Sales Analysis and Forecasting

An analysis of the sales of Walmart stores was done with the objective to provide an insight on how the stores are performing in respect to sales and also to check the effect of various given factors in the performance of the given stores. Further, a prediction of the performance of the stores was made for the next 12 weeks to get insights on the required inventories.

The dataset provides insights on 8 features for the stores which are described below:

Feature name	Description
Store	Store number
Date	Week of Sales
Weekly_Sales	Sales for the given store in that week
Holiday_Flag	If it is a holiday week
Temperature	Temperature on the day of the sale
Fuel_Price	Cost of the fuel in the region
CPI	Consumer Price Index
Unemployment	Unemployment Rate

Data Pre-processing Steps

Various information of 45 stores spread around 143 weeks are available.

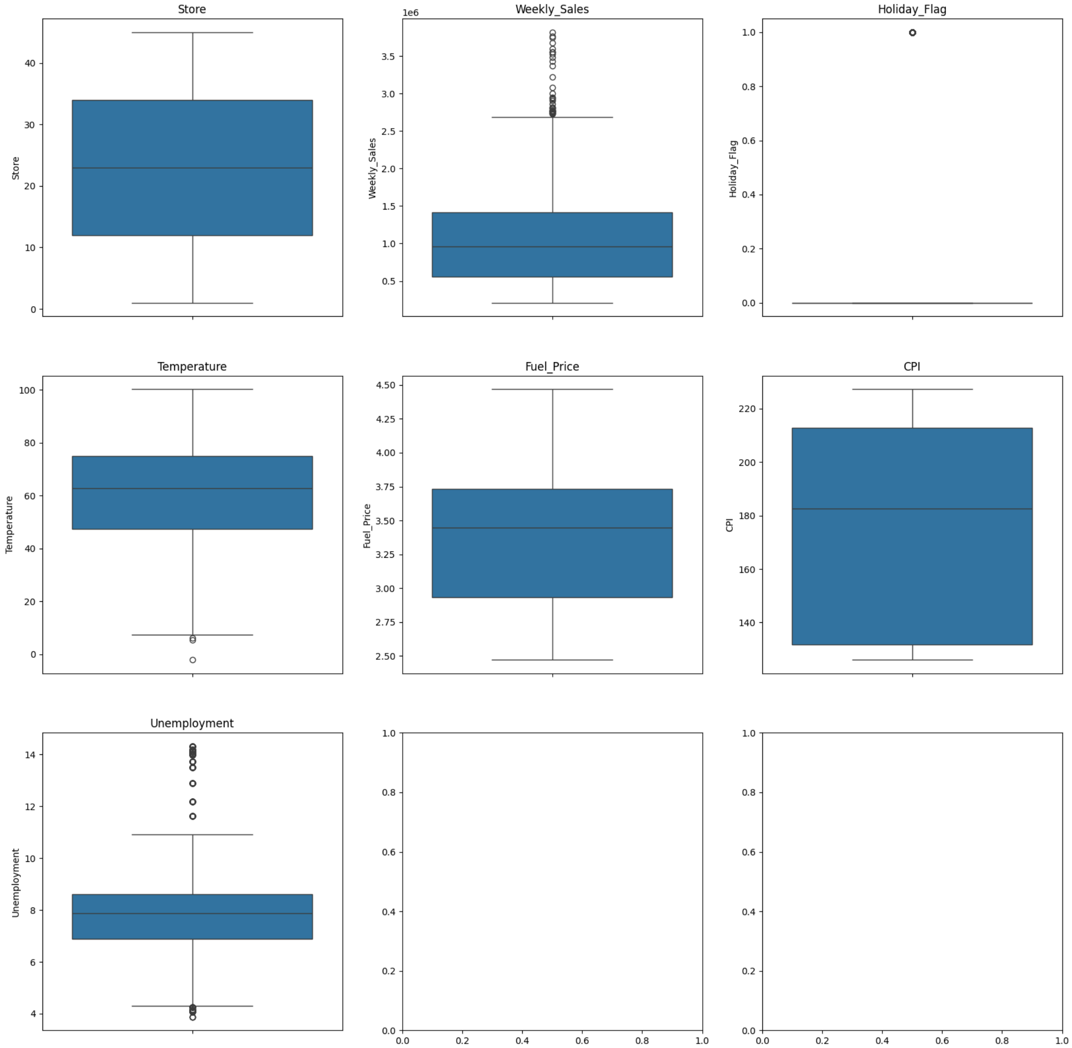
Checking for Data Types: All features have either integer or float value except the date feature which is an object type.

