Capstone Project - Walmart

Table of Contents

Problem Statement

The dataset for this project, walmart.csv, represents weekly sales data from various Walmart outlets across the country. These retail stores, which have multiple outlets nationwide, are facing challenges in managing their inventory to effectively match supply with demand. The dataset contains 6435 rows and 8 columns. The columns represent the following features:

Store: The store number
Date: The week of sales
Weekly_Sales: The sales for the given store in that week
Holiday_Flag: A flag indicating if it is a holiday week
Temperature: The temperature on the day of the sale
Fuel_Price: The cost of fuel in the region
CPI: The Consumer Price Index
Unemployment: The Unemployment Rate

Project Objective

The objective of this project is to analyze and understand the sales performance of a retail store with multiple outlets across the country, which is currently facing challenges in managing its inventory to match supply with demand.

The project aims to:

1. Conduct a comprehensive statistical analysis and Exploratory Data Analysis (EDA) of the provided dataset, handle missing values, and identify any outliers.
2. Derive insights on how factors such as unemployment rate, seasonality, temperature, and Consumer Price Index (CPI) affect weekly sales.
3. Identify the top and worst-performing stores based on historical data and quantify the difference between them.
4. Use predictive modeling techniques to forecast the sales for each store for the next 12 weeks.

The ultimate goal is to provide the retail store with actionable insights and accurate sales forecasts that can aid in effective inventory management and overall business decision-making.

Data Description

The dataset for this project, walmart.csv, contains weekly sales data for multiple outlets of a retail store. Here's a description of the data:

1. **Store:** This is the unique identifier for each store. There are 45 stores in total.
2. **Date:** This represents the week of sales. The data is recorded on a weekly basis, and each store has 143 records, representing 143 weeks of data.
3. **Weekly_Sales:** This represents the sales for the given store in that week. It's the target variable we want to predict with our model.
4. **Holiday_Flag:** This is a binary variable that indicates whether the week is a holiday week (1) or not (0). Holidays are expected to have more sales.
5. **Temperature:** This represents the temperature on the day of the sale. The temperature can affect what products are sold.
6. **Fuel_Price:** This represents the cost of fuel in the region. The fuel price can affect transportation costs, which can indirectly influence product prices and sales.
7. **CPI:** This stands for the Consumer Price Index. It reflects the average spending power of consumers in a particular region. Stores located in regions with higher CPI values tend to have higher sales figures.
8. **Unemployment:** This represents the unemployment rate in the region. If the unemployment rate is high, it might indicate that fewer people have disposable income to spend, which could lead to lower sales.

The dataset contains 6435 rows, where each row represents a week of sales data for a particular store. The goal is to use this data to understand the factors affecting sales and to predict future sales.

Data Preprocessing Steps And Inspiration

1) Data Overview

- 1.1 First 5 Rows
- 1.2 Last 5 Rows
- 1.3 Dataset Shape
- 1.4 Data Information

2) Checking for Duplicates

3) Summary Statisitc

- 3.1 describe():
- 3.2 Frequency counts for the store column:
- 3.3 Frequency counts for the 'Holiday_Flag' column:

4) Handling missing values

5) Outlier Analysis

6) Univariate Analysis

- 6.1 Examine the distribution of Variables using histograms
- 6.2 Feature Engineering
  - 6.2.1 Converting 'Date' column to datetime:
  - 6.2.2 Temporal Features:
- 6.3 Analyzing categorical variable
  - 6.3.1 Comparison of Weekly Sales between Holidays and Non-Holidays:
  - 6.3.2 Distribution of Observations by Year:
  - 6.3.3 Distribution of Observations by Season:
  - 6.3.4 Distribution of Observations by Quarter:
  - 6.3.5 Distribution of Observations by Month:

7) Bivariate Analysis

- 7.1 Mean Weekly Sales for Each Holiday Flag:
- 7.2 Total Weekly Sales for Each Holiday Flag:
- 7.3 Sum of Weekly Sales for Each Store:
- 7.4 Scatter Plot Analysis
  - 7.4.1 Temperature vs Weekly Sales:
  - 7.4.2 Fuel Price vs Weekly Sales:
  - 7.4.3 CPI vs Weekly Sales:
  - 7.4.4 Unemployment vs Weekly Sales:
- 7.5 Analysis of Seasonal Trends in Weekly Sales:
  - 7.5.1 Total sales in each year:
  - 7.5.2 Total Sales in each Season:
  - 7.5.3 Total Sales in each Quarter:
  - 7.5.4 Total Sales in each Week:
  - 7.5.5 Visualize Weekly Sales Over Time:
  - 7.5.6 Weekly_sales over the years:

1) Data Overview

It involves visually inspecting the data, understanding its structure, and getting a general sense of its contents.

1.1 First 5 Rows:

- Displaying the first 5 rows of the dataset provides an initial glimpse into the structure and content of the data.

1.2 Last 5 Rows:

Examining the last 5 rows allows us to confirm the completeness of the dataset and observe any potential patterns or trends towards the end.

1.3 Dataset Shape:

The dataset comprises 6435 rows and 8 columns, providing a comprehensive view of the data's size and dimensions.

1.4 Data Information:

- Utilizing the **df.info()** function, we extracted valuable insights into the data types present in each column, ensuring a clear understanding of the attributes and their characteristics.
- From the info() method, we can deduce that our dataset encompasses three different data types: int64, float64, and object.
- Datatype of column:
  - The Store and Holiday_Flag columns are of type int64.
  - The Date column is categorized as object, typically indicative of string or categorical data.
  - If the Date column represents dates, it might be advantageous to convert it to a datetime data type for facilitating date-related operations.
  - The remaining columns have a data type of float64.
- About missing values:
  - By employing this method, we can effortlessly detect the presence of any missing values.
  - Here since each column contains 6435 observations, corresponding to the identical number of rows we previously observed using the shape attribute. So we can say that none of the column has missing values.

This initial exploration serves as a foundation for further analysis, enabling us to delve deeper into the dataset and extract meaningful insights to address the outlined problem statement.

2) Checking for duplicates

Identify and remove duplicate rows in the dataset to ensure data integrity.

```
1 df.duplicated().sum()
```

```
0
```

- Sum is 0, meaning there are no duplicate rows in the DataFrame.

3) Summary statistics

It provides a concise overview of the numerical attributes within the dataset, offering insights into their central tendencies, dispersion, and distributional characteristics.

3.1 describe():

The describe method in pandas compute descriptive statistics for numerical variables like count, min, max etc.

```
1 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

Store Distribution: There are 45 unique stores represented in the dataset, ranging from Store 1 to Store 45. The distribution of data across stores is relatively balanced, with each store having a similar number of observations on average.

Weekly Sales Distribution: The average weekly sales across all stores is approximately $1,046,965. The minimum and maximum weekly sales are approximately $209,986 and $3,818,686 respectively. This indicates a significant variation in weekly sales.

Holiday Flag Distribution: Approximately 7% of the data points are from holiday weeks.

Temperature Distribution: The average temperature is about 60.66°F, with a minimum of -2.06°F and a maximum of 100.14°F. The distribution of temperatures appears to be relatively normal.

Fuel Price Distribution: The average fuel price is approximately $3.36 per gallon, with prices ranging from around $2.47 to $4.47 per gallon. The distribution of fuel prices appears to be relatively normal.

Consumer Price Index (CPI) Distribution: The average CPI is approximately 171.58, with values ranging from approximately 126.06 to 227.23. The distribution of CPI values may vary across different regions and time periods.

Unemployment Rate Distribution: The average unemployment rate is approximately 8%, with rates ranging from approximately 3.87% to 14.31%. The distribution of unemployment rates may reflect economic conditions and regional variations.

Overall, these insights provide a comprehensive overview of the distribution and variability of each feature in the dataset. Further analysis, such as correlation analysis and visualization, can help uncover additional patterns and relationships within the data.

3.2 Frequency counts for the store column:

```
store_counts = df['Store'].value_counts()
store_counts
```

```
Store
1     143
24    143
26    143
```

So from above data we can tell that each store have 143 records, suggesting consistent data collection across all locations.

3.3 Frequency counts for the 'Holiday_Flag' column:

```
holiday_counts = df['Holiday_Flag'].value_counts()
holiday_counts
```

```
Holiday_Flag
0    5985
1     450
Name: count, dtype: int64
```

There are 5985 instances (or weeks) where Holiday_Flag is 0, indicating regular weeks without any holidays. Additionally, there are 450 instances where Holiday_Flag is 1, signifying these weeks include at least one holiday.

4) Handling missing values

Checking if any column having null values:

```
1 df.isnull().sum()
```

```
Store           0
Date            0
Weekly_Sales    0
Holiday_Flag    0
Temperature     0
Fuel_Price      0
CPI             0
Unemployment    0
dtype: int64
```

- Previously, we determined that our dataset has no missing values by using the info() method. We reverified this by using the isna() function, confirming that our dataset does not contain any null values.
- We also checked for any irregularities in our data and found none.

5) Outlier Analysis:

Detecting Outliers Using Boxplots: Boxplots are a useful visualization tool for identifying outliers within numerical columns. By visually inspecting the distribution of data points in each column, we can identify any observations that lie significantly outside the typical range of values.

I have generated a boxplot for the following columns: 'Store', 'Weekly_Sales', 'Holiday_Flag', 'Temperature', 'Fuel_Price', 'CPI', and 'Unemployment'.

Note:1

- For categorical variables with only two unique values (often represented as 1 and 0) like here in 'Holiday_Flag', the concept of outliers doesn't apply in the traditional sense. Outliers are typically defined as data points that significantly deviate from the rest of the observations in a numerical distribution.
- so we have outliers in 'Weekly_sales', 'Temperature' and 'Unemployment' columns

Note:2

- In many cases, 'Weekly_Sales' is considered as the target or dependent variable in predictive modeling tasks, such as sales forecasting.
- Removing outliers from the target variable may lead to biased model predictions.
- If outliers represent genuine data points (e.g., exceptionally high sales during holiday seasons), removing them can distort the model's ability to capture such patterns.

- However, it's essential to preprocess other independent variables (such as 'Temperature' and 'Unemployment') to mitigate the impact of outliers on model training and performance.

Note3 :

- Here 'Temperature' and 'Unemployment' columns have outliers, it could be due to natural fluctuations in weather and economic conditions.
- In such cases, these outliers are actual representations of the variability in the data, and removing them might lead to loss of information.
- I have checked before if we remove these outliers
- So before removing outlier, we have 6435 rows
- but after removing outlier we got only 5951 rows,
- means we have removed 484 rows and i.e. nothing but loss of 7.5% of data and i.e not a small amount. which could potentially include important, meaningful data.
- Therefore i am thinking filling data with median is good way to resolve this issue.

Imputation: Replace the outlier values with statistical measures such as mean, median or mode.

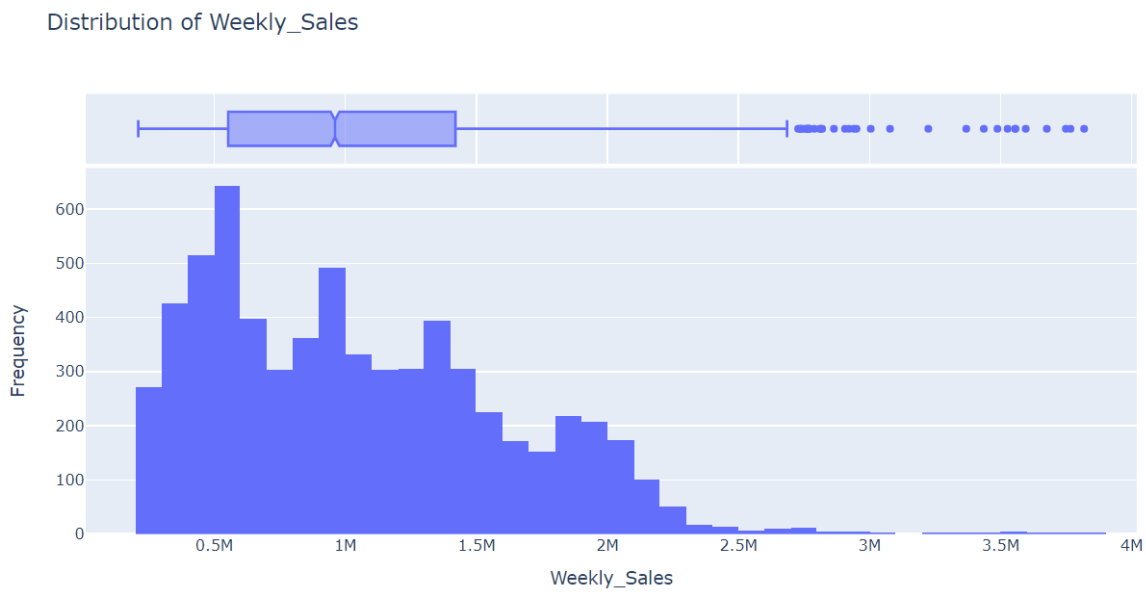
6) Univariate Analysis:

It focuses on understanding the distribution and characteristics of a single variable.

6.1 Examine the Distribution of Variables Using Histograms:

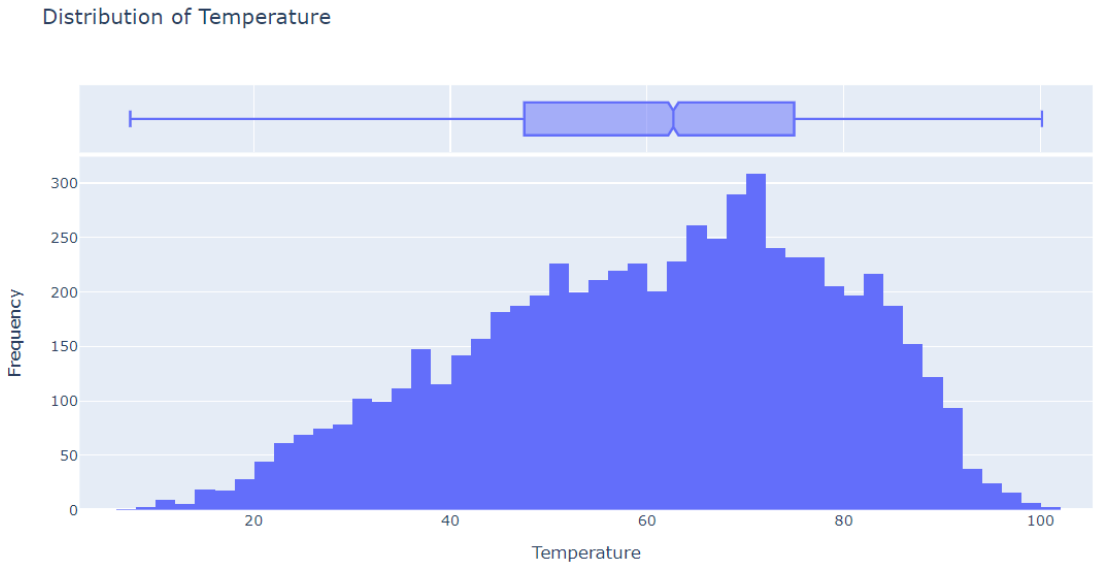
Understanding the distribution of key variables is relevant for understanding the factors that influence weekly sales and overall store performance. By visualizing their distributions through histograms, we can gain insights into the variability, central tendency, and potential relationships within the data, aiding in the analysis and decision-making process for inventory management strategies.

Weekly_Sales: This is the target variable, and understanding its distribution is crucial for analyzing sales patterns across different stores and time periods. A histogram of weekly sales can provide insights into the range of sales values, the presence of outliers, and the overall distribution shape.



- The distribution of weekly sales is right-skewed, meaning there are a few weeks with exceptionally high sales.
- The majority of the weekly sales values fall within the box in the box plot, while there are some outliers represented as individual points above the box.
- The most common range of weekly sales is around 0.5M, as indicated by the highest bar in the histogram.

Temperature: Analyzing the distribution of temperature values can help identify seasonal patterns and understand how temperature variations might impact sales of certain products, such as seasonal items or weather-dependent goods.

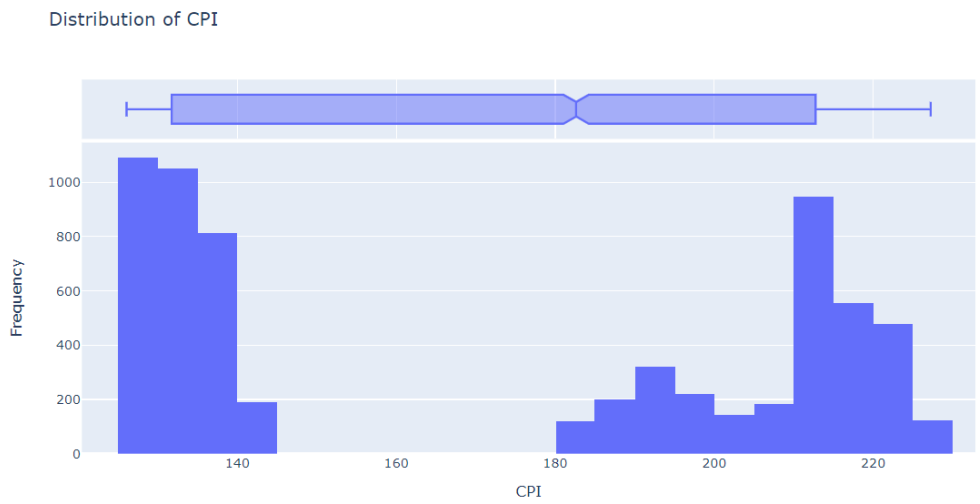


- The shape of the distribution appears to be somewhat bell-shaped, peaking around a temperature of 60. This suggests that most of the temperature values in your dataset are around this value.
- There are fewer occurrences of very low and very high temperatures, as indicated by the lower bars at the ends of the histogram.

Fuel_Price: Examining the distribution of fuel prices can reveal trends in fuel costs over time and their potential impact on transportation expenses, which may influence sales patterns and store profitability.

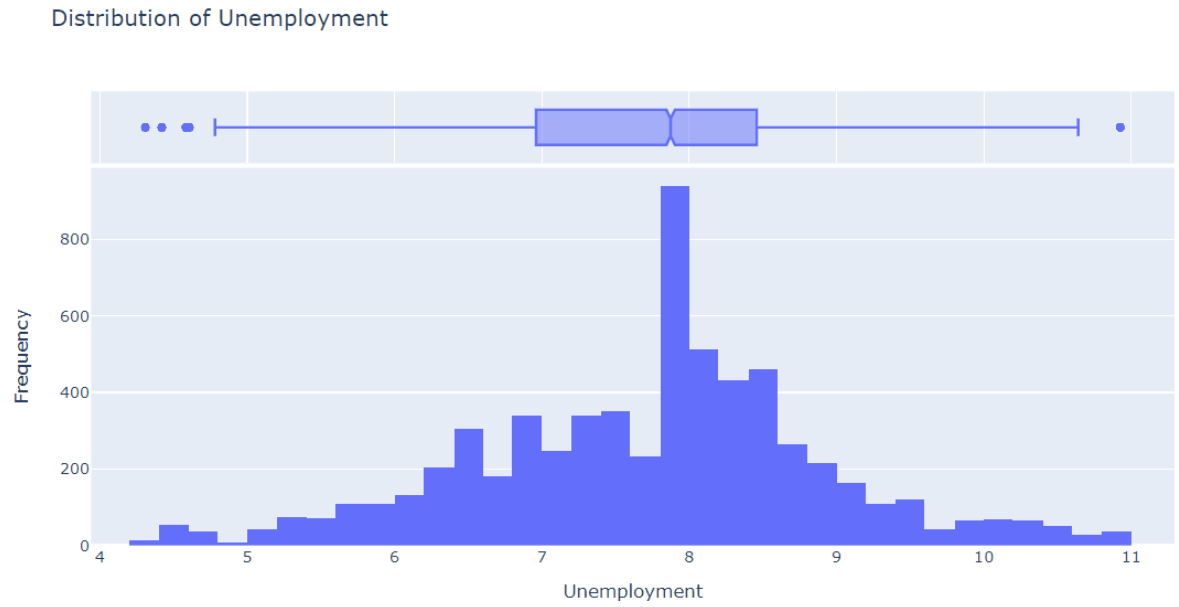
- The shape of the distribution appears to be somewhat bell-shaped, peaking around a fuel price of 3.75 to 3.799. This suggests that most of the fuel prices in your dataset are around this value.
- The histogram shows that the frequency of fuel prices between 2.5 and 3.5 is relatively lower compared to the peak around 3.75 to 3.799.
- There are fewer occurrences of very low and very high fuel prices, as indicated by the lower bars at the ends of the histogram.

CPI (Consumer Price Index): Understanding the distribution of CPI values can provide insights into regional economic conditions and consumer spending power, which may affect purchasing behavior and sales trends.



- The shape of the distribution appears to be bimodal, with two peaks: one around a CPI of 140 and another just above 200.
- This suggests that most of the CPI values in your dataset are around these two values.
- There are fewer occurrences of very low and very high CPI values, as indicated by the lower bars at the ends of the histogram.

Unemployment: Analyzing the distribution of unemployment rates can help assess the economic environment in different regions and understand how employment trends might influence consumer confidence and spending patterns.



- The shape of the distribution appears to be somewhat bell-shaped, peaking around an unemployment rate of around 8. This suggests that most of the unemployment rates in your dataset are around this value.

- There are fewer occurrences of very low and very high unemployment rates, as indicated by the lower bars at the ends of the histogram.

In summary:

- Weekly_Sales is right-skewed
- Temperature, Fuel Price, and Unemployment Rate appear to be somewhat normally distributed
- CPI appears to be bimodal, indicating that there are two distinct peaks or modes in the distribution

6.2) Feature Engineering

Feature engineering plays a crucial role in data preprocessing, enabling us to extract valuable information from raw data and enhance the performance of analytical models. In this section, we detail the feature engineering techniques applied to our dataset to derive meaningful features for our analysis.

6.2.1 Converting 'Date' column to datetime:

- During info() method we have seen 'Date' column is of object type, !!!Right
- So it is good to have our 'Date' column in a standard date format.across the dataset, enabling seamless manipulation and analysis of temporal data.

