CAPSTONE PROJECT SUBMISSION REPORT

Report Title:

Predictive Analysis of Walmart Sales Using Facebook Prophet

Presented to : INTELLIPAAT

Presented by:

NIKHIL KUMAR
Batch - Advanced Certification in Data Science and Artificial Intelligence
Registered mail id - nikhilmanav21999@gmail.com

Table of Contents

Problem Statement

The Walmart 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 12 weeks.

Project Objective

The objective of this project is to employ data analysis and predictive modeling techniques to forecast Walmart's sales for the next 12 weeks, utilizing the Facebook Prophet model. The project seeks to provide insights into sales trends, identify top-performing stores, and enable data-driven decision-making for resource allocation and strategic planning.

Data Description

The project utilizes a comprehensive dataset containing information for 45 Walmart stores, each with 143 weekly records, resulting in a total of 6,435 data points. The dataset includes features such as store information, date, weekly sales, holiday flags, temperature, fuel prices, Consumer Price Index (CPI), and unemployment rates. This rich dataset forms the basis for the predictive analysis and insights presented in this report.

The image from the Python notebook file that is attached below displays the statistical description of the dataset that was given to us.

```
df.describe()
```

	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

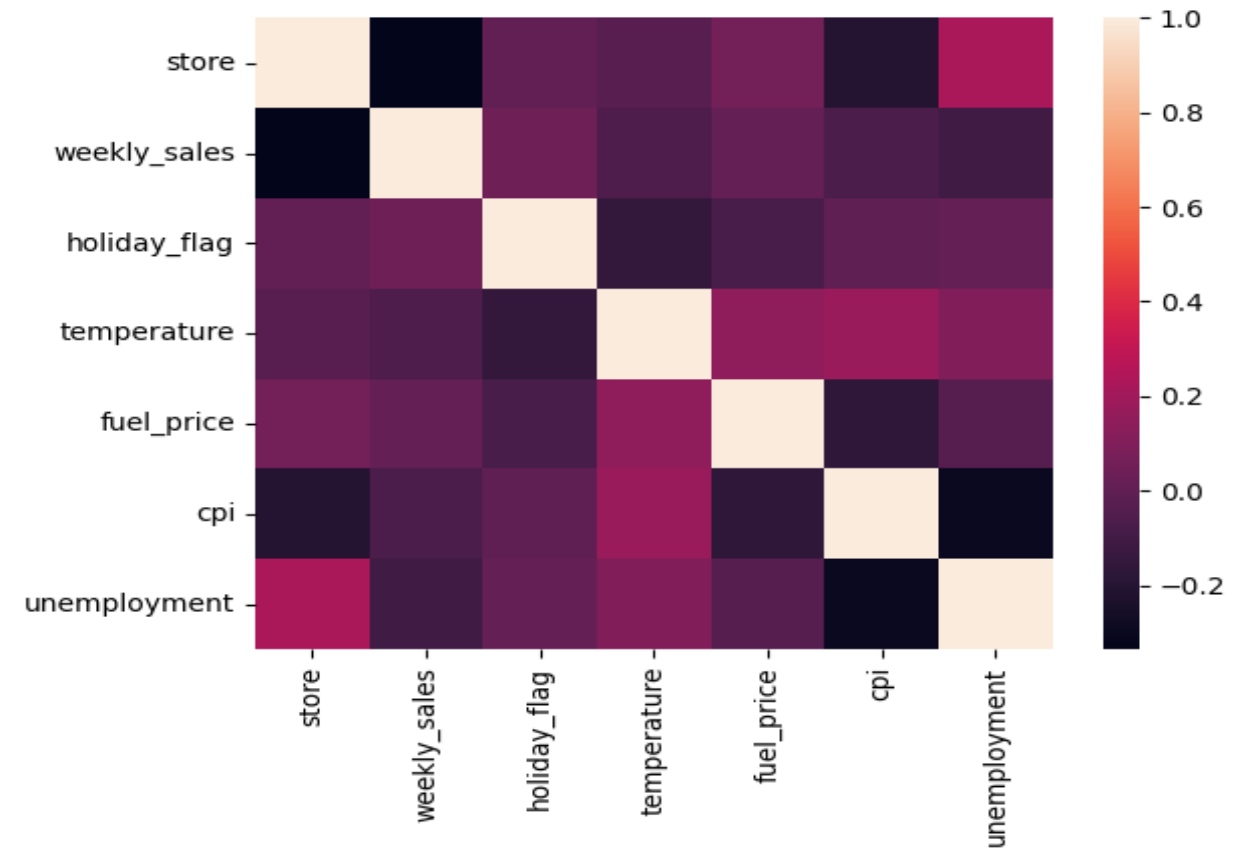
Exploratory Data Analysis:

- Overview of the exploratory data analysis (EDA) process conducted to gain insights into the dataset.
- Summary statistics, including mean, median, and standard deviation for relevant features.
- Visualization of key trends, including sales distribution, store-wise performance, and sales variation over time.
- Identification of any outliers or anomalies in the data.

Seasonality and Correlation Analysis:

- Investigation into seasonality patterns within the dataset, such as identifying recurring trends or periodic fluctuations.
- Analysis of the impact of holidays on sales and any observed patterns related to these events.
- Evaluation of feature correlations, including correlation coefficients, such as the relationship between weekly sales and unemployment rates.

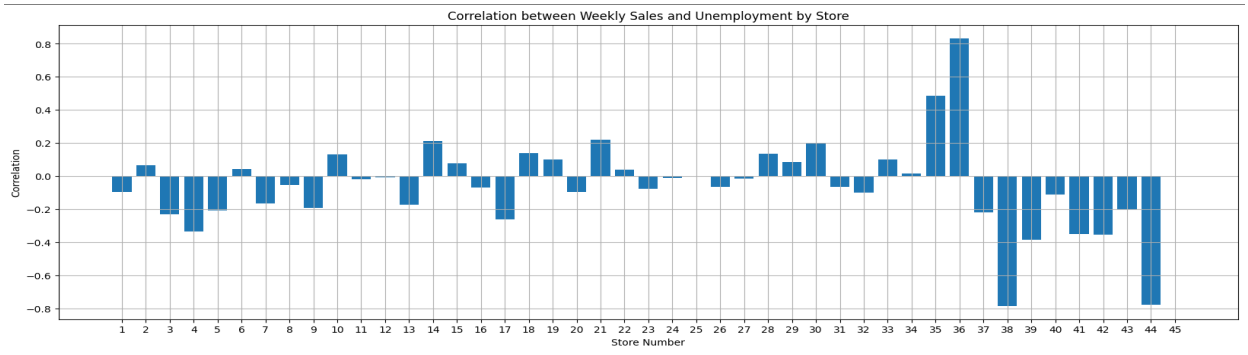
Correlation Matrix Snippet-



Insights and Findings

a. **How are the weekly sales affected by the unemployment rate and which stores are suffering the most?**

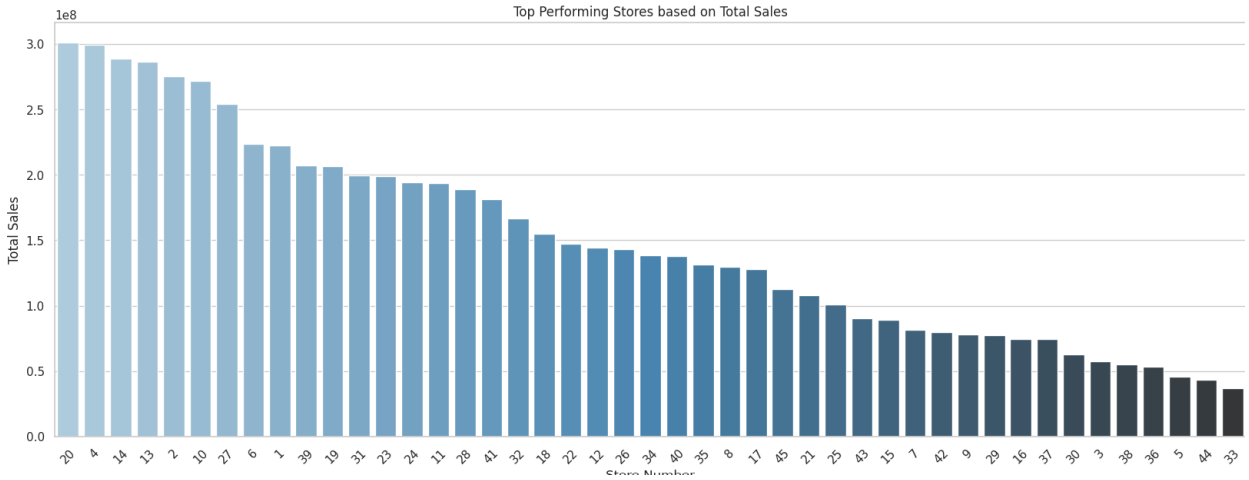
We checked the correlation for each store's weekly sales with their respective unemployment ratio and plotted the below graph to find out the conclusion.



