**1.Problem Statement**

The retail store chain is encountering challenges in effectively managing inventory across its numerous outlets throughout the country. They are struggling to align the demand for products with the available supply, resulting in inventory imbalances and potential revenue losses. As a data scientist, the objective is to leverage the available dataset, which includes time-related data and a column indicating weekly sales, in order to generate valuable insights and develop a robust predictive model that can forecast the weekly sales for each outlet.

**2.Objective**

The primary objective of this project is to create a predictive model that accurately forecasts the weekly sales for each outlet of the retail store chain. By utilizing the available time-related data, the model will enable the company to proactively manage inventory levels and ensure a more efficient supply chain**.**

The specific goals are as follows:

1.Perform exploratory data analysis (EDA) on the dataset and visualize the data to gain valuable insights and identify patterns. This will involve analysing the distribution of weekly sales across outlets and time periods, detecting any outliers or missing values, and exploring correlations with other variables such as employment rate, holidays, or demographics.

2.Choose a suitable machine learning algorithm that is capable of handling the predictive task of forecasting weekly sales. Consider time series forecasting algorithms such as ARIMA, SARIMA, or Prophet, as they are specifically designed to capture seasonality and trend patterns in time-dependent data.

3.Conduct appropriate preprocessing steps to make the data suitable for the chosen machine learning algorithm. This may include handling missing values, smoothing or detrending the time series, identifying and removing seasonality or trends, and transforming variables if required.

4. Develop and train a predictive model using the selected machine learning algorithm. Utilize historical sales data, along with relevant features such as promotions, holidays, or outlet-specific characteristics, to train the model. Split the dataset into training and test sets to evaluate the model's performance.

5.Compare the performance of multiple models using suitable evaluation metrics such as mean absolute error (MAE), mean squared error (MSE), or root mean squared error (RMSE). This comparison will help determine the best-performing model for the task of weekly sales forecasting, ensuring accurate predictions and reliable decision-making.

6. Choose the most appropriate model or algorithm based on the evaluation results and further fine-tune it if necessary.

7. Utilize the selected algorithm to forecast future weekly sales based on the trained model. Use the available data to make predictions for specific time periods, such as months or years ahead. This will provide valuable insights for inventory management, allowing the retail store chain to optimize their supply chain, anticipate demand fluctuations, and make informed business decisions.

By following these steps, the project aims to leverage data analysis, machine learning, and forecasting techniques to provide actionable insights and improve the retail store chain's inventory management and overall business performance.

**3.Data Description**

The dataset is stored in a CSV file format.It contains 6435 rows, each representing a specific data point or observation. There are 8 columns in the dataset, with the following names:

1.Store: Represents the unique identifier for each store.

2.Date: Represents the date of the observation.

3.Weekly\_Sales: Represents the target variable, which indicates the sales for a particular week.

4.Holiday\_Flag: Represents a binary indicator (0 or 1) that denotes whether the week contains a holiday or not.

5.Temperature: Represents the temperature during the week in the respective store's location.

6.Fuel\_Price: Represents the fuel price during the week.

7.CPI: Represents the Consumer Price Index during the week, which measures changes in the price level of a basket of consumer goods and services.

8.Unemployment: Represents the unemployment rate during the week.

**4.Data Preprocessing steps**

In the Walmart sales prediction project, I performed several preprocessing steps to prepare the data for time series forecasting using the selected algorithm. Here's a summary of the steps I took:

Data Quality Check:

* Checked for duplicate values and null values in the dataset.
* Verified that there were no duplicate values or nulls present.

Data Formatting:

* Selected the relevant columns for analysis, including the date column and weekly sales.
* Converted the date column to a datetime format for time-based analysis.

Data Ordering and Aggregation:

* Noticed that the date values were not in chronological order, which is essential for time series analysis.
* Used resampling with a weekly frequency ('W') and computed the sum of sales for each week.This process brought the sales values into an organized weekly format.

Handling Zero Sales:

* Observed that some weeks had zero sales, which could affect the prediction accuracy.
* Checked for seasonality using ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots.Determined that there was no clear seasonality in the data.
* Employed line decomposition techniques to further visualize seasonality, confirming its absence.
* Recognized that accurate seasonality wasn't evident from the data.
* Decided against using seasonal interpolation due to the lack of captured seasonality.
* Opted for padding interpolation (filling with recent values) to handle zero sales and maintain data continuity.