Converting to datetime: The date feature is converted to datetime format for proper use in upcoming steps.

Null and Duplicate values: There are no null values, and no duplicate values are present for any store.

Outlier inspection: After checking for outliers through the IQR method, these columns had outliers

- Weekly Sales
- Unemployment

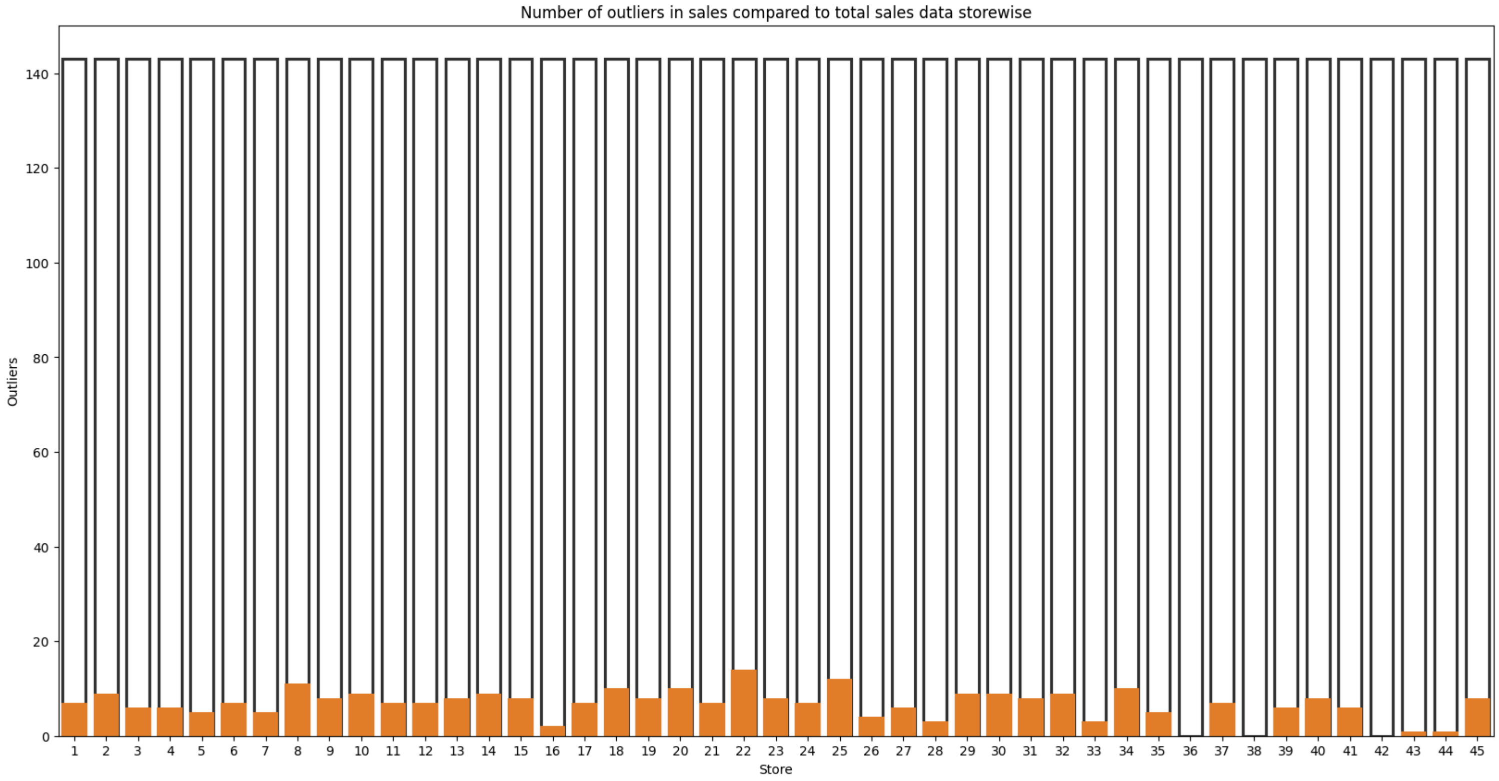


For the weekly sales, there are 34 outliers and the distribution of outliers store wise are:

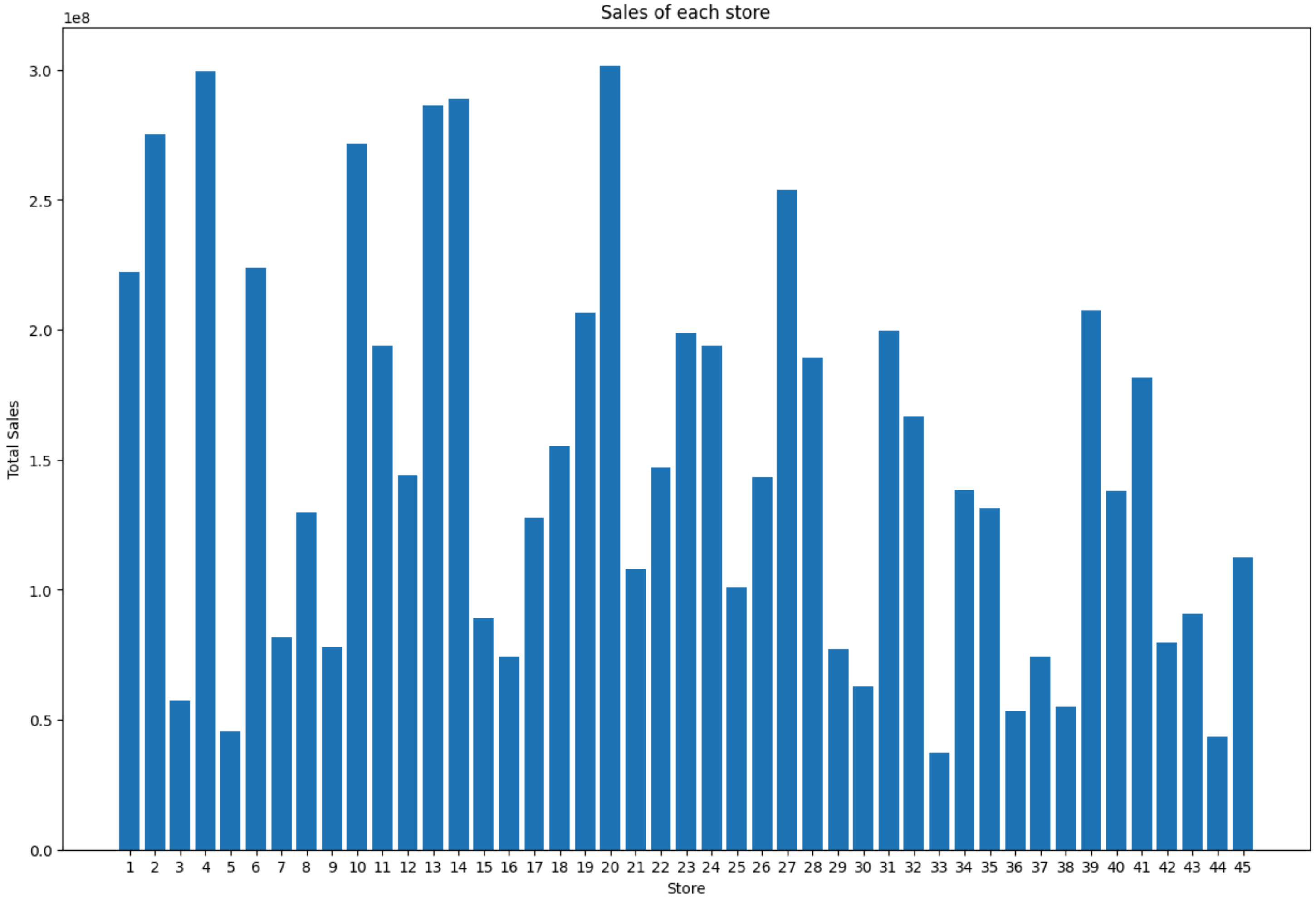
Store no.	No. of outliers
20	7

4	6
13	6
10	5
14	4
2	2
27	2
6	1
23	1

Since the number of data is less per store and the number of outliers are very less compared to the total data. Considering the fact that the outliers are less in number store wise, **no changes are done to the outliers to maintain integrity of the overall data.**



