**Capstone Project - Walmart**

# **Table of Contents**

[Table of Contents 2](#_u5lh77q68k3n)

[Problem Statement 3](#_6yswno636dha)

[Project Objective 3](#_lvyowqtf293b)

[Data Description 4](#_yfm2yprawmj9)

[Data Preprocessing Steps 5](#_5s69224q2x)

[Choosing the Model for the Project 7](#_cvd30j3h0pxm)

[Assumptions 9](#_iu2wpf284z9p)

[Model Evaluation Techniques 10](#_kxyifkbd4v6e)

[Inferences from the Project 11](#_8ihmi3a2y5gm)

[Future Possibilities 18](#_em5ivh25r67k)

[Conclusion 19](#_b8ncva6fz7ua)

# 

# **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.

# **Project Objective**

1. 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.
2. Use predictive modeling techniques to forecast the sales for each store for the next 12 weeks.

# 

# **Data Description**

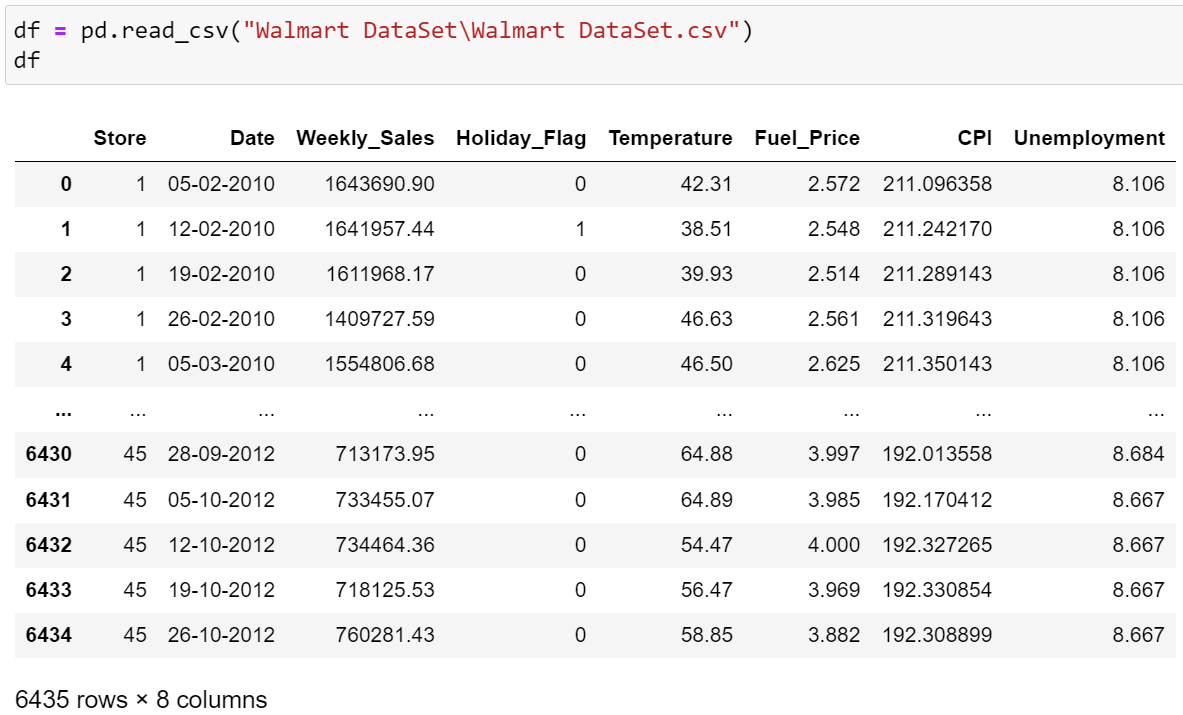
The walmart.csv contains 6435 rows and 8 columns.

| **Feature Name** | **Description** |
| --- | --- |
| Store | Store number |
| Date | Week of Sales |
| Weekly\_Sales | Sales for the given store in that week |
| Holiday\_Flag | If it is a holiday week |
| Temperature | Temperature on the day of the sale |
| Fuel\_Price | Cost of the fuel in the region |
| CPI | Consumer Price Index |
| Unemployment | Unemployment Rate |

# **Data Preprocessing Steps**

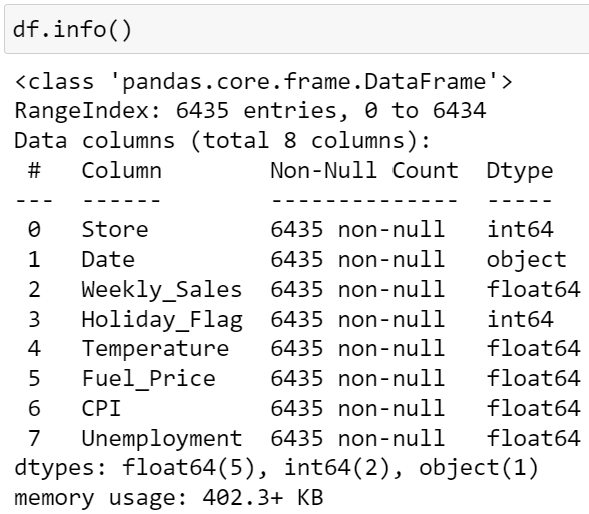
The preprocessing of the data included the following steps:

1. The data is first imported into a DataFrame ‘df’ and is displayed for having a general idea of the what our data looks like