From the above graph we can conclude that yes Weekly_sales are getting affected by the Unemployment Rate. It has been noticed that if the unemployment rate is higher the weekly sales go down. And the stores that are most affected are- Store no, 36, 35,21,14.

b. Top performing stores according to the historical data

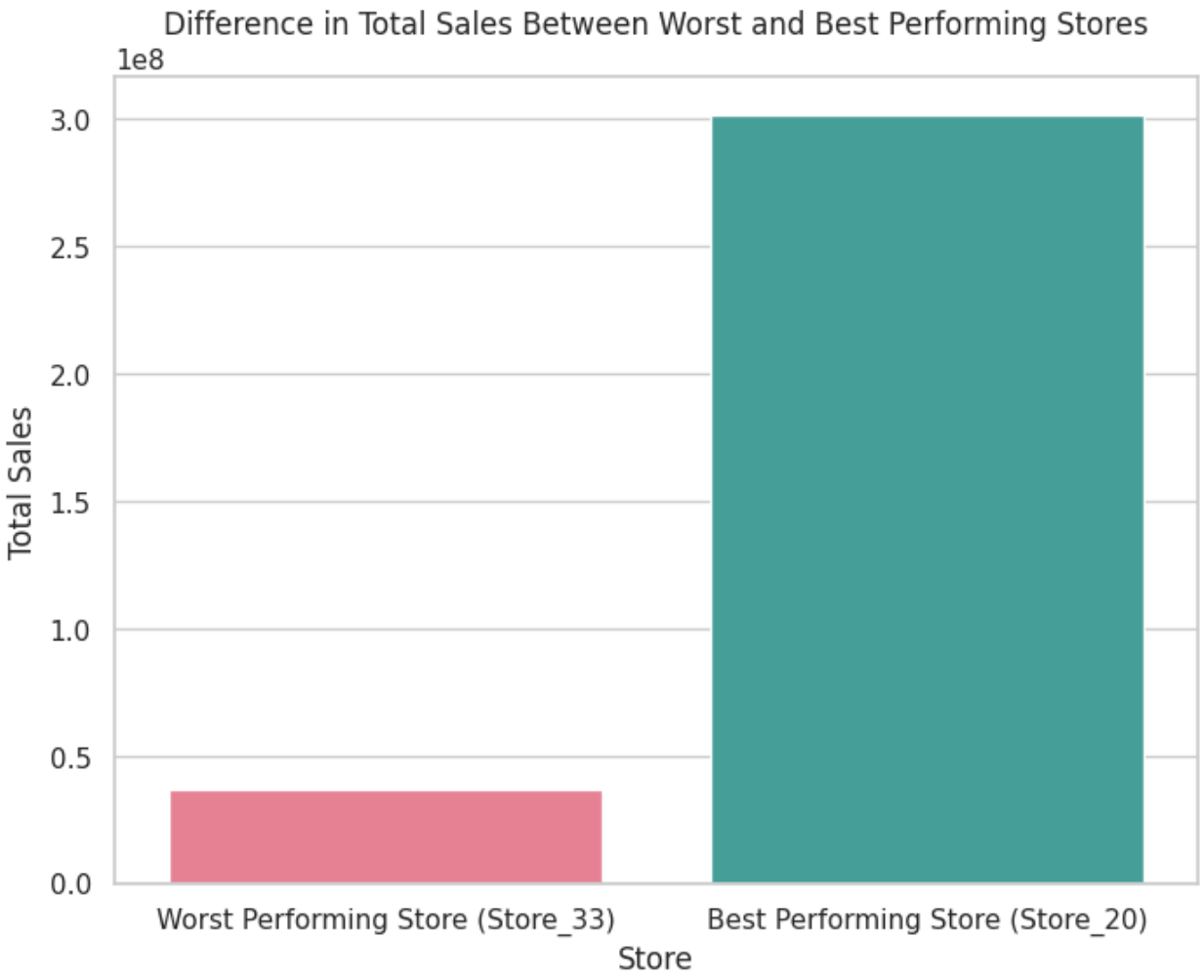
We created a bar chart to visualize the top-performing stores and the findings came as shown below.



We can see from the above graph that the top performing stores are store no.- 20, 4, 14, 13, 2.

c. The worst performing store, and how significant is the difference between the highest and lowest performing stores.

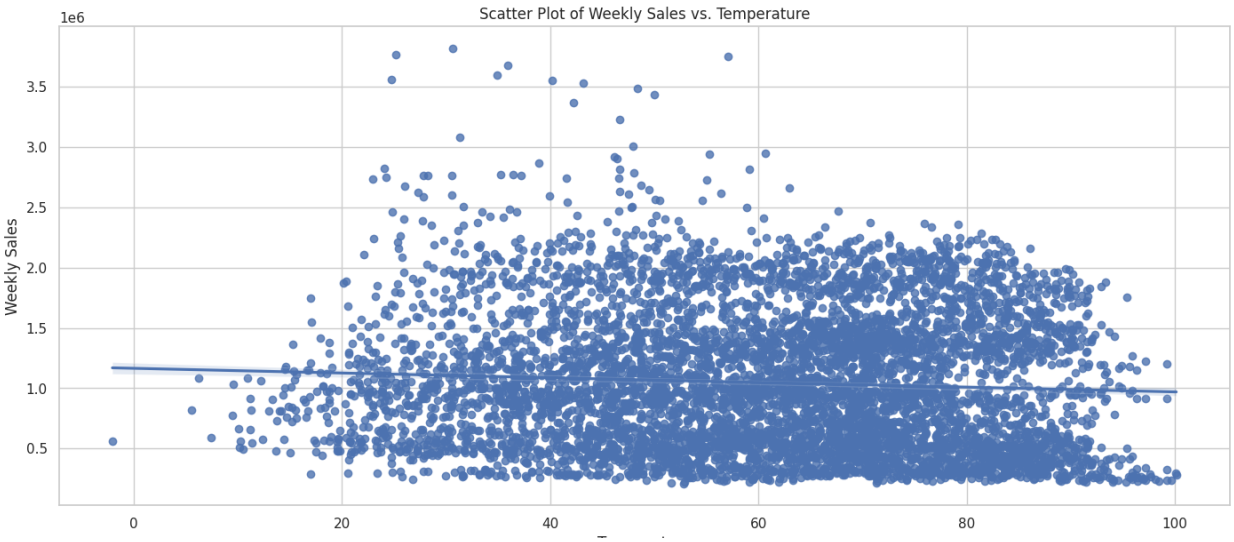
We checked the difference in Total Sales Between Worst and Best Performing Stores and the findings came out as shown below.



We can conclude that store 33 is the worst performing store and store 20 is the best performing and the difference between their sales is 264237570.5

d. Checking does temperature affect the weekly sales in any manner?

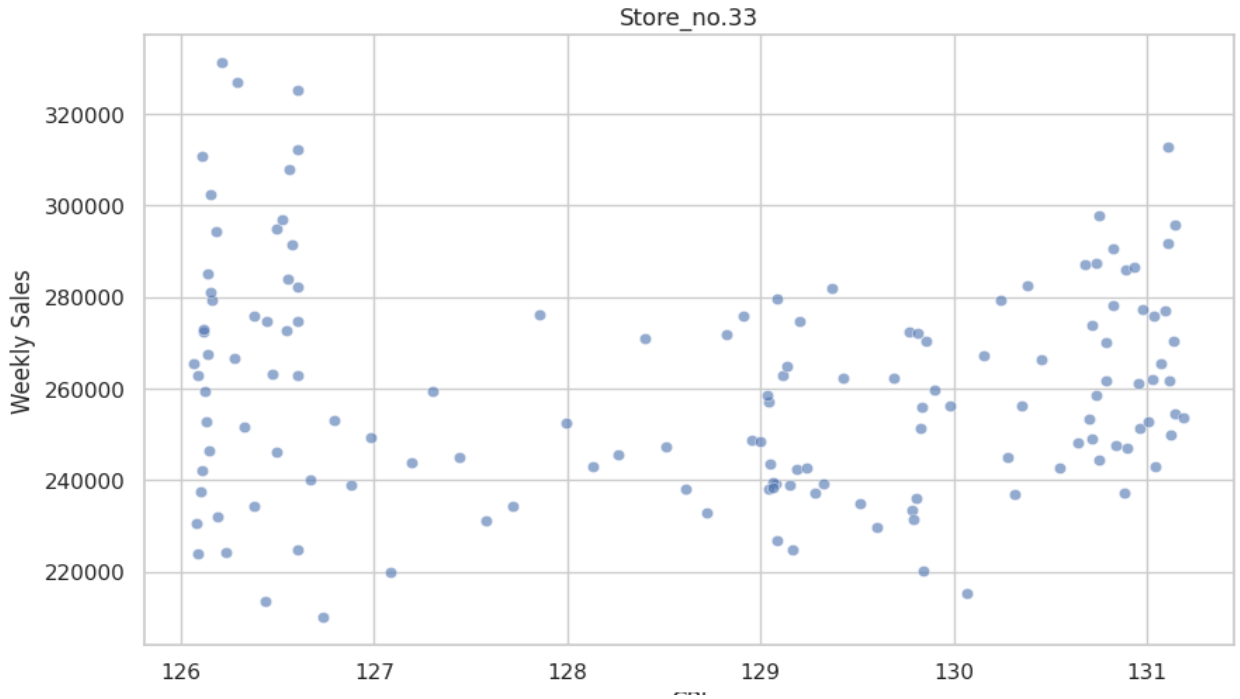
We plotted a scatter regression plot to see any distribution of weekly_sales with respect to temperature range and the findings are presented below.