The distribution of total sales for each store is shown below:



The top 5 performing stores and their total sales are given below:

Store no.	**Total sales**
20	301397792
4	299543953
14	288999911
13	286517703
2	275382440

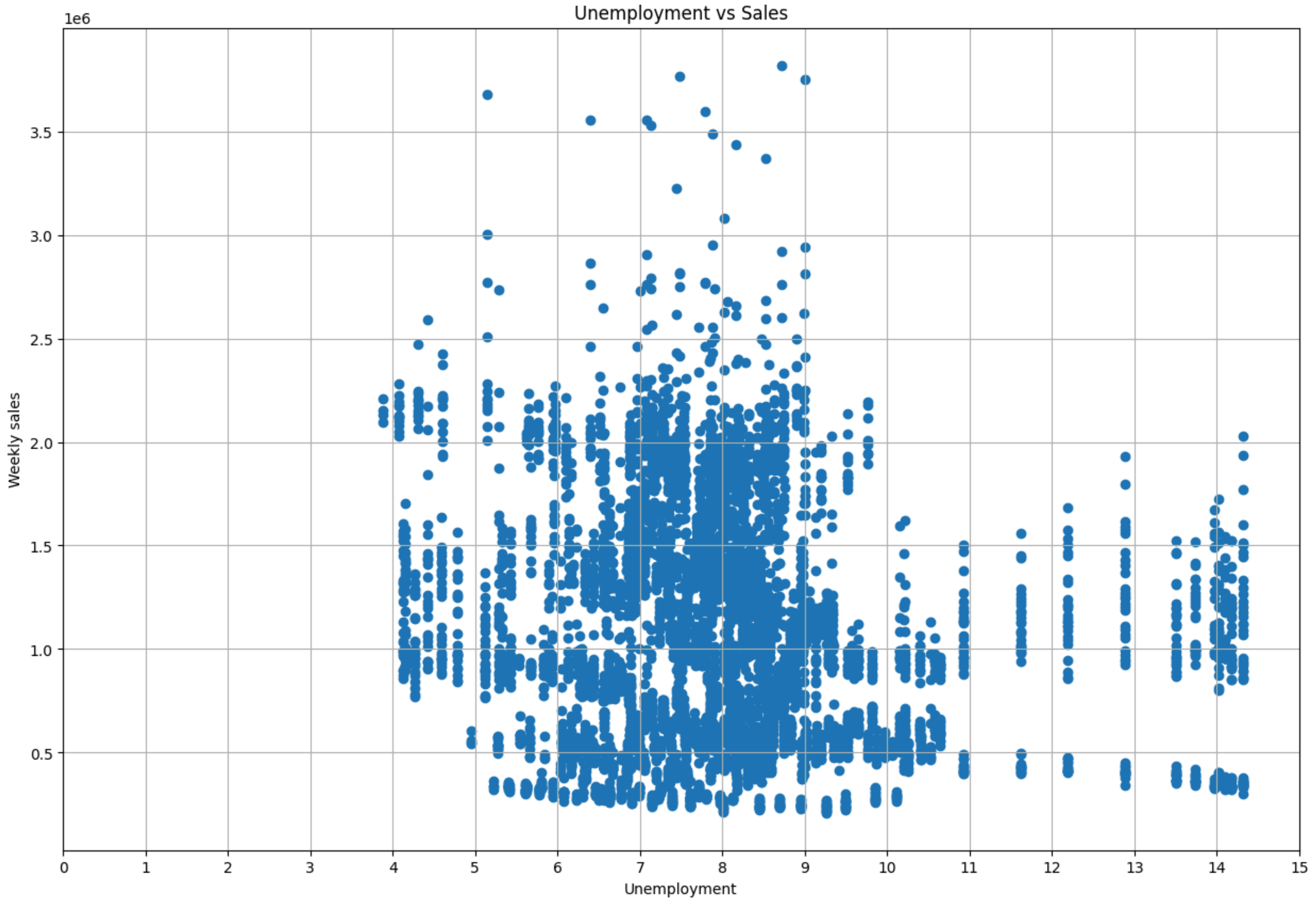
The bottom 5 performing stores and their total sales are given below:

Store no.	Total sales
38	55159626
36	53412214
5	45475688
44	43293087
33	37160221

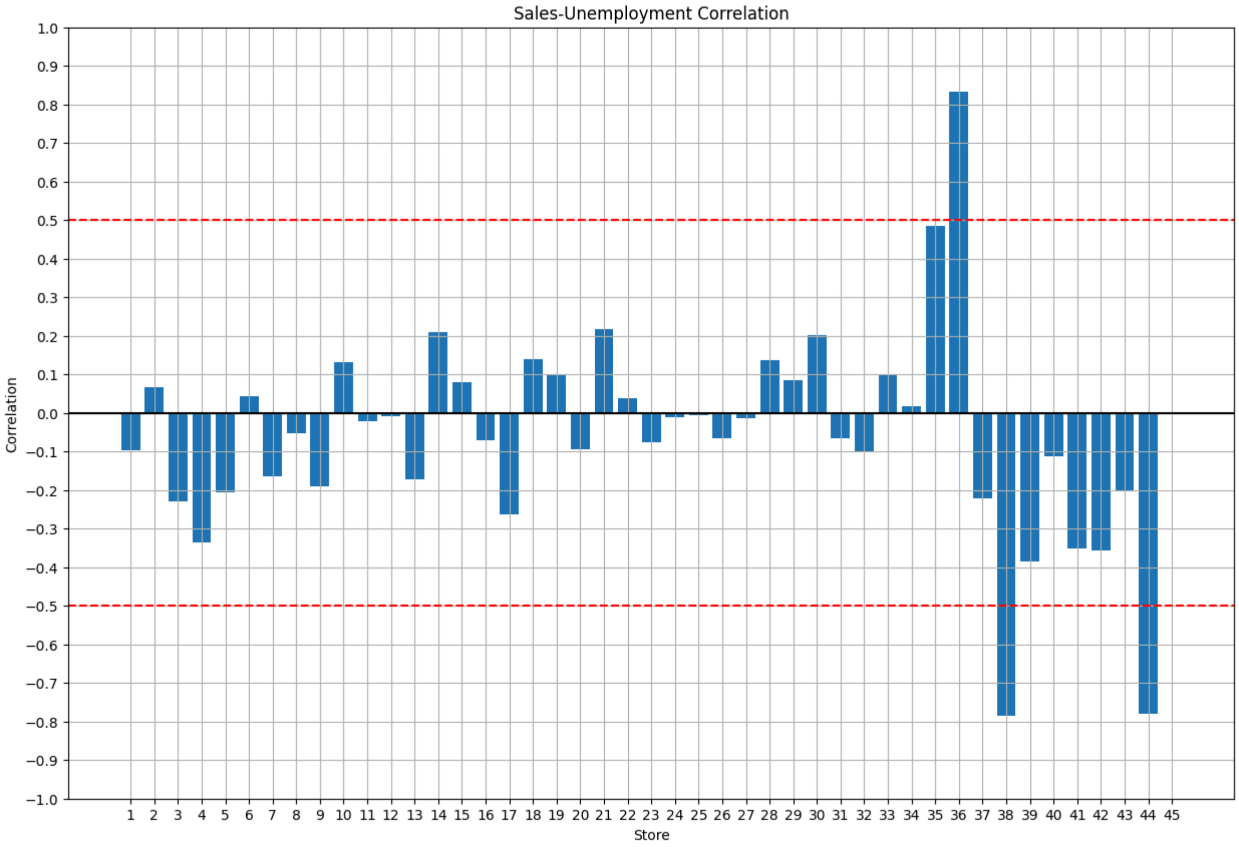
The difference between sales of the highest and lowest performing store is 264237571, which is greatly significant as it lies in the range of 10^8(tens of million).

Insights from analysis

Weekly sales vs Unemployment rate:



