Capstone Project (Walmart)

Problem Statement

The world's largest company by revenue, Walmart, sells everything from groceries, home furnishings, body care products to electronics, clothing, etc. and generates a large amount of consumer data that it utilizes to predict customer buying patterns, future sales, and promotional plans and creating new and innovative in-store technologies. The employment of modern technological approaches is crucial for the organization to survive in today's cutting-edge global market and create products and services that distinguish them from its competitors.

The main focus of this research is to predict Walmart's sales based on the available historic data and identify whether factors like temperature, unemployment, fuel prices, etc affect the weekly sales of particular stores under study. This aims to understand whether sales are relatively higher during holidays like Christmas and Thanksgiving than normal days so that stores can work on creating promotional offers that increase sales and generate higher revenue. It is also essential for Walmart to understand user requirements and user buying patterns to create higher customer retention, increasing their demand adding to their profits. The findings from this can help the organization understand market conditions at various times of the year and allocate resources according to regional demand and profitability.

Project Objective

The objective of the project is to develop a statistical Model based on the dataset available. The historical sales data for 45 Walmart stores located in different regions to predicting the weekly sales for each store and come up with various insights that can give them a clear perspective on the following:

a. Which store has maximum sales?

b. Provide a monthly and semester view of sales in units and give insights.

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

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

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

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

g. Top performing stores according to the historical data.

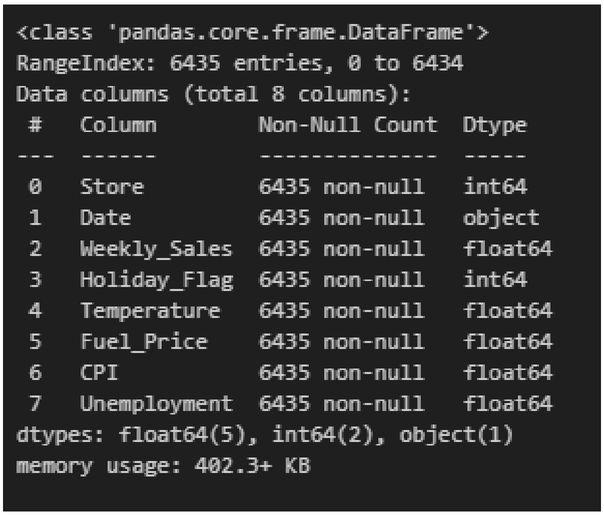
h. The worst performing store, and how significant is the difference between the highest and lowest performing stores.

g. Use predictive modelling techniques to forecast the sales for each store for the next 12 weeks.

Data Description

The dataset contains 6435 rows and 8 columns with the following features:

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

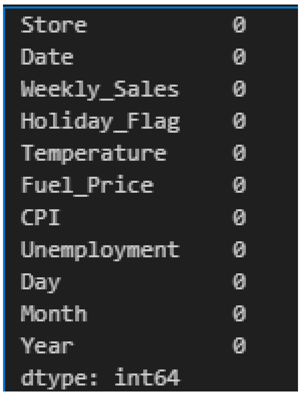


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6435 entries, 0 to 6434
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Store         6435 non-null   int64  
 1   Date          6435 non-null   object 
 2   Weekly_Sales 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

Data Pre-processing Steps and Inspiration:

About the Dataset

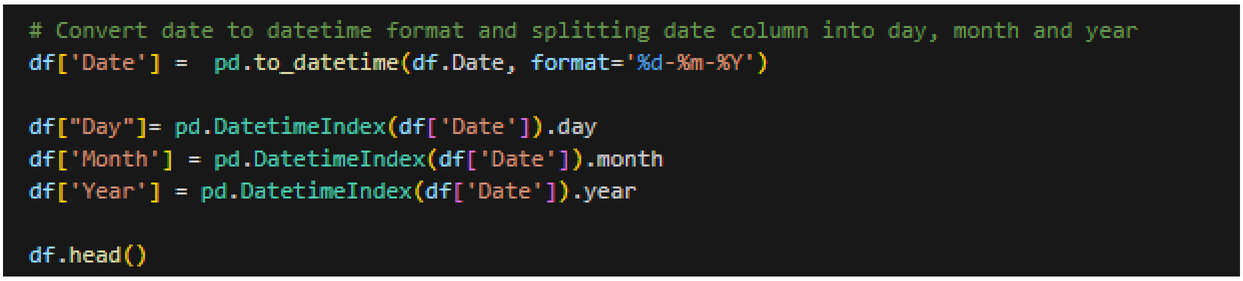
- **Handling Missing Values**: Inspect and impute any missing or null values.



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

- There are no missing values

- **Data Transformation**: Convert the date feature into a more usable format (e.g., separating the month and year or extracting useful time features like seasonality).