```
1  df["Date"] = pd.to_datetime(df["Date"], format="%d-%m-%Y")
```

```
1  df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6435 entries, 0 to 6434
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype         
---  ------        --------------  -----         
 0   Store         6435 non-null   int64         
 1   Date          6435 non-null   datetime64[ns]
 2   Weekly_Sales  6435 non-null   float64       
 3   Holiday_Flag  6435 non-null   int64         
 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: datetime64[ns](1), float64(5), int64(2)
memory usage: 402.3 KB
```

So now, we have one column of 'datetime' data type, and the remaining columns are either of 'int64' or 'float64' data types, which represent numerical values. Therefore, there is no need for label or one-hot encoding to convert categorical variables into numerical format

6.2.2 Temporal Features:

Now the 'Date' column is in a standard date format, so now we are going to derive new temporal features to capture various aspects of time-related information. This included features such as day of the month, day of the week, week number, month, month name, quarter, season, and year.

These features include:

- **Day:** Represents the day of the month.
- **Day_Of_Week:** Indicates the day of the week (e.g., Monday, Tuesday).
- **Week:** Identifies the week number within the year.
- **Month:** Represents the numeric month (e.g., January = 1, February = 2).
- **Month_Name:** Specifies the name of the month.
- **Quarter:** Identifies the quarter of the year (e.g., Q1, Q2, Q3, Q4).
- **Season:** Categorizes the data into seasons (e.g., Winter, Spring, Summer, Fall) based on the month.
- **Year:** Represents the year component of the date.

These engineered features provide valuable insights into temporal patterns and seasonal variations, which are crucial factors influencing store performance and sales trends. By incorporating these features into our analysis, we aim to enhance the predictive capability of our models and gain deeper insights into the underlying dynamics of the datase

```
1  df.info()
```

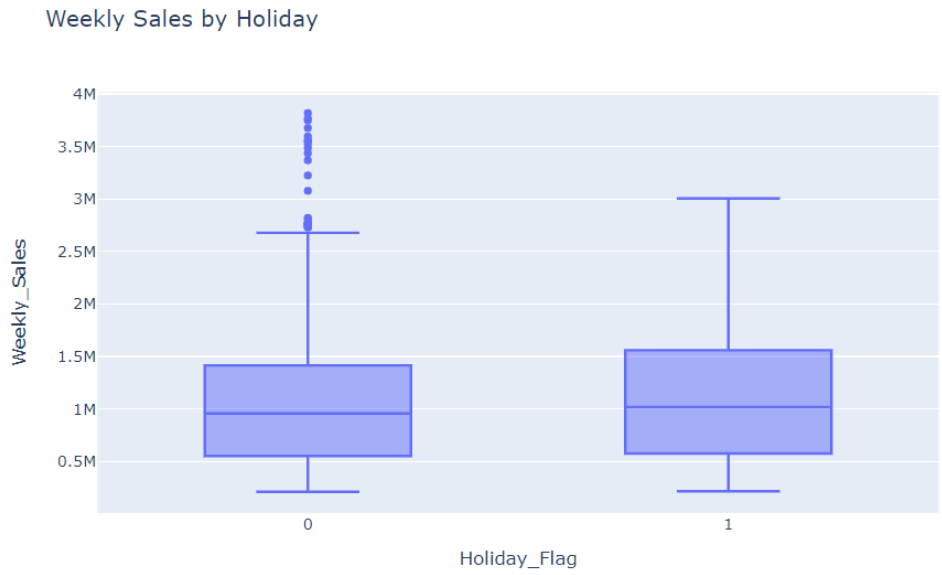
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6435 entries, 0 to 6434
Data columns (total 16 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Store         6435 non-null   int64
 1   Date          6435 non-null   datetime64[ns]
 2   Weekly_Sales  6435 non-null   float64
 3   Holiday_Flag  6435 non-null   int64
 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
 8   Day           6435 non-null   int32
 9   Day_Of_Week   6435 non-null   object
 10  Week          6435 non-null   UInt32
 11  Month         6435 non-null   int32
 12  Month_Name    6435 non-null   object
 13  Quarter       6435 non-null   int32
 14  Season        6435 non-null   object
 15  Year          6435 non-null   int32
dtypes: UInt32(1), datetime64[ns](1), float64(5), int32(4), int64(2), object(3)
memory usage: 685.1+ KB
```

6.3) Analyzing categorical variable

Categorical Features such as Holiday_Flag, Day_Of_Week, Month_Name and Season

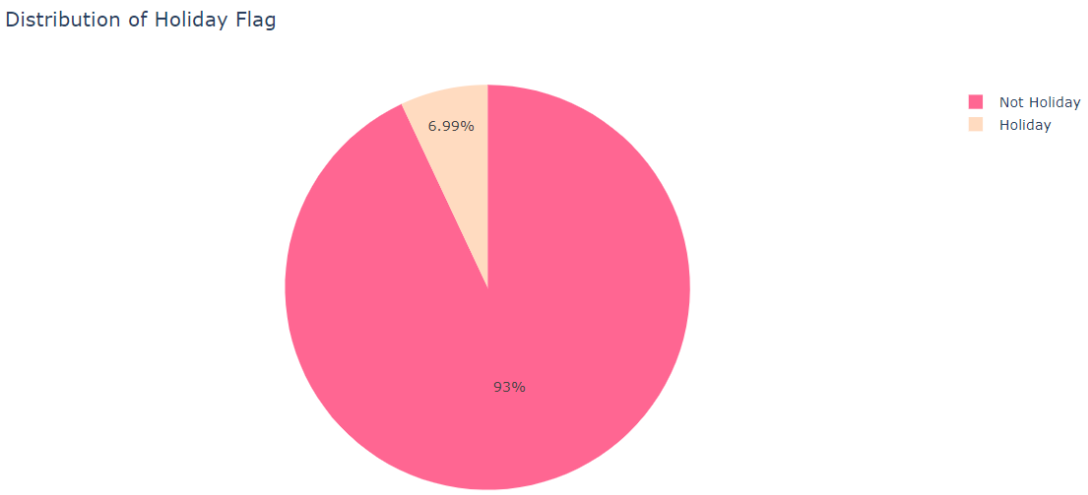
6.3.1 Comparison of Weekly Sales between Holidays and Non-Holidays:

Using Boxplot:



The boxplot visualization reveals that the average sales during holidays are higher compared to non-holiday days. This suggests that holidays have a significant impact on sales volume, potentially driven by increased consumer spending during festive seasons or promotional events.

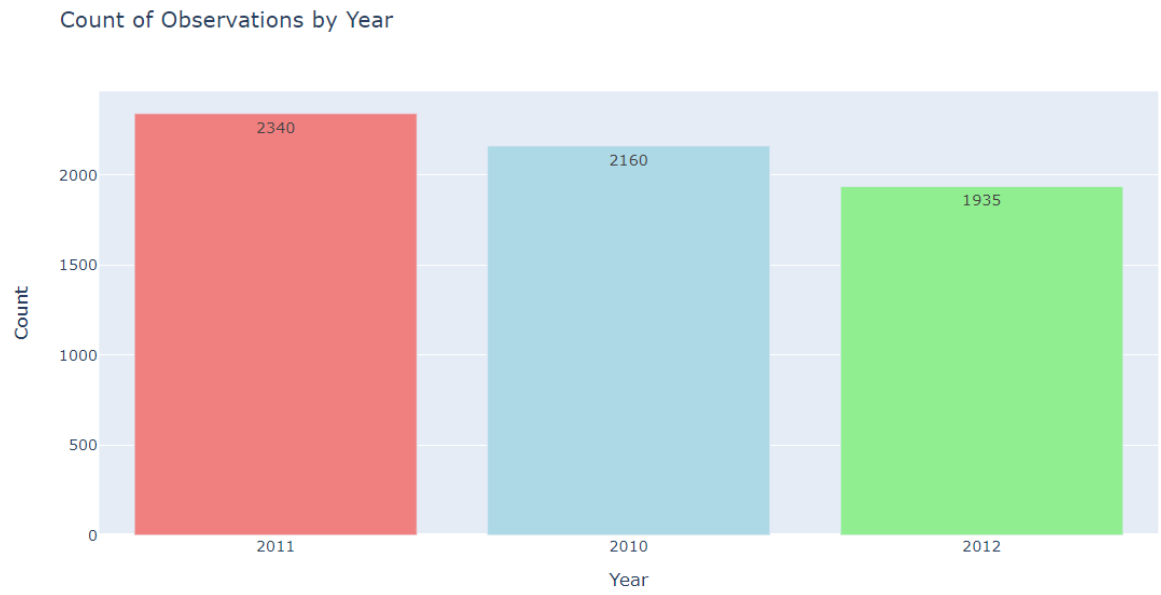
Using Pie Chart:



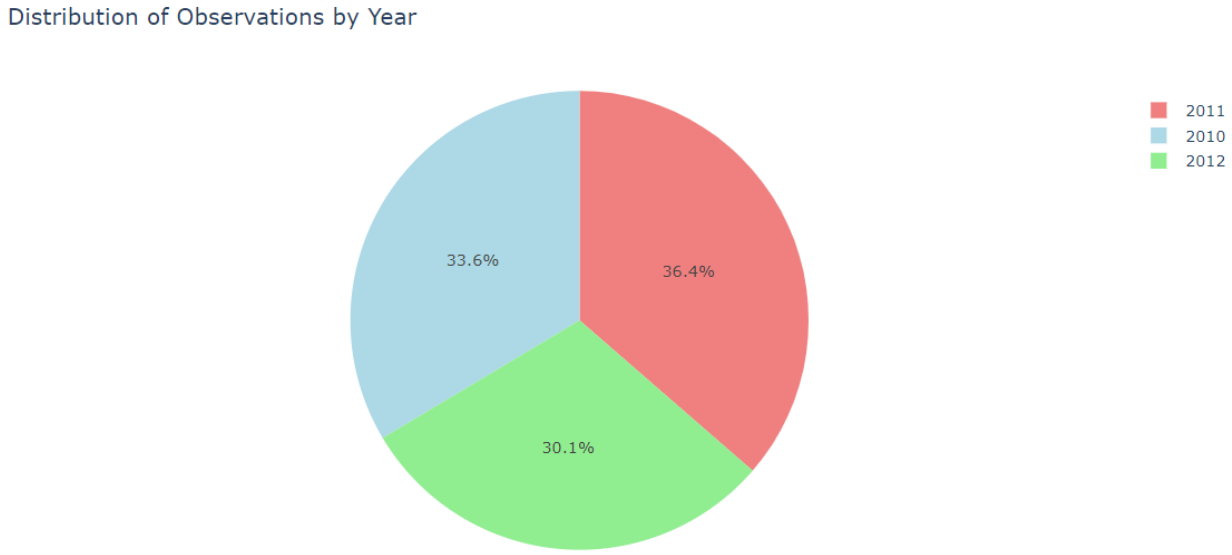
The pie chart illustrates the distribution of sales between holiday and non-holiday periods. It clearly indicates the dominance of non-holiday sales, with the bulk of sales occurring during non-holiday periods. However, it's noteworthy that approximately 6.99% of sales are generated during holidays, indicating a smaller yet potentially significant increase in sales during these periods. This could be attributed to seasonal promotions, special offers, or holiday shopping trends.

6.3.2 Distribution of Observations by Year:

Using Barplot:



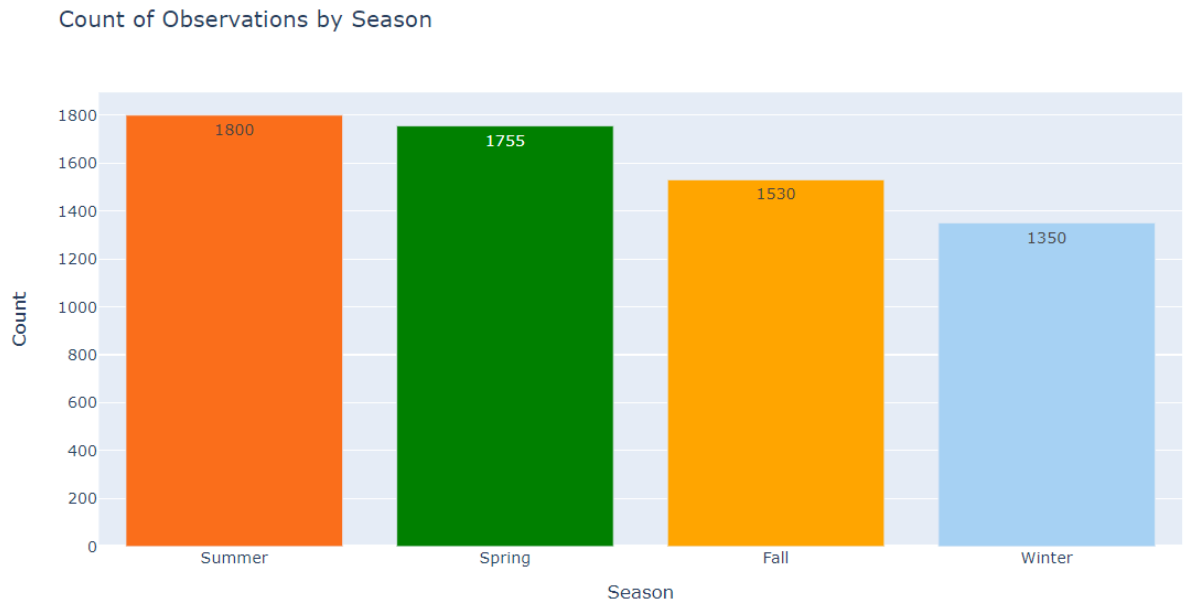
Using Pie chart



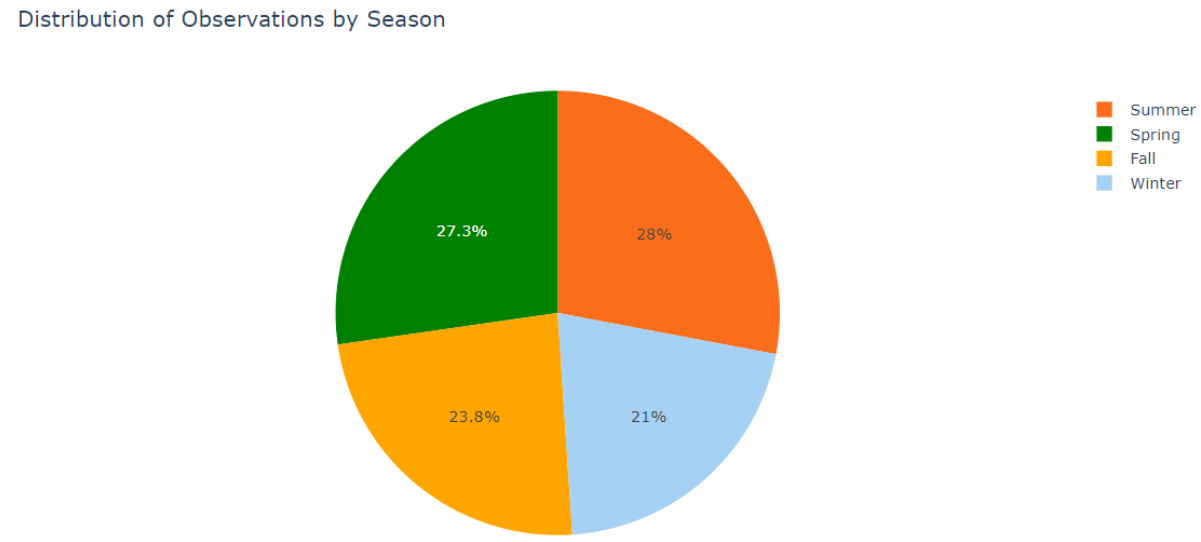
Both visualizations demonstrate an even distribution of observations across the three years, with slight variations in counts. Notably, 2011 exhibits a modest increase in the number of observations compared to the other years. However, overall consistency suggests stable trends over the three-year period.

6.3.3 Distribution of Observations by Season:

Using Barplot:



Using pie chart

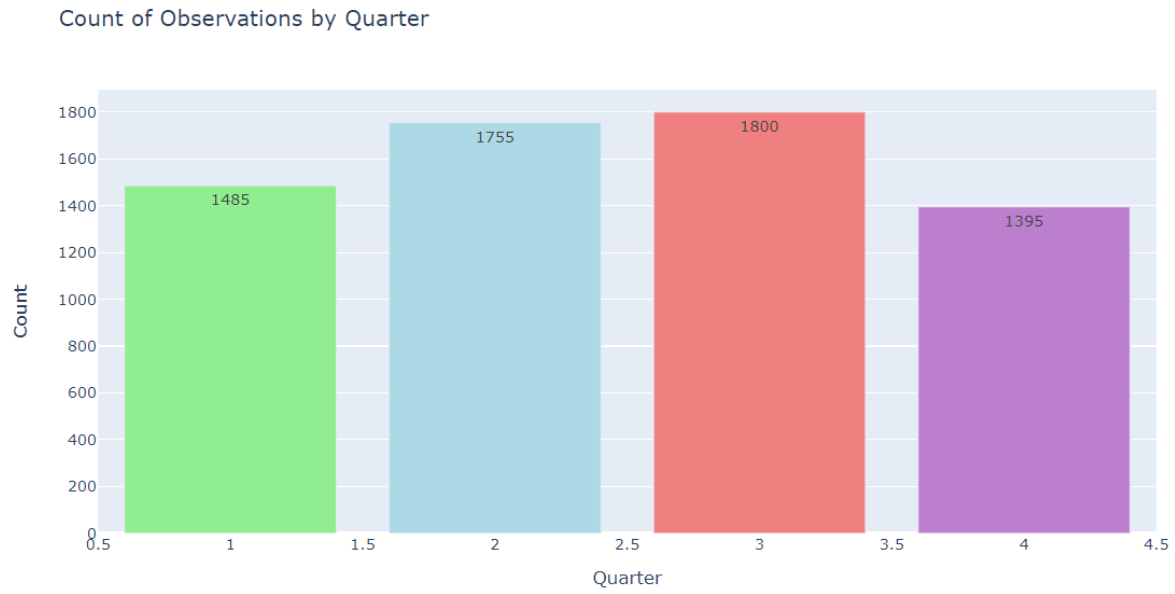


Utilizing both bar plot and pie chart visualizations, we examined the distribution of observations across seasons. The analysis revealed that summer had the highest count of observations, closely trailed by spring. Conversely, fall and winter showed fewer

observations, indicating potential decreases in sales activity during these seasons. The visual representations underscore the significance of summer and spring as peak sales periods, while highlighting potential seasonal variations in consumer behavior.

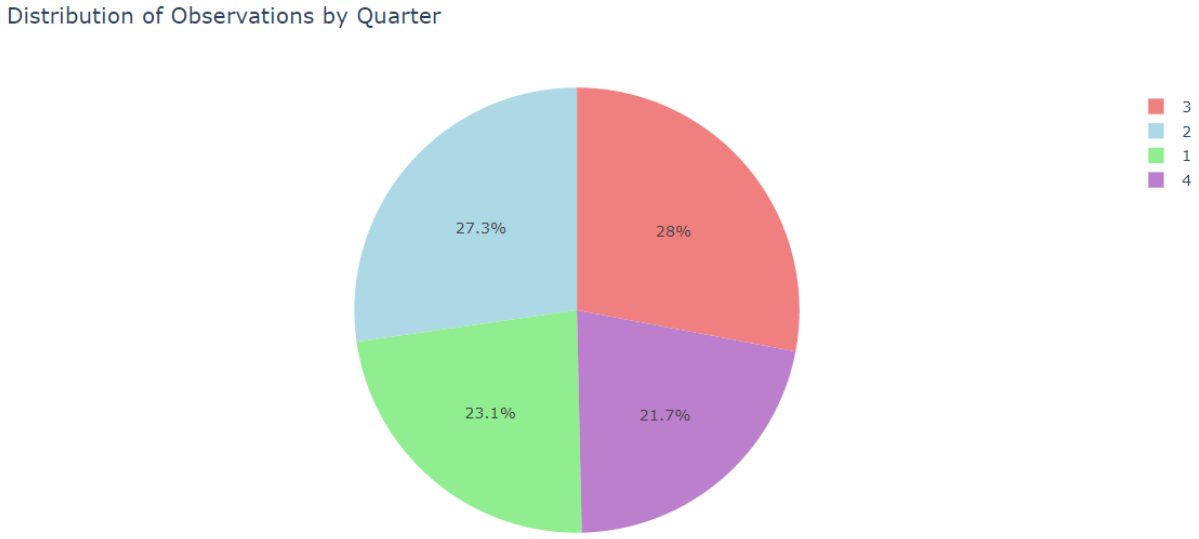
6.3.4 Distribution of Observations by Quarter:

About barplot :



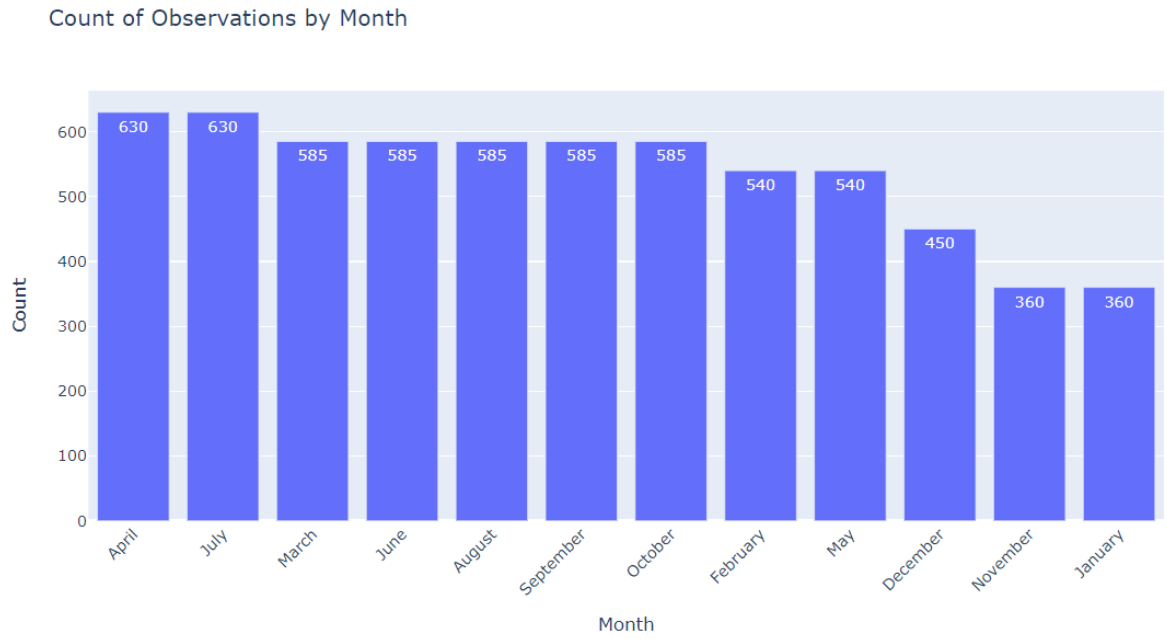
The barplot graphically represents the counts of observations by quarter. Quarter 3 exhibits the highest count of observations, closely followed by Quarter 2 and Quarter 1, while Quarter 4 demonstrates the lowest count.

