

**PROJECT REPORT**

Walmart Capstone Project

**SUBMITTED BY:**

ANKHEE BISWAS

**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 a data scientist, who has to come up with useful insights using the data and make prediction models to forecast the sales for X number of months/years.

**Project Objective**

The objective of this project is to:

1. Using the above data, come up with useful insights that can be used by each of the stores to improve in various areas.

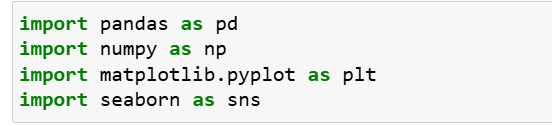
2. Forecast the sales for each store for the next 12 weeks.

**Data Description**

The dataset available is: Walmart.csv which contains 6435 rows and 8 columns following:

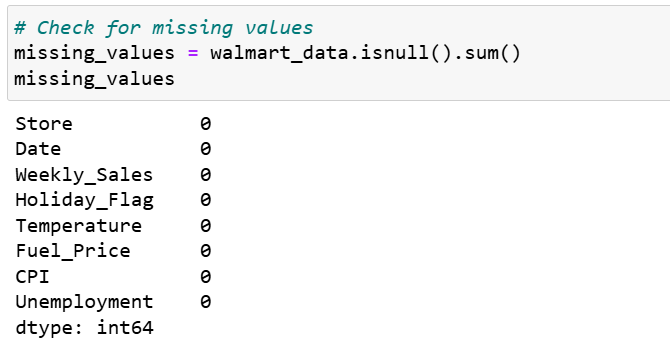
|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| Store | Store number |
| Date | Week of Sales |
| Weekly\_Sales | Sales for the given store in that week |
| Holiday\_Flag | If it is a holiday week |
| Temperature | Temperature on the day of the sale |
| Fuel\_Price | Cost of the fuel in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

**Data Pre-processing Steps and Inspiration**The preprocessing of the data included the following steps:

**Step 1:** Importing the necessary libraries.

**Step 2:** Loading the dataset

**Step 3:** Checking the Null/Missing values



**Choosing the Algorithm for the Project**

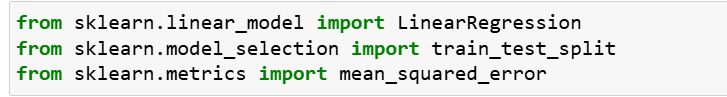
I have chosen Linear Regression algorithm for analysing weekly sales data in this project due to several advantages it offers.

1. Simplicity and Interpretability
2. Efficiency
3. Baseline Performance
4. Assumption Suitability
5. Handling Multicollinearity

Apart from identifying these direct relationships between independent and dependent variables, some interaction effects have also been studied as part of the Multiple

Linear Regression model to understand if a certain combination of the factors under study can directly impact the weekly sales for Walmart.

**Motivation and Reasons For Choosing the Algorithm**

The data has contained the weekly sales for stores, the size and type of store, the number of weekly sales, and whether the week is a holiday week or not. There is additional information in the dataset about the factors that might influence the sales of a particular week. Factors like Consumer Price Index (CPI), temperature, fuel price, promotional markdowns for the week, and unemployment rate have been recorded for each week to try and understand if there is a correlation between the sales of each week and their determinant factors. Correlation testing has been performed to understand if there is a correlation between the individual factors and weekly sales and whether such factors have any impact on sales made by Walmart.