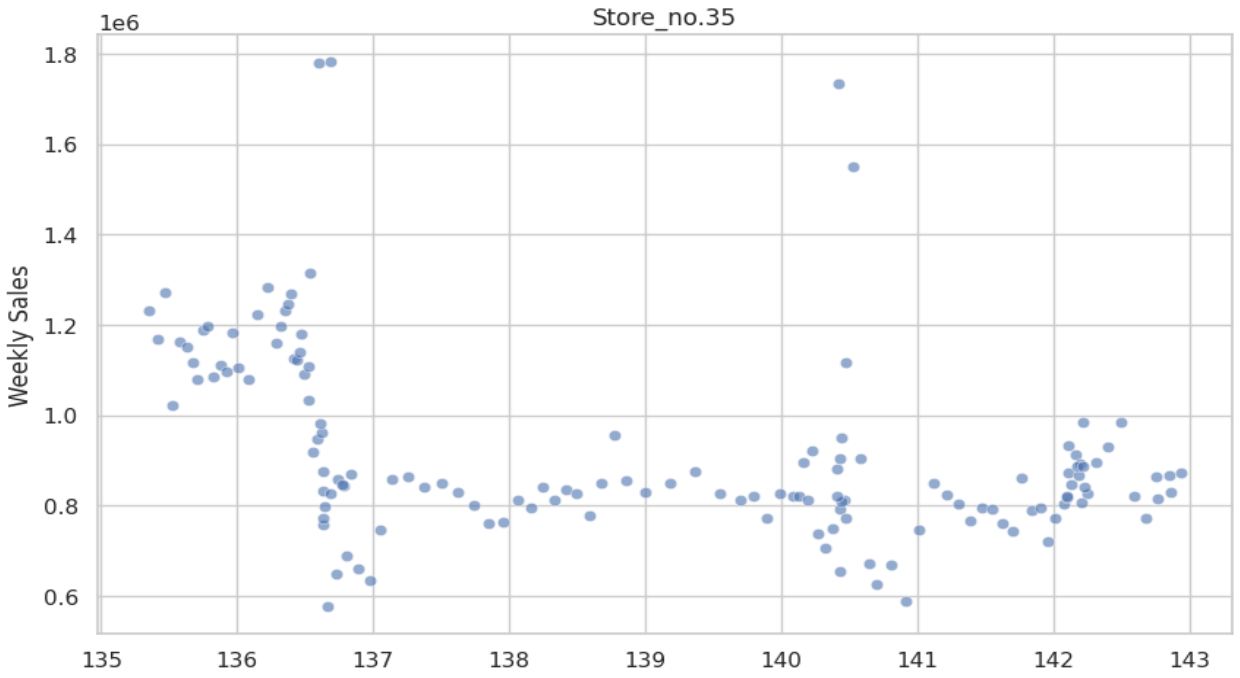


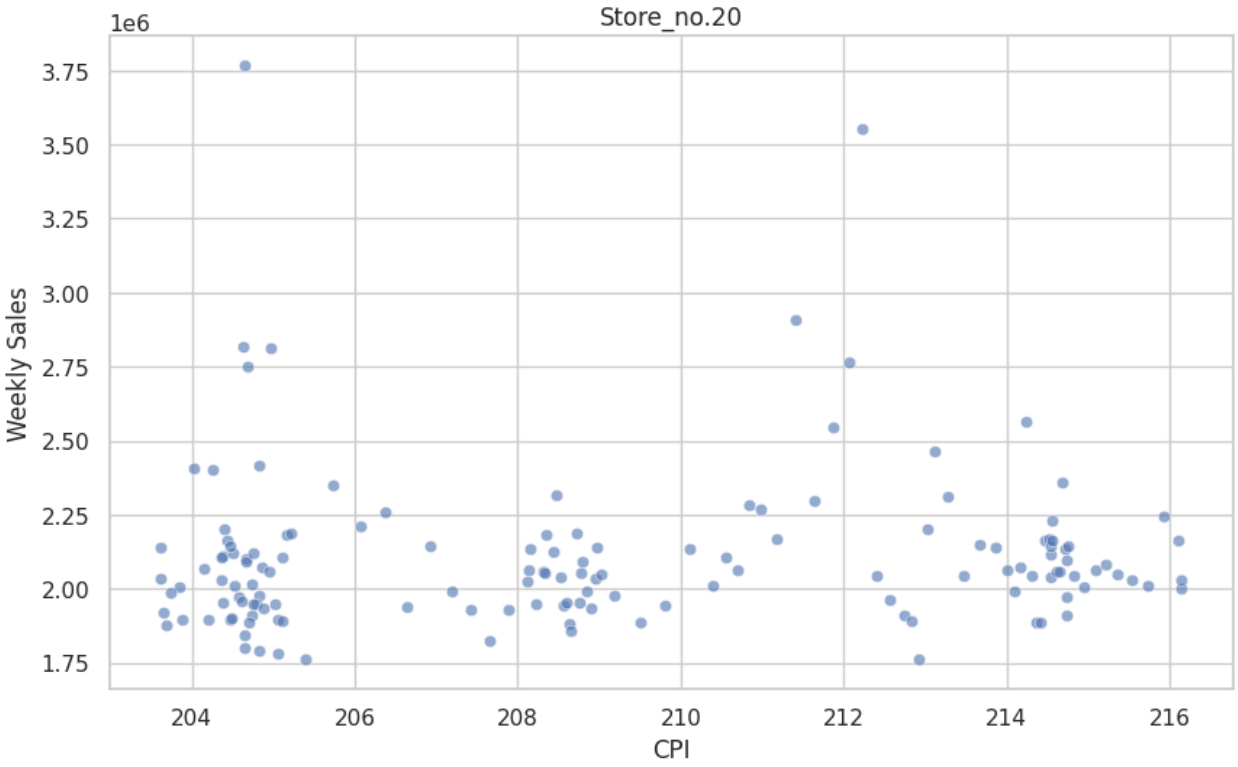
We can see from the aforementioned graph that the weekly sales distribution is very stable for a temperature range of 40 to 80 F. However, the weekly sales are low if the temperature falls below 20 F or rises beyond 80 F.

e. **Checking how the Consumer Price index affects the weekly sales of various stores?**

Below is the snippet of a few stores showing distribution of CPI in their respective region with their weekly sales.