About piechart



The piechart illustrates the percentage distribution of observations across quarters. Notably, Quarter 3 has the highest percentage of observations at 28%, suggesting a potential increase in activity or sales during this period. Quarter 1 closely follows with 23.1%, while Quarters 2 and 4 exhibit lower percentages of observations at 27.3% and 21.7%, respectively.which helps to do strategic planning for resource allocation throughout the year.

6.3.5 Distribution of Observations by Month:



April and July emerge as peak sales months, indicating heightened activity during these periods. In contrast, November and January experience declines, potentially influenced by seasonal or post-holiday effects. Understanding these fluctuations in sales can inform strategic planning and resource allocation for optimal performance throughout the year.

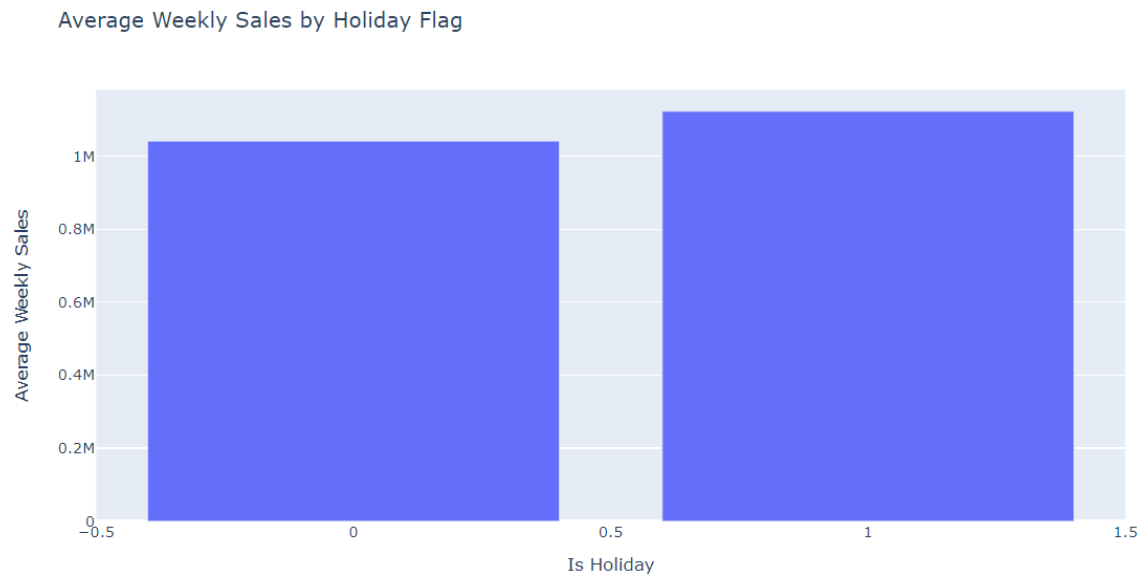
7) Bivariate Analysis:

In bivariate analysis, we explore the relationship between two variables to uncover patterns and associations within the data. One such analysis involves calculating the mean weekly sales for each holiday flag.

7.1 Mean of Weekly Sales for Each Holiday Flag:

We calculated the mean of weekly sales for observations categorized by the holiday flag.

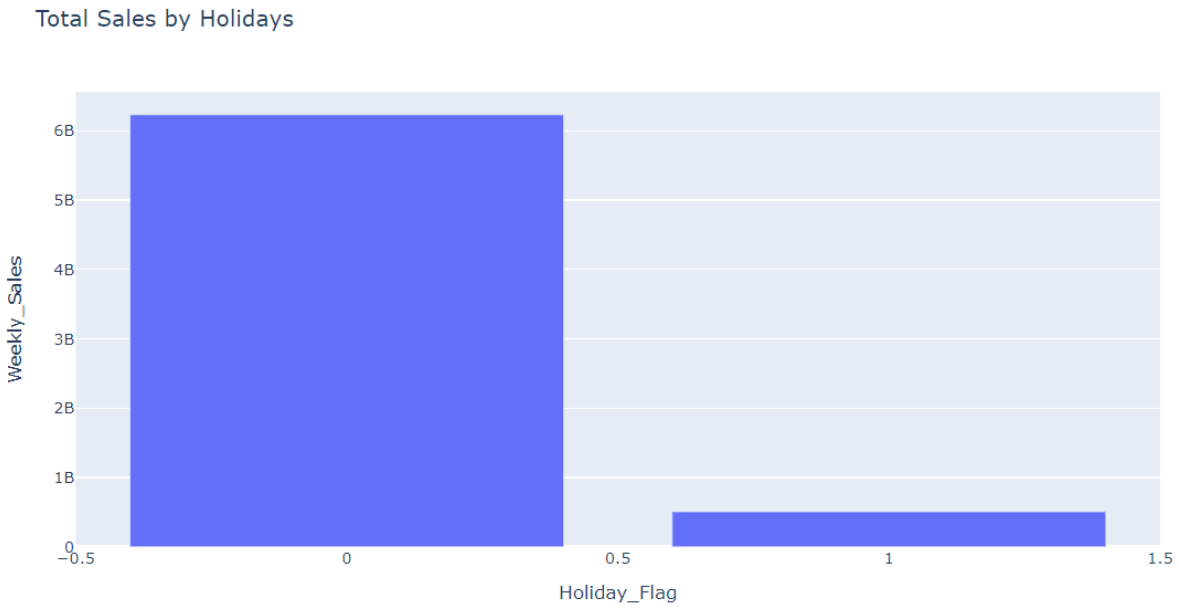
- Holiday Flag 0 (Non-Holiday): Mean Weekly Sales = 1.041256e+06
- Holiday Flag 1 (Holiday): Mean Weekly Sales = 1.122888e+06



So from above plot we can easily deduce that higher average sales during holiday periods (1 on x-axis) compared to non-holidays (0), suggesting a positive impact of holidays on sales due to increased consumer spending. This insight enables businesses to strategize inventory and marketing efforts to capitalize on holiday shopping seasons, reflecting a common trend of sales spikes during holidays driven by promotions and festive shopping behaviors.

7.2 Total Weekly Sales for Each Holiday Flag:

In this analysis, we calculated the total weekly sales for observations categorized by the holiday flag. The results are summarized as follows:
Holiday Flag 0 (Non-Holiday): Total Weekly Sales = 6.231919e+09
Holiday Flag 1 (Holiday): Total Weekly Sales = 5.052996e+08

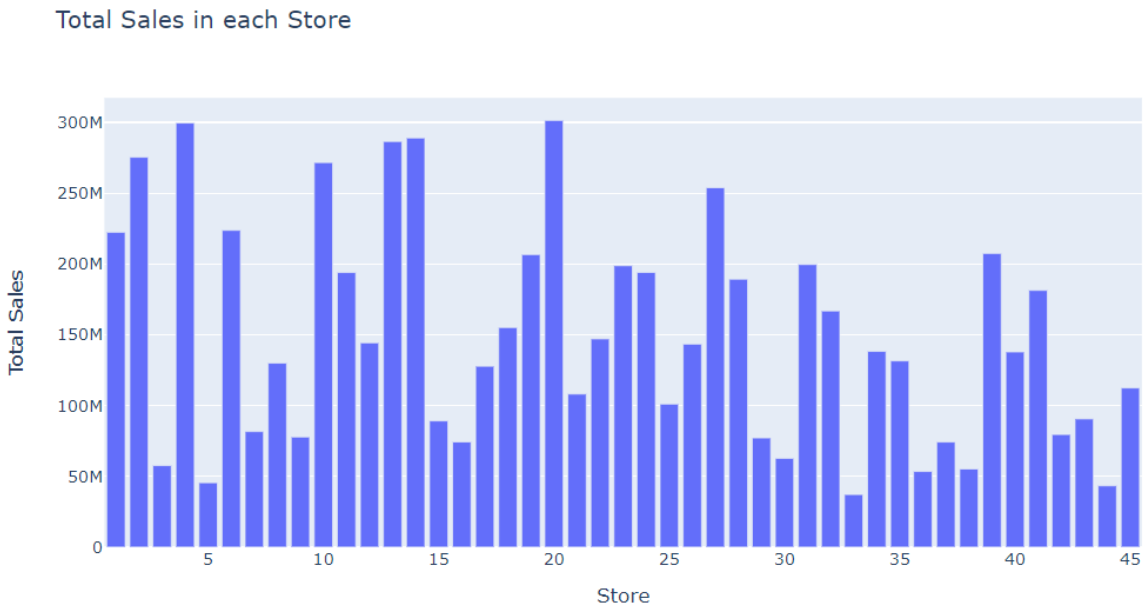


Total sales are higher during non-holiday periods, although holidays also see a sales spike. Businesses can use this data to strategize inventory and marketing, focusing on both holiday and non-holiday periods.

This data provides valuable insights for businesses to strategize their inventory management and marketing efforts. By recognizing the differences in sales patterns between holiday and non-holiday periods, businesses can tailor their strategies to effectively address the unique demands and opportunities presented by each period. This holistic approach enables businesses to maximize sales revenue and optimize resource allocation throughout the year.

7.3 Sum of Weekly Sales for Each Store:

In this analysis, we calculated the sum of weekly sales for each store.



- The analysis reveals significant variation in weekly sales across different stores. Based on the total sales figures, the following insights are observed:

- Top Performing Stores: Stores 20, 4, 14, and 13 emerge as the top-performing stores based on total weekly sales. These stores have consistently demonstrated high sales figures, indicating strong performance and customer engagement.

- Worst Performing Stores: Stores 36, 5, 44, and 33 are identified as the worst performing stores based on total weekly sales. These stores exhibit notably lower sales figures compared to others, suggesting areas for improvement and optimization.

- The accompanying bar graph visually represents the total weekly sales for each store. Store No. 20 leads with the highest sales, closely followed by Store No. 4. In contrast, Store No. 33 has the lowest sales, indicating differing performance levels among the stores.

- This analysis highlights the importance of understanding and addressing the varying performance levels among stores. By identifying top-performing stores and areas for improvement, businesses can implement targeted strategies to enhance sales performance and drive overall growth and profitability.

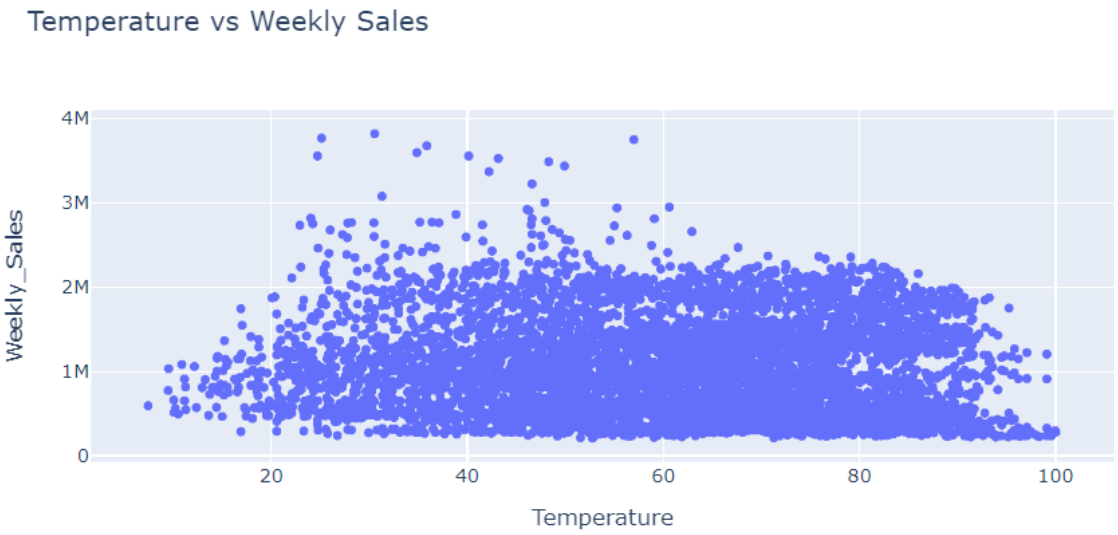
7.4 Scatter Plot Analysis

In this analysis, we explore the relationship between various factors (Temperature, Fuel_Price, CPI, Unemployment, and Holiday_Flag) and Weekly_Sales through scatter plots. These visualizations provide insights into how each factor influences Weekly_Sales.

Scatter Plots:

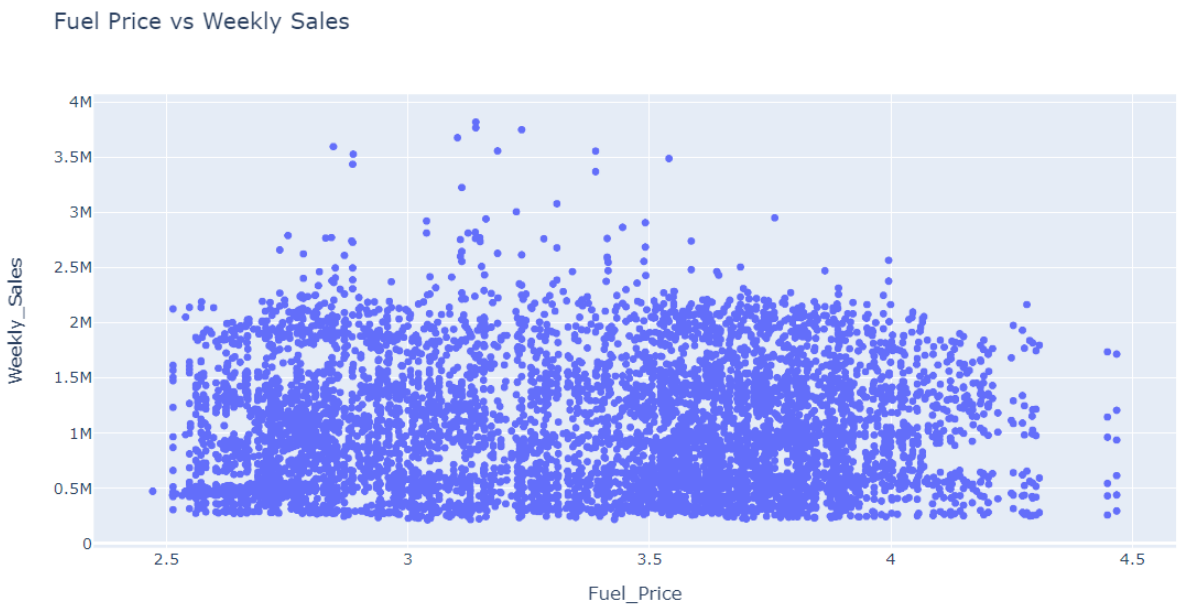
- Temperature vs Weekly Sales
- Fuel Price vs Weekly Sales
- CPI vs Weekly Sales
- Unemployment vs Weekly Sales

7.4.1 Temperature vs Weekly Sales:



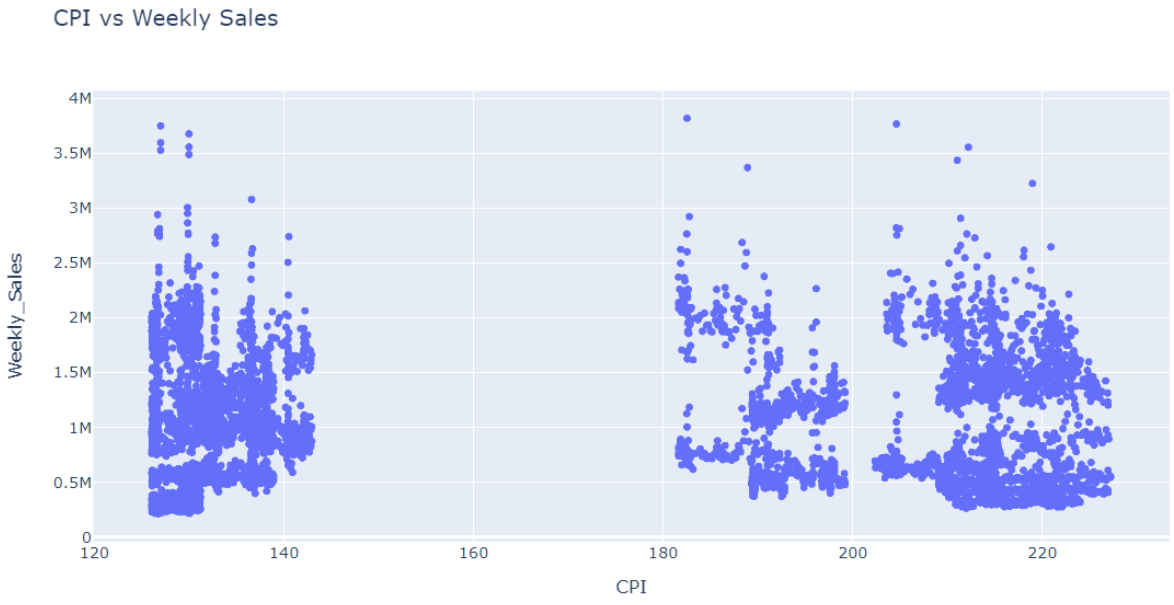
There appears to be a moderate positive correlation between temperature and weekly sales, indicating that higher temperatures may lead to increased sales.

7.4.2 Fuel Price vs Weekly Sales:



There is no clear trend between fuel price and weekly sales, suggesting that fuel price may not have a significant impact on sales.

7.4.3 CPI vs Weekly Sales:



Consumer Price Index (CPI) has little impact on sales. Based on the distribution of typical consumer prices in the figure above, clients can be categorized into two groups: clients that pay from 120 and 150 are considered middle-class clients. consumers who pay between 180 and 230 are considered high-class consumers.

7.4.4 Unemployment vs Weekly Sales:

There appears to be a slight negative correlation between unemployment rate and weekly sales, indicating that lower unemployment rates may be associated with higher sales.

7.5 Analysis of Seasonal Trends in Weekly Sales:

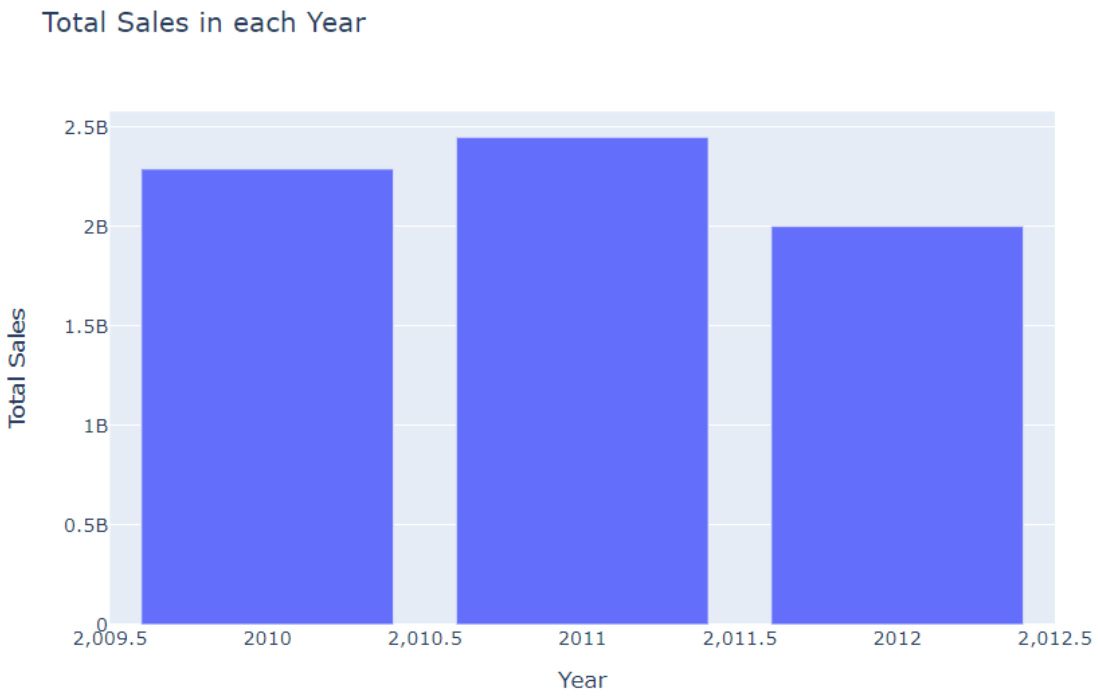
Identifying seasonal trends in weekly sales involves analyzing the data to observe patterns that repeat over specific periods, such as weeks, months, or quarters.