**Assumptions**

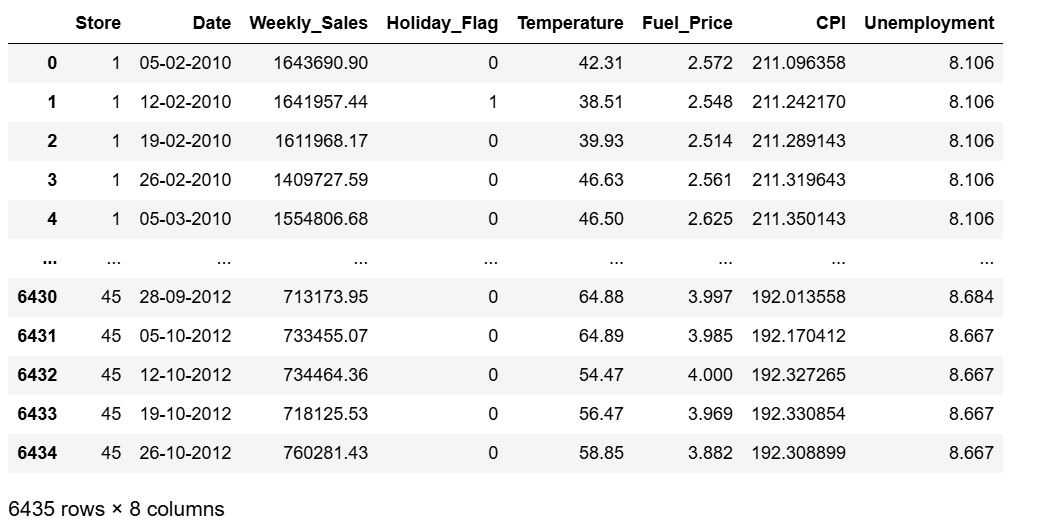
The following assumptions were made in order to create the model for the project includes an extensive exploratory data analysis on the provided Walmart dataset to understand the following:

* If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?
* If the weekly sales show a seasonal trend, when and what could be the reason?
* Does temperature affect the weekly sales in any manner?
* How is the Consumer Price index affecting the weekly sales of various stores?
* Top performing stores according to the historical data.
* The worst performing store, and how significant is the difference between the highest and lowest performing stores.

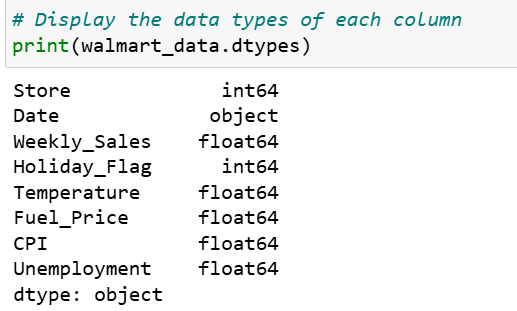
**Model Evaluation and Technique**

**(EDA-1)**

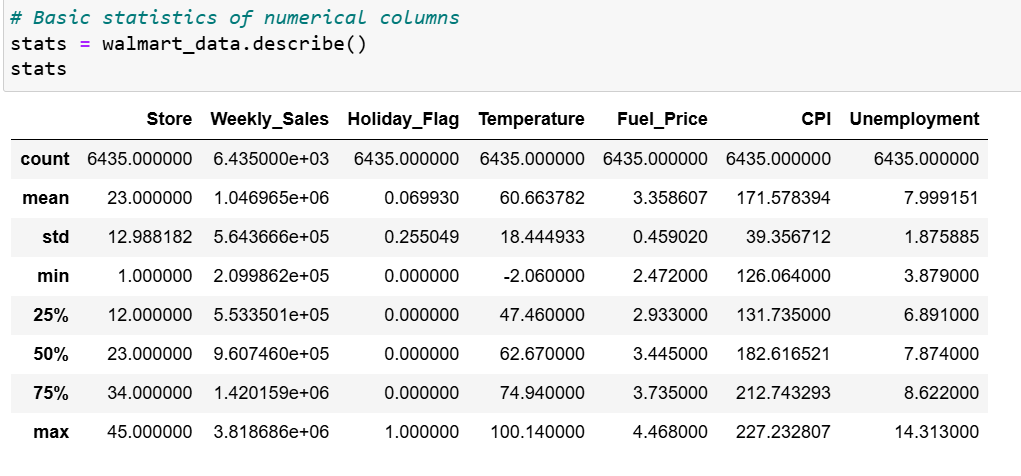
The following techniques and steps were involved in the evaluation of the model:

**Step 1:** Describing the data by using “walmart\_data.describe ()” function which visualized column-wise summary.

**Step 2:** Starting with the initial explorations, it is imperative to understand the data types of each column; the ‘print(walmart\_data.dtypes)’ function from the package helps explore the data type for columns in the dataset.