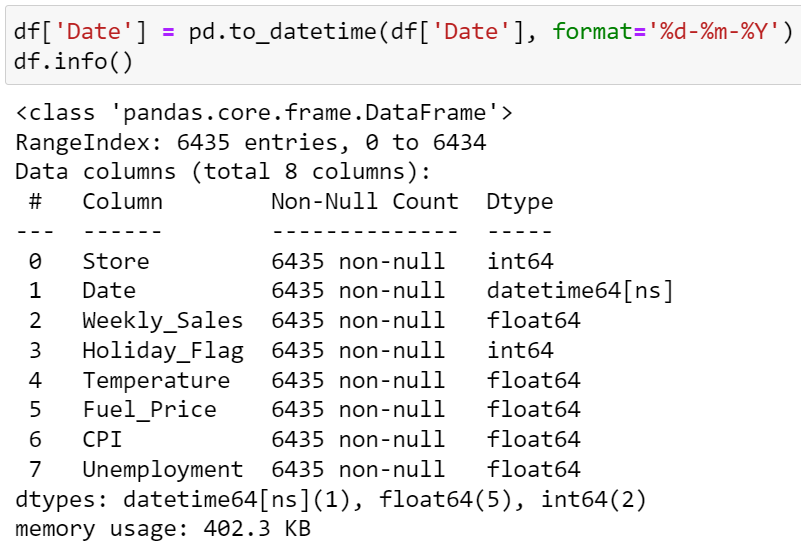
* 1. This gives a birds eye view of the data on how our data generally looks across all the columns, the column names, number of rows and columns, etc…

1. We get some more information about our data by using the following code;

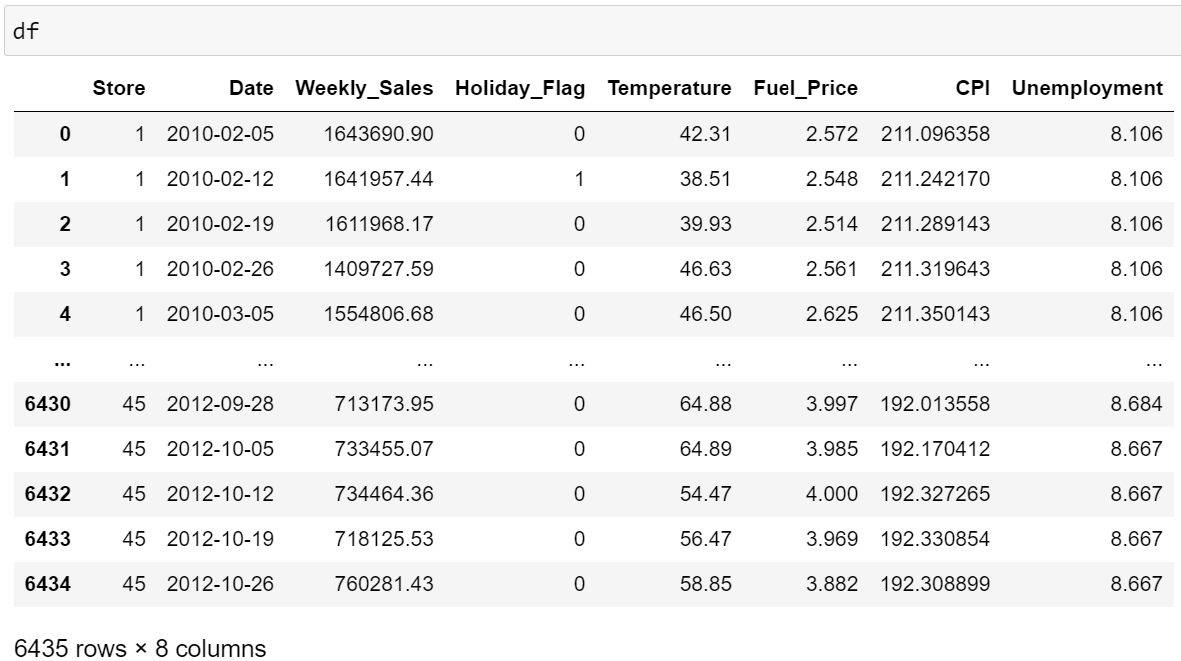


* 1. None of the columns have any null values (non-null count matches total entries)
  2. Date column data type is object instead of datetime

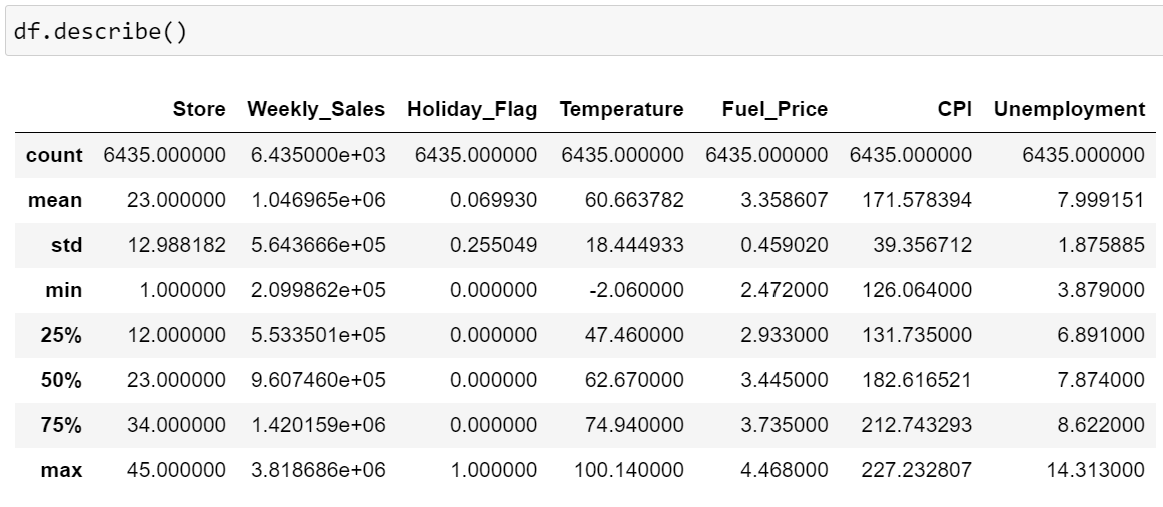
1. Convert the ‘Date’ column data type to datetime for better operability