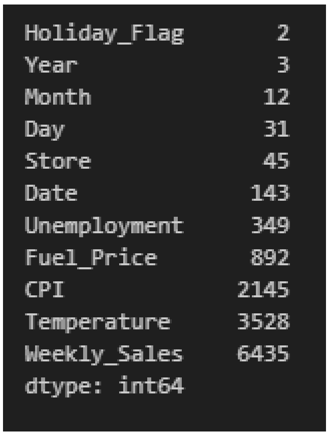
```
# Convert date to datetime format and splitting date column into day, month and year
df['Date'] =  pd.to_datetime(df.Date, format='%d-%m-%Y')

df["Day"]= pd.DatetimeIndex(df['Date']).day
df['Month'] = pd.DatetimeIndex(df['Date']).month
df['Year'] = pd.DatetimeIndex(df['Date']).year

df.head()
```

- Data Shape: 6435 rows and 11 columns.

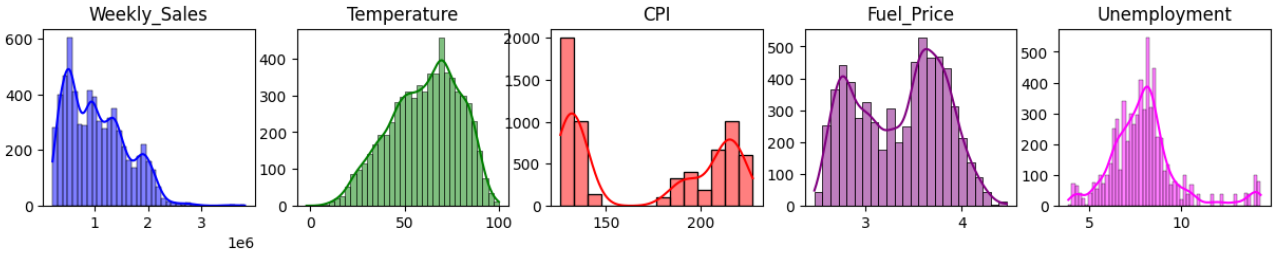
- Check for unique values:



```
Holiday_Flag       2
Year               3
Month             12
Day               31
Store             45
Date             143
Unemployment     349
Fuel_Price       892
CPI             2145
Temperature     3528
Weekly_Sales    6435
dtype: int64
```

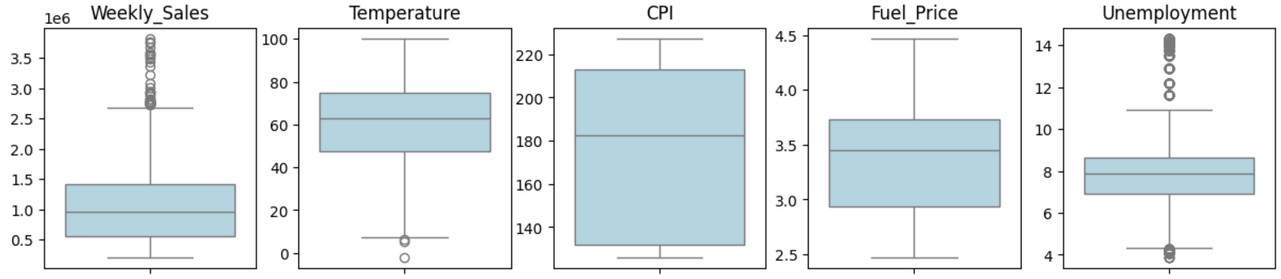
- Duplicated Values: No duplicated values

- The histograms with KDE curves show the distributions of Weekly_Sales, Temperature, CPI, Fuel_Price, and Unemployment.

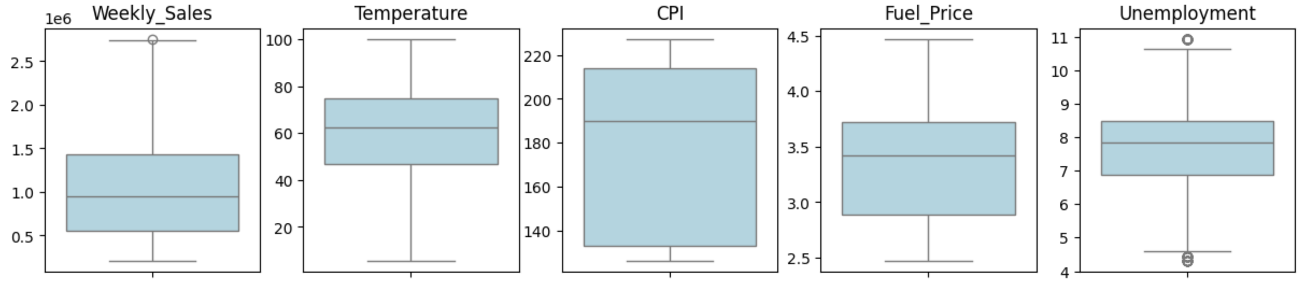


- **Outlier Detection**: Identify and handle outliers in the data.

  Box plots for Weekly_Sales, Temperature, CPI, Fuel_Price, and Unemployment to show detect any outliers in these variables.