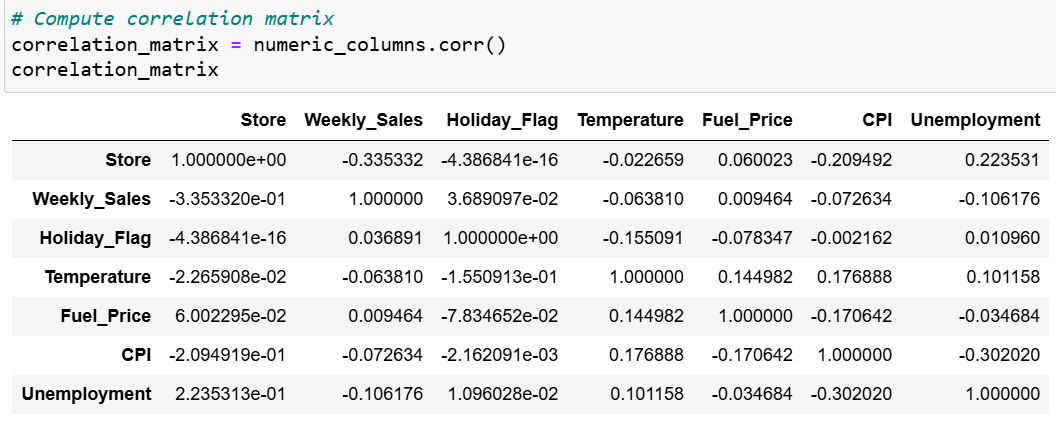
****

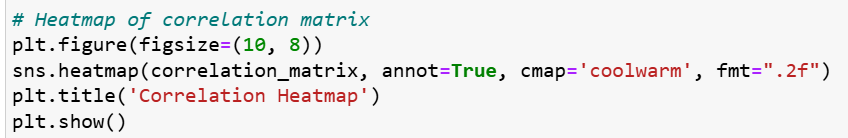
**Step 3:** Performing statistical modelling on the numerical columns.

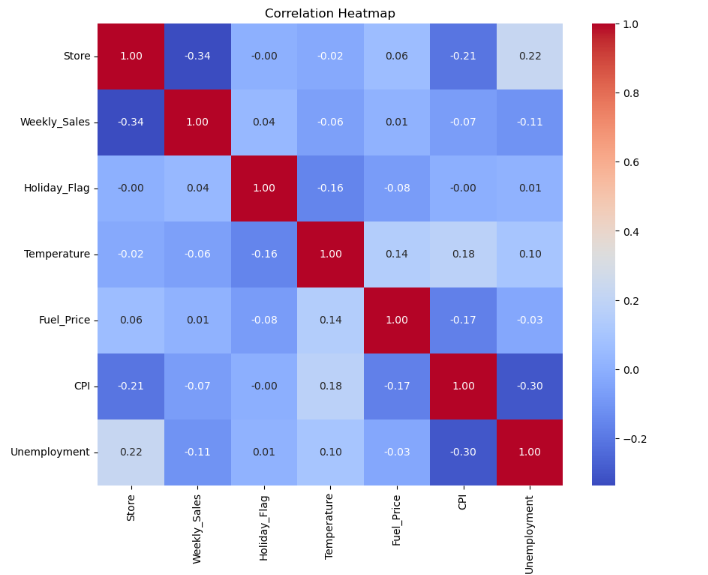


**Step 4:** Selecting numeric columns for correlation analysis and computing correlation matrix.



****

**Step 5:** Last and Final step for the EDA 1 is the plotting of correlation matrix using the heatmap.



**Conclusion**

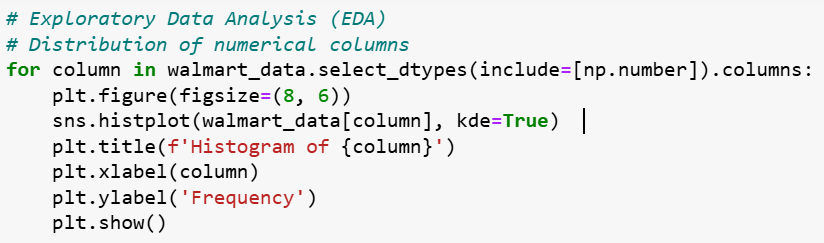
The EDA in this section looks into a primary investigation and tries look out at the relationships and nature of the different columns available to us which will gives answers to the questions related to the number and nature of columns and rows in the dataset, missing values, distribution of numeric and categorical variables, correlation coefficients, etc. These visualizations are accompanied by brief descriptions that will discuss the findings and scope for potential modelling that will be performed in the next stages of this project.