1. Cross check if the datetime conversion was done correctly



1. Get some statistical information about the numbers



# **Choosing the Model for the Project**

**Seasonal Autoregressive Integrated Moving Average (SARIMA)** is a statistical technique used for time series forecasting. It extends the Autoregressive Integrated Moving Average (ARIMA) model to account for seasonality in the data. SARIMA models are particularly useful when the time series exhibits patterns that repeat at known intervals, such as daily, weekly, or monthly seasonality.

The SARIMA model is defined by three main components:

**Seasonal Component:** SARIMA incorporates seasonal patterns by introducing additional parameters that capture seasonal changes. These parameters include seasonal autoregressive (SAR) terms, seasonal differences, and seasonal moving average (SMA) terms. The seasonal component helps the model capture recurring patterns over specific intervals.

**Autoregressive Component:** This component models the relationship between an observation and a number of lagged observations (autoregressive terms). It represents the linear relationship between the variable and its own past values.

**Moving Average Component:** This component models the error of the model as a linear combination of error terms observed at previous time points (moving average terms). It helps in capturing short-term fluctuations and noise in the data.

The SARIMA model is specified using the notation SARIMA(p, d, q)(P, D, Q)m, where:

p: Autoregressive order

d: Degree of differencing (to make the time series stationary)

q: Moving average order

P: Seasonal autoregressive order

D: Seasonal differencing

Q: Seasonal moving average order

m: Number of time steps in each seasonal cycle (e.g., 12 for monthly data, 7 for weekly data)

By adjusting these parameters, SARIMA can effectively capture the complex dynamics of seasonal time series data and provide accurate forecasts. However, selecting appropriate values for these parameters often requires iterative analysis, including visual inspection of the data, autocorrelation and partial autocorrelation plots, and model diagnostics.

There are several compelling reasons that made me choose the SARIMA model for the time series forecasting on the provided dataset:

1. **Seasonality:** SARIMA is specifically designed to handle time series data with seasonal patterns. As our dataset exhibits clear seasonal repeating patterns, SARIMA can effectively capture and model these seasonal variations.
2. **Flexibility:** SARIMA offers flexibility in modeling both non-seasonal and seasonal components of time series data. It allows incorporation of autoregressive, differencing, and moving average terms for both non-seasonal and seasonal patterns, making it suitable for a wide range of time series datasets.
3. **Robustness:** SARIMA is robust against common time series issues such as trend, seasonality, and autocorrelation. By incorporating differencing and moving average terms, SARIMA can handle non-stationary data and ensure that the model residuals are stationary.
4. **Interpretability:** SARIMA models provide interpretable parameters that describe the underlying dynamics of the time series. The autoregressive, differencing, and moving average parameters offer insights into the relationship between past and current observations, as well as the short-term fluctuations in the data.
5. **Forecast Accuracy:** When properly tuned, SARIMA models can provide accurate forecasts for seasonal time series data. By capturing both non-seasonal and seasonal patterns, SARIMA can effectively capture the underlying structure of the data and produce reliable forecasts.
6. **Diagnostic Tools:** SARIMA models come with diagnostic tools such as autocorrelation and partial autocorrelation plots, residuals analysis, and goodness-of-fit tests. These tools help in selecting appropriate model parameters and assessing the model's goodness of fit to the data.
7. **Widely Used:** SARIMA is a well-established and widely used time series forecasting method, with extensive literature and software support available. This makes it easier to implement and apply SARIMA models in practice, with plenty of resources and expertise available for guidance.

# **Assumptions**

When utilizing the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, several assumptions are made to ensure the model's validity and applicability to the dataset. These assumptions include:

1. **Stationarity:** The time series data should be stationary or transformable to achieve stationarity through differencing. Stationarity implies that the statistical properties of the series (such as mean, variance, and autocorrelation) remain constant over time. SARIMA assumes stationarity or enforces it through differencing.
2. **Linearity:** SARIMA assumes that the relationships between variables in the model are linear. This assumption means that the effects of changes in the predictor variables on the response variable are constant over time.
3. **Independence:** Observations in the time series should be independent of each other. While this assumption may not hold in all cases (especially in some types of financial or spatial data), it is generally assumed in SARIMA modeling. Seasonal patterns, autocorrelation, and other dependencies should be adequately captured by the model.
4. **Constant Variance:** The variance of the residuals (errors) across the entire series should be constant. This assumption ensures that the model's performance remains consistent across different segments of the data.
5. **Normality of Residuals:** SARIMA assumes that the residuals of the model are normally distributed. This assumption allows for reliable estimation of confidence intervals and hypothesis testing.
6. **No Perfect Multicollinearity:** In SARIMA, predictors should not be perfectly correlated with each other. While some degree of correlation is expected and may even be desirable for capturing complex patterns in the data, perfect multicollinearity can lead to estimation problems and unreliable model results.
7. **No Outliers or Influential Observations:** SARIMA assumes that the time series data does not contain outliers or influential observations that could unduly influence the estimation of model parameters.