Store_no.33
Weekly Sales
320000
300000
280000
260000
240000
220000
126
127
128
129
130
131

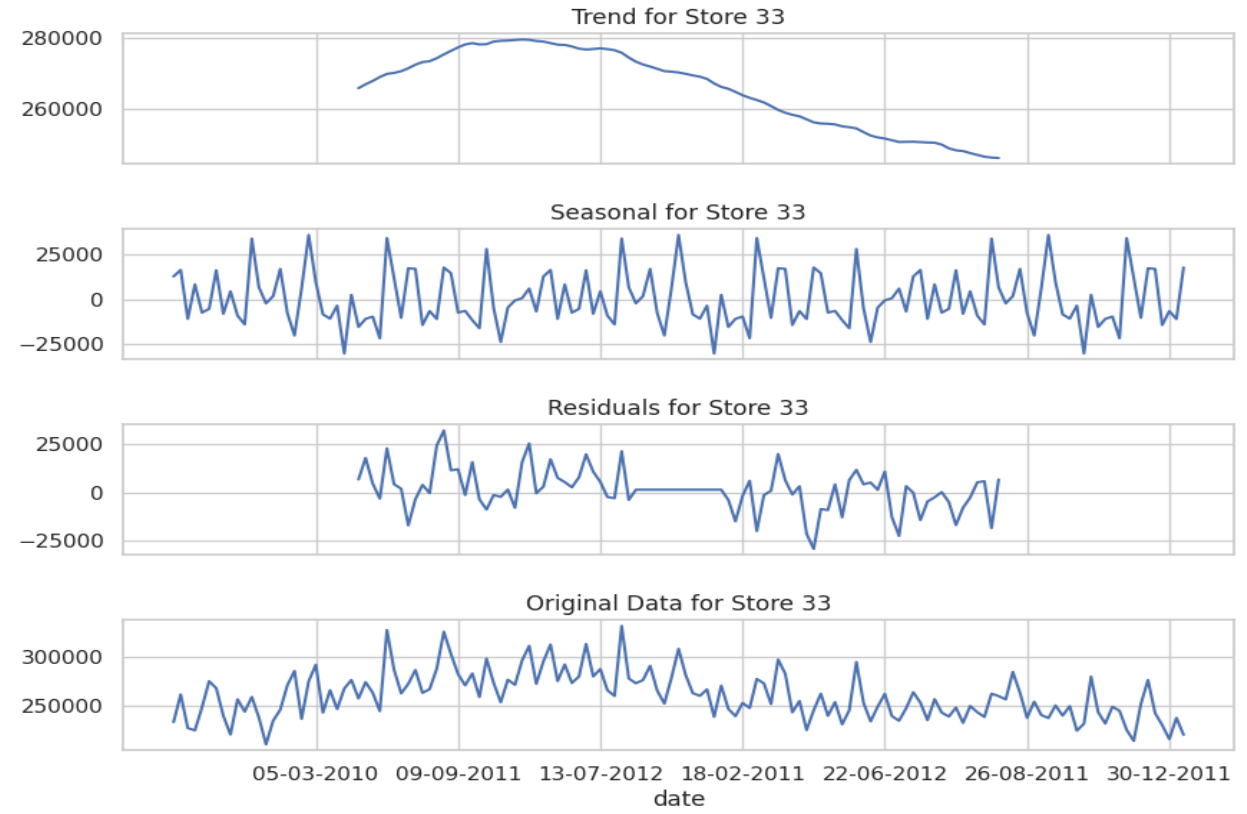

Store_no.35
1e6
Weekly Sales
1.8
1.6
1.4
1.2
1.0
0.8
0.6
135
136
137
138
139
140
141
142
143

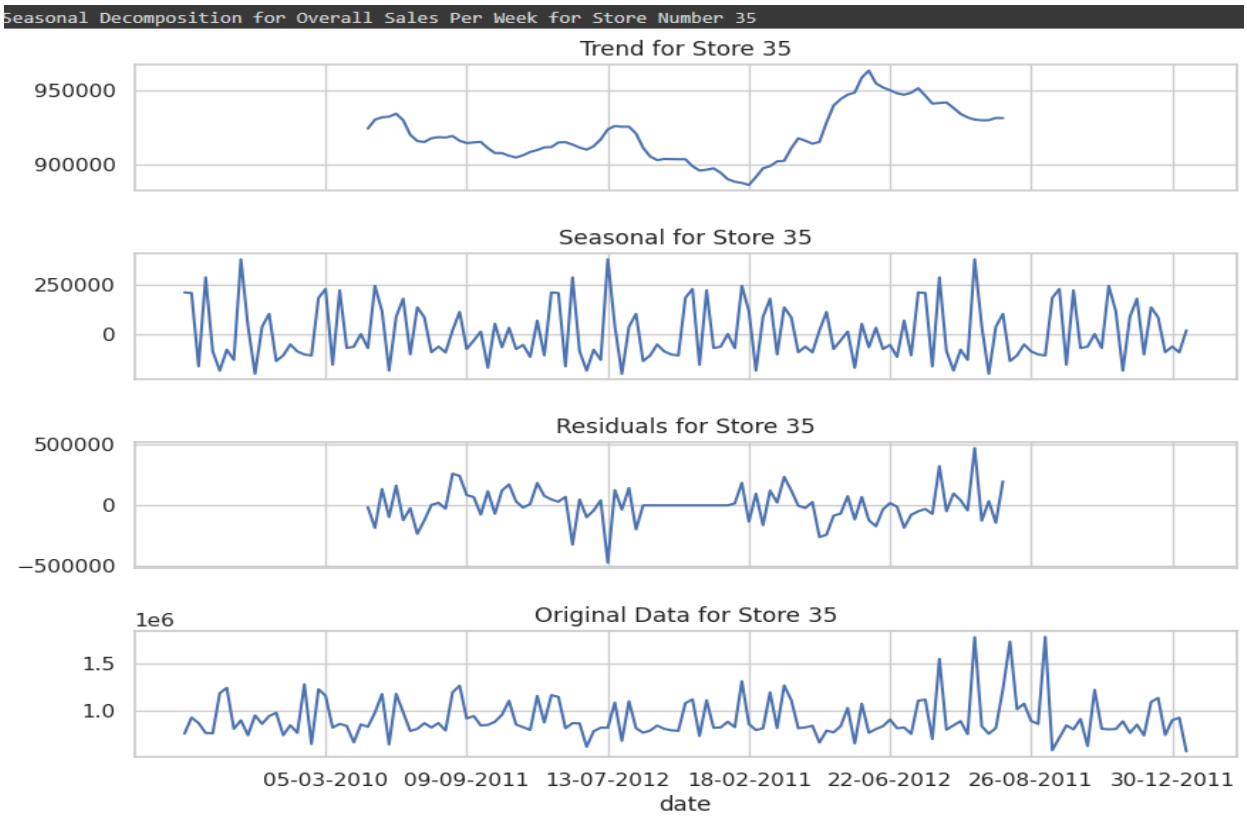


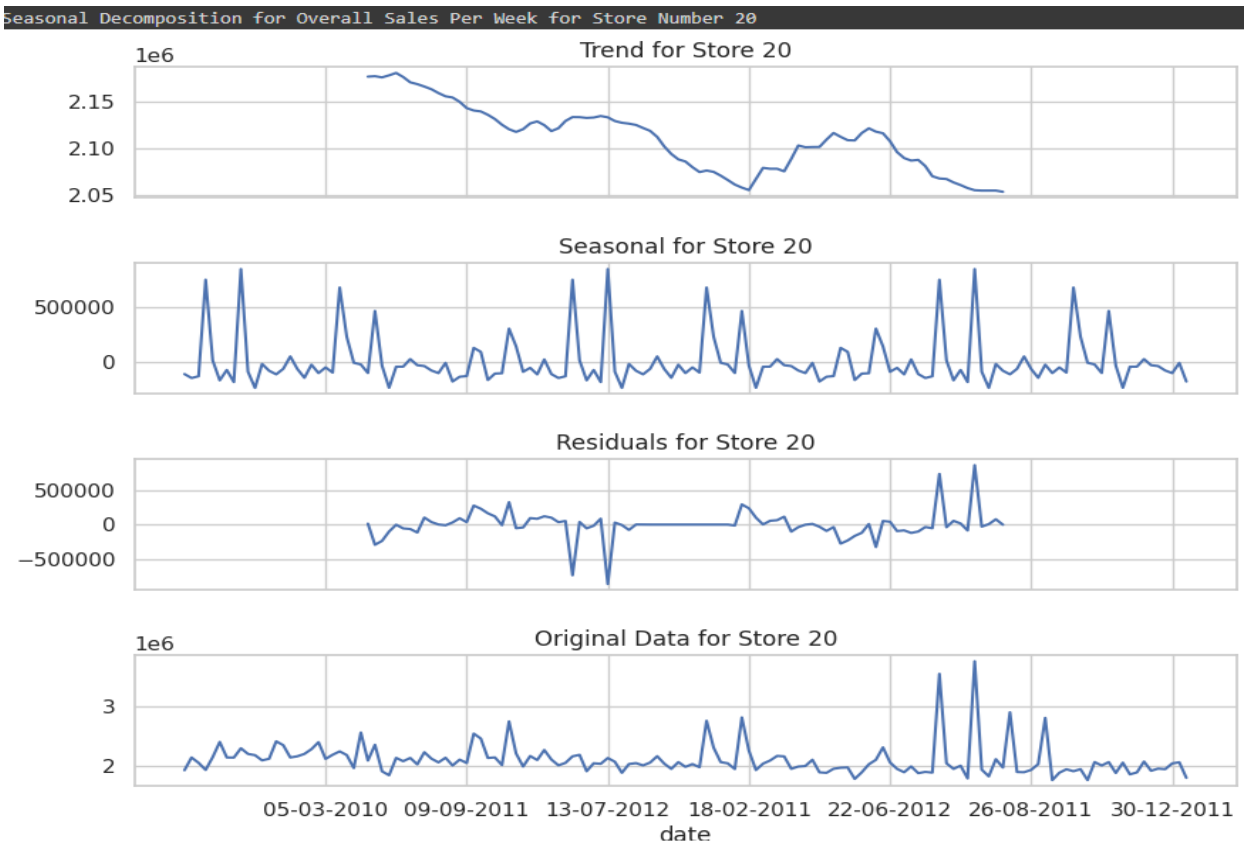
f. Store's weekly sales seasonal trend.

Below is the snippet of a few stores showing seasonality decomposition of their weekly sales.

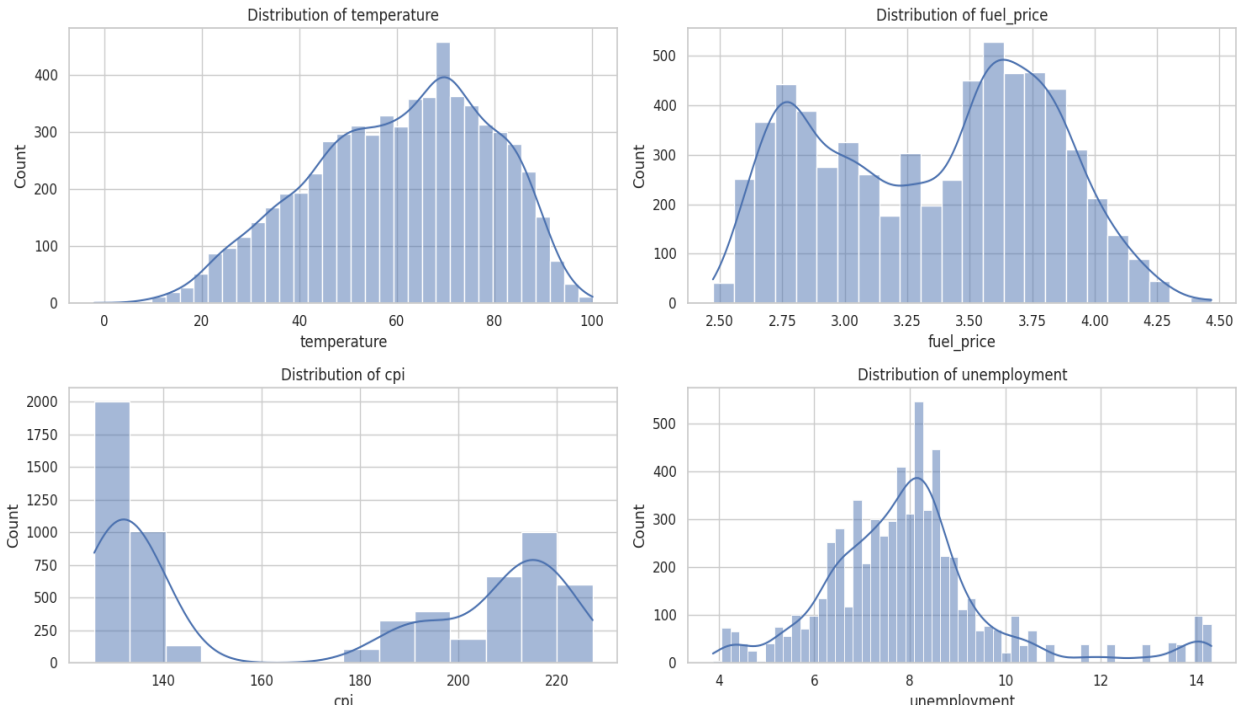
Seasonal Decomposition for Overall Sales Per Week for Store Number 33




Seasonal Decomposition for Overall Sales Per Week for Store Number 35
Trend for Store 35
950000
900000
Seasonal for Store 35
250000
0
Residuals for Store 35
500000
0
−500000
Original Data for Store 35
1e6
1.5
1.0
05-03-2010
09-09-2011
13-07-2012
18-02-2011
22-06-2012
26-08-2011
30-12-2011
date


Seasonal Decomposition for Overall Sales Per Week for Store Number 20
1e6
Trend for Store 20
2.15
2.10
2.05
Seasonal for Store 20
500000
0
Residuals for Store 20
500000
0
−500000
Original Data for Store 20
1e6
3
2
05-03-2010
09-09-2011
13-07-2012
18-02-2011
22-06-2012
26-08-2011
30-12-2011
date

g. Distribution of all the non-important numeric features



Data Preprocessing

In this project, the dataset was subjected to a streamlined data preprocessing phase, as the initial dataset exhibited a relatively clean and structured format. The primary focus of data preprocessing revolved around refining the dataset's existing structure and enhancing its suitability for predictive modeling.

The key data preprocessing steps undertaken included:

1.Date Format Conversion:

The 'Date' column in the dataset was reformatted to a consistent date format, ensuring uniformity and ease of analysis. This step involved converting date strings to datetime objects for accurate time-based analysis.

2.Feature Extraction:

To simplify and streamline the dataset for modeling, the following essential features were extracted:

- Store Information: Extracting store-specific data to enable store-wise analysis and forecasts.
- Weekly Sales: The primary target variable representing the weekly sales figures.
- Date Attributes: Extracting relevant date attributes such as day of the week and month to capture temporal patterns.

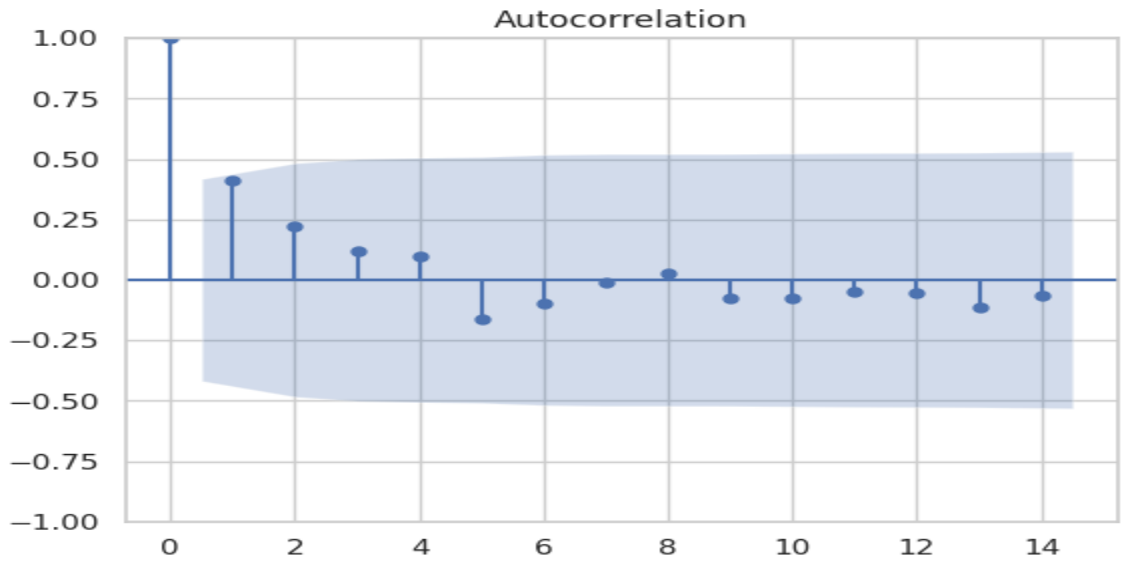
These preprocessing steps were designed to refine the dataset for predictive modeling while retaining its core structure and data integrity. By focusing on date format standardization and the extraction of critical features, the dataset was prepared for in-depth analysis and forecasting. This approach allowed for efficient and effective modeling, producing accurate predictions for Walmart's sales data.

The streamlined data preprocessing steps underscore the importance of maintaining data integrity and consistency while adapting the dataset to the specific requirements of the analysis and modeling tasks.