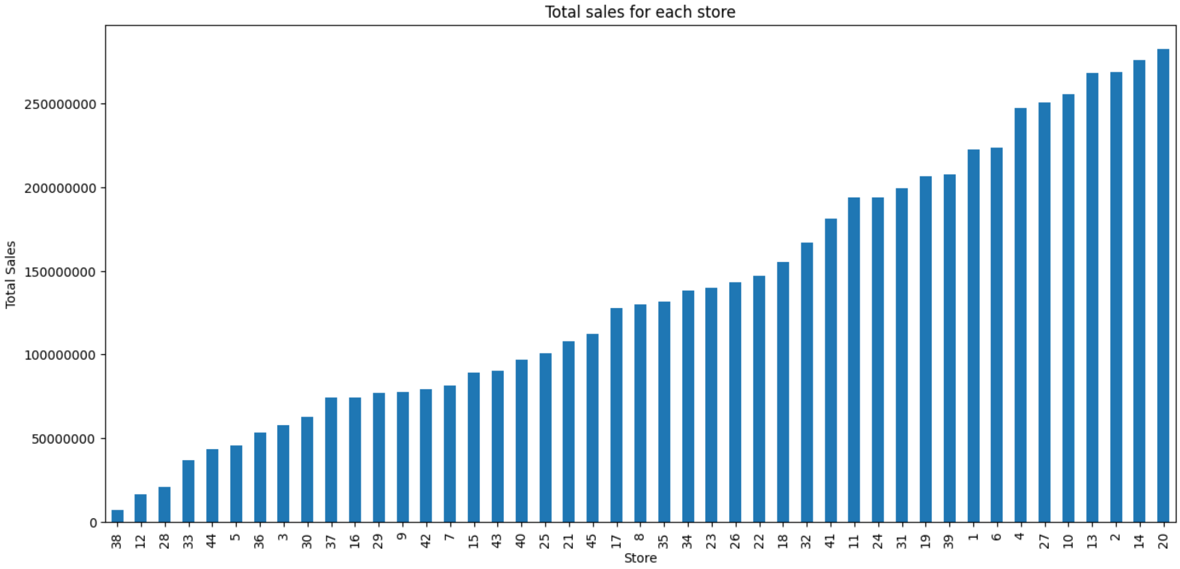


Dropping the outliers: Outliers present in Weekly_Sales, Temperature and Unemployment columns, after droping the outliers.



Insights from the dataset:

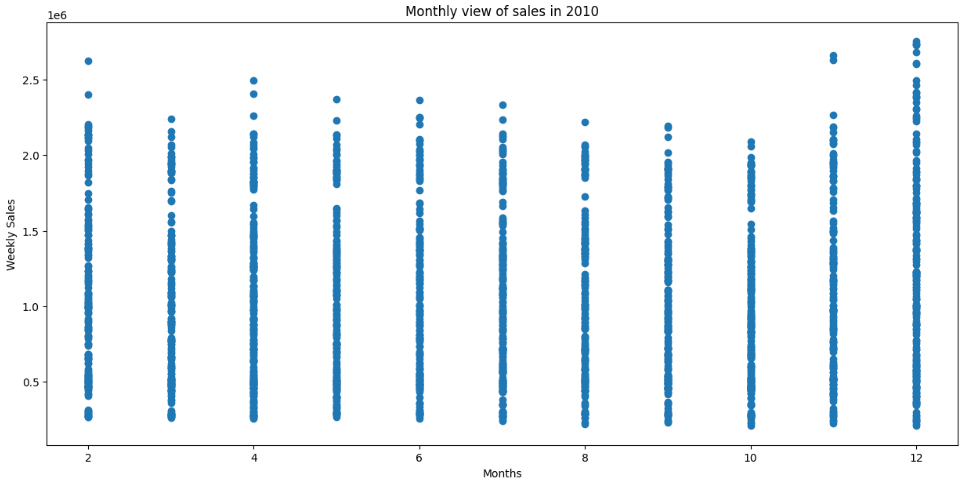
a) Which store has maximum sales?