# **Model Evaluation Techniques**

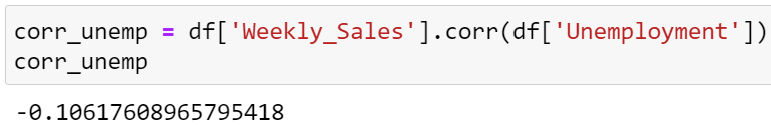
Model evaluation techniques for SARIMA (Seasonal Autoregressive Integrated Moving Average) models typically involve assessing the model's ability to accurately forecast future values based on historical data. Following are the techniques that were used to evaluate our SARIMA model;

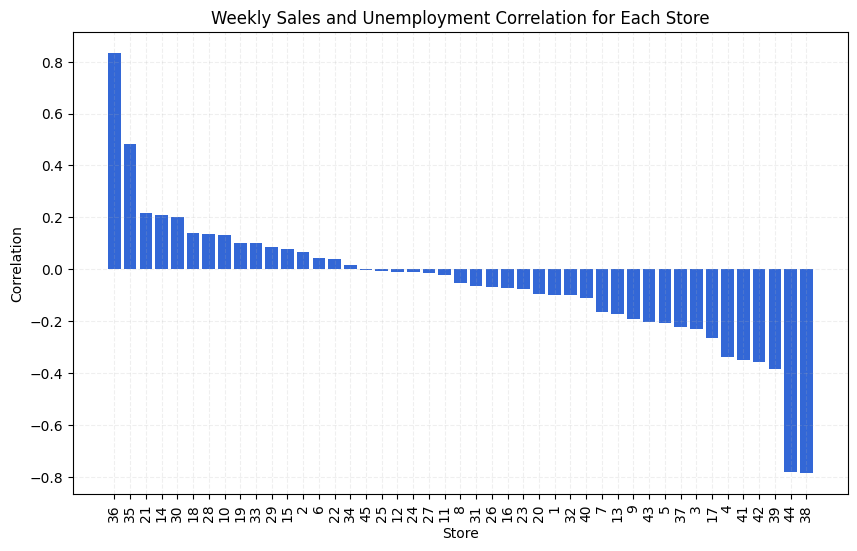
1. **Cross-Validation:** Splitting the data into training and testing sets and evaluating the model's performance on the testing set can provide a more robust assessment of forecasting accuracy, particularly for models trained on smaller datasets.
2. **Visual Inspection:** Plotting the observed values against the predicted values can provide a straightforward assessment of the model's performance. If the predicted values closely follow the observed values, it indicates a good fit.
3. **Residual Analysis:** Analyzing the residuals (the differences between observed and predicted values) is crucial for assessing model adequacy. Residuals should ideally exhibit randomness, with no discernible patterns or trends.
   1. The technique for residual analysis used on our model was Mean Absolute Error (MAE). It measures the average absolute difference between the predicted values and the actual values. It provides a straightforward indication of how close the forecasts are to the actual values on average. A lower MAE indicates better performance, as it means the model's forecasts are closer to the actual values on average.

# **Inferences from the Project**

Following are the inferences we derived from the project;