Stationarity Check and Transformation:

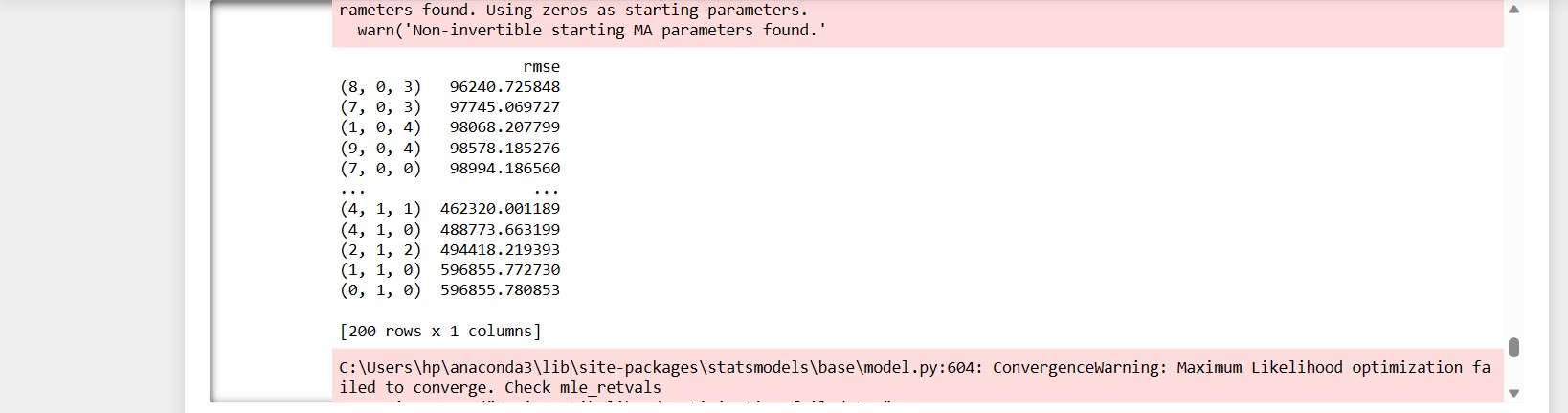
* Utilized Augmented Dickey-Fuller (ADF) test to assess data stationarity.Identified that all stores except one had stationary data.
* Applied first-order differencing to handle non-stationary data, making it suitable for time series analysis.

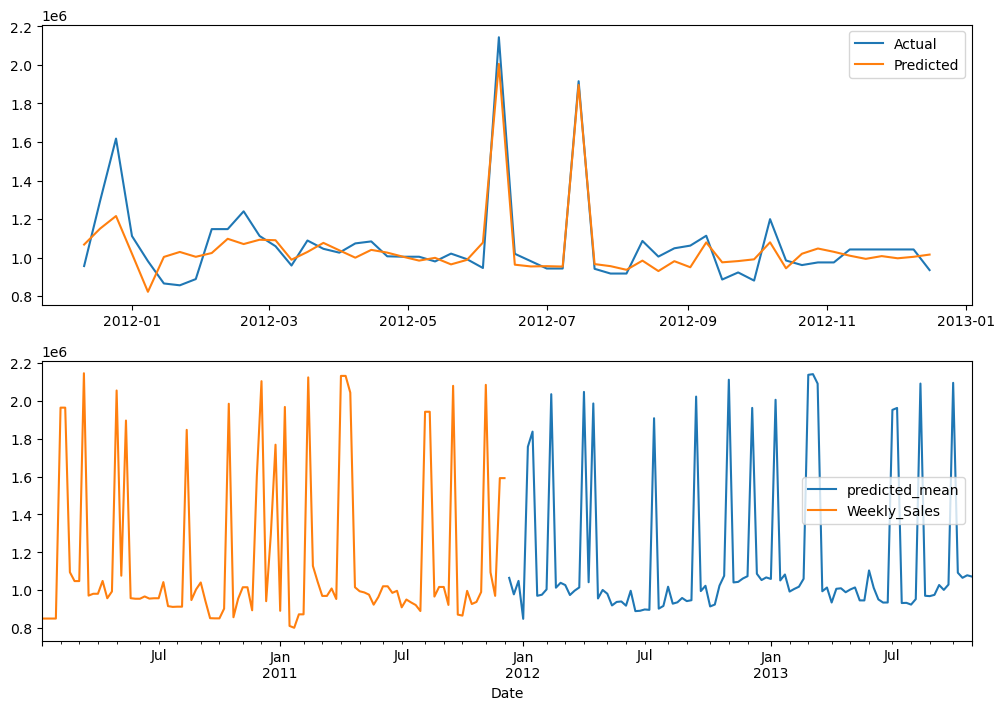
**5.Choosing the algorithm for the project**

Given the nature of the data being time-based and requiring predictions for future periods, the choice of a time series forecasting algorithm was deemed appropriate. While regression could also be an option, time series forecasting accounts for the sequential dependency inherent in time-based data.For this project, I experimented with two popular time series forecasting algorithms: ARIMA (AutoRegressive Integrated Moving Average) and SARIMAX (Seasonal ARIMA with Exogenous Variables). Both algorithms are well-suited for capturing the underlying patterns and trends in time series data.