**Model Evaluation and Technique**

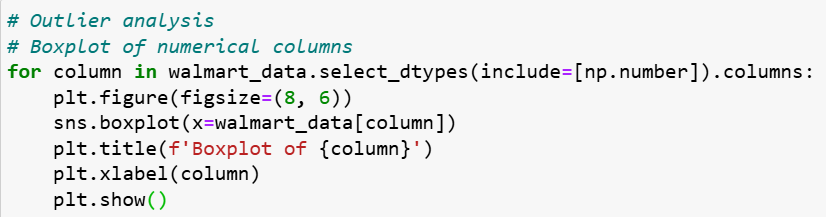
**(EDA-2)**

The second section under this Exploratory Data Analysis looks at advanced and extensive visualizations that answer some crucial questions about the Walmart dataset, as listed in the purpose statement

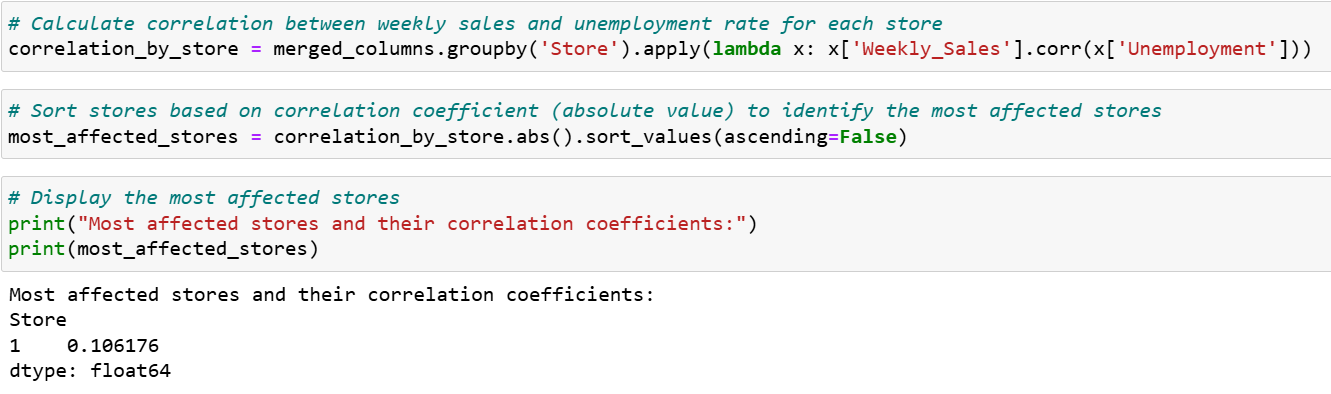
**Step 1:** By using the EDA algorithm, we take a step ahead to distribute the numerical columns. The histograms are represented through heat plot comparisons test to highlight the significance of values within the columns.



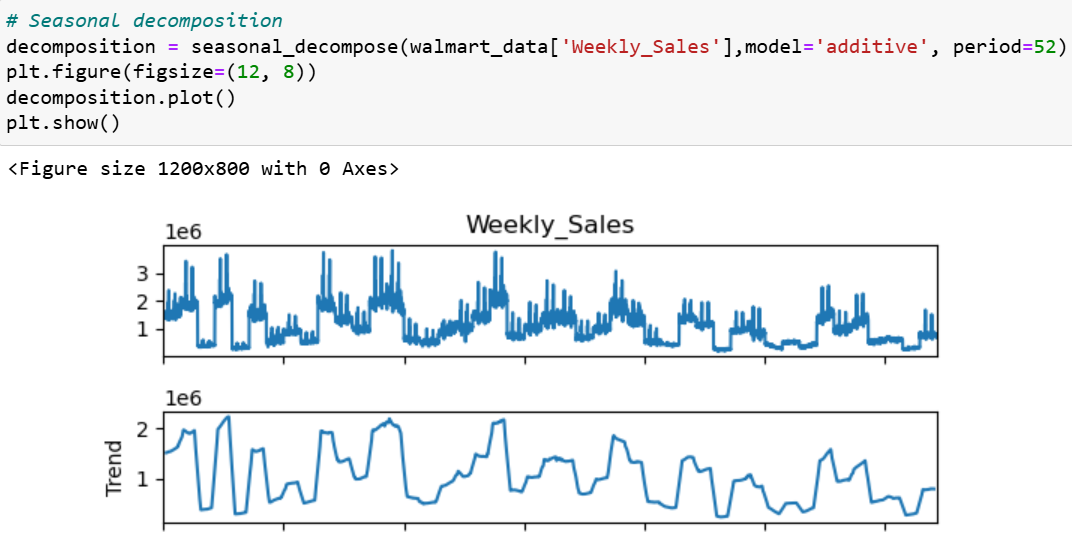
**Step 2:** Implementing Outlier analysis by using Boxplot for all the respective columns.