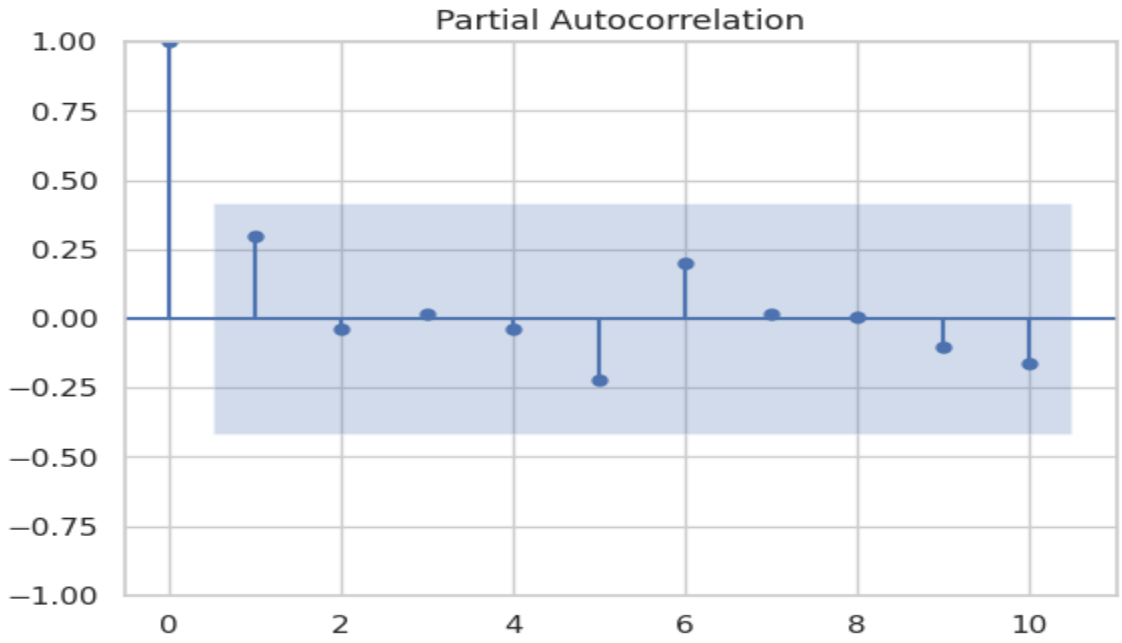
Choosing the Algorithm and Model Building

Exploration with ARIMA:
To employ the ARIMA model effectively, we embarked on the crucial task of identifying the model's hyperparameters, notably the order parameters (p, d, q). We conducted an in-depth analysis using the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) to determine the appropriate values. However, despite our rigorous efforts, the ARIMA model yielded unsatisfactory results, failing to provide promising sales forecasts.

ACF Plot (q) -

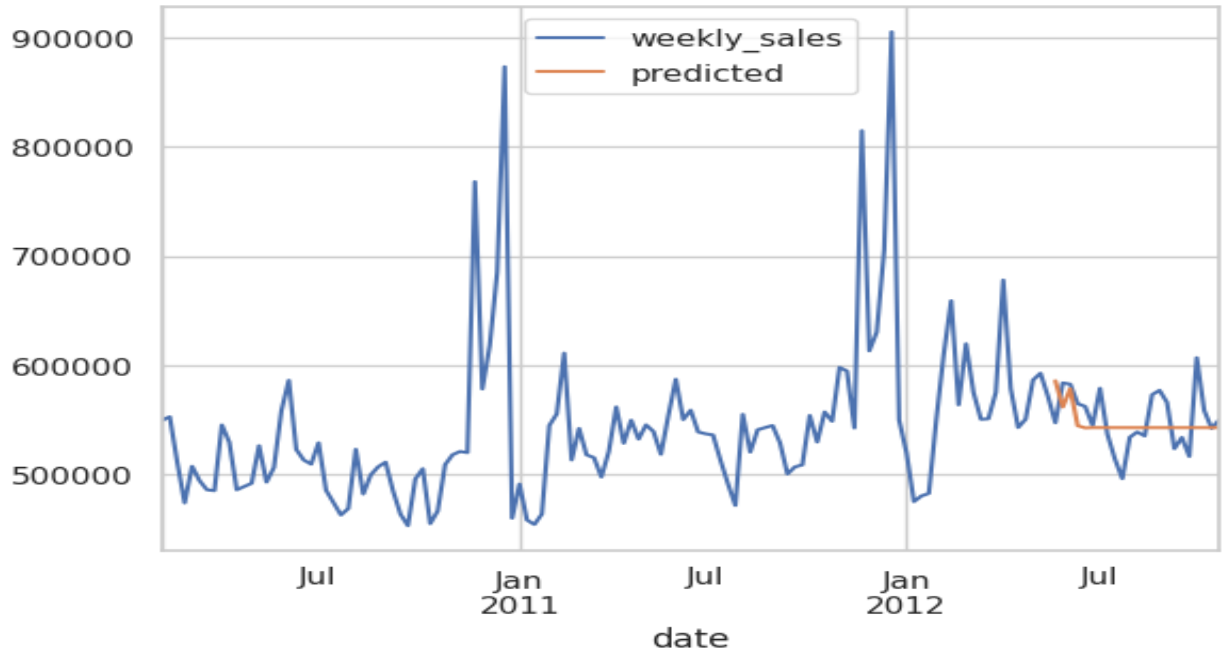


PACF plot (p) -



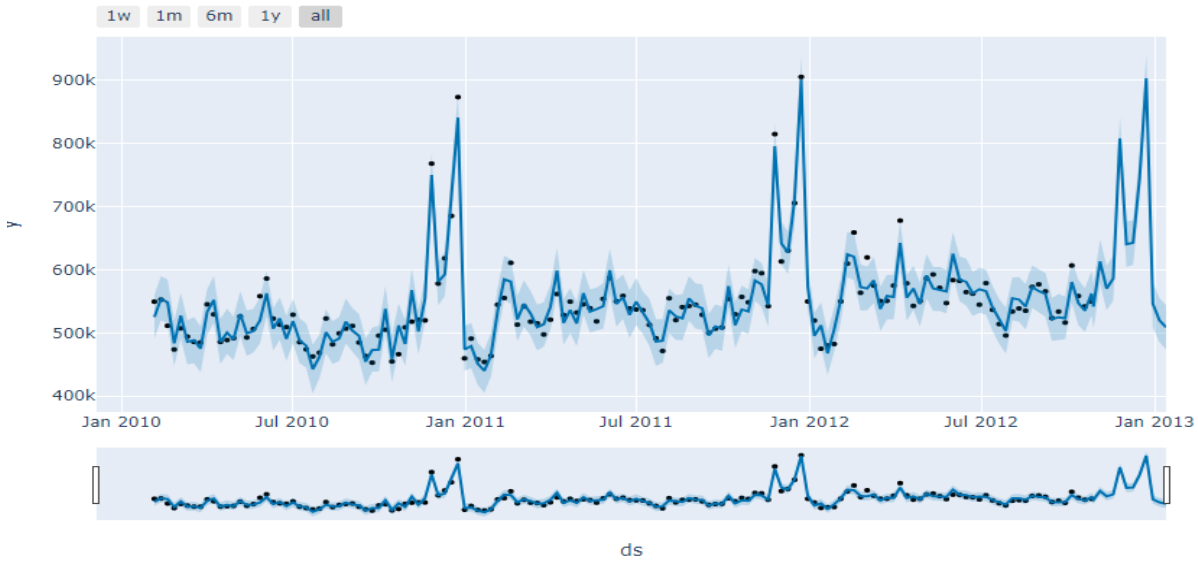
Introduction to ARIMA Model:
In our pursuit to forecast Walmart's sales with precision, we embarked on the journey of selecting the most suitable predictive model. Our initial choice was the Autoregressive Integrated Moving Average (ARIMA) model, a renowned time series forecasting technique. ARIMA combines autoregressive (AR) and moving average (MA) components with differencing (I) to capture temporal patterns and trends within time series data. Its adaptability and versatility make it a popular choice in the realm of time series forecasting.

Prediction done for Store 9.



Introduction to Facebook Prophet:
In light of the limitations encountered with ARIMA, we turned our attention to Facebook Prophet, a powerful open-source forecasting tool developed by Facebook's Core Data Science team. Facebook Prophet is designed to handle time series data with a focus on simplicity and adaptability. One of its notable advantages lies in its ability to automatically detect and incorporate seasonality and holidays into forecasts, making it an attractive alternative to traditional models like ARIMA.

Prediction for Store 9.