7.5.1 Total sales in each year:

In this analysis, we examine the total sales in each year to identify any significant trends or fluctuations.

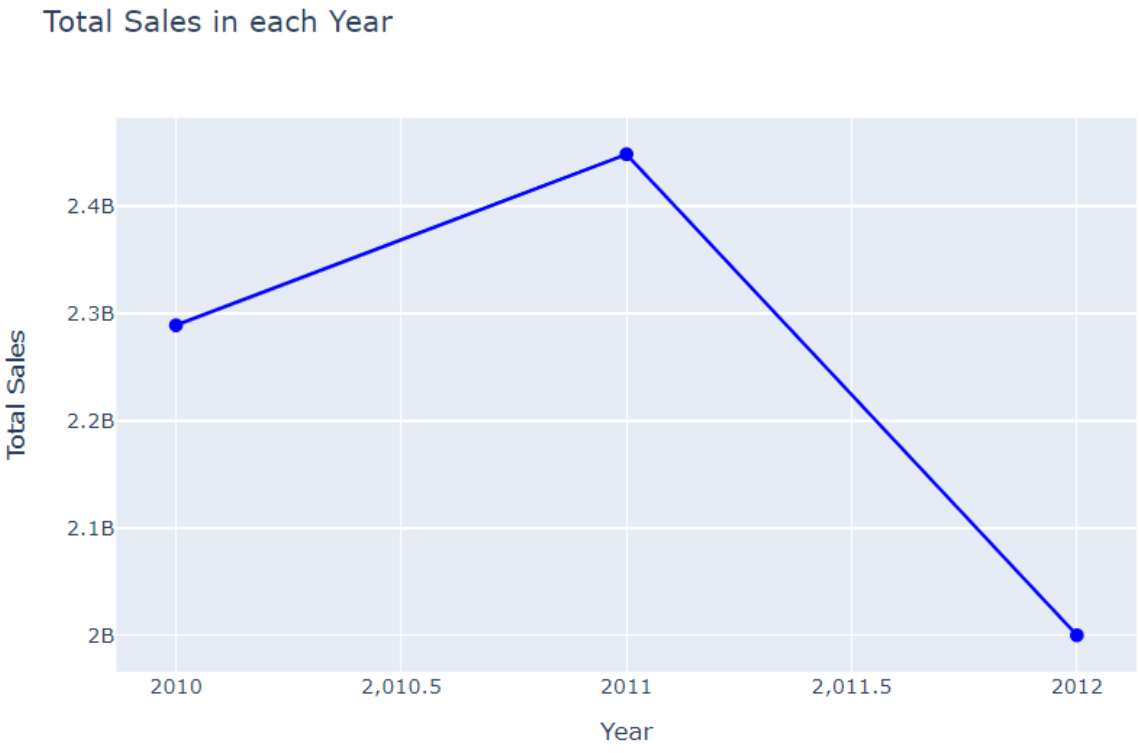
Barplot:

The barplot below displays the total sales in each year:



In 2011, total sales peaked at $2.4482 billion, indicating a period of robust sales performance. However, there was a significant drop in total sales in 2012 compared to the previous year. This decline suggests the influence of certain factors, such as specific events, marketing campaigns, or economic conditions, that may have impacted sales performance negatively.

Lineplot



The lineplot visually depicts the trend in total sales over the two years, highlighting the peak in sales in 2011 followed by a decline in 2012. This trend underscores the importance of investigating the underlying variables that contributed to the decline in 2012 to strategize effectively for future sales growth.

Conclusion:

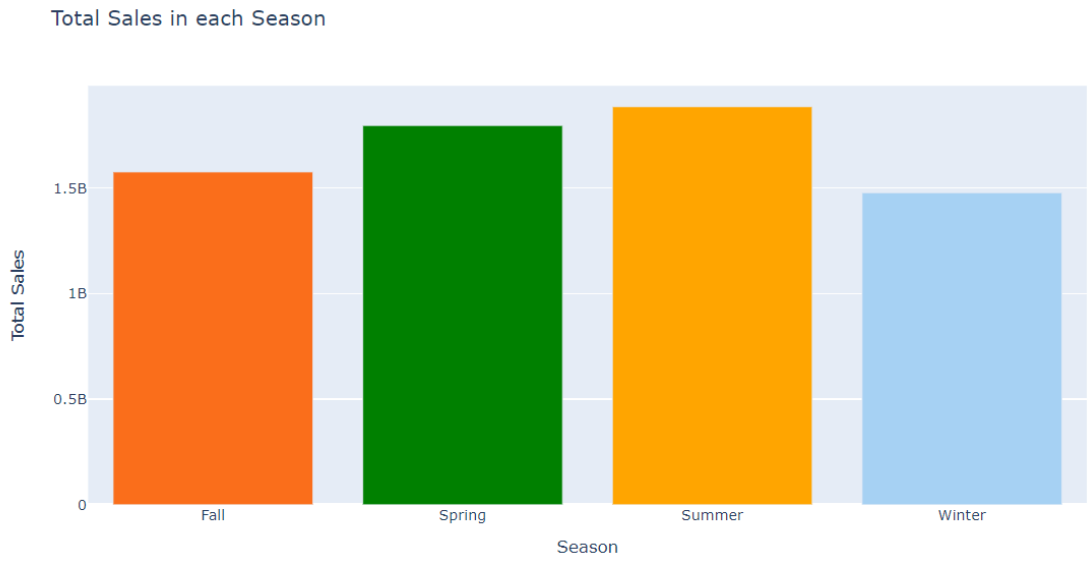
Understanding the seasonal trends in weekly sales, as evidenced by the analysis of total sales in each year, is crucial for businesses to adapt their strategies and optimize sales performance. By investigating the factors influencing sales fluctuations and implementing targeted interventions, businesses can mitigate risks and capitalize on opportunities for sustainable growth.

7.5.2 Total Sales in each Season:

This analysis examines the total sales in each season to understand the seasonal trends in sales performance.

Barplot:

The barplot below displays the total sales in each season:

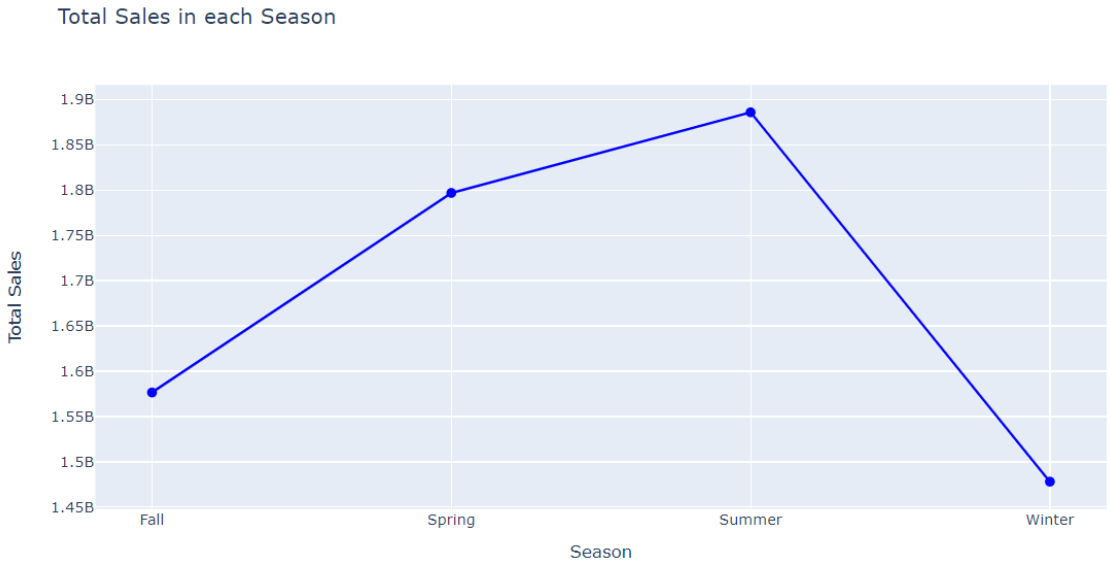


- Spring and Summer seasons experience the highest sales, indicating a peak in sales during these periods. This trend may be attributed to seasonal shopping habits, such as purchasing for holidays or outdoor activities, which drive increased consumer spending.
- Despite slight variations, total sales remain relatively consistent across all seasons, indicating a stable customer base and steady demand throughout the year.

- Lower sales in Fall and Winter seasons suggest potential opportunities for Walmart to implement targeted marketing strategies or promotions to stimulate sales during these periods.

Lineplot:

Additionally, the lineplot illustrates the trend in total sales across the four seasons:



The lineplot visually depicts the trend in total sales across the four seasons, highlighting the variations in sales performance throughout the year. The consistent pattern of higher sales in Spring and Summer seasons, followed by lower sales in Fall and Winter, underscores the importance of seasonal factors in driving consumer behavior and purchasing patterns.

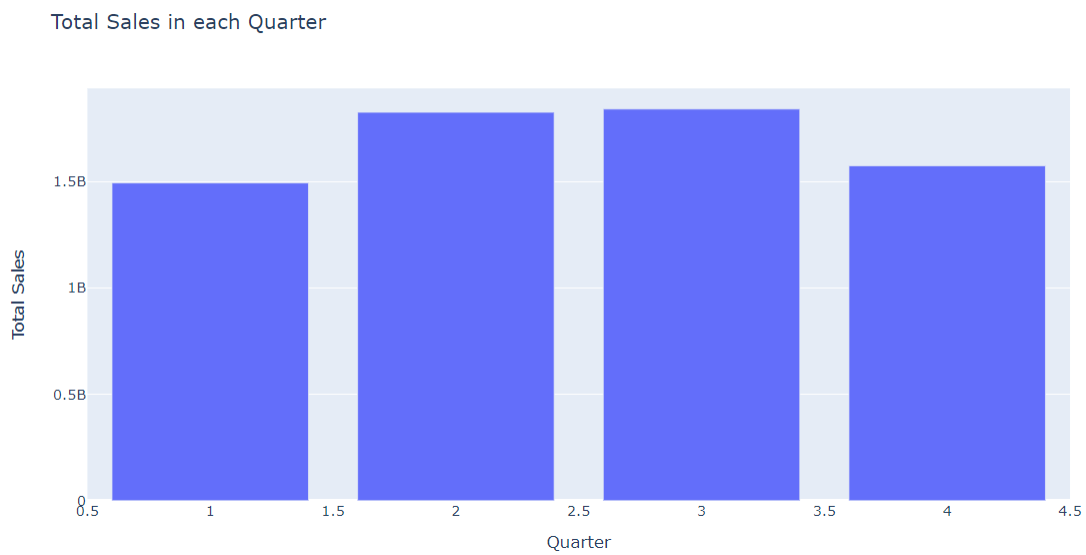
Conclusion:

Understanding the seasonal trends in total sales enables Walmart to tailor its marketing strategies and promotions to capitalize on peak sales periods while addressing potential challenges during slower seasons. By leveraging insights from seasonal sales data, Walmart can optimize its sales performance and enhance customer satisfaction year-round.

7.5.3 Total Sales in each Quarter:

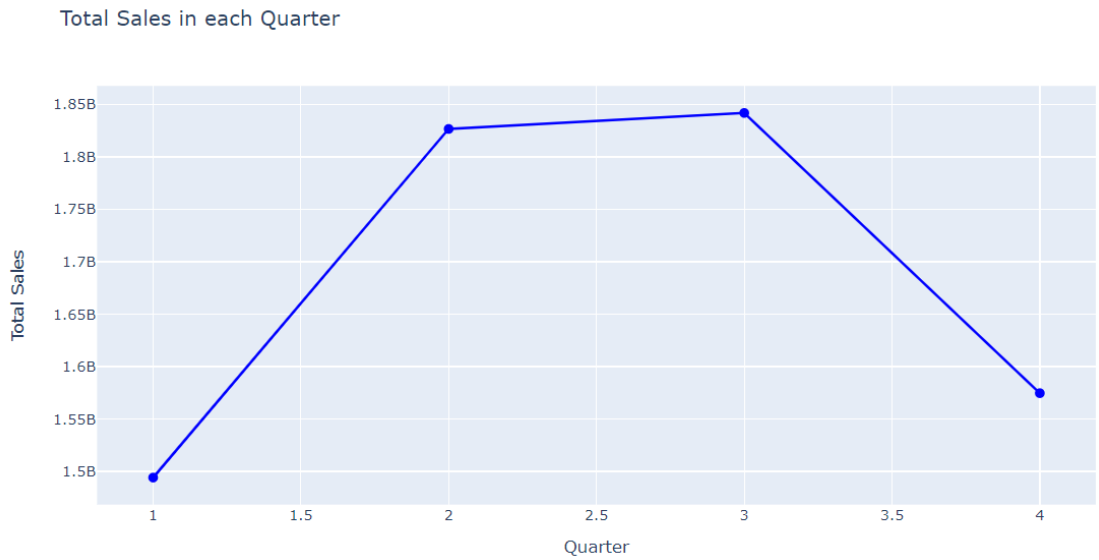
Barplot:

The barplot illustrates the total sales in each quarter:



Lineplot:

The lineplot visually depicts sales trends across quarters:



- Peak Sales in Quarters 2 and 3: The graph reveals that Walmart experiences peak sales in Quarters 2 and 3, particularly during the middle of the year. This surge in sales can be attributed to factors such as summer holidays and back-to-school shopping.
- Lower Sales in Quarters 1 and 4: Conversely, sales are lower in Quarters 1 and 4. Quarter 1 records the least sales, likely due to reduced consumer spending after the holidays and towards the end of the year.

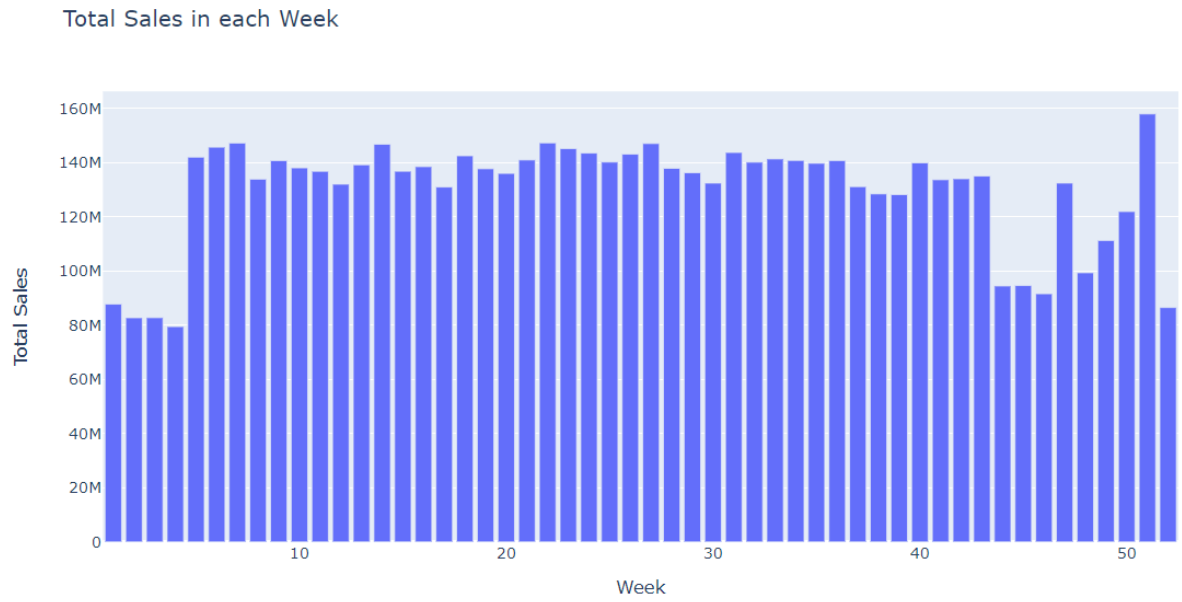
Conclusion:

Understanding the timing of peak sales periods is crucial for effective planning at Walmart. By focusing on promoting sales during busy times and implementing strategies to boost sales during slower periods, Walmart can enhance customer satisfaction and maximize revenue throughout the year.

7.5.4 Total Sales in each Week:

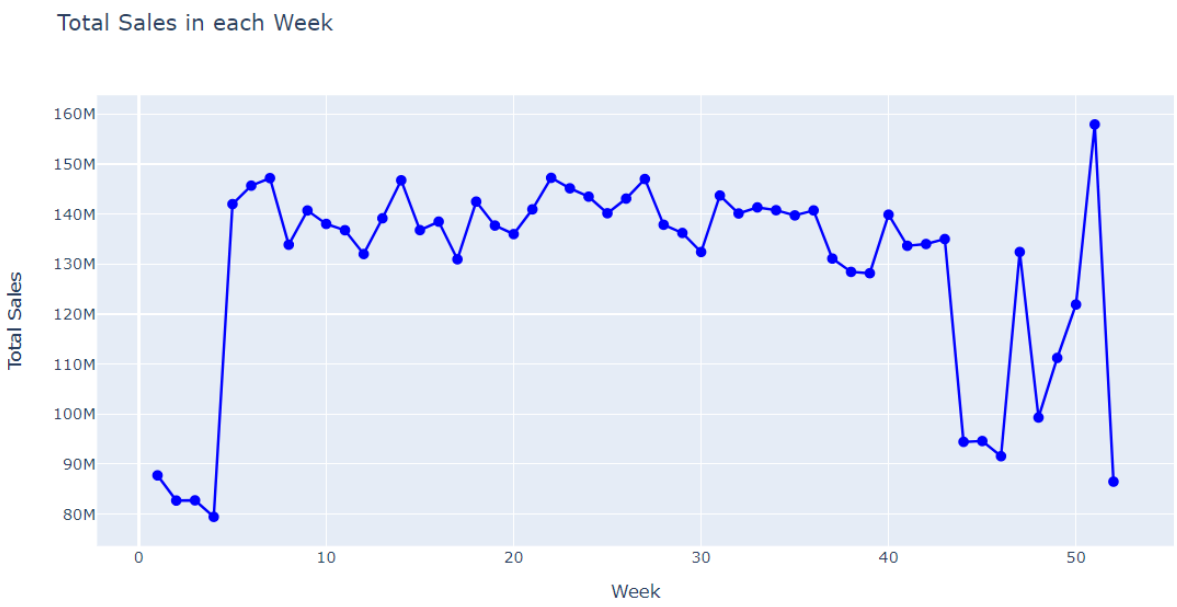
Barplot:

The barplot illustrates the weekly sales at Walmart.



Lineplot:

The lineplot depicts the trends in weekly sales.



The graph shows that Walmart's weekly sales start strong, dip sharply around week 10, stabilize with minor fluctuations until week 40, and then exhibit significant volatility with sharp declines and spikes towards week 50. This pattern could reflect various external factors such as holidays, promotions, or seasonal changes affecting consumer behavior and sales trends.

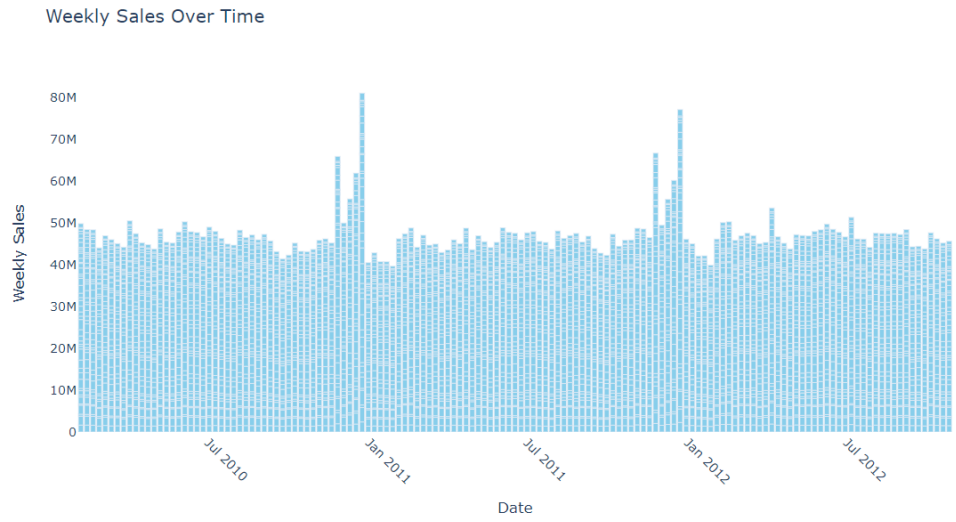
Conclusion:

Understanding these patterns is essential for Walmart to adapt its strategies effectively. By analyzing the factors contributing to sales fluctuations, Walmart can optimize its inventory management, marketing initiatives, and promotional campaigns to ensure sustained sales performance and enhance customer satisfaction throughout the year.