**6.Motivation and Reasons for choosing the algorithm & Model Evaluation Technique**

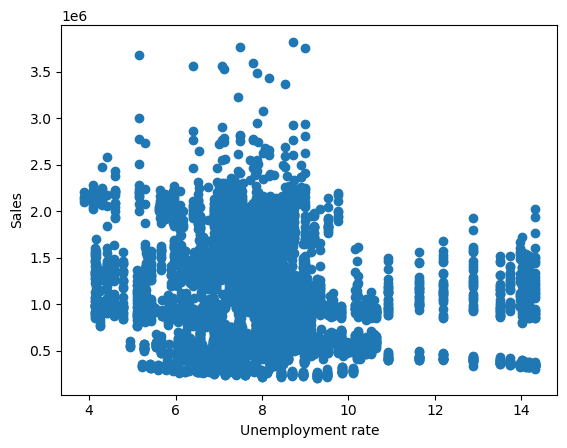
Given the temporal nature of the data, two prominent time series forecasting algorithms, ARIMA (AutoRegressive Integrated Moving Average) and SARIMAX (Seasonal ARIMA with Exogenous Variables), were experimented with.During experimentation, ARIMA emerged as a promising choice in terms of predictive accuracy, as indicated by the Root Mean Squared Error (RMSE) metric. ARIMA's ability to capture underlying trends without imposing seasonality made it suitable for the dataset. SARIMAX, on the other hand, struggled due to the absence of a clear seasonal pattern in the data. Despite achieving lower RMSE values, both ARIMA and SARIMAX encountered challenges in properly fitting the forecasting line to the validation data. The weekly sales line did not align adequately with observed values, raising concerns about the models' overall efficiency .Upon closer analysis, it became evident that weekly sales were influenced by external factors such as temperature, fuel price, and Consumer Price Index (CPI). These exogenous variables were not considered in the initial modelling attempts. Recognizing their potential impact, a new approach was devised .To address the influence of exogenous factors, ARIMA was enhanced by incorporating external columns using the exog parameter. This allowed the model to consider additional variables beyond the sales history. The inclusion of temperature, fuel price, and CPI data enabled the model to capture more nuanced relationships affecting weekly sales.The inclusion of exogenous variables in the ARIMA model led to a noticeable improvement in forecasting accuracy. The model's predictions now closely aligned with the validation data, resulting in a more accurate representation of weekly sales trends. This was particularly evident when visualizing the predicted line alongside the actual sales data.





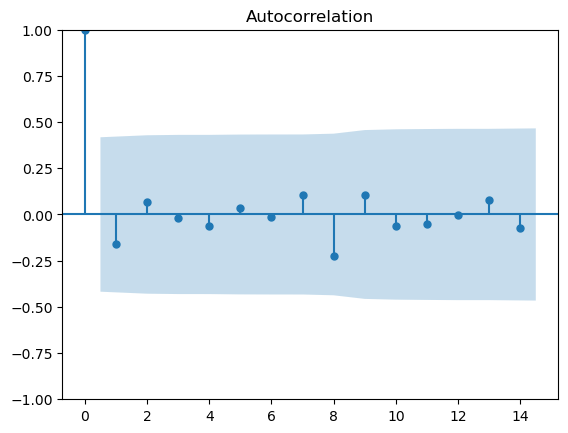
**7.Inferences and Insights Derived**

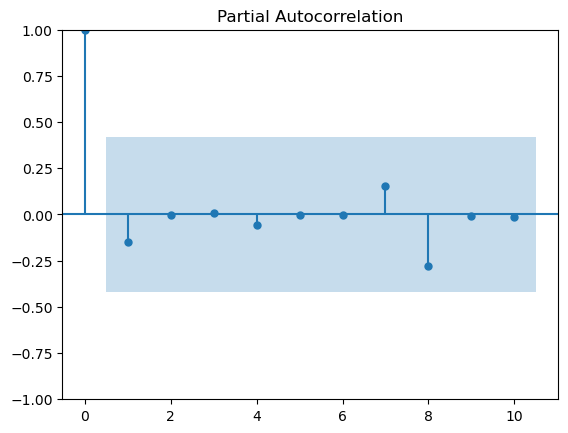
a. Weekly sales and Unemployment



The unemployment rate column has a correlation of -0.10617608965795419 with Sales column