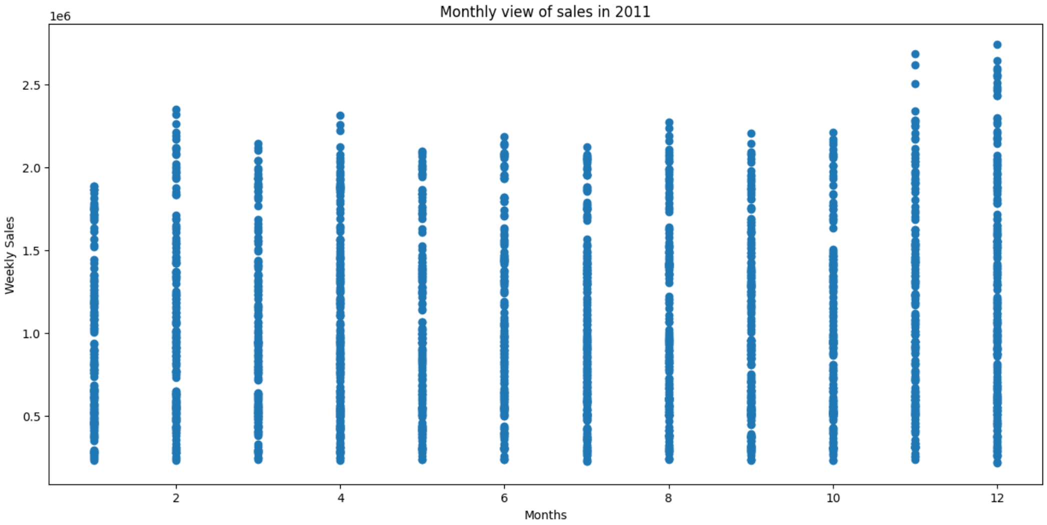


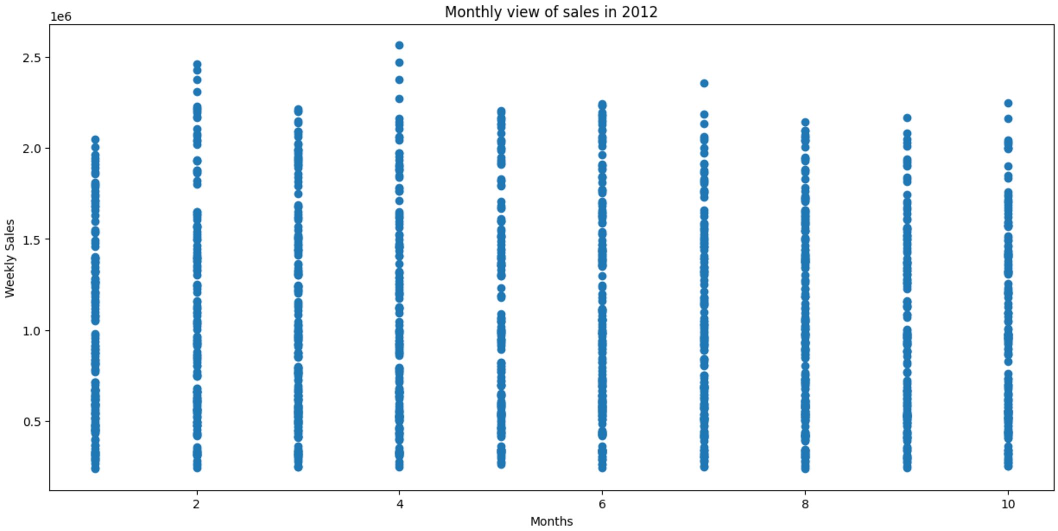
From the above graph, it is visible that the store which has maximum sales is store number 20 and the store which has minimum sales is the store number 33.

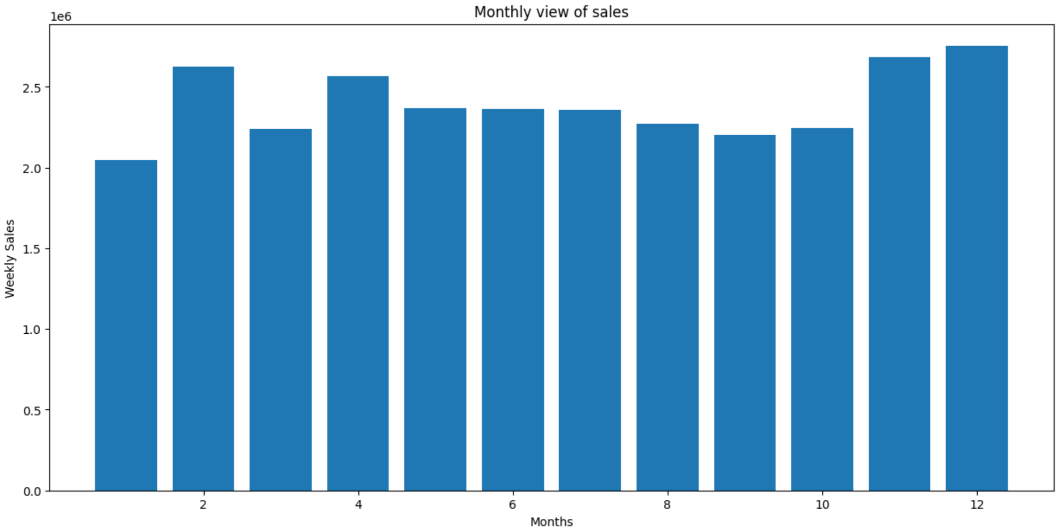
b) Provide a monthly and semester view of sales in units and give insights.