7.5.5 Visualize Weekly Sales Over Time:

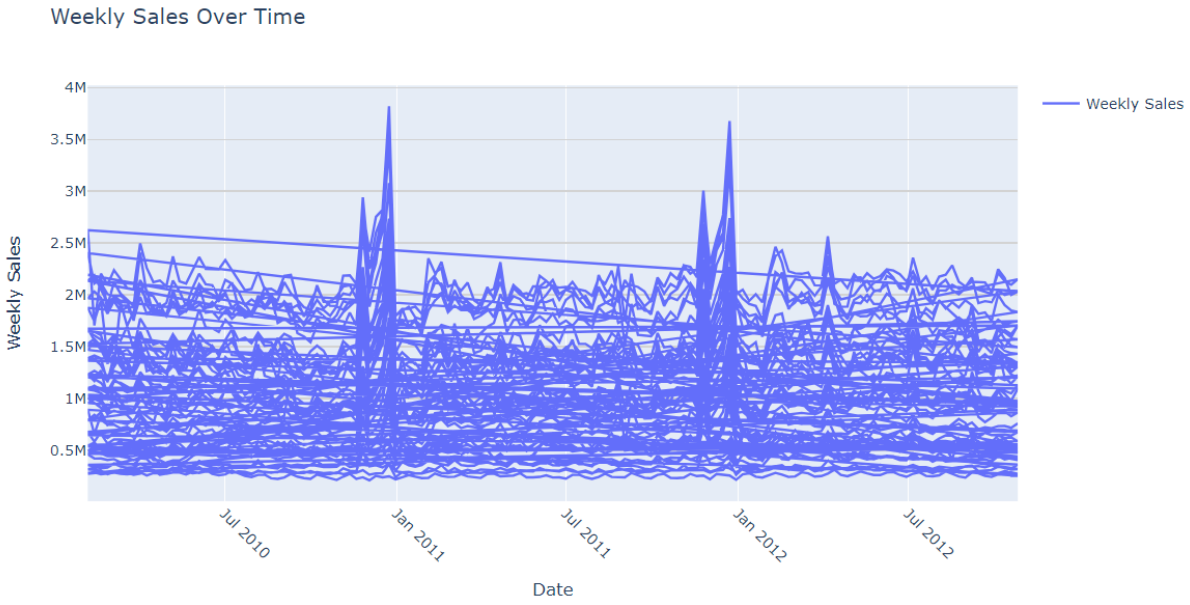
Barplot:

The barplot illustrates the weekly sales over time.



Lineplot:

The lineplot visually depicts the trends in weekly sales from February 2010 to October 2012.



The graph spanning from February 2010 to October 2012 illustrates the following insights:

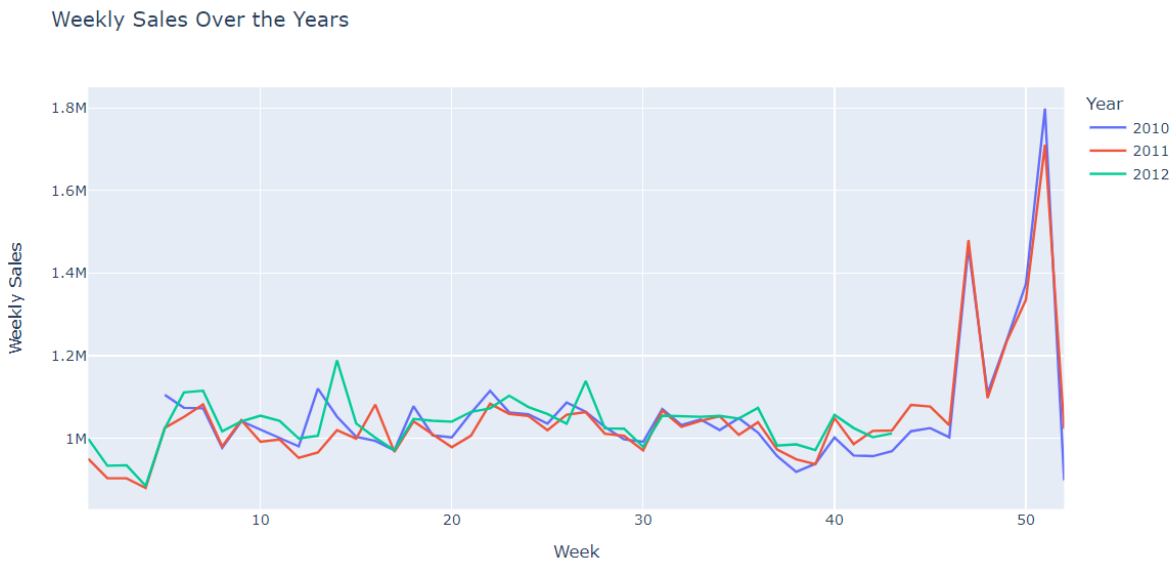
- Stable Sales with Spikes: Weekly sales demonstrate relative stability over the period, with two notable spikes occurring around January 2011 and January 2012. These spikes likely correspond to seasonal events or promotions that drove higher sales volumes.

- Consistent Fluctuation: The data exhibits consistent fluctuation between approximately 30M and 50M for the majority of the time, indicating a steady demand. Occasional increases in sales may be attributed to specific high-traffic events or holidays.

Conclusion:

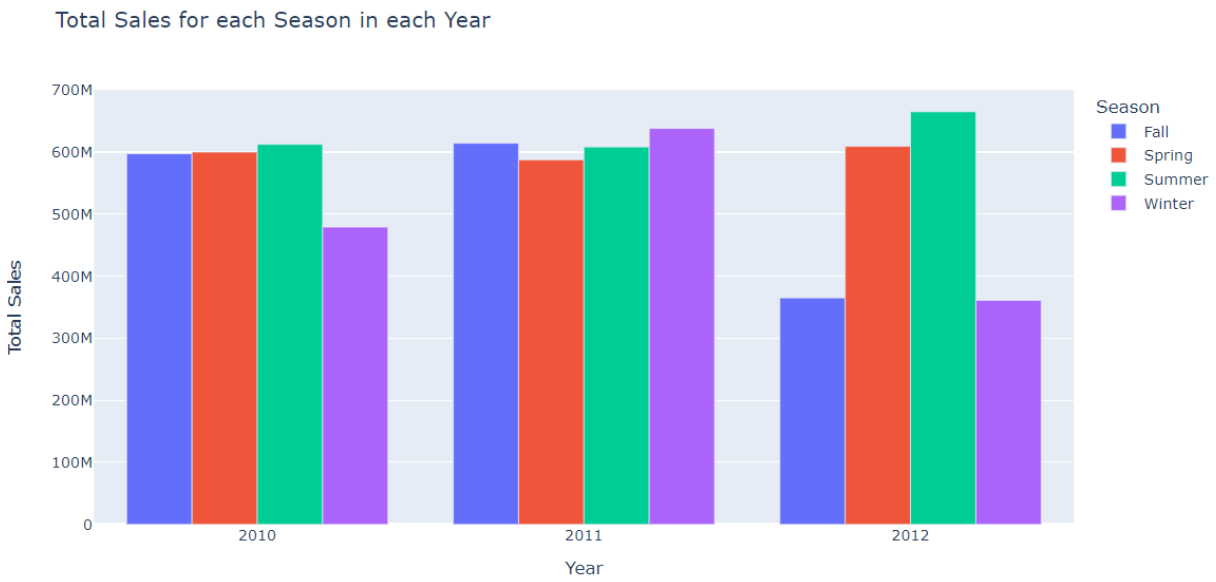
Understanding these sales patterns is crucial for planning inventory and sales strategies. By capitalizing on peak times and ensuring a steady supply during regular demand periods, Walmart can optimize its sales performance and enhance customer satisfaction.

7.5.6 Weekly_sales over the years:



Conclusion:Total sales for all years in week 51 are the highest from any week, with $157,929,657

7.5.7 Total Sales for each Season in each Year:

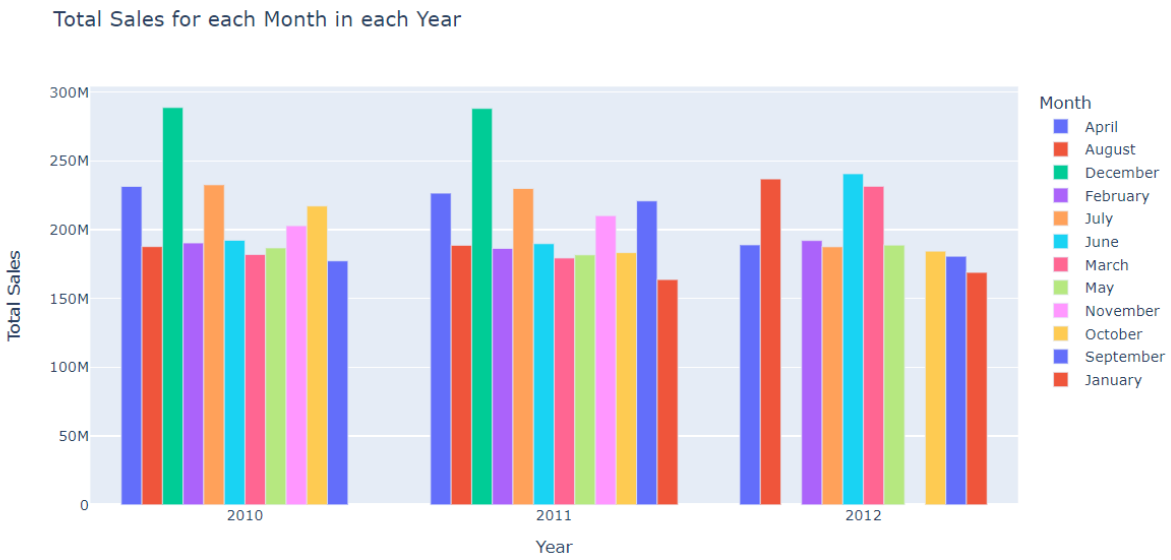


Conclusion:

- Seasonal Sales Trends: The graph shows that sales are highest in the Fall season each year, which could be due to back-to-school shopping or holiday sales events.
- Spring Sales Dip: There was a noticeable drop in sales during the Spring of 2011, which may require further investigation to understand the cause.
- Consistent Winter Sales: Sales during Winter have been consistent over the three years, indicating stable consumer behavior during this season.
- Growing Summer Sales: There is an upward trend in Summer sales from 2010 to 2012, suggesting increasing consumer activity or successful marketing strategies in that period.

These insights can help in understanding consumer behavior and planning inventory and marketing strategies accordingly.

7.5.8 Total Sales for each Month in each year



Conclusion:

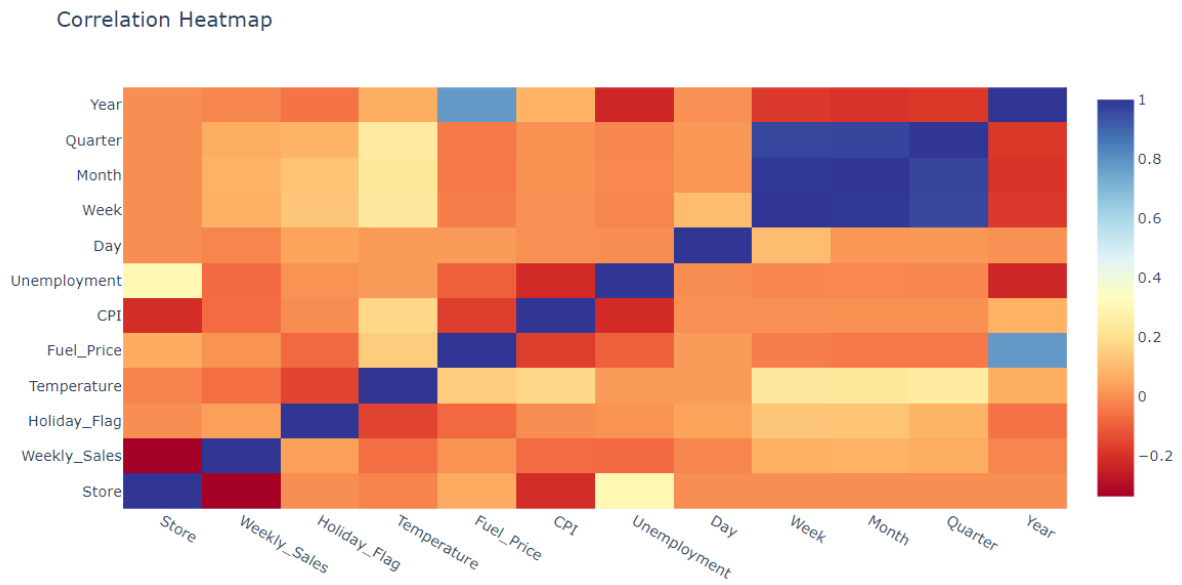
- Monthly Sales Trends: The graph shows trends in total sales for each month over the years 2010, 2011, and 2012, highlighting which months had higher sales and allowing comparison of annual growth or decline.
- Consistent High-Performing Months: April and December consistently show higher total sales compared to other months across all three years, suggesting these months may have key events or holidays that drive sales.
- Sales Variability: The fluctuating monthly sales indicate that certain factors, possibly including seasonal changes, holidays, or promotions, significantly impact sales.
- No Clear Growth Pattern: The absence of a consistent pattern of growth or decline suggests that sales are influenced by a variety of factors, rather than a steady market change.

These insights can inform strategic decisions for inventory management, marketing campaigns, and sales forecasting.

8) Multivariate Analysis:

Now we are going to understand the relationship between all the different columns numerically to check how they correlate with the weekly sales in order to confirm the inferences we have gathered from the above EDA study.

8.1 Correlation Heatmap :



The correlation matrix offers valuable insights into how different variables may influence weekly sales at Walmart. Here's a summary of the insights:

- Date and Fuel Price: There's a strong positive correlation (0.77), suggesting that as time progresses, fuel prices also tend to increase.
- Weekly Sales and Store: A negative correlation (-0.34) indicates that higher store numbers might be associated with lower weekly sales, which could suggest a pattern in store performance based on their numbering.
- Weekly Sales and Unemployment: The negative correlation (-0.074) suggests that higher unemployment rates might be associated with slightly lower weekly sales.
- CPI: The Consumer Price Index doesn't show significant correlations with Holiday_Flag or Temperature, indicating that CPI's impact on weekly sales might be independent of these factors.

These insights can help Walmart understand the dynamics affecting sales and make informed decisions on inventory management, pricing, and promotions. However, it's important to remember that correlation does not imply causation, and further analysis would be needed to establish direct relationships between these variables.

9) Data Correlation:

Correlation: A measure of the extent of interdependence between variables.

Pearson Correlation: The Pearson Correlation measures the linear dependence between two variables X and Y. The resulting coefficient is a value between -1 and 1 inclusive, where:

- 1: Perfect positive linear correlation.
- 0: No linear correlation, the two variables most likely do not affect each other.
- -1: Perfect negative linear correlation.

P-value: What is this P-value? The P-value is the probability value that the correlation between these two variables is statistically significant. Normally, we choose a significance level of 0.05, which means that we are 95% confident that the correlation between the variables is significant.

- By convention, when the
- p-value is < 0.001: we say there is strong evidence that the correlation is significant.
- p-value is < 0.05: there is moderate evidence that the correlation is significant.
- p-value is < 0.1: there is weak evidence that the correlation is significant.
- p-value is > 0.1: there is no evidence that the correlation is significant.

9.1 Pearson Correlation Coefficient and P-value of 'Fuel_Price' and 'Weekly_Sales':

The Pearson Correlation Coefficient between 'Fuel_Price' and 'Weekly_Sales' is 0.009463786314475135, with a corresponding p-value of P = 0.44782874894858093.

Scatter Plot with Regression Line:

Conclusion:

Based on the analysis, the p-value of 0.4478 indicates that the correlation between fuel price and weekly sales is not statistically significant. Additionally, the regression line in the scatter plot appears to be nearly horizontal, suggesting a weak relationship between fuel price and weekly sales. Therefore, fuel price may not be a reliable predictor of weekly sales, and other factors should be considered when forecasting sales.

9.2 Unemployment vs. Weekly Sales:

The Pearson Correlation Coefficient between 'Unemployment' and 'Weekly_Sales' is -0.07392603084823524, with a corresponding p-value of P = 2.899228577082314e-09.

Scatter Plot with Regression Line:

Conclusion:

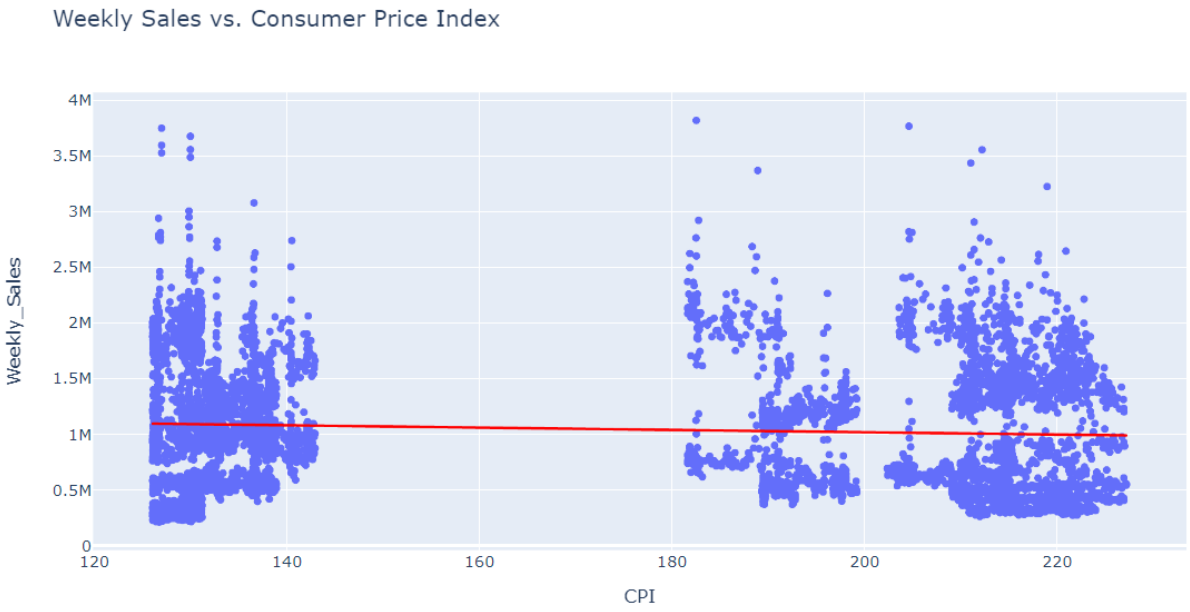
The low p-value of < 0.001 indicates a significant correlation between unemployment and weekly sales. This suggests that unemployment is a strong predictor of weekly sales, with a negative correlation observed; as the unemployment rate increases, weekly sales tend to decrease. Therefore, unemployment rate can be considered as an important factor affecting weekly sales.

Your section looks good as well! It provides clear information about the correlation between CPI and weekly sales, along with a concise conclusion. Here's the revised version:

9.3 CPI vs. Weekly Sales

The Pearson Correlation Coefficient between 'CPI' and 'Weekly_Sales' is -0.07263416204017627, with a corresponding p-value of P = 5.438292612176735e-09.

Scatter Plot with Regression Line:



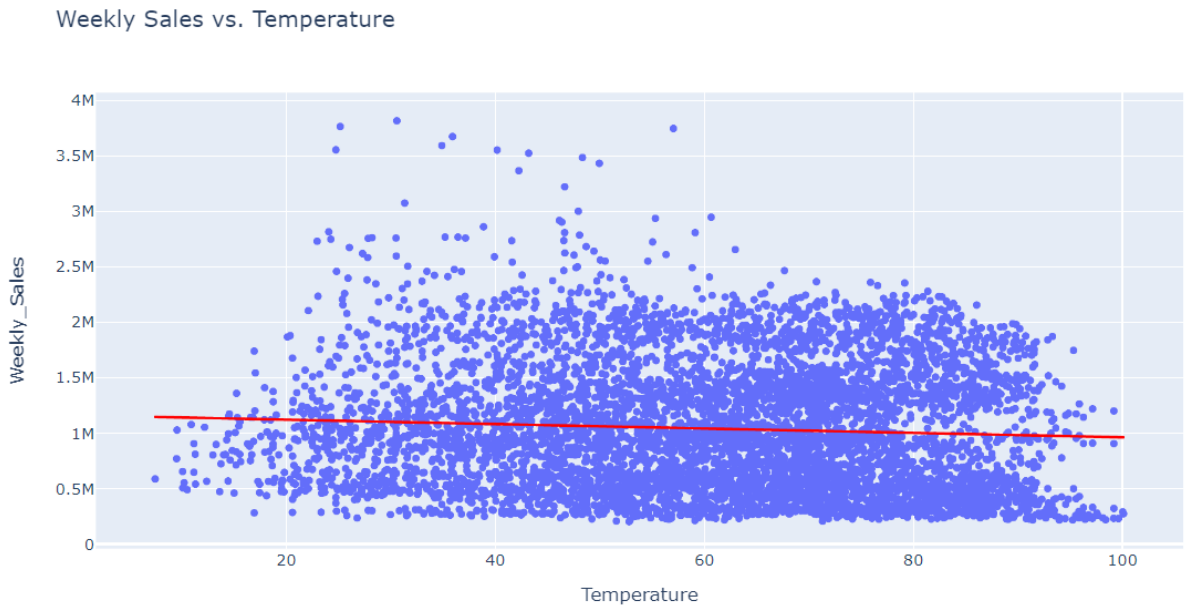
Conclusion:

With a p-value of < 0.001, there is strong evidence to suggest a significant correlation between CPI and weekly sales. This indicates that CPI is a relevant factor influencing weekly sales. The negative correlation observed suggests that as the Consumer Price Index increases, weekly sales tend to decrease. Therefore, CPI can be considered as an important determinant of weekly sales.

9.4 Temperature vs. Weekly Sales

The Pearson Correlation Coefficient between 'Temperature' and 'Weekly_Sales' is -0.06459453485480003, with a corresponding p-value of P = 2.1476279300634197e-07.