1. **Effect of Unemployment Rate on Weekly Sales:**

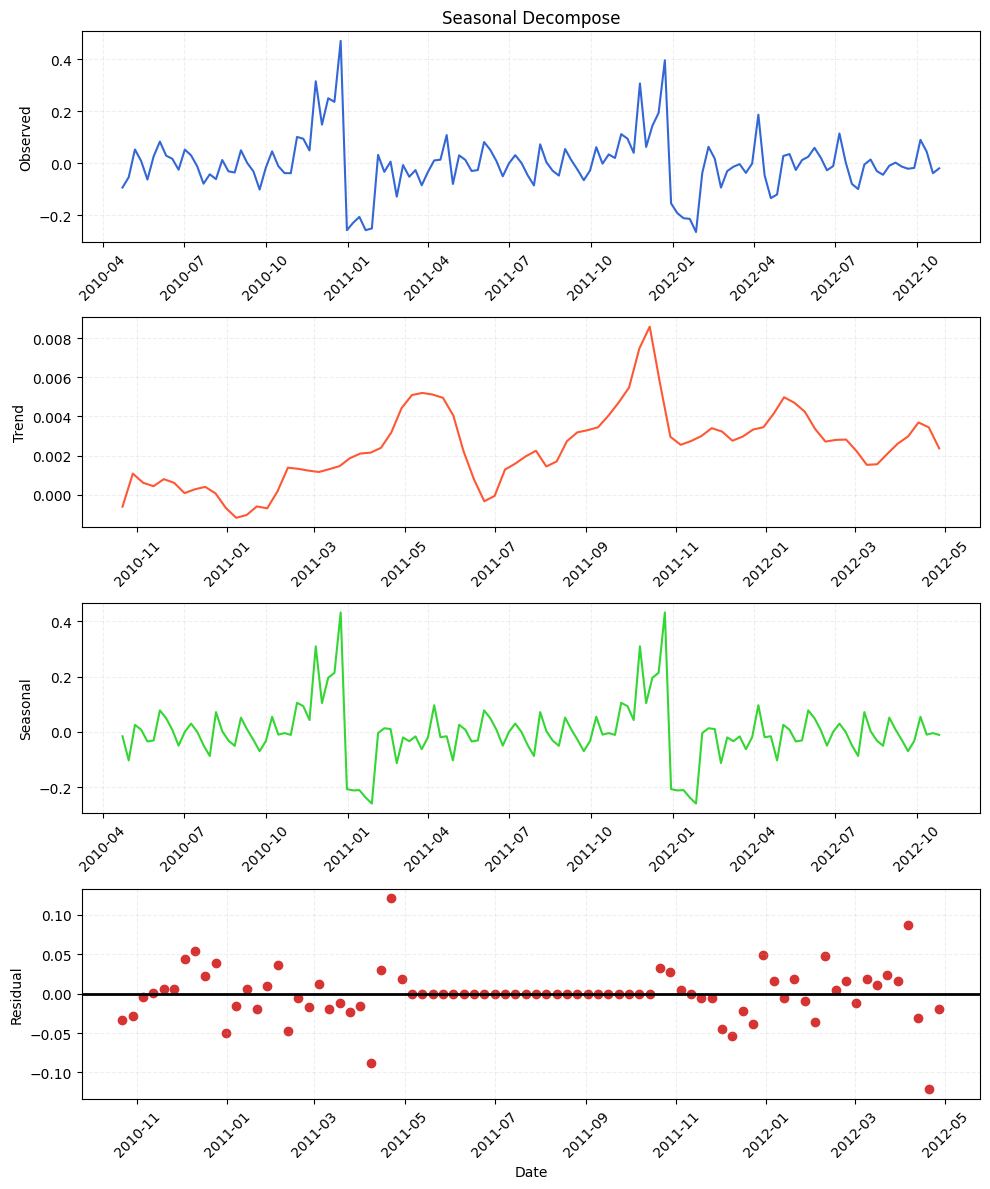
****



* Overall, there was no significant correlation between weekly sales and unemployment rate.
* However, at a store level, certain stores exhibited notable correlations:
  + Stores 44 and 38 showed significant negative correlation, meaning as unemployment rises, weekly sales decrease.
  + Stores 36 and 35 showed significant positive correlation, indicating that as unemployment rises, weekly sales also increase. This trend is counterintuitive and requires further investigation.

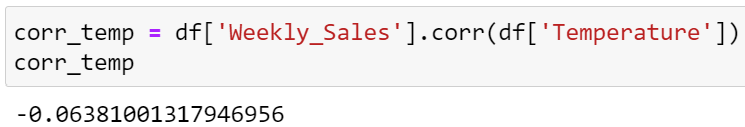
1. **Seasonal Trends in Weekly Sales:**

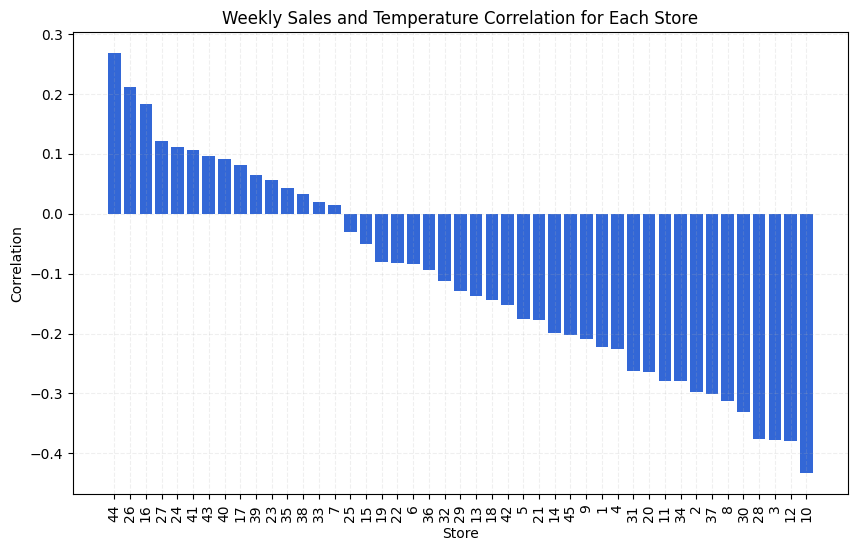
* To get this, we first made the data stationary and normalized, then performed seasonal decompose, which gave us the following chart;



* Seasonal decomposition revealed a clear seasonal trend in weekly sales, with peaks during the holiday season every year.
* Seasonal behavior in sales throughout the year can be influenced by a variety of factors, both internal and external to the business. Some possible reasons for seasonal patterns in sales include:
  + **Holidays and Festivals:** Sales often surge during holiday seasons such as Christmas, New Year, Thanksgiving, Eid, Diwali, or other culturally significant festivals. Consumers tend to increase spending on gifts, decorations, food, and other items associated with these occasions.
  + **Seasonal Weather Changes:** Weather patterns can significantly affect consumer behavior and purchasing decisions. For example, sales of seasonal items like winter clothing, air conditioners, or outdoor equipment tend to peak during specific seasons depending on climatic conditions.
  + **School and College Schedules:** Sales may vary with academic calendars due to back-to-school shopping, graduation ceremonies, or other school-related events.