-Year wise monthly sales



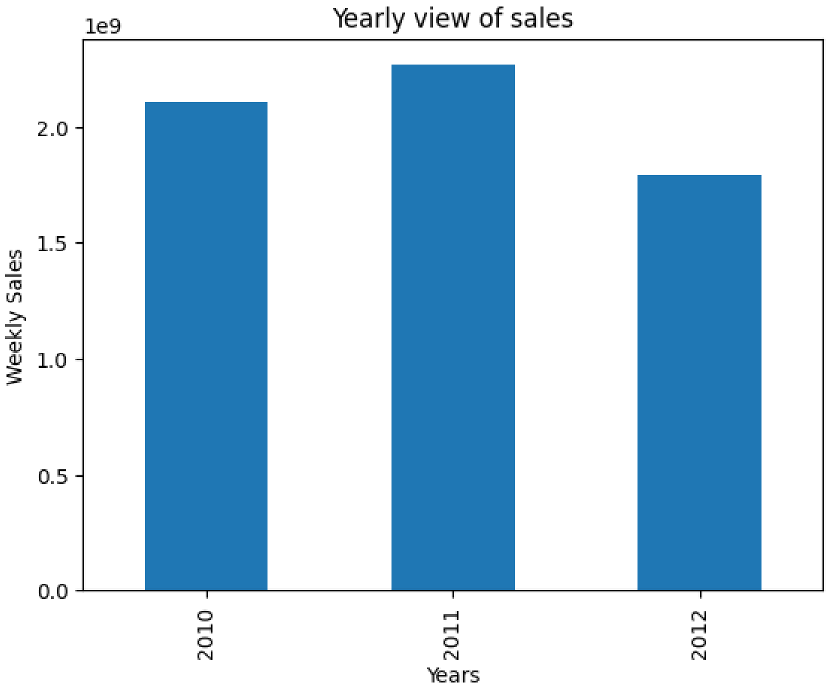




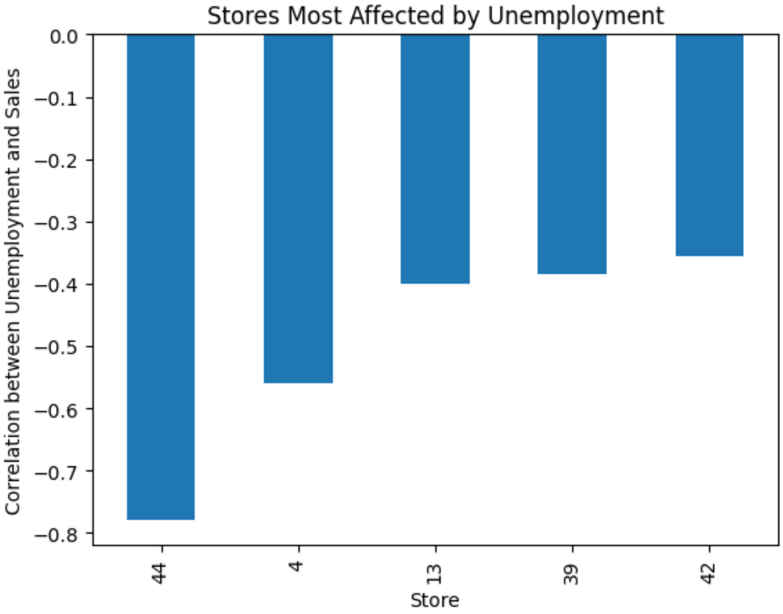


overall monthly sales are higher in the month of December while the yearly sales in the year 2011 are the highest.

- **Yearly Sales**

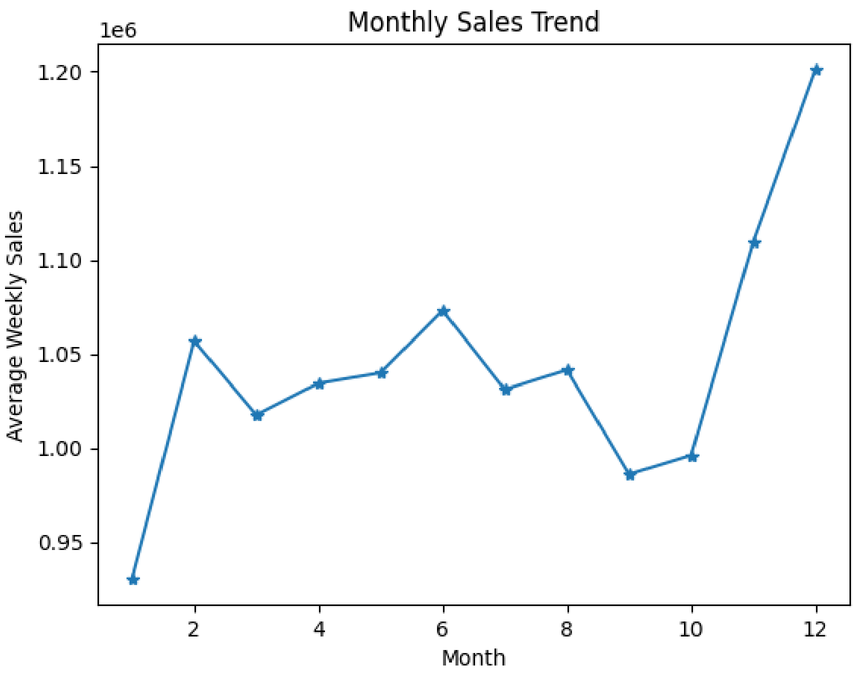


c. If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?