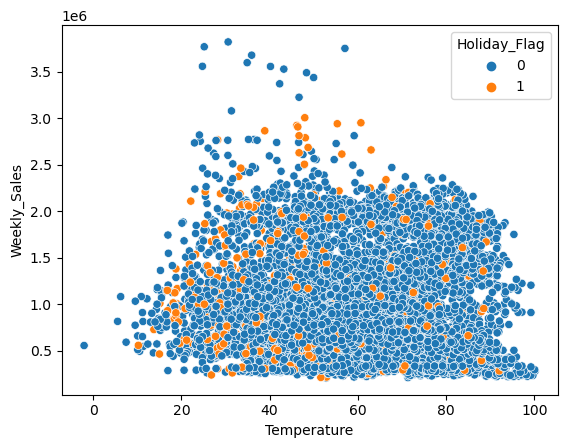
b. Seasonality check





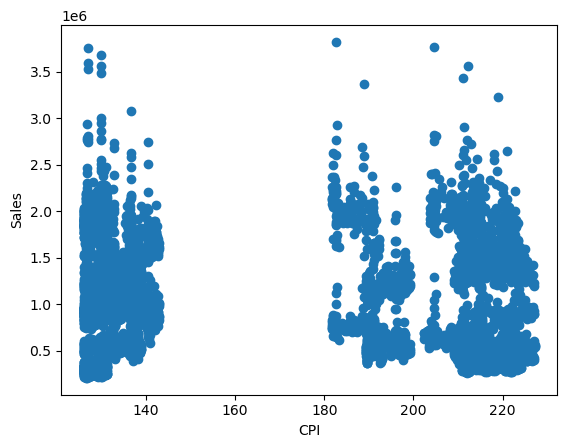
It is evident from the graph that there exists no seasonality within the particular dataframe

c. Tempurature with Sales



The temperature column has a correlation of -0.06381001317946962 with Sales column