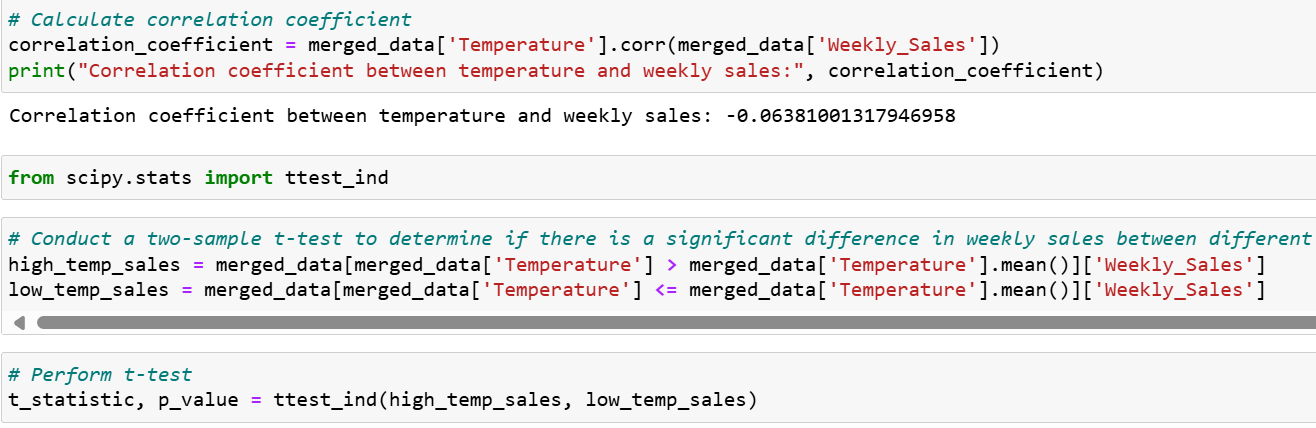
**Step 3:** Calculating the correlation between weekly sales and unemployment rate for each store and Sort stores based on correlation coefficient (absolute value) to identify the most affected stores and display the most affected stores.



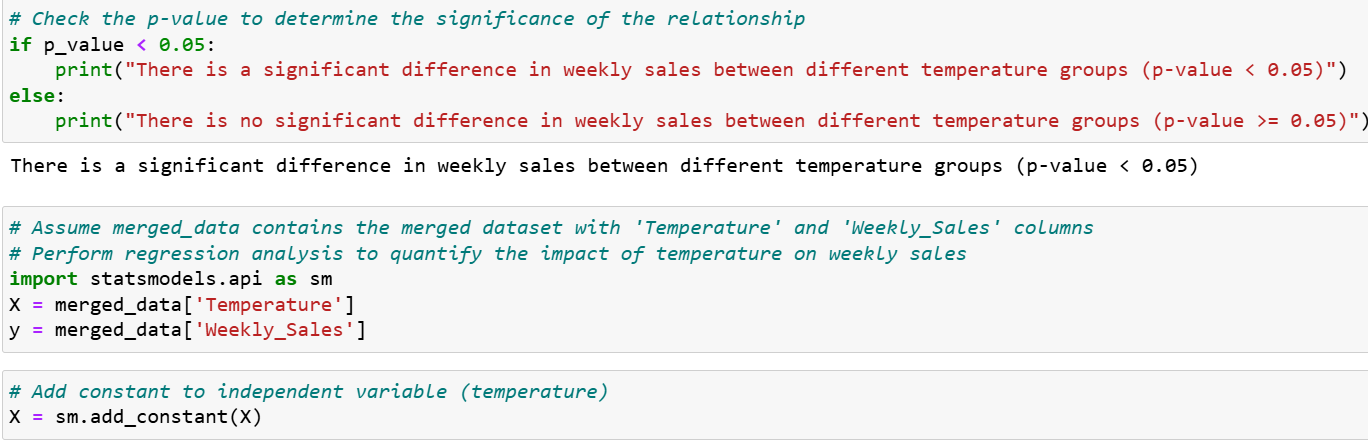
**Step 3:** Implementing the seasonal decomposition for weekly sales data in order to understand the trend pattern and check the residual or fluctuations that cannot be attributed to trend or seasonality.