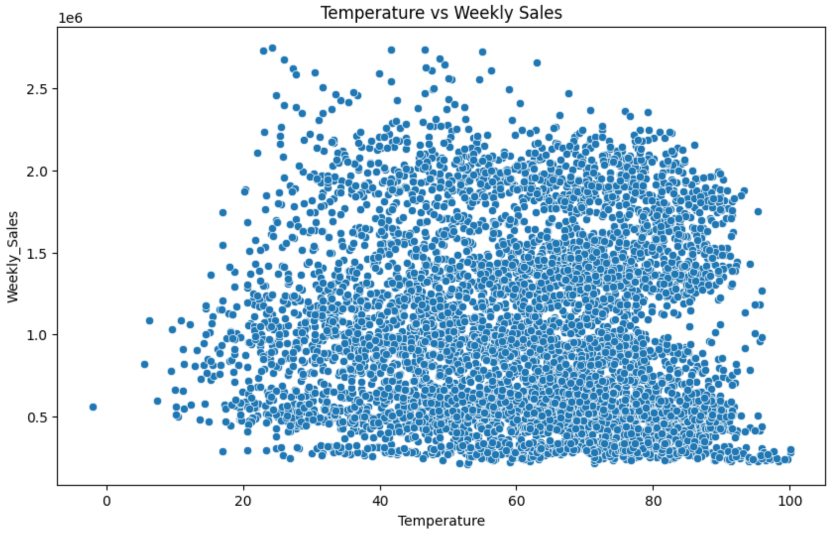
Correlation value ranges from approximately -0.4 to -0.8, which is a negative correlation. So higher unemployment in the region correlation with lower sales in these stores.

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



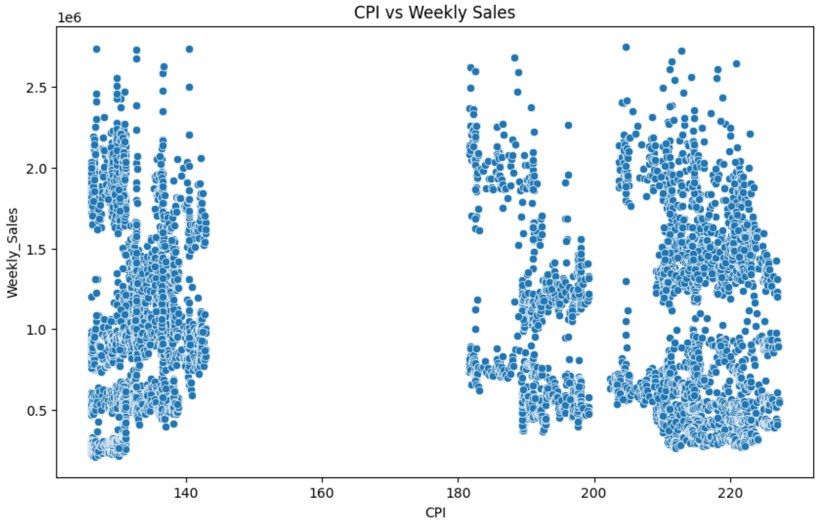
- There is a **significant spike in December** (Month 12), where the average weekly sales sharply increase compared to all other months.
- From **January to November**, sales are relatively stable, fluctuating between 0.95M and 1.10M.
- There is a slight dip around August (Month 8) and a moderate increase in sales during mid-year (June and July).
- Yes, the plot clearly shows a seasonal trend, especially towards the end of the year:
  - In December due to Christmas and New Year sales are high.
  - In June it could be due to summer shopping or after vacation back to school shopping there is a slight increase in sales.
  - In September there is a dip, which shows there is a less shopping because there is no holiday season.

e) Does temperature affect the weekly sales in any manner?



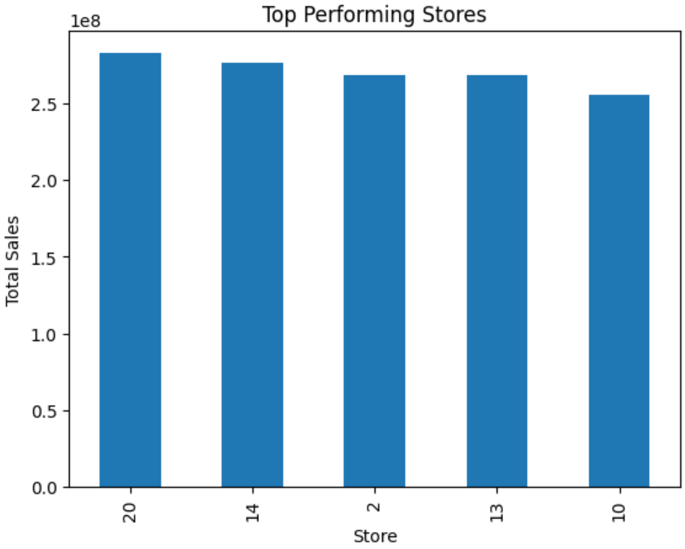
Temperature does not significantly affect weekly sales based on the scatter plot

f) How is the Consumer Price index affecting the weekly sales of various stores?