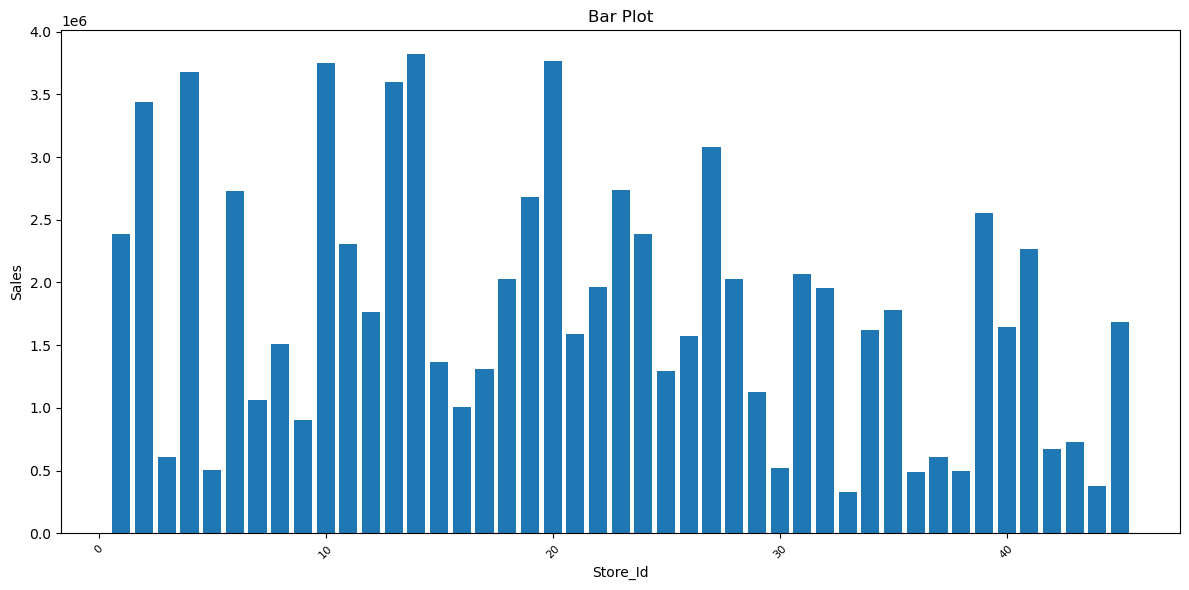
d. Consumer Price Index with Sales



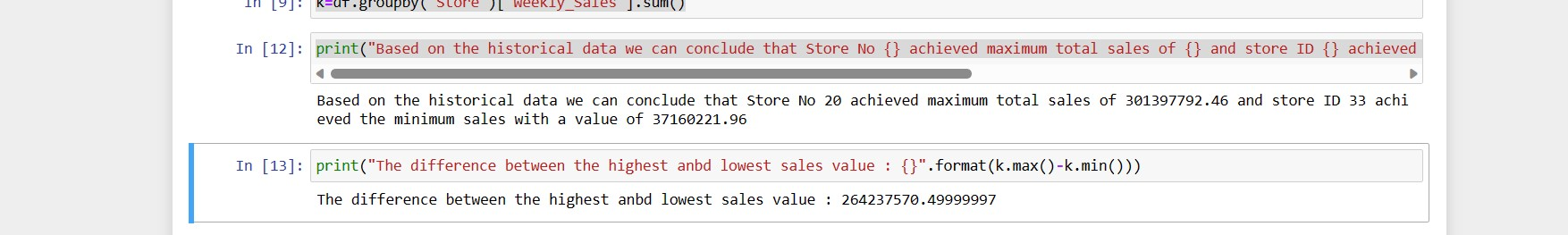
The Consumer Price Index column has a correlation of -0.07263416204017631 with Sales

column

e . Performance of stores



f. Highest and lowest performing store and the difference in sales



Based on the historical data we can conclude that Store No 20 achieved maximum total sales of 301397792.46 and store ID 33 achieved the minimum sales with a value of 37160221.96

The difference between the highest and lowest sales value : 264237570.49999997

**8.Future Possibilities of project**

Testing the Model with Various External Columns:

An exciting avenue for further exploration in this project involves the rigorous testing of the developed ARIMA model with exogenous variables. While the initial implementation

incorporated external factors such as temperature, fuel price, and CPI, there are numerous other potential external variables that might influence weekly sales. Future possibilities include:

1.Economic Indicators: Incorporating economic indicators such as unemployment rates, GDP growth, and consumer sentiment indices could provide valuable insights into the correlation between economic conditions and sales trends.

2.Promotions and Holidays: Factoring in information about promotional events, holidays, and special sales periods could lead to a more accurate model, considering the spikes in consumer spending during these periods.

3.Competitor Data: Gathering data on competitor activities, such as pricing strategies or marketing campaigns, might reveal how the company's sales are affected by industry dynamics.

4.Social Media Engagement: Exploring correlations between social media engagement metrics (likes, shares, comments) and sales patterns could provide a new dimension of understanding customer behavior.

5.Local Events: Including data on local events, such as festivals, community gatherings, or sporting events, could highlight regional fluctuations in sales due to varying consumer interests.

Solving the Problem Statement using Regression Algorithm:

Another intriguing future possibility lies in solving the same sales prediction problem using

regression algorithms. While time series forecasting models like ARIMA are tailored for temporal patterns, regression algorithms can capture complex relationships between independent variables and the dependent variable.

**9.Conclusion**

In conclusion, the Walmart sales prediction project showcased the intricacies of working with time-based data and the challenges of accurately forecasting sales trends. Through careful

experimentation and analysis, it was determined that while ARIMA initially outperformed

SARIMAX in terms of RMSE, the absence of a seasonal pattern posed challenges in

achieving optimal line fitting.

However, the introduction of exogenous variables, including temperature, fuel price, and CPI, into the ARIMA model proved to be a decisive enhancement. This adaptation effectively

captured external influences on weekly sales, resulting in improved predictions that align

closely with observed values. This experience underscores the importance of considering a

variety of factors, both internal and external, in developing robust time series forecasting

models.The project serves as a valuable case study in data-driven decision-making and the

iterative process of model refinement. By combining algorithmic selection, parameter tuning, and domain-specific insights, the final ARIMA model with exogenous variables showcases

the power of blending statistical techniques with real-world context to achieve accurate

predictions in the realm of time series forecasting.

**10.References**

[**https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp**](https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp)

**https://towardsdatascience.com/time-series-forecasting-with-arima-sarima-and-sarimax-ee61099e78f6**

[**https://www.youtube.com/watch?v=wJASx5-mnuc&list=LL&index=36**](https://www.youtube.com/watch?v=wJASx5-mnuc&list=LL&index=36)

[**https://www.youtube.com/watch?v=pLHm4cvoZiY&list=LL&index=17**](https://www.youtube.com/watch?v=pLHm4cvoZiY&list=LL&index=17)

[**https://www.youtube.com/watch?v=gDwx3RPUfPw&list=LL&index=18**](https://www.youtube.com/watch?v=gDwx3RPUfPw&list=LL&index=18)