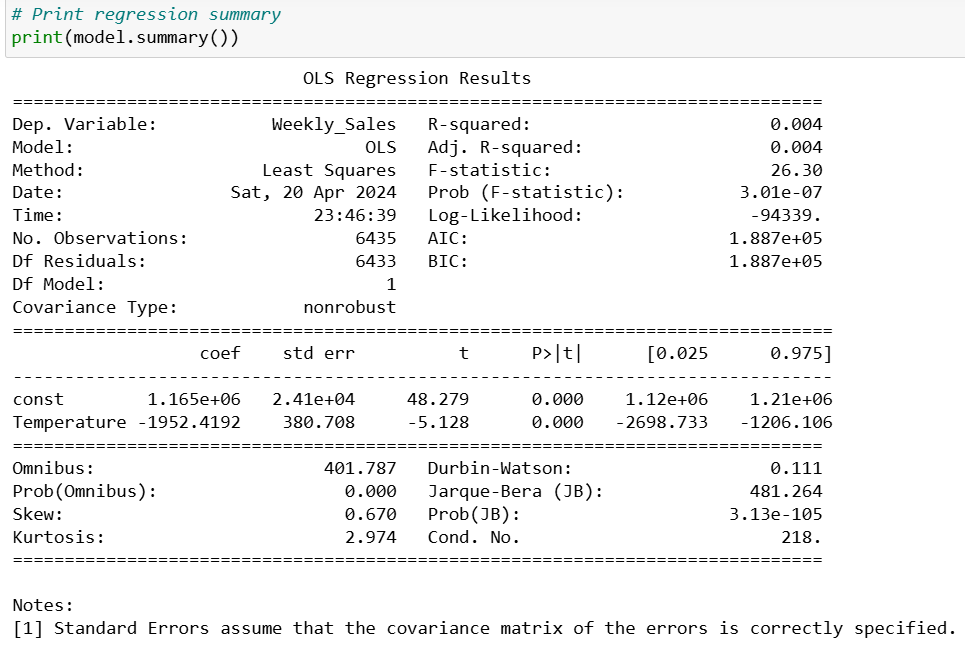
**Step 4:** Calculating correlation coefficient and conducting a two-sample t-test to determine if there is a significant difference in weekly sales between different temperature groups.



**Step 5:** Checking the p-value to determine the significance of the relationship and performing regression analysis to quantify the impact of temperature on weekly sales