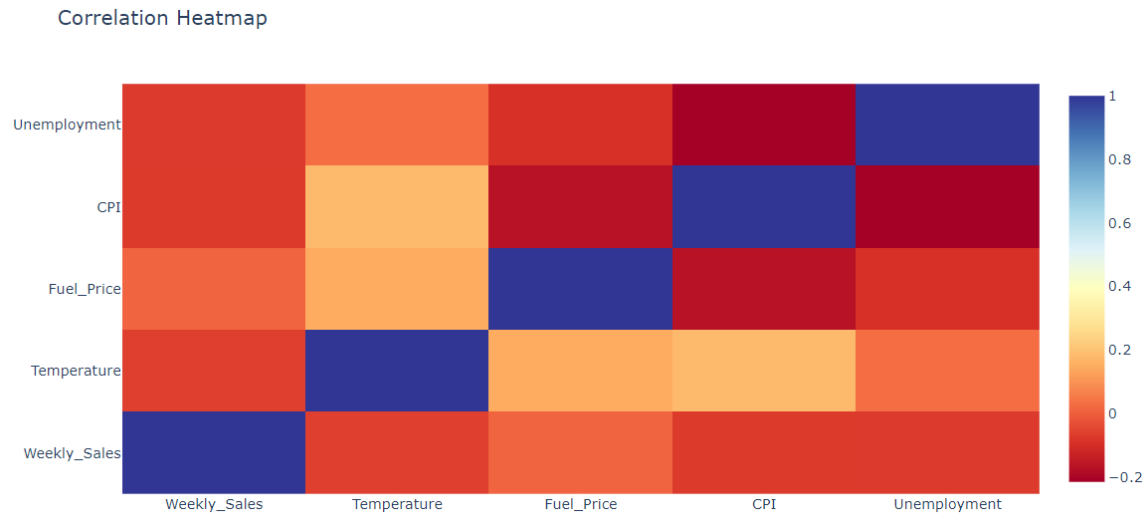
Scatter Plot with Regression Line:



Conclusion:

- With a p-value of < 0.05, there is moderate evidence to suggest a significant correlation between temperature and weekly sales.
- Temperature appears to be a relevant predictor of weekly sales, with a negative correlation observed; as the temperature increases, weekly sales tend to decrease.

9.5 Calculate correlation coefficients between Weekly_Sales and other numerical variables to assess the strength and direction of relationships.



9.5.1 Correlations with weekly sales:

	Weekly_Sales
Weekly_Sales	1.000000
Fuel_Price	0.009464
Temperature	-0.064595
CPI	-0.072634
Unemployment	-0.073926

- Fuel Price and Weekly Sales: There's a slight positive correlation, indicating that as fuel prices increase, weekly sales might also show a slight increase.
- Temperature and Weekly Sales: There's a negative correlation, suggesting that higher temperatures might lead to a decrease in weekly sales.
- CPI and Weekly Sales: The negative correlation here implies that as the Consumer Price Index rises, weekly sales might decrease.
- Unemployment and Weekly Sales: Similarly, a negative correlation indicates that higher unemployment rates might be associated with lower weekly sales.

In summary, the dataset indicates that higher fuel prices have a very slight positive effect on weekly sales, while increased temperatures, a higher CPI, and higher unemployment rates might negatively impact weekly sales. However, all these effects are relatively small according to the correlation coefficients.

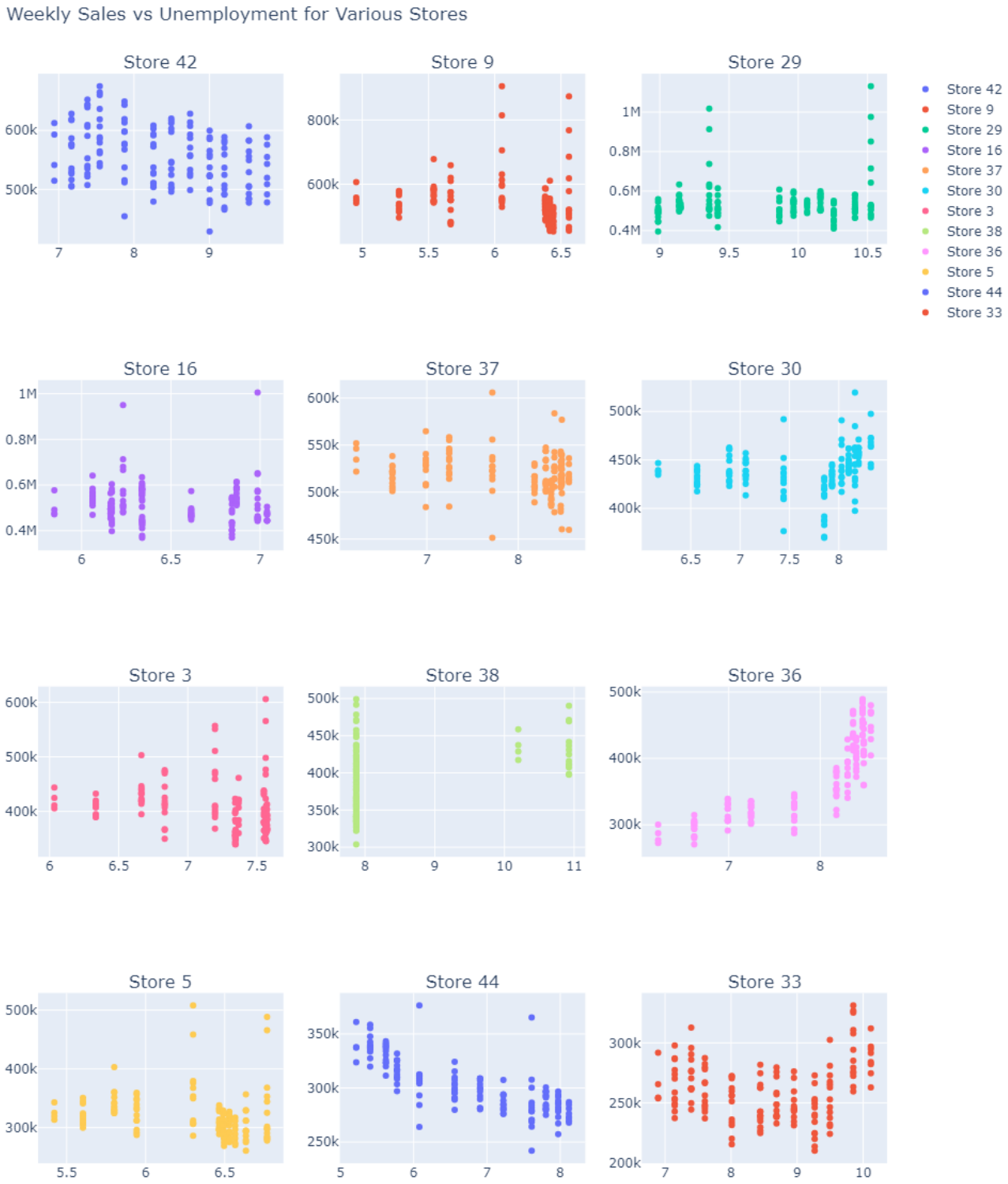
10) Project Questions

Que.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:

a. If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?

To analyze the impact of the unemployment rate on weekly sales, we focused on identifying stores with lower weekly sales, indicating potential suffering due to economic factors. We selected 12 stores with the least weekly sales for this analysis:
Stores with the least Weekly Sales: [42, 9, 29, 16, 37, 30, 3, 38, 36, 5, 44, 33]

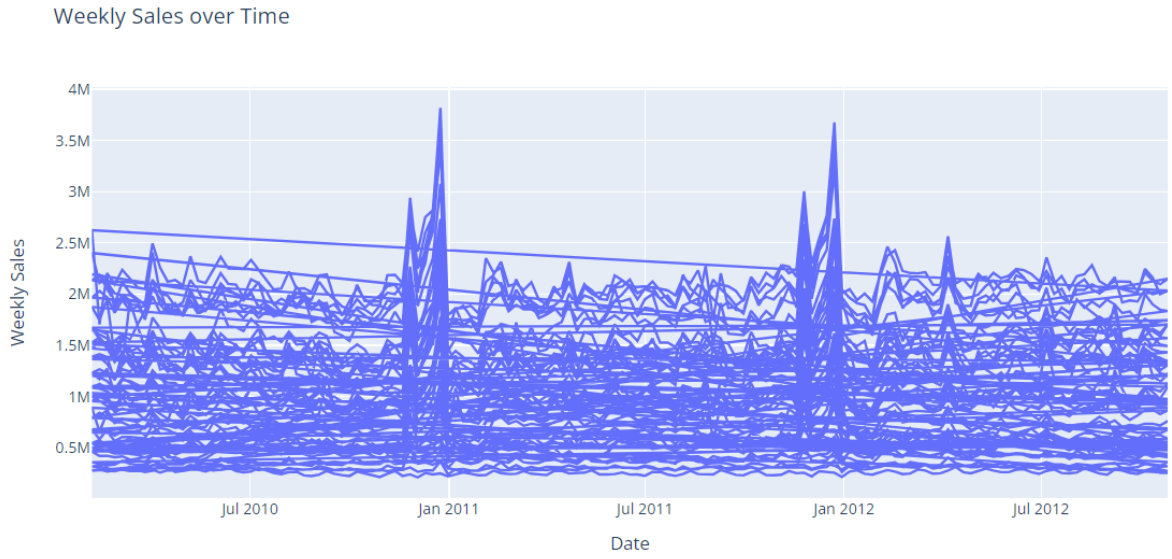
Scatter plot for Weekly Sales vs Unemployment for Various Stores:



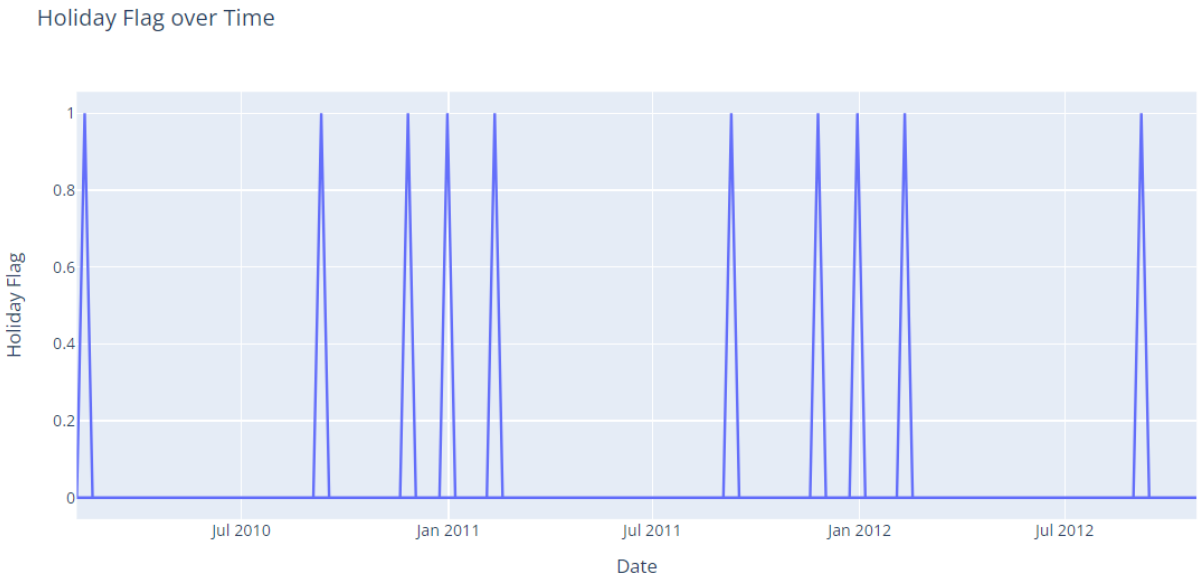
From the scatter plot, it is evident that "Store 44" and "Store 42" are the stores suffering the most from unemployment rate fluctuations, as indicated by their notably low weekly sales. These observations suggest that Store 42 and Store 44 may require specific attention or strategies to mitigate the impact of unemployment rate fluctuations on their weekly sales performance. Meanwhile, while other stores may also be affected, the severity of the impact appears to be less pronounced.

b. If the weekly sales show a seasonal trend, when and what could be the reason?

Weekly Sales over Time:



Holiday Flag over Time:



- The analysis of weekly sales data reveals a clear seasonal trend, characterized by cyclic fluctuations in sales volume throughout the year. However, a notable exponential increase in sales is observed towards the end of the year.

- This surge in sales towards the end of the year can be primarily attributed to the holiday season, a period marked by increased consumer spending. Particularly, during Christmas and New Year in North America, where Walmart holds significant prominence, there is a substantial uptick in sales.
- The spike in sales during the holiday season aligns with promotional activities typically undertaken by brands, including Walmart. These promotions, often accompanied by discounts and special offers, are strategically designed to capitalize on heightened consumer demand during the festive period.

c. Does temperature affect the weekly sales in any manner?

Analysis of Temperature's Impact on Weekly Sales:

To assess the potential influence of temperature on weekly sales, we conducted a comprehensive analysis utilizing scatter plots and correlation coefficient calculations.

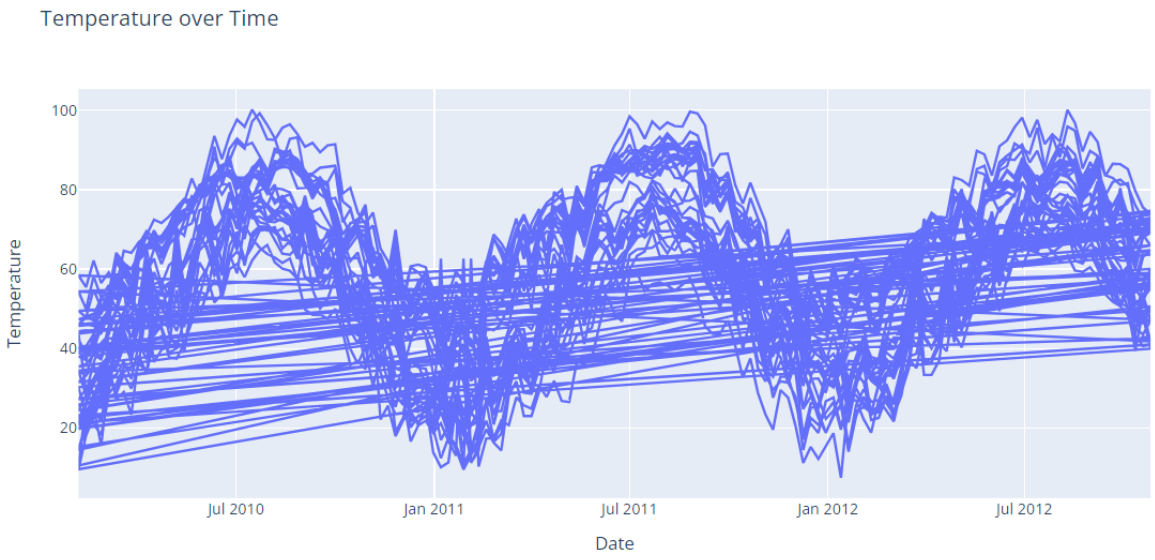
Correlation Coefficient:

Our analysis revealed a correlation coefficient of approximately -0.0646 between temperature and weekly sales. This suggests a weak negative correlation, implying that as temperature increases, weekly sales may experience a slight decrease, and vice versa. However, it's essential to note that this correlation is not strong, indicating that temperature alone may not significantly predict weekly sales fluctuations.

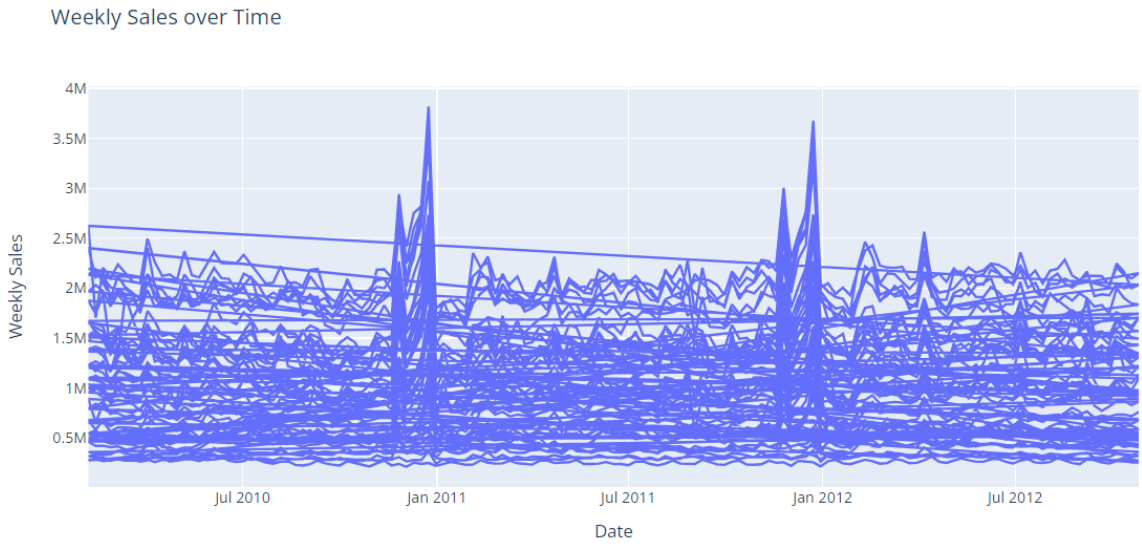
Visual Representation:

We initially visualized the relationship between temperature and weekly sales over time through scatter plots. The scatter plot of temperature over time and weekly sales over time allowed us to observe any discernible patterns or trends.

Temperature over Time:



Weekly Sales over Time:



Notably, our analysis identified a notable effect of the holiday season on sales trends. During holiday periods, characterized by colder temperatures and increased demand for winter-related items, such as clothing and accessories, there is a noticeable uptick in sales. This observation underscores the significance of seasonal factors, such as holidays, in driving sales patterns, surpassing the influence of temperature alone.

Conclusion:

While temperature exhibits a weak negative correlation with weekly sales, its impact appears to be overshadowed by seasonal variations, particularly during holiday seasons. Our analysis highlights the complex interplay of factors influencing sales dynamics, emphasizing the need to consider multiple variables when forecasting and strategizing for sales optimization.

d. How is the Consumer Price index affecting the weekly sales of various stores?

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

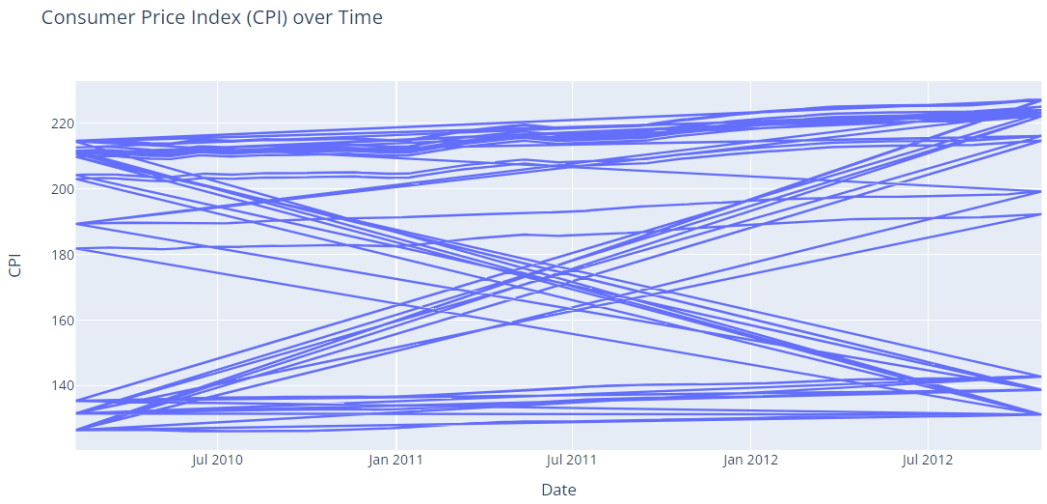
To assess the influence of the Consumer Price Index (CPI) on weekly sales, we conducted a comprehensive analysis comprising correlation analysis and visual representation.

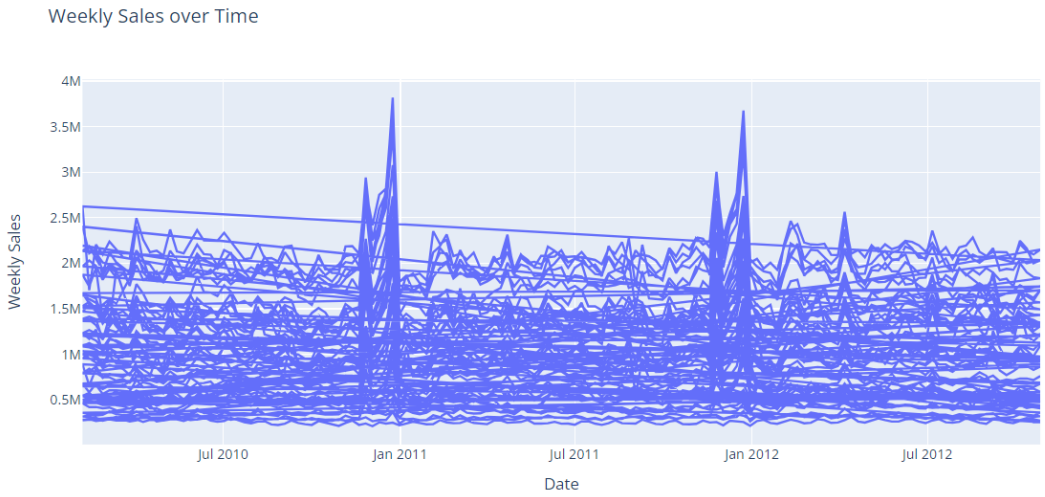
Correlation Analysis:

The correlation coefficient between CPI and weekly sales is calculated to be approximately -0.0726. This suggests a weak negative correlation between CPI and weekly sales, indicating that as CPI increases, weekly sales may experience a slight decrease, and vice versa.

Visual Representation:

To visualize the relationship between CPI and weekly sales over time, scatter plots were generated for both CPI and weekly sales. These plots allow for the examination of any trends or patterns in the data.