1. **Impact of Temperature on Weekly Sales:**

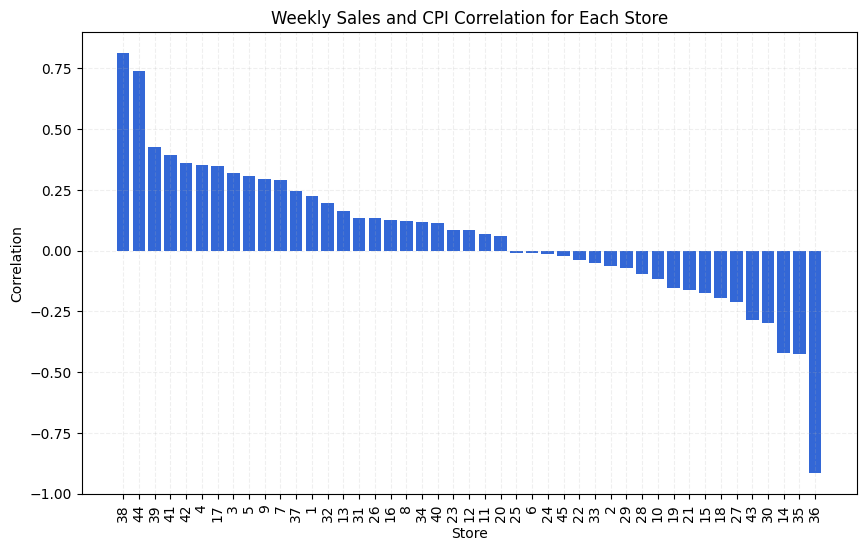




* There was no significant correlation between weekly sales and temperature at both overall and store levels.
* Thus, temperature does not appear to have a significant effect on weekly sales.