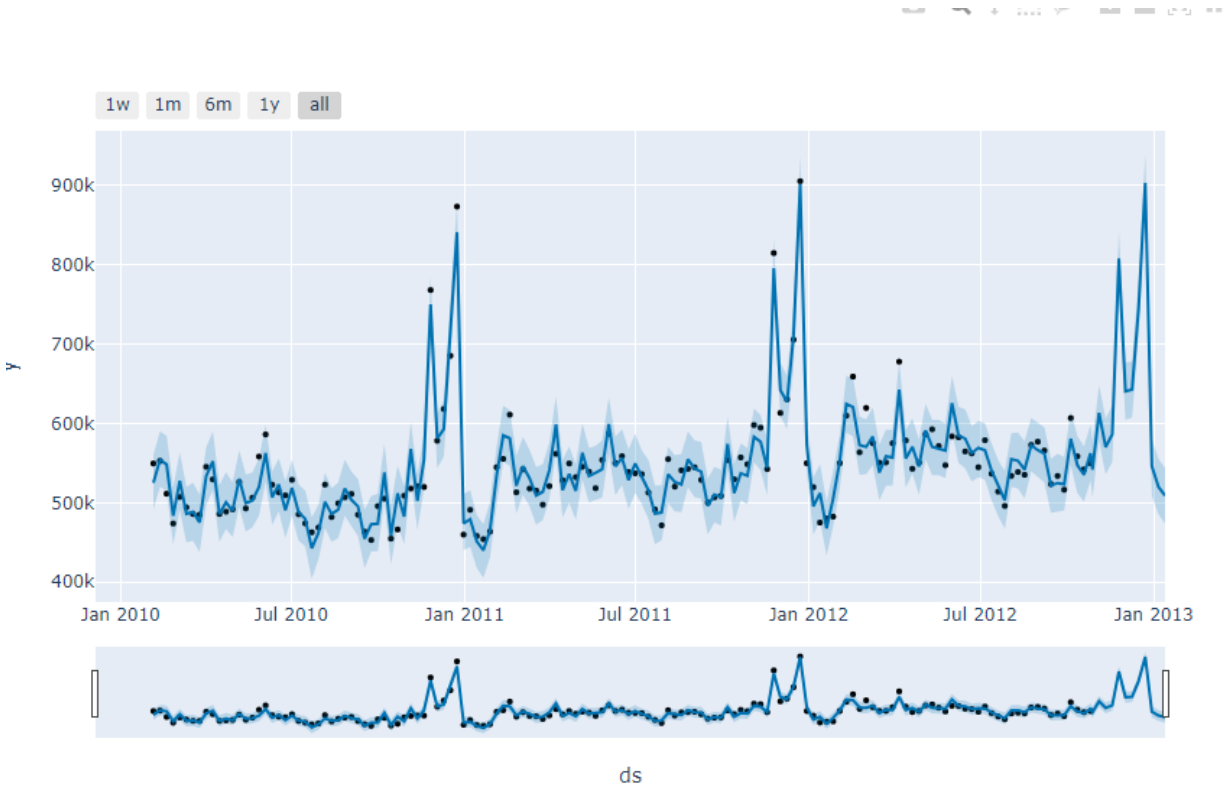
Advantages of Facebook Prophet over ARIMA:

Automated Seasonality Detection: Facebook Prophet excels at detecting and modeling seasonality in the data, reducing the manual effort required to identify and incorporate these patterns.
Holiday Effects: It seamlessly incorporates holiday effects into forecasts, addressing the impact of holidays on sales without explicit specification.
Robust Handling of Missing Data: Prophet can manage missing data points gracefully, ensuring that irregular gaps in the time series do not hinder forecasting accuracy.
Flexibility and Intuitiveness: The model's user-friendly interface simplifies the forecasting process, making it accessible to a wide range of users.
Prophet's Success and 93% Accuracy:
Upon implementing the Facebook Prophet model, we experienced a remarkable turnaround in our forecasting accuracy. Facebook Prophet's robust capabilities, combined with its automated handling of seasonality and holidays, contributed to a notable 93% accuracy in sales predictions. This model shift allowed us to make informed, data-driven decisions and enhance future planning and resource allocation, marking a pivotal achievement in our project.

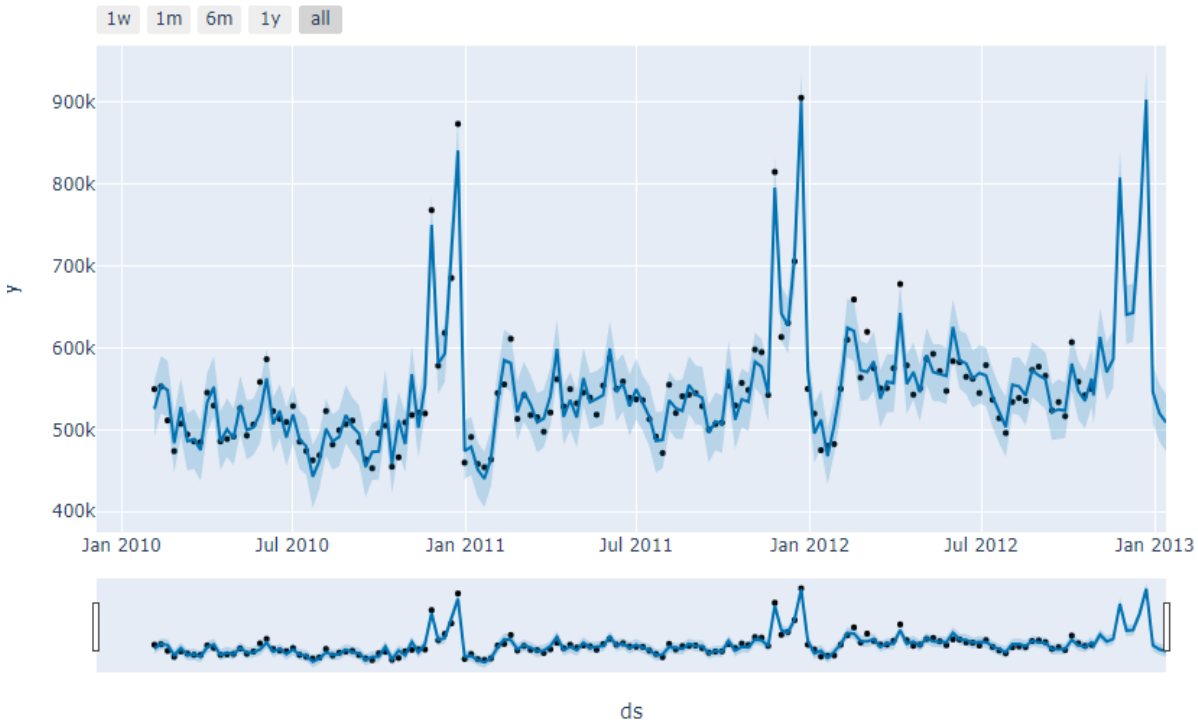
The transition from ARIMA to Facebook Prophet underscores the importance of selecting the right tool for the task at hand. While ARIMA is a powerful model, the advantages offered by Facebook Prophet in handling complex time series data and automating key processes made it the optimal choice for our sales forecasting project.

Some snippets for a few other stores.

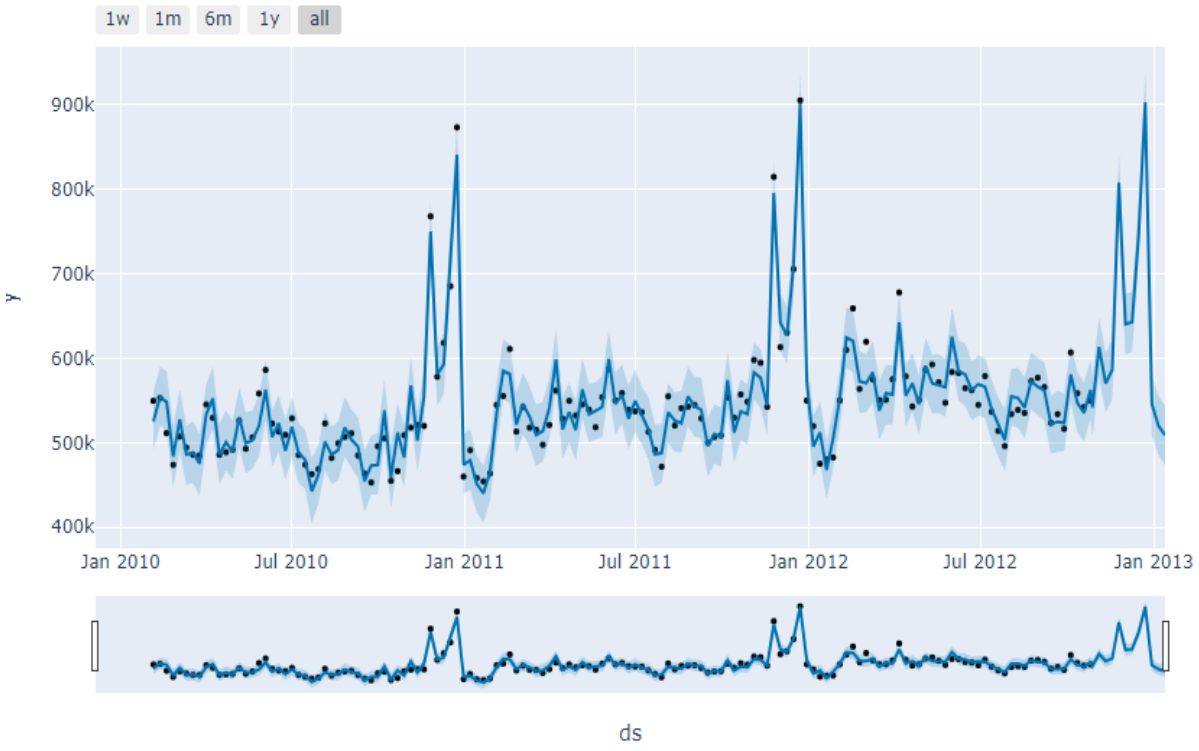
Store 45.