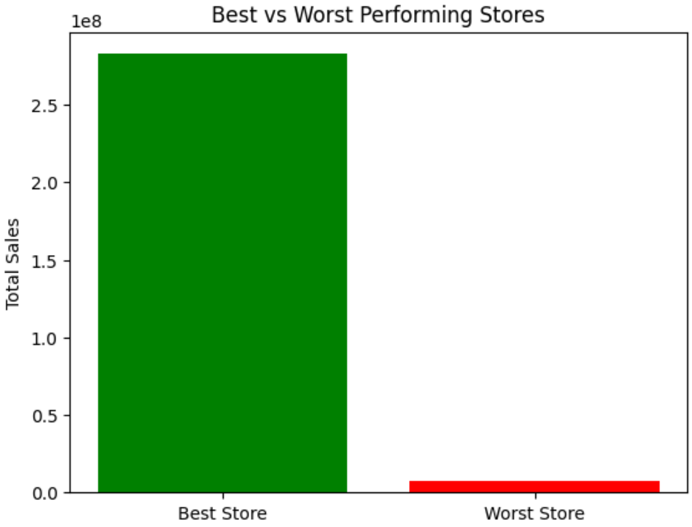
The scatter plot suggests that CPI has a **weak or non-linear correlation** with weekly sales, based on the variability in sales for different CPI ranges.

g) Top performing stores according to the historical data.



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

Difference between highest and lowest performing stores: 275428478.0



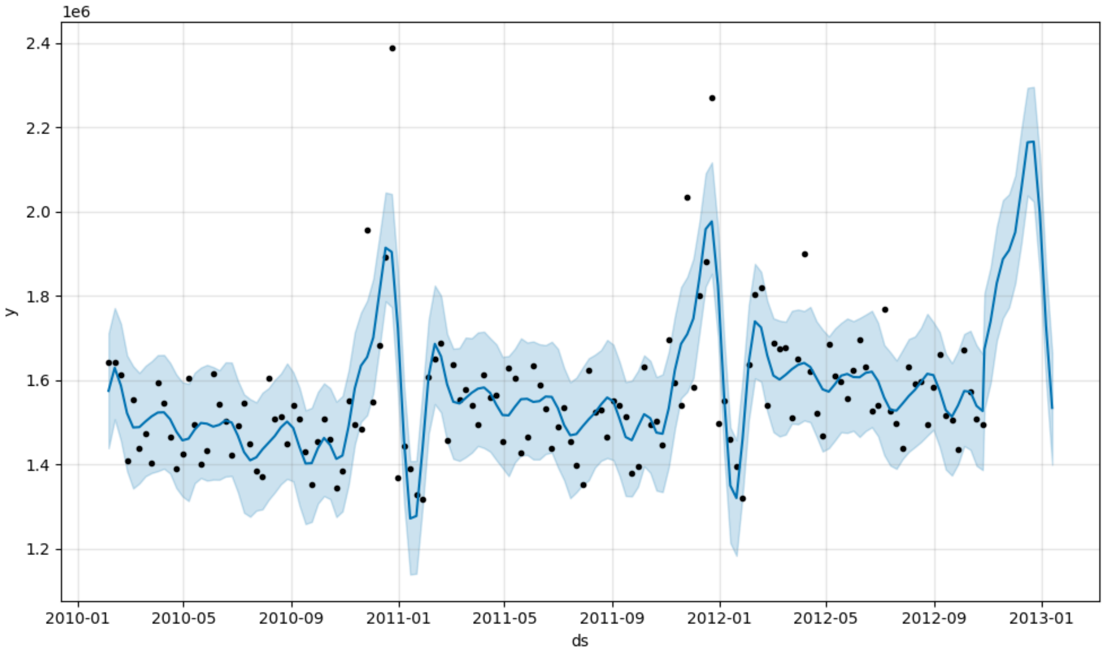
The chart shows the huge difference between the best-performing and worst-performing stores.