1. **Impact of Consumer Price Index (CPI) on Weekly Sales:**

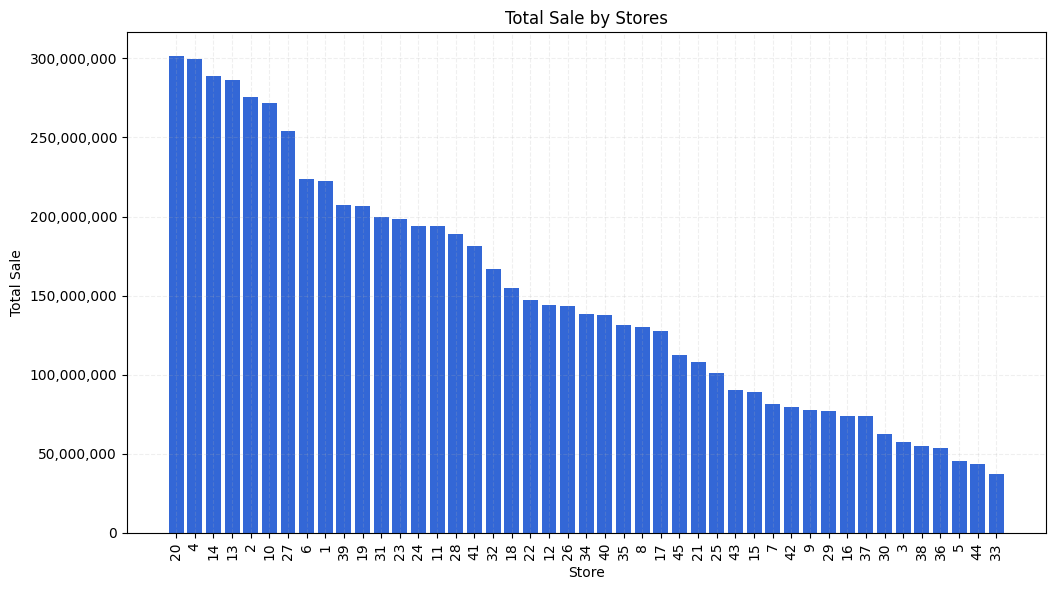


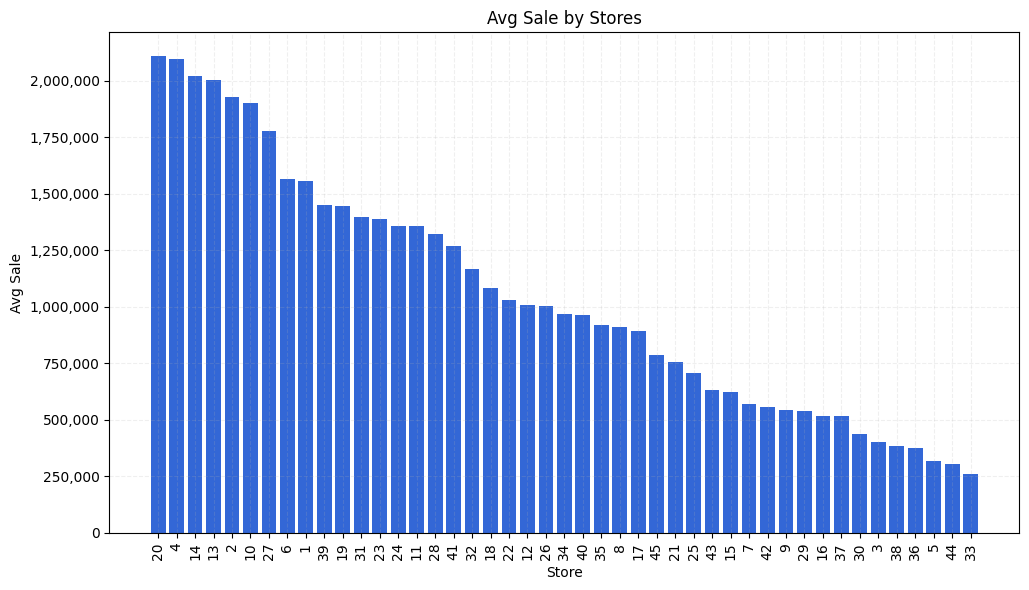


* Overall correlation between weekly sales and CPI was not significant.
* At a store level:
  + Stores 38 and 44 showed a significant positive correlation between weekly sales and CPI.
  + Store 36 showed a significant negative correlation between weekly sales and CPI.
  + Other stores didn't exhibit significant correlations.

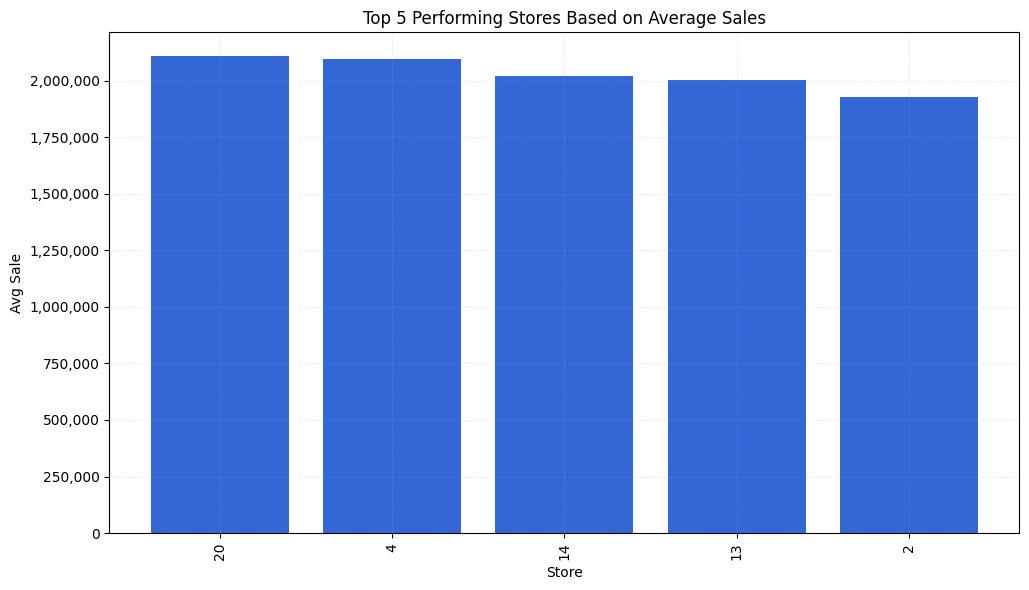
1. **Top Performing Stores:**

* To find this we plotted the total and average weekly sales of all the stores

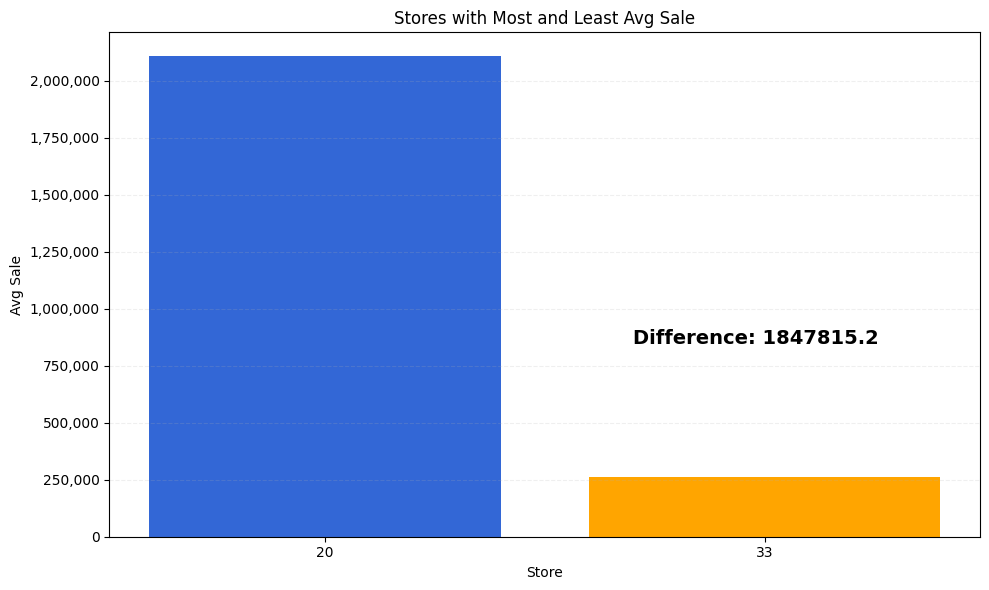




* The total and average weekly sales show a very similar trend. Hence, we can pick any one and be sure it will represent the complete picture
* As per the historic data, following are the top 5 performing stores