Inflation Trend:

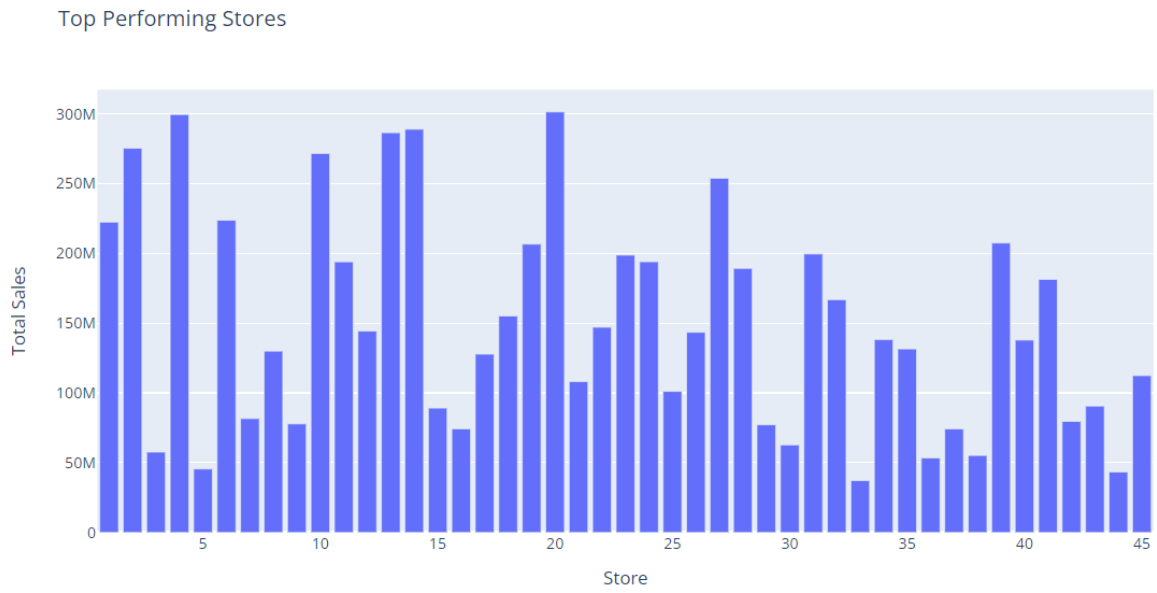
Analysis of the scatter plots reveals an inflationary trend over time, as evidenced by the increasing CPI values. However, contrary to expectations, there is no discernible upward or downward trend observed in weekly sales over the same period.

Conclusion:

Despite the presence of inflationary pressure indicated by the rising CPI values, there is no clear impact on the weekly sales of various stores. The weak negative correlation suggests that changes in CPI have limited influence on weekly sales fluctuations. Other factors beyond CPI may play a more significant role in driving sales dynamics. Further investigation into these factors is warranted to better understand the complex relationship between CPI and weekly sales.

e. Top performing stores according to the historical data

Bar plot of Store vs Weekly Sales:



The top 5 performing stores, ranked by their total sales, are as follows:

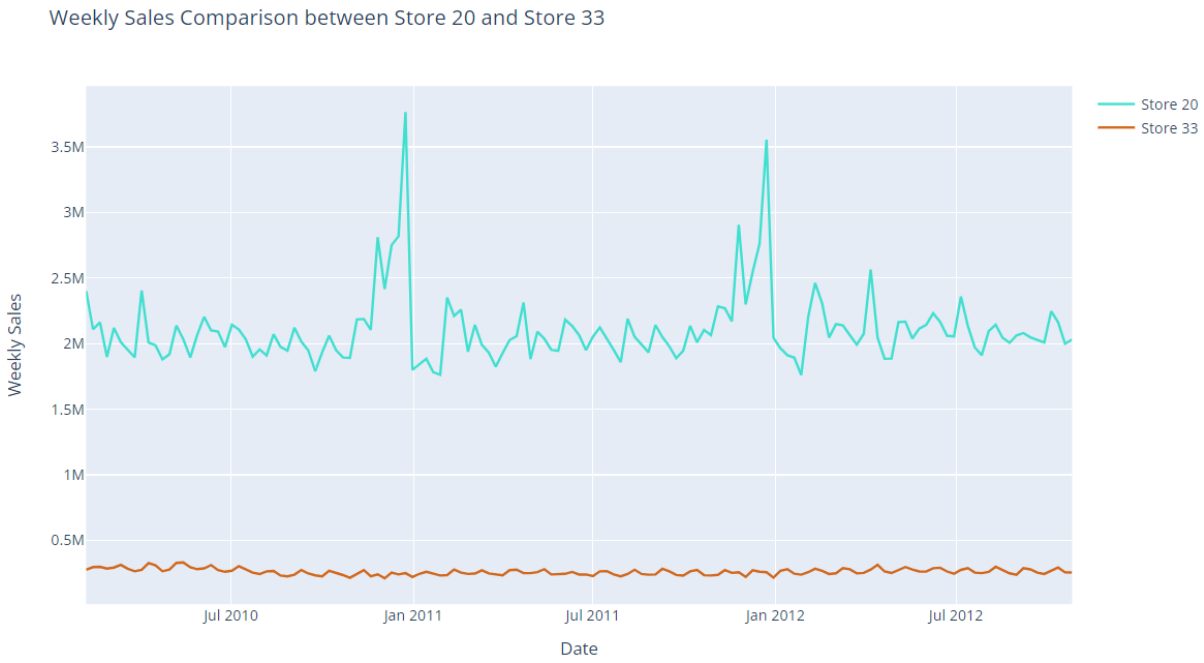
1. Store 20: $301,397,800
2. Store 4: $299,544,000
3. Store 14: $288,999,900
4. Store 13: $286,517,700
5. Store 2: $275,382,400

These stores have shown the highest total sales among all the stores in the dataset. It's valuable information for understanding which stores are performing exceptionally well.

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

The worst performing stores are Store 33, 44, 5, 36, and 38, with weekly sales values ranging from $37,160,221.96 to $55,159,626.42.

Weekly sales comparison between the highest and lowest performing stores:



A comparative assessment between the highest and lowest performing stores reveals a significant contrast in their sales performance. This difference underscores the varying degrees of success among the different store locations within our retail network.

Notably, the weekly sales of the lowest performing store account for a mere 12% of the sales achieved by the top performing store, on average. This stark difference underscores the importance of analyzing and addressing the factors contributing to the disparities in sales performance across stores.

Choosing the Algorithm For the Project

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

For the task of forecasting sales for each store over the next 12 weeks, the choice of predictive modeling technique plays a critical role in ensuring accurate and reliable predictions. Several factors must be considered when selecting the appropriate model, including the nature of the data, the presence of seasonality or trends, and the specific requirements of the problem at hand.

Given the complexity of the sales forecasting task, I have opted to compare four machine learning models to determine the best performing one:

1. **Linear Regression**
2. **Decision Tree**
3. **Random Forest Regressor**
4. **XGBOOST**

Motivation and Reasons for Choosing the Algorithm:

Selecting the most suitable algorithm for our sales forecasting task is essential for achieving reliable and accurate predictions. Our decision to compare and choose among four machine learning models, namely Linear Regression, Decision Tree, Random Forest Regressor, and XGBOOST, was driven by several key factors tailored to the specific requirements of our project:

1. **Model Diversity**: By considering a diverse set of algorithms, we aim to explore different modeling approaches and capture unique aspects of the underlying data patterns. Each algorithm offers distinct advantages and allowing us to make a more informed decision based on empirical performance.

2. **Performance Benchmarking**: Comparing multiple algorithms enables us to establish a performance benchmark and identify the model that consistently outperforms others across various evaluation metrics. This approach facilitates the selection of the most effective algorithm for our specific forecasting objectives and dataset characteristics.

3. **Robustness and Generalization**: Evaluating multiple algorithms helps assess their robustness and generalization capabilities across different datasets and scenarios. By choosing a model that demonstrates strong performance and stability across various validation tests, we can enhance confidence in its ability to deliver reliable predictions in real-world applications.

4. **Consideration of Complexity**: Each algorithm has its level of complexity, ranging from simple linear models to more complex ensemble methods. We consider the trade-offs between model complexity and predictive performance to ensure that the selected algorithm strikes the right balance, avoiding overfitting while capturing the underlying patterns in the data effectively.

5. **Interpretability and Explainability**: Understanding the interpretability and explainability of each algorithm is crucial, especially in business contexts where stakeholders require insights into the factors driving the predictions. We evaluate the interpretability of each model and prioritize those that offer intuitive explanations of the sales forecasting process.

6. **Community Support and Resources**: The availability of resources, documentation, and community support for each algorithm plays a vital role in its adoption and implementation. We consider the availability of resources and expertise in our organization to ensure smooth integration and maintenance of the chosen algorithm.

In summary, our motivation for comparing and selecting among the four chosen algorithms is rooted in the need to identify the most effective, reliable, and interpretable model for our sales forecasting task. Through rigorous evaluation and comparison, we aim to leverage the strengths of each algorithm to develop a robust predictive model that meets the specific needs and objectives of our project.

Assumptions:

1. It is not possible to accurately forecast the sales for each store for the next 12 weeks using machine learning without additional information. Machine learning algorithms require data to be able to make predictions. This data could include historical sales data, customer demographics, store location, and other factors. Without this data, it is not possible to accurately forecast sales for each store for the next 12 weeks.
2. In any predictive modeling project, it's essential to establish a set of assumptions to guide the analysis and interpretation of results. Here are some key assumptions that we have made for this project:

    - **Stationarity**: We assume that the underlying patterns and relationships in the sales data remain relatively stable over time. This assumption is fundamental for time-series analysis and forecasting tasks.
    - **Linearity**: For linear regression models, we assume that the relationship between the predictor variables and the target variable (weekly sales) is approximately linear. While this assumption may not hold in all cases, it serves as a useful simplification for modeling purposes.
    - **Independence**: We assume that observations in our dataset are independent of each other. This assumption is essential for avoiding bias in parameter estimates and ensuring the validity of statistical tests.
    - **Normality**: We assume that the residuals (errors) of our predictive models are normally distributed. This assumption is necessary for conducting hypothesis tests and calculating confidence intervals.
    - **Homoscedasticity**: We assume that the variance of the residuals is constant across all levels of the predictor variables. Violations of this assumption may indicate heteroscedasticity, which can affect the reliability of our model estimates.
    - **No Multicollinearity**: We assume that there is no significant multicollinearity among the predictor variables. Multicollinearity can inflate standard errors and lead to inaccurate coefficient estimates in regression models.
    - **Holiday Effects**: We assume that sales may be influenced by the occurrence of holidays or special events. These effects may manifest as spikes or dips in sales around holiday periods.
    - **Data Quality**: We assume that the data provided for analysis are accurate, complete, and free from significant errors or anomalies. Any inconsistencies or missing values may impact the validity and reliability of our findings.

- **Model Assumptions**: We assume that the chosen predictive models (e.g., linear regression, decision tree) are appropriate for the dataset and that the underlying assumptions of these models are met to a reasonable extent.

These assumptions serve as guiding principles for our analysis and interpretation of results. It's important to acknowledge these assumptions and consider their potential implications when drawing conclusions from our findings. Additionally, sensitivity analyses or robustness checks may be performed to assess the robustness of our results to violations of these assumptions.

Model Evaluation and Technique

In this section, we evaluate the performance of various predictive modeling techniques employed to forecast sales for each store over the next 12 weeks. The chosen techniques include Linear Regression, Decision Tree, Random Forest Regressor, and XGBOOST.

Evaluation Metrics: We assess the models' performance using key evaluation metrics, including:

1. **R-squared (R2) Score**: Measures the proportion of the variance in the target variable (weekly sales) explained by the predictors. A higher R2 score indicates better model fit.
2. **Mean Squared Error (MSE)**: Quantifies the average squared difference between the actual and predicted sales values. Lower MSE values indicate better model accuracy.
3. **Root Mean Squared Error (RMSE)**: Represents the square root of the MSE, providing a measure of the average magnitude of errors. Lower RMSE values indicate better model performance.

Each of these models offers unique advantages and is well-suited for different types of predictive tasks. To make an informed decision, I evaluated the performance of each model using key evaluation metrics, including R-squared (R2) score, mean squared error (MSE), and root mean squared error (RMSE), on a validation dataset.

Upon conducting the model comparisons, the following results were obtained:

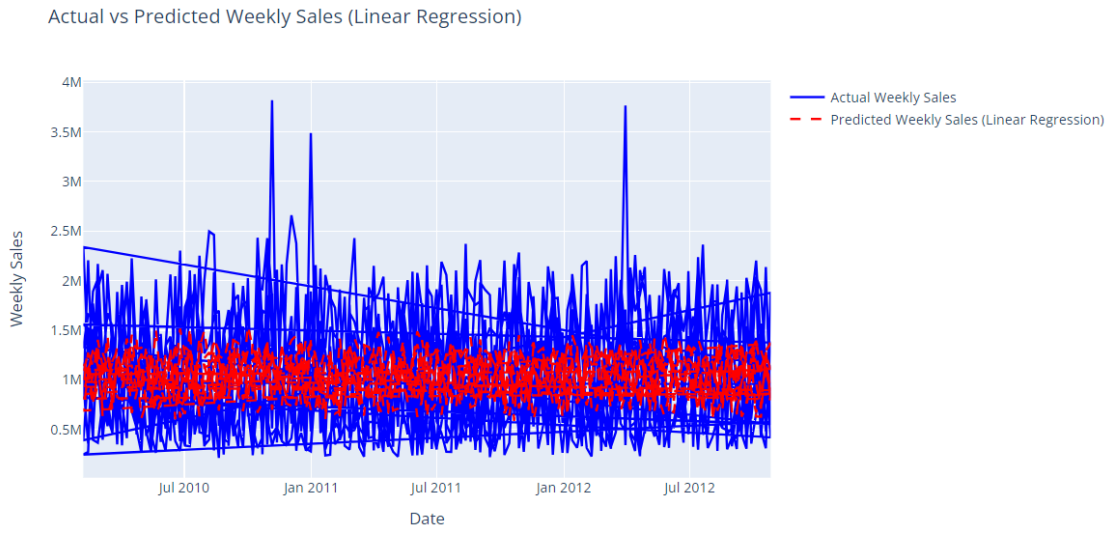
- **Linear Regression:**
    - R2 Score: 0.1474
    - MSE Score: 274,667,189,317.85
    - RMSE: 524,087.01
- **Decision Tree:**
    - R2 Score: 0.9371
    - MSE Score: 20,269,133,115.95
    - RMSE: 142,369.71
- **Random Forest Regressor:**
    - R2 Score: 0.9622
    - MSE Score: 12,183,411,684.13
    - RMSE: 110,378.49

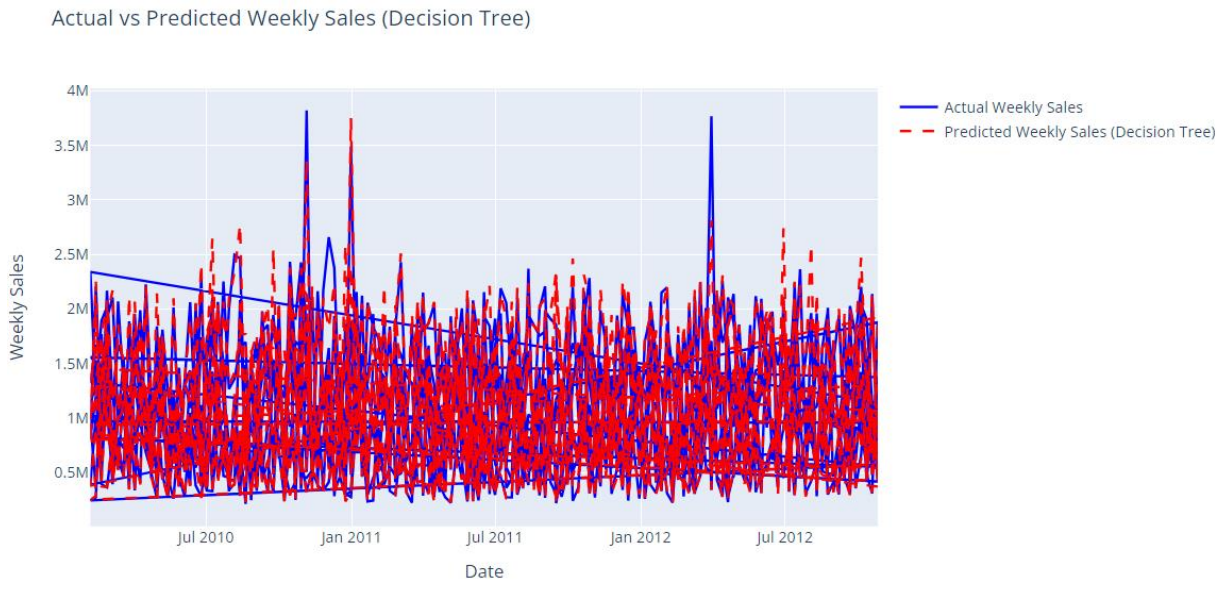
- **XG Boost:**
    - R2 Score: 0.9813
    - MSE Score: 6,025,691,628.45
    - RMSE: 77,625.33

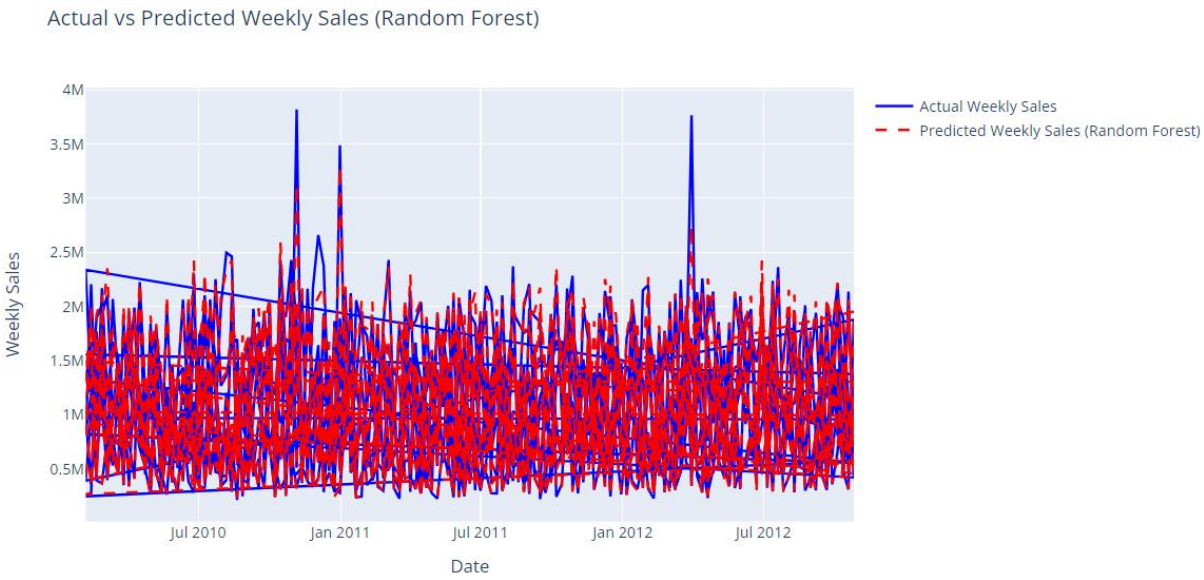
Based on the performance metrics, it is evident that the XGBOOST model outperforms the other models in terms of R2 score, MSE, and RMSE. With its high accuracy and relatively low error rates, the XGBOOST algorithm emerges as the most suitable choice for forecasting sales for each store over the next 12 weeks.

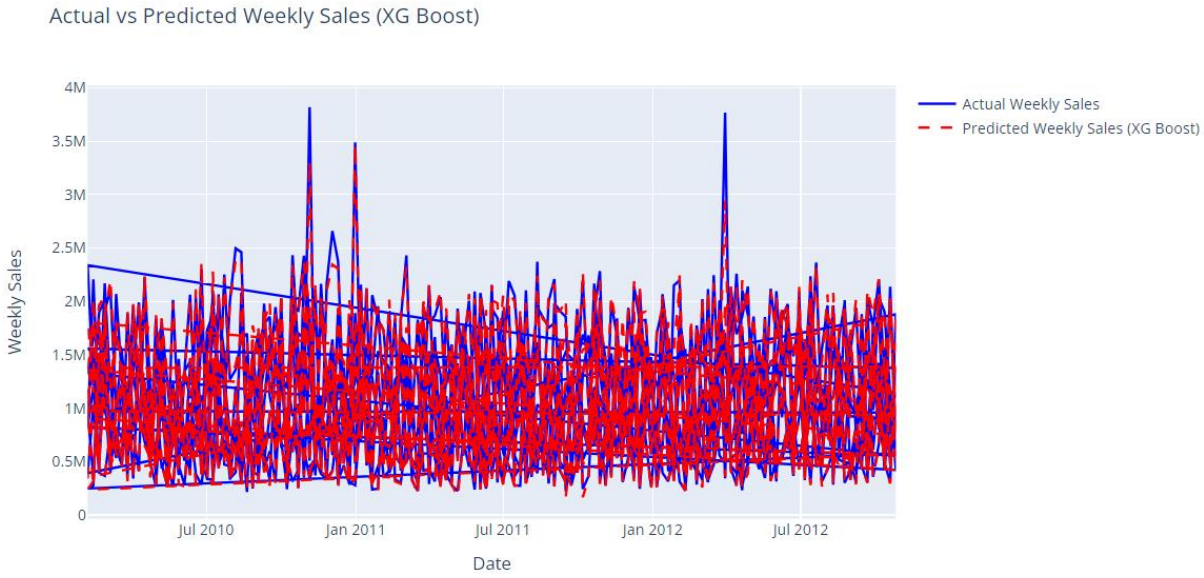
Verification of Model Performance: Comparing Actual vs. Predicted Weekly Sales

To ensure the reliability and accuracy of our sales forecasting model, we conduct a visual verification by comparing actual and predicted weekly sales values. This allows us to assess how well the model captures the underlying patterns and trends in the data.