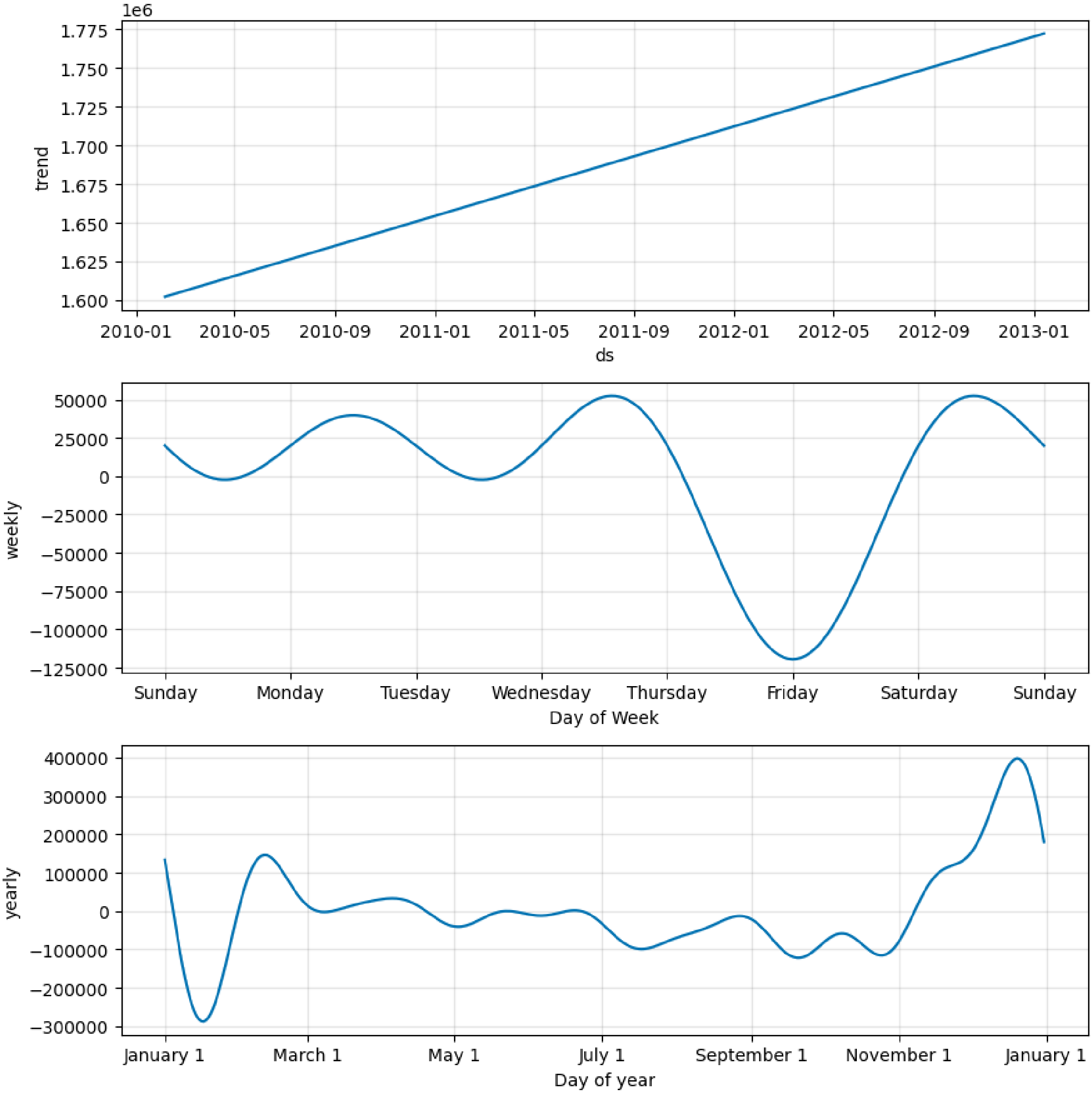
Model Evaluation and Techniques

i) Use predictive modeling techniques to forecast the sales for each store for the next 12 weeks.

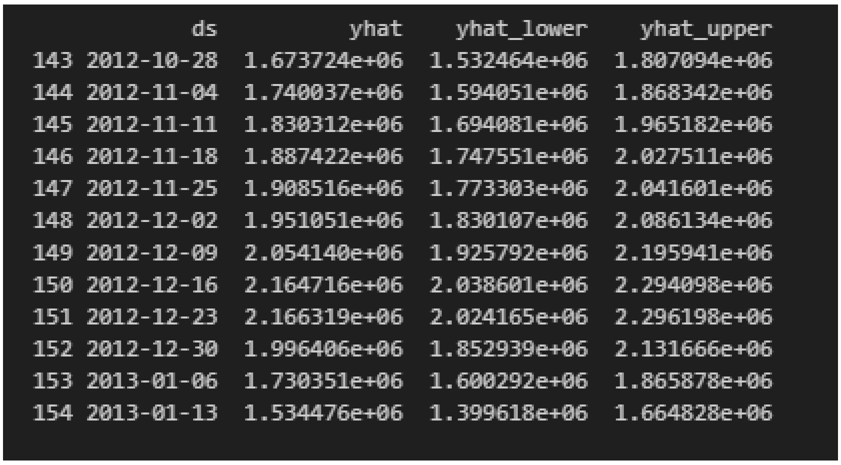
Predicting next 12 weeks for Store 1

To forecast the sales for each store for the next 12 weeks, a common approach is to use a time series forecasting model. Here, I'll demonstrate how to use the **Prophet** model, developed by Facebook, which is particularly well-suited for datasets that exhibit strong seasonal effects





Predict future sales



```
             ds          yhat    yhat_lower    yhat_upper
143 2012-10-28  1.673724e+06  1.532464e+06  1.807094e+06
144 2012-11-04  1.740037e+06  1.594051e+06  1.868342e+06
145 2012-11-11  1.830312e+06  1.694081e+06  1.965182e+06
146 2012-11-18  1.887422e+06  1.747551e+06  2.027511e+06
147 2012-11-25  1.908516e+06  1.773303e+06  2.041601e+06
148 2012-12-02  1.951051e+06  1.830107e+06  2.086134e+06
149 2012-12-09  2.054140e+06  1.925792e+06  2.195941e+06
150 2012-12-16  2.164716e+06  2.038601e+06  2.294098e+06
151 2012-12-23  2.166319e+06  2.024165e+06  2.296198e+06
152 2012-12-30  1.996406e+06  1.852939e+06  2.131666e+06
153 2013-01-06  1.730351e+06  1.600292e+06  1.865878e+06
154 2013-01-13  1.534476e+06  1.399618e+06  1.664828e+06
```