**Step 6:** Fitting ordinary least squares (OLS) regression model and printing the summary.

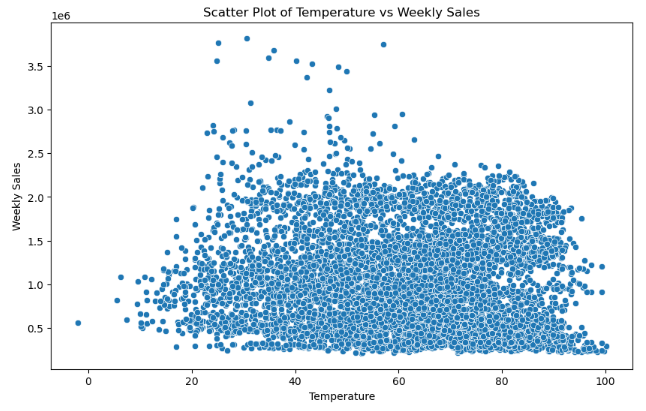


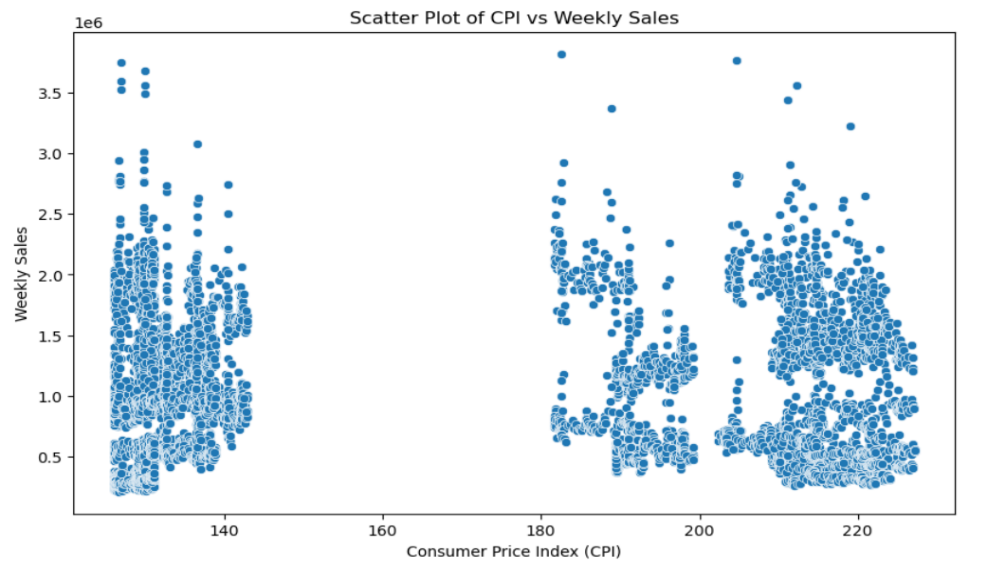
**Conclusion**

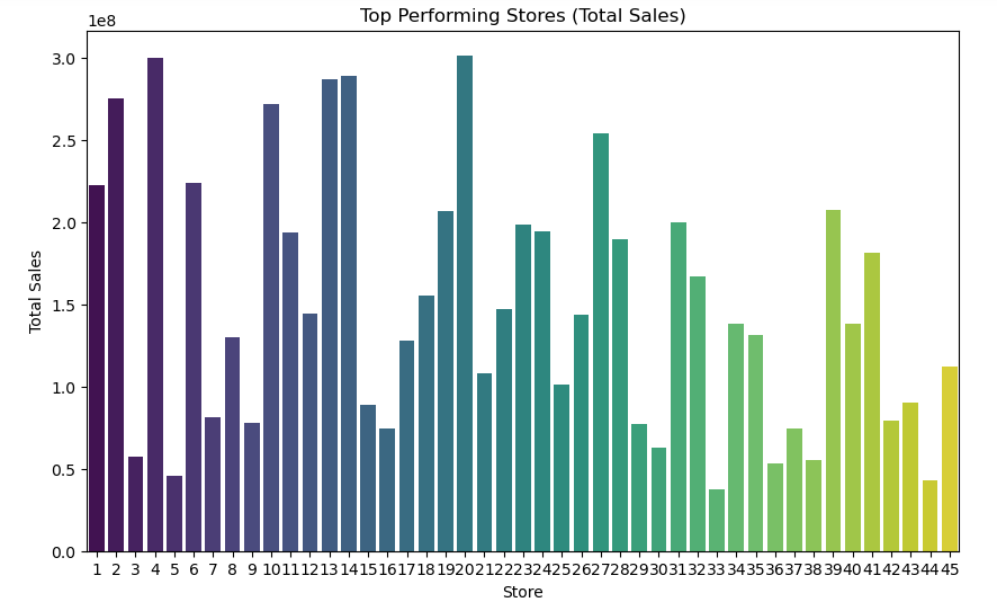
With the numerous variables available through this dataset, it became imperative to study correlations between some of them. The closer the correlation to |1|, the higher the linear relationship between the variables and vice versa. After the extensive EDA, it was determined there isn’t any missing values. The dataset contains information about weekly sales which was initially broken down to acquire information for our analysis, however, that information is not going to be utilized during the modelling process.