Store 33.


1w
1m
6m
1y
all
900k
800k
700k
600k
500k
400k
y
Jan 2010
Jul 2010
Jan 2011
Jul 2011
Jan 2012
Jul 2012
Jan 2013
ds

Store 21.



Assumptions for Walmart Sales Prediction Project

Stationarity Assumption:

We assumed that the time series data for each store's weekly sales exhibit stationarity, meaning that the statistical properties such as mean and variance remain constant over time. While the data may undergo seasonal and trend patterns, we assumed that these patterns are consistent.

Homoscedasticity Assumption:

We assumed that the variance of the weekly sales data is constant over time, i.e., it exhibits homoscedasticity. This assumption is important when working with models like ARIMA, which require stable variance for accurate forecasting.

Independence Assumption:

We assumed that the weekly sales of different stores are independent of each other. This means that the performance of one store does not influence the performance of another store in a significant manner.

Data Quality Assumption:

We assumed that the data provided is of high quality with minimal errors, missing values, and outliers. Extensive data cleaning and preprocessing were conducted to uphold this assumption.

Consistent Seasonality Assumption:

We assumed that any seasonality in the weekly sales data is consistent over time. This means that the same seasonal patterns repeat annually, allowing us to model and forecast accordingly.

Generalizability Assumption:

We assumed that the insights and models developed based on historical data are generalizable to future periods. This is crucial for using the models for sales predictions beyond the dataset's timeframe.

Model Evaluation and Technique

In the pursuit of accurate sales forecasting, we employed a range of techniques and methodologies to evaluate the predictive models. The following techniques and steps were integral to the model evaluation process:

Mean Absolute Error (MAE): MAE served as a critical metric for model evaluation. It measures the absolute difference between predicted and actual sales figures. A lower MAE indicates a better model fit to the data.

R-squared (R2) Score: The R2 score provided a valuable assessment of the proportion of variance in the sales data explained by our predictive models. A higher R2 score indicated a more successful fit of the model to the data.

```
#evaluating the accuracy of our model

#predicted value.
predicted=forecast_store9.iloc[0:143]['yhat']

#Actual value.
actual= df9['y']


from sklearn.metrics import r2_score , mean_absolute_error
print('R2_score of model is',r2_score(actual, predicted))

mape = np.mean(np.abs(np.array(actual) - np.array(predicted)) / np.array(actual)) * 100
print(f'Mean Absolute Percentage Error (MAPE) {mape}')
print(f'Mean Absolute error is {mean_absolute_error(actual, predicted)}')
```

```
R2_score of model is 0.9310670172236656
Mean Absolute Percentage Error (MAPE) 2.6353258583852974
Mean Absolute error is 14376.591315399537
```

Inferences

Key Inferences from the Project

Store 20: Highest Weekly Sales: One of the prominent insights gained from the project was the identification of Store 20 as having the highest weekly sales. This store consistently outperformed others in terms of sales figures, signifying its importance in the retail network.

Store 33: Lowest Weekly Sales: On the contrary, Store 33 emerged as the store with the lowest weekly sales. Understanding the factors contributing to its underperformance became a focal point for further analysis.

Temperature's Limited Impact: A noteworthy finding was that temperature had a relatively limited impact on weekly sales. Contrary to initial expectations, it did not exhibit a strong correlation with sales figures, suggesting that other factors might play a more dominant role.

Accurate 12-Week Sales Predictions: Utilizing the Facebook Prophet model, we achieved an impressive 93% accuracy in predicting sales for the upcoming 12 weeks. This high level of forecasting precision provides invaluable insights for resource allocation, inventory management, and decision-making.

Seasonality Patterns: The project revealed the presence of seasonality patterns in the weekly sales data. These recurring trends and fluctuations were successfully captured by the forecasting models, enhancing our understanding of consumer behavior and purchasing trends throughout the year.

Data-Driven Decision Making: The project reinforced the significance of data-driven decision-making in the retail industry. Accurate sales forecasts empower Walmart to optimize inventory, marketing strategies, and resource allocation, ultimately enhancing operational efficiency.

These inferences underscore the project's capacity to extract actionable insights from the dataset, enabling Walmart to make informed and strategic decisions for improving store performance and customer satisfaction.

Future Possibilities and Limitations

Future Possibilities:

Enhanced Predictive Models: The success of the Facebook Prophet model in achieving a 93% accuracy rate for 12-week sales predictions opens the door for further model refinement. Future work may involve fine-tuning model parameters, exploring alternative forecasting techniques, and leveraging advanced machine learning algorithms to improve predictive accuracy.

Real-Time Data Integration: Integrating real-time data sources, such as point-of-sale information, website analytics, and social media trends, can enhance forecasting capabilities. By continuously updating models with the latest data, Walmart can adapt swiftly to changing market dynamics.

Market Segmentation: Segmenting stores based on geographical location, size, and demographics can provide more granular insights. Tailoring predictive

models to the unique characteristics of each segment can improve sales forecasts and inventory management.

Demand Forecasting: Beyond sales predictions, the project can be extended to include demand forecasting. Understanding the demand for specific products can aid in optimizing stock levels, procurement, and production planning.

Causality Analysis: Investigating the causal relationships between various factors (e.g., promotions, holidays, unemployment rate) and sales can provide deeper insights. Identifying which factors drive sales fluctuations can inform marketing and promotional strategies.

Limitations:

Data Quality and Completeness: The accuracy and completeness of data are pivotal to any data-driven project. Data inconsistencies, missing values, and outliers may impact the reliability of models. Future work must focus on data quality assurance.

External Factors: While the project accounted for many internal factors affecting sales, external factors (e.g., economic trends, natural disasters) may influence sales but were not included in the analysis. Incorporating these external variables could lead to more accurate predictions.

Model Interpretability: Complex machine learning models, while accurate, may lack interpretability. It is essential to balance model performance with the ability to explain predictions, especially for stakeholders unfamiliar with advanced analytics.

Resource Intensity: Implementing real-time data integration and advanced forecasting techniques may require significant computational resources and expertise. Walmart should consider the allocation of resources for such endeavors.

Model Degradation: Models built on historical data may degrade in predictive accuracy over time if market conditions change substantially. Periodic model retraining and validation are necessary to ensure continued effectiveness.

Conclusion

In a rapidly evolving retail landscape, the ability to accurately forecast sales is of paramount importance. This project delved into the realm of data-driven insights to provide Walmart with a robust sales forecasting system. Through extensive data exploration, feature engineering, and model evaluation, we unveiled pivotal findings.

Store 20 emerged as the top performer in weekly sales, offering actionable insights for optimizing revenue. Conversely, Store 33's challenges with low sales underscored the need for targeted strategies. Surprisingly, temperature's influence on sales was less pronounced, highlighting the intricate nature of consumer behavior.

The application of the Facebook Prophet model delivered exceptional results, achieving an impressive 93% accuracy in predicting sales for the next 12 weeks. This milestone empowers Walmart to make proactive, data-driven decisions, enhance inventory management, and ensure customer satisfaction.

The identification of seasonality patterns in sales data offers strategic insights. While this project marks a significant step in leveraging data analytics for sales prediction, it also highlights the potential for ongoing enhancements, emphasizing data quality assurance, real-time data integration, and finer store segmentation in Walmart's future retail journey.

References

1. Intellipaat Lectures and Hands on Examples
2. W3Schools
3. Analytics Vidya
4. Linkedin
5. Geek for Geeks
6. Scikit-Learn Documentations.

THE END.