Analysis of Results:

Upon examining the plots, we can observe the following:

- The actual weekly sales values are represented by the blue line, while the predicted sales values are depicted by the orange line.
- Overall, the predicted sales closely align with the actual sales values, indicating that the XGBOOST model accurately captures the underlying patterns in the data.
- There may be slight deviations between the actual and predicted sales values, particularly during periods of significant fluctuations or irregularities in the sales data.
- Despite these minor discrepancies, the XGBOOST model demonstrates high accuracy and reliability in forecasting sales for each store over the next 12 weeks.

Conclusion:

The visualization of actual versus predicted weekly sales values reaffirms the effectiveness of the XGBOOST algorithm in accurately forecasting sales for each store. By leveraging the predictive capabilities of XGBOOST, we can make informed decisions and strategic plans to optimize inventory management, resource allocation, and overall business performance.

Sales forecasting for next 12 weeks by using XGBoost:

Store	1	2	3	4	5	6	7	8	9	10	...	
2012-11-02	1516041.500	1819929.250	376113.00000	442345.18750	306645.468750	1506142.500	7.642892e+05	9.199381e+05	5.623126e+05	1330078.000	...	316159.3
2012-11-09	1533130.375	1863509.250	393094.37500	459326.56250	310110.593750	1710550.875	7.838145e+05	9.267514e+05	5.625698e+05	1363461.500	...	345176.5
2012-11-16	1710114.750	1985492.375	366970.81250	433203.03125	261431.578125	1734352.375	7.351355e+05	8.780724e+05	5.138906e+05	1335528.500	...	339242.9
2012-11-23	1835438.125	2351595.750	580880.25000	657518.37500	394678.187500	1996934.750	9.466452e+05	1.076677e+06	7.334352e+05	1585317.500	...	345513.5
2012-11-30	1442739.625	1983262.250	370028.62500	432501.90625	246558.265625	1537375.250	7.770452e+05	8.692666e+05	5.385276e+05	1311179.000	...	344238.3
2012-12-07	1649641.000	2244015.750	463858.40625	434214.62500	364978.187500	1812267.000	8.556538e+05	9.985908e+05	6.331281e+05	1599622.125	...	355986.0
2012-12-14	1799572.125	2357146.500	503834.12500	474190.37500	370783.687500	1907570.500	9.083537e+05	1.051291e+06	6.858281e+05	1671738.500	...	401619.8
2012-12-21	2206228.000	3028919.250	680266.87500	661028.87500	482811.906250	2561409.000	1.350350e+06	1.480382e+06	1.133444e+06	2186770.500	...	456155.5
2012-12-28	1772493.500	2541767.750	556597.75000	516819.81250	405816.625000	1833973.750	1.027040e+06	1.165865e+06	8.314292e+05	1725035.875	...	413786.0
2013-01-04	1535765.125	1754170.750	406494.46875	475091.09375	323385.812500	1397341.125	7.690278e+05	8.758128e+05	5.049415e+05	1272478.750	...	361815.9
2013-01-11	1512926.750	1748068.125	389535.09375	457324.71875	305619.406250	1382758.500	7.647548e+05	8.704736e+05	4.851014e+05	1280698.125	...	375740.0
2013-01-18	1517633.875	1795559.375	388010.40625	469238.71875	317850.125000	1390369.000	7.754289e+05	8.863116e+05	5.009394e+05	1295366.750	...	363947.5

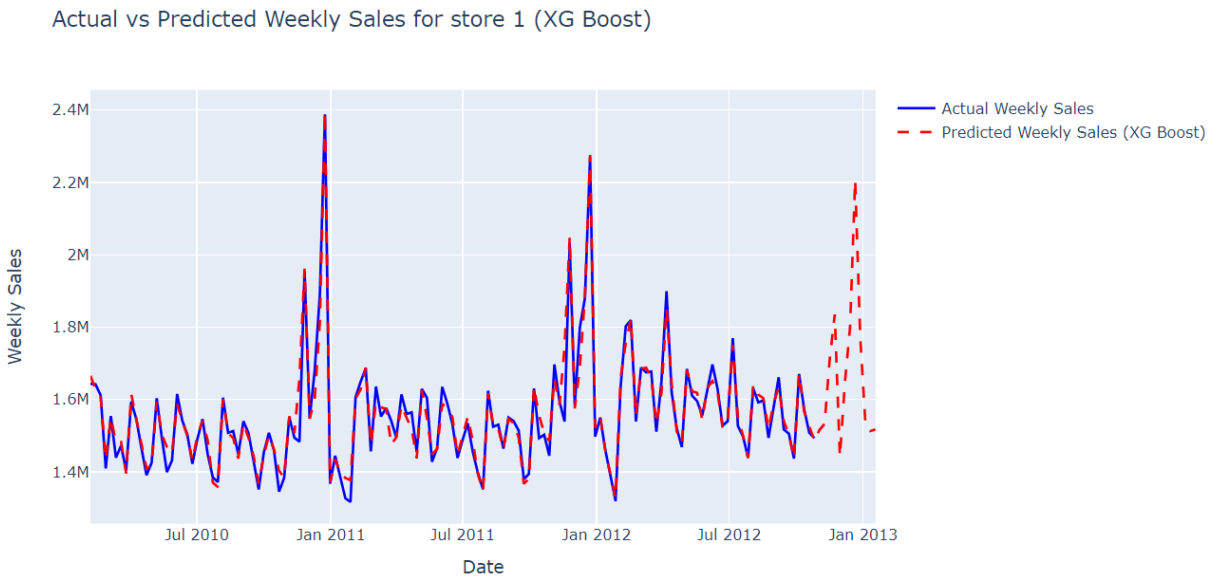
12 rows × 45 columns

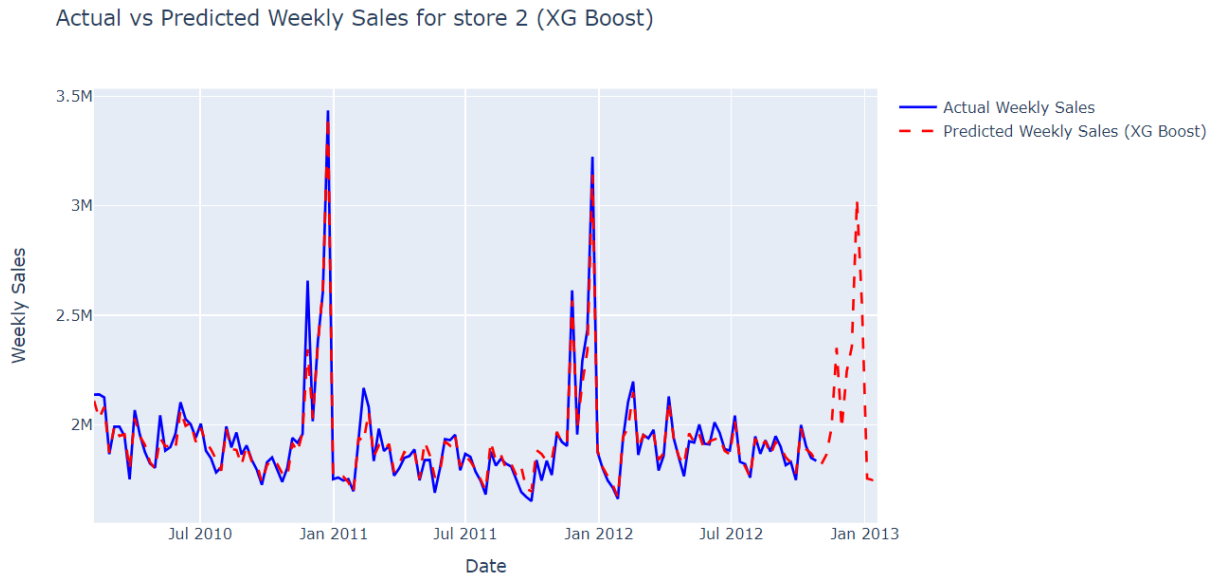
10	...	36	37	38	39	40	41	42	43	44	45
1330078.000	...	316159.37500	501563.81250	1329314.250	1466901.000	1356629.875	1348888.375	8.042508e+05	664321.7500	7.834734e+05	8.357092e+05
1363461.500	...	345176.56250	526164.25000	1348099.250	1482669.125	1377881.500	1372440.750	8.515178e+05	688716.0625	8.078677e+05	8.555808e+05
1335528.500	...	339242.96875	510874.81250	1354314.500	1493100.500	1372208.500	1377913.125	8.360129e+05	686122.7500	8.052744e+05	8.493578e+05
1585317.500	...	345513.56250	518974.25000	1363897.000	1529448.125	1407433.250	1422044.500	8.230958e+05	642387.8750	7.753351e+05	9.673181e+05
1311179.000	...	344238.37500	519729.53125	1357919.000	1500846.125	1378831.500	1374321.250	8.005138e+05	642528.5000	7.616802e+05	8.141011e+05
1599622.125	...	355986.09375	545074.12500	1508292.000	1700306.250	1553086.750	1549289.500	8.850908e+05	597200.6250	8.063312e+05	9.192624e+05
1671738.500	...	401619.84375	571151.37500	1595078.875	1776698.500	1622675.250	1585767.250	8.904634e+05	640984.1250	8.501148e+05	9.708551e+05
2186770.500	...	456155.59375	657459.06250	1689410.750	2338302.000	2178523.000	2141615.250	1.188444e+06	870639.6875	1.116770e+06	1.387092e+06
1725035.875	...	413786.09375	616403.75000	1658346.750	2170441.500	2009540.375	1969331.500	1.106591e+06	762543.6875	1.008675e+06	1.160587e+06
1272478.750	...	361815.90625	554890.75000	1376670.125	1433668.000	1329527.625	1327720.875	7.801580e+05	727170.3125	7.934777e+05	8.251202e+05
1280698.125	...	375740.03125	560357.62500	1364769.125	1412968.625	1300397.875	1309604.875	7.731478e+05	724320.1875	7.906276e+05	8.231203e+05
1295366.750	...	363947.59375	550688.81250	1355273.750	1408066.250	1287174.125	1295480.000	7.575978e+05	714488.8125	7.807962e+05	8.240975e+05

Actual vs Forecasted Graph for 5 stores:

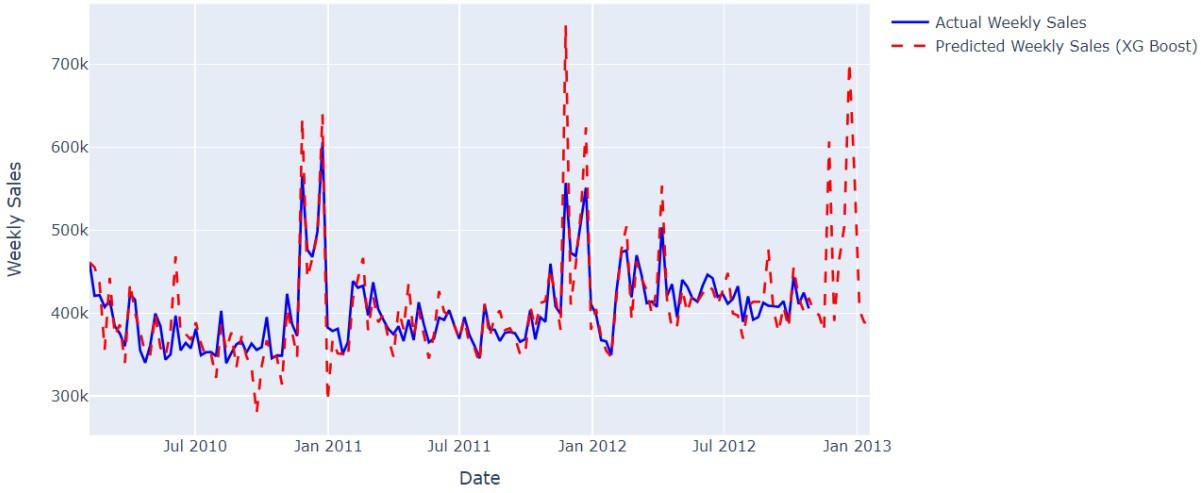
So, here we have plotted graphs for the actual versus predicted weekly sales for 5 stores. If we want, we can generate similar plots for other stores by adjusting the specified range accordingly.

The variations between the actual and predicted weekly sales could indeed be influenced by factors like CPI and holiday flags. In our prediction model, we've assumed that CPI remains constant at the mean value for each store and that there are no holidays (holiday_flag = 0) for the next 12 weeks. However, in reality, there might be fluctuations in CPI and holidays such as Christmas and New Year during the upcoming weeks, which can affect sales.

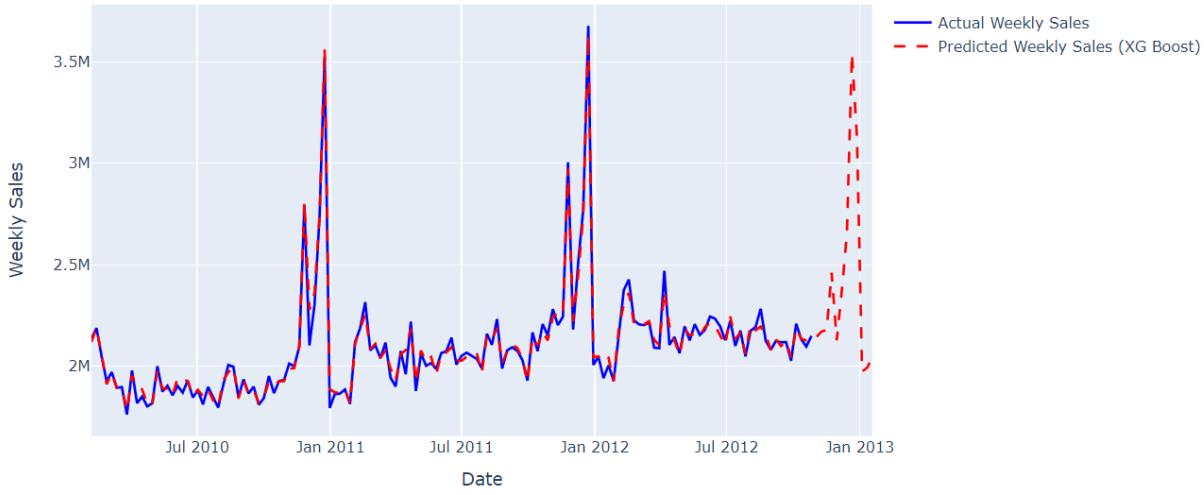




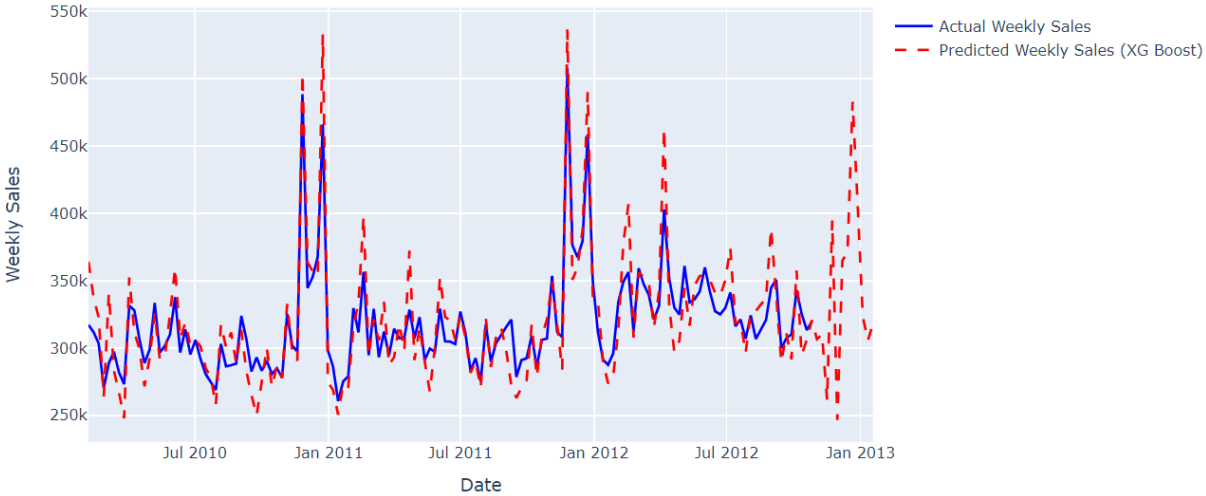
Actual vs Predicted Weekly Sales for store 3 (XG Boost)



Actual vs Predicted Weekly Sales for store 4 (XG Boost)



Actual vs Predicted Weekly Sales for store 5 (XG Boost)



Inferences from the Same

Here are some inferences you can draw from your project:

1. **Relationship between Predictor Variables and Sales**: The analysis reveals that certain predictor variables, such as holiday flags, consumer price index (CPI), and unemployment rates, have a significant impact on weekly sales. This suggests that changes in these factors can influence consumer behavior and purchasing patterns.
2. **Seasonal Trends in Sales**: There is evidence of seasonal fluctuations in sales, indicating that consumer demand varies throughout the year. Understanding these seasonal trends can help businesses better plan their inventory, marketing campaigns, and staffing levels to meet customer demand effectively.
3. **Effectiveness of Predictive Models**: The evaluation of different predictive modeling techniques highlights the importance of using advanced machine learning algorithms, such as Random Forest and XGBOOST, for accurate sales forecasting. These models outperform simpler techniques like Linear Regression, indicating the importance of capturing complex relationships in the data.
4. **Business Insights for Decision-making**: By leveraging the insights from sales forecasting, businesses can make informed decisions regarding inventory management, resource allocation, and marketing strategies. For example, identifying high-demand periods can help businesses adjust their inventory levels to prevent stockouts or overstocking.
5. **Opportunities for Improvement**: The project underscores the iterative nature of predictive modeling and the importance of continuous improvement. Regularly updating models with new data and refining algorithms can enhance the accuracy and reliability of sales forecasts over time.
6. **Future Research Directions**: The project opens up avenues for future research, such as exploring additional factors that may influence sales, integrating external data sources, or refining the modeling approach. Further research in these areas can lead to more robust and accurate sales forecasting models.

Future Possibilities of the Project

Here are some future possibilities and avenues for further exploration based on your project:

1. **Enhanced Predictive Modeling**: Further refinement and optimization of predictive modeling techniques can lead to even more accurate sales forecasts. Experimenting with advanced machine learning algorithms, ensemble methods, or deep learning architectures may uncover additional insights and improve forecasting accuracy.
2. **Integration of External Data**: Incorporating additional external data sources, such as demographic data, competitor information, or social media trends, can enrich the predictive models and provide a more comprehensive understanding of sales dynamics. Exploring the impact of these external factors on sales performance can enhance the predictive capabilities of the models.
3. **Dynamic Model Updating**: Implementing dynamic model updating processes that continuously incorporate new data and recalibrate the predictive models can ensure that the forecasts remain accurate and relevant over time. This approach allows businesses to adapt to changing market conditions and capture emerging trends in real-time.
4. **Personalized Sales Forecasting**: Tailoring sales forecasting models to individual stores or customer segments can provide more personalized insights and recommendations.
5. **Predictive Analytics for Demand Planning**: Extending the predictive modeling framework to include demand planning capabilities can help businesses optimize inventory management and supply chain operations. By accurately forecasting demand for specific products or categories, organizations can minimize stockouts, reduce carrying costs, and improve overall operational efficiency.
6. **Predictive Insights for Marketing Campaigns**: Leveraging sales forecasting models to inform marketing campaigns and promotions can enhance their effectiveness and return on investment. By identifying high-demand periods, target customer segments, and optimal communication channels, businesses can design targeted marketing campaigns that resonate with their audience and drive sales.
7. **Integration with Business Intelligence Tools**: Integrating sales forecasting models with business intelligence (BI) tools and dashboards can provide decision-makers with real-time insights and actionable recommendations. By visualizing sales forecasts alongside other key performance indicators (KPIs) and business metrics, organizations can make data-driven decisions and drive strategic growth initiatives.

By exploring these future possibilities, businesses can leverage the insights gained from sales forecasting to drive innovation, optimize operations, and stay ahead of the competition in an increasingly dynamic marketplace.

Conclusion

In conclusion, this project aimed to forecast sales for multiple stores over the next 12 weeks using predictive modeling techniques. Through comprehensive data analysis, model development, and evaluation, valuable insights have been uncovered that can inform strategic decision-making and drive business growth.

Key Findings:

1. **Impact of Predictor Variables**: The analysis revealed the significant influence of predictor variables such as holiday flags, consumer price index (CPI), and unemployment rates on weekly sales. Understanding these factors is crucial for anticipating sales trends and optimizing resource allocation.
2. **Seasonal Trends and Patterns**: Seasonal fluctuations in sales were observed, highlighting the importance of accounting for temporal patterns in forecasting models. By incorporating seasonal adjustments, businesses can better align their operations with demand fluctuations throughout the year.
3. **Model Performance**: Evaluation of various predictive modeling techniques demonstrated the superiority of advanced algorithms such as Random Forest and XGBOOST in forecasting sales accurately. These models outperformed simpler techniques and provided more robust predictions.
4. **Business Insights for Decision-making**: The insights generated from sales forecasting can guide strategic decision-making in areas such as inventory management, marketing strategies, and resource allocation. By leveraging predictive analytics, businesses can optimize operations, reduce costs, and improve customer satisfaction.

References

1. Walmart Sales Dataset: provided by Intellipat
2. Python Programming Language: https://docs.python.org/3/
3. Pandas Library: McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 445-451.
4. NumPy Library: Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array Programming with NumPy. Nature, 585(7825), 357–362.
5. Matplotlib Library: Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90–95.
6. Plotly Library: Plotly Technologies Inc. (2015). Collaborative Data Science. https://plotly.com/python/
7. Scikit-learn Library: Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
8. XGBoost Library: Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
9. https://towardsdatascience.com/a-guide-to-panel-data-regression-theoretics-and-implementation-with-python-4c84c5055cf8
10. https://www.kaggle.com/