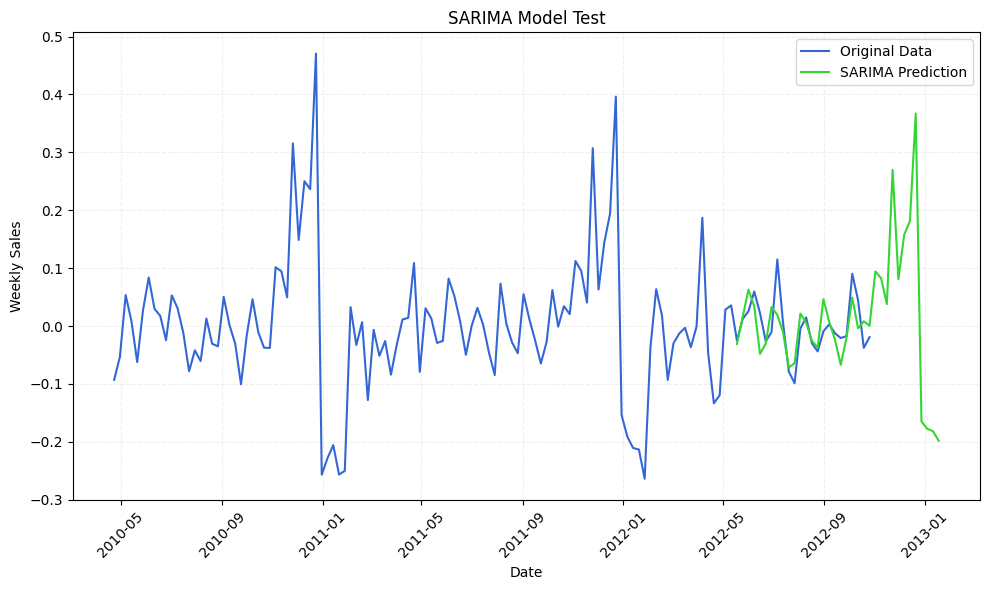


1. **Worst Performing Store and Disparity in Performance:**



1. **Sales Forecasting Using SARIMA Model:**

* A SARIMA model was trained on historical data from one store (store 20) to forecast sales for the next 12 weeks.



* Similar predictive modeling techniques could be applied to each store individually or at an overall level to predict weekly sales.

# **Future Possibilities**

There are several future possibilities and directions for this project:

1. **Further Analysis and Investigation:**
   1. Conduct deeper analysis to understand the underlying factors contributing to the observed correlations and trends. For example, investigate why certain stores show counterintuitive correlations between unemployment rate and weekly sales.
   2. Explore additional variables or external factors that might influence sales, such as promotional activities, competitor analysis, economic indicators, or local events.
2. **Refinement of Predictive Models:**
   1. Refine and optimize predictive models for more accurate forecasting of weekly sales. Experiment with different time series forecasting techniques beyond SARIMA, such as Prophet, LSTM networks, or ensemble methods, to improve prediction performance.
   2. Develop models that can capture and account for the seasonal and cyclical nature of sales data more effectively.
3. **Segmentation and Targeted Strategies:**
   1. Segment stores based on their sales patterns, geographical location, customer demographics, or other relevant criteria.
   2. Develop targeted marketing and sales strategies for different store segments to maximize revenue and profitability.
4. **Integration with Business Operations:**
   1. Integrate insights and forecasts from the analysis into decision-making processes within the organization.
   2. Implement data-driven strategies to optimize inventory management, staffing levels, pricing strategies, and marketing campaigns.
5. **Real-Time Monitoring and Adaptive Strategies:**
   1. Develop real-time monitoring systems to track sales performance and external factors continuously.
   2. Implement adaptive strategies that can dynamically adjust based on changes in market conditions, customer behavior, or other relevant variables.

# **Conclusion**

In conclusion, this project has provided valuable insights into the factors influencing weekly sales for a set of stores. Key findings include:

1. **Unemployment Rate Impact:** While overall correlation between weekly sales and unemployment rate was not significant, at a store level, notable correlations were observed. Stores 44 and 38 exhibited a significant negative correlation, whereas stores 36 and 35 showed a counterintuitive positive correlation.
2. **Seasonal Trends:** Seasonal decomposition revealed a clear seasonal trend in weekly sales, with peaks during holiday seasons. Various factors such as holidays, weather changes, and school schedules could contribute to this seasonal variation.
3. **Temperature Influence:** There was no significant correlation between weekly sales and temperature, indicating that temperature does not have a significant impact on sales.
4. **Consumer Price Index (CPI) Impact:** While overall correlation between weekly sales and CPI was not significant, certain stores showed significant positive or negative correlations between weekly sales and CPI.
5. **Top Performing Stores:** Based on historical data, the top performing stores were identified to be store 20, 4, 14, 13, and 2, providing insights into which stores are driving higher sales volumes.
6. **Worst Performing Store:** Based on historical data, the worst performing store was identified to be store 33, performing on average with a difference in weekly sales of about 1.8 million, when compared with the best performing store.
7. **Sales Forecasting:** A SARIMA model was employed to forecast sales for one store, demonstrating the potential for predictive modeling to anticipate future sales trends.

Moving forward, there are several future possibilities for this project, including deeper analysis, refinement of predictive models, targeted strategies, integration with business operations, and real-time monitoring. By leveraging these possibilities, the organization can make informed decisions to optimize sales performance and drive sustainable growth.

Here is the [link](https://drive.google.com/file/d/14hW5H6Db6hnW8Hp2SAsjwvAdDItDI9cP/view?usp=sharing) to the complete analysis