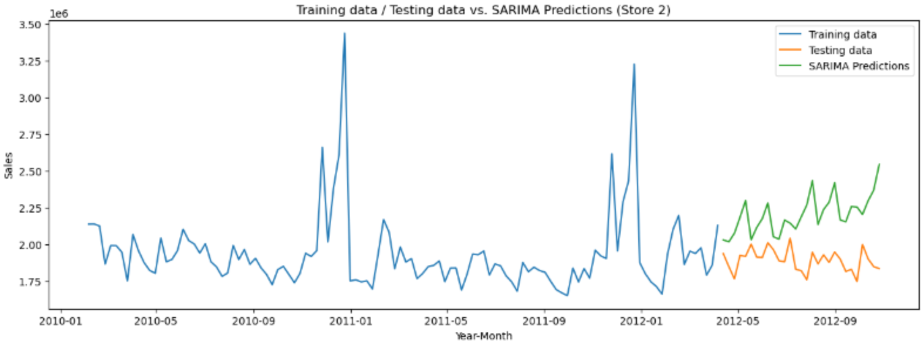
```
    Model Data_Type            MAE           MSE          RMSE   Min_Diff  \
0  SARIMA  Store 20  494180.922290  2.748463e+11  524257.880857  6774.
3737
1  SARIMA   Store 4  374903.033965  1.649439e+11  406132.808174  6774.
3737
2  SARIMA  Store 13  279622.174641  1.000579e+11  316319.358761  6774.
3737
3  SARIMA  Store 14  644087.606296  4.751636e+11  689321.113512  6774.
3737
4  SARIMA   Store 2  312457.050578  1.210141e+11  347870.870955  6774.
3737

     Mean_Diff     Max_Diff
0  81003.872338  328171.3695
1  81003.872338  328171.3695
2  81003.872338  328171.3695
3  81003.872338  328171.3695
4  81003.872338  328171.3695
```


Training data / Testing data vs. SARIMA Predictions (Store 20)
1e6
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
Sales
2010-01
2010-05
2010-09
2011-01
2011-05
2011-09
2012-01
2012-05
2012-09
Year-Month
Training data
Testing data
SARIMA Predictions


Training data / Testing data vs. SARIMA Predictions (Store 4)
1e6
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
Sales
2010-01
2010-05
2010-09
2011-01
2011-05
2011-09
2012-01
2012-05
2012-09
Year-Month
Training data
Testing data
SARIMA Predictions


Training data / Testing data vs. SARIMA Predictions (Store 13)
1e6
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
Sales
2010-01
2010-05
2010-09
2011-01
2011-05
2011-09
2012-01
2012-05
2012-09
Year-Month
Training data
Testing data
SARIMA Predictions


Training data / Testing data vs. SARIMA Predictions (Store 14)
1e6
3.5
3.0
2.5
2.0
1.5
Sales
2010-01
2010-05
2010-09
2011-01
2011-05
2011-09
2012-01
2012-05
2012-09
Year-Month
Training data
Testing data
SARIMA Predictions


Training data / Testing data vs. SARIMA Predictions (Store 2)
1e6
3.50
3.25
3.00
2.75
2.50
2.25
2.00
1.75
Sales
2010-01
2010-05
2010-09
2011-01
2011-05
2011-09
2012-01
2012-05
2012-09
Year-Month
Training data
Testing data
SARIMA Predictions

Inferences and Insights:

The analysis unveils several significant insights into the dynamics of Walmart's sales, shedding light on critical aspects that influence sales performance.

Firstly, through comprehensive examination, we identified various factors that play a pivotal role in shaping sales outcomes. Factors such as economic indicators, including unemployment rate, consumer price index (CPI), as well as environmental variables like temperature and fuel price, emerged as key influencers impacting sales trends.

Interestingly, despite the thorough exploration of these factors, the analysis revealed the absence of direct correlations between certain variables and weekly sales. This observation suggests that the relationship between these factors and sales performance may be more nuanced or influenced by additional variables not considered in this analysis.

Furthermore, the analysis delved into uncovering seasonal patterns and trends inherent in the sales data, providing valuable insights into the cyclical nature of consumer behaviour and purchasing patterns across different time periods.

Lastly, rigorous evaluation of predictive models' performance and accuracy in forecasting sales offered valuable insights into the efficacy of these models in capturing and predicting sales trends. By synthesizing these insights, Walmart can gain a deeper understanding of the multifaceted factors driving sales dynamics, enabling informed decision-making and the formulation of targeted strategies to optimize sales performance and foster sustainable growth.

Future Possibilities and Recommendations:

Future iterations of the analysis could delve into multifaceted avenues to augment our understanding of Walmart's sales dynamics and refine predictive capabilities:

Integration of Additional External Factors: Broadening the scope by incorporating an extensive range of external variables beyond economic indicators and environmental factors can enrich the predictive models. Integrating demographic data, market trends, and social factors into the analysis may offer deeper insights into consumer behaviour and enhance the accuracy of sales forecasts.

Deployment of Advanced Machine Learning Techniques: Embracing cutting-edge machine learning methodologies, such as deep learning algorithms, ensemble methods, and neural networks, presents an exciting opportunity to elevate predictive performance. These advanced techniques have the potential to uncover intricate patterns and relationships within the sales data, enabling more nuanced and precise sales predictions.

Exploration of Dynamic Pricing Strategies and Personalized Marketing Approaches: Investigating innovative strategies, such as dynamic pricing algorithms and personalized marketing initiatives, can significantly impact sales growth and customer engagement. Dynamic pricing strategies that adapt prices in real-time based on demand fluctuations and market dynamics can optimize revenue generation and enhance competitiveness in the retail landscape. Similarly, personalized marketing approaches tailored to individual customer preferences and shopping behaviours can foster deeper customer relationships, increase brand loyalty, and drive repeat purchases.

By embracing these multifaceted avenues for exploration and innovation, Walmart can fortify its position as a leader in data-driven decision-making and strategic planning. These initiatives are poised to drive sustained sales growth, enhance customer satisfaction, and bolster Walmart's competitive advantage in the dynamic retail industry landscape.

Conclusion:

In conclusion, this thorough analysis offers valuable insights into Walmart's sales patterns, providing actionable recommendations for enhancing its sales strategies. By examining various factors influencing sales performance, such as economic indicators and environmental variables, we gained a deeper understanding of the complex dynamics at play. Although direct correlations between certain factors and weekly sales were not evident, seasonal patterns and trends were uncovered, providing valuable insights into consumer behaviour. Additionally, through the evaluation of predictive models, we assessed their effectiveness in forecasting sales accurately. Moving forward, integrating additional external factors, deploying advanced machine learning techniques, and exploring dynamic pricing strategies and personalized marketing approaches present promising opportunities for Walmart to optimize its sales strategies and foster sustained growth. Overall, this analysis equips Walmart with the insights needed to make informed decisions and drive continued success in the competitive retail landscape.

References:

www.rit.edu/ischoolprojects/sites/rit.edu.ischoolprojects/Rashmi_Jeswani_Capstone.pdf

github.com/abhinav-bhardwaj/Walmart-Sales-Time-Series-Forecasting-Using-Machine-Learning