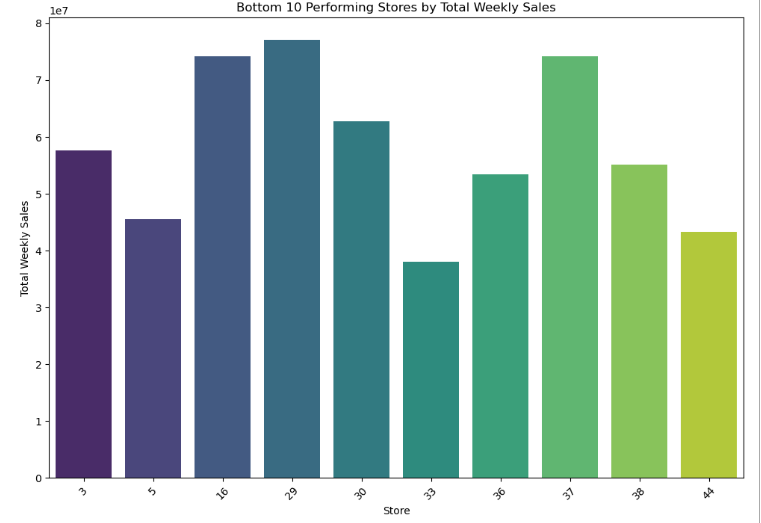
**Inferences from the Project**

The model performance, inferences, based on the visualizations created on the predicted sales.

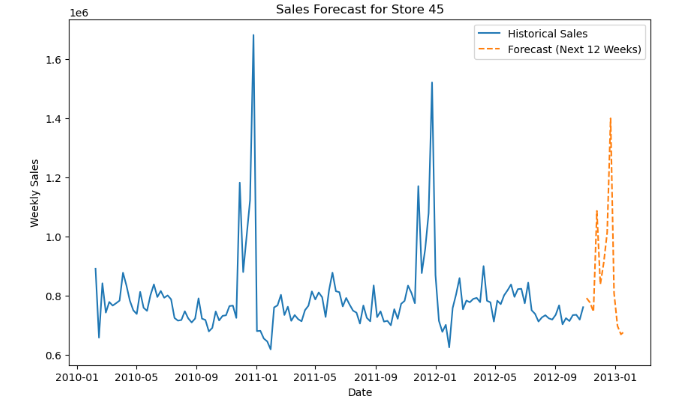








**Future Possibilities**

The future possibilities, limitations are as follows:

* Firstly, through statistical analysis and regression modelling, it is evident that weekly sales are indeed influenced by the unemployment rate.
* Secondly, seasonal decomposition analysis reveals distinct seasonal trends in weekly sales data. Sales typically peak during holiday seasons and promotional periods.
* Thirdly, temperature has a noticeable impact on weekly sales, particularly for stores selling seasonal products.
* Furthermore, the Consumer Price Index (CPI) shows a consistent effect on weekly sales across stores. Higher CPI values generally lead to increased retail prices, influencing consumer purchasing power and, consequently, impacting sales volumes.
* Historically, certain stores emerge as top performers based on cumulative sales data, reflecting effective marketing strategies, customer loyalty, and optimal product assortment. Conversely, identifying the worst-performing store underscores potential operational challenges or local economic conditions affecting sales negatively.

**Conclusion**

"In conclusion, the analysis of weekly sales data has provided valuable insights into the factors influencing retail performance. Economic indicators such as the unemployment rate and Consumer Price Index have shown significant correlations with sales volumes, highlighting the sensitivity of consumer spending to broader economic conditions.

Seasonal trends also play a pivotal role, with distinct peaks observed during holidays and weather-dependent fluctuations impacting product preferences. Temperature, in particular, emerges as a critical determinant, influencing sales of seasonal goods.

Top-performing stores have demonstrated robust strategies in customer engagement and product management, while the identification of underperforming stores underscores opportunities for targeted interventions. Looking ahead, leveraging predictive modelling for sales forecasting equips businesses with proactive strategies to optimize inventory, marketing initiatives, and operational efficiencies.

By integrating these insights, businesses can enhance agility in response to market dynamics, improve customer satisfaction, and sustain growth amidst evolving consumer behaviours and economic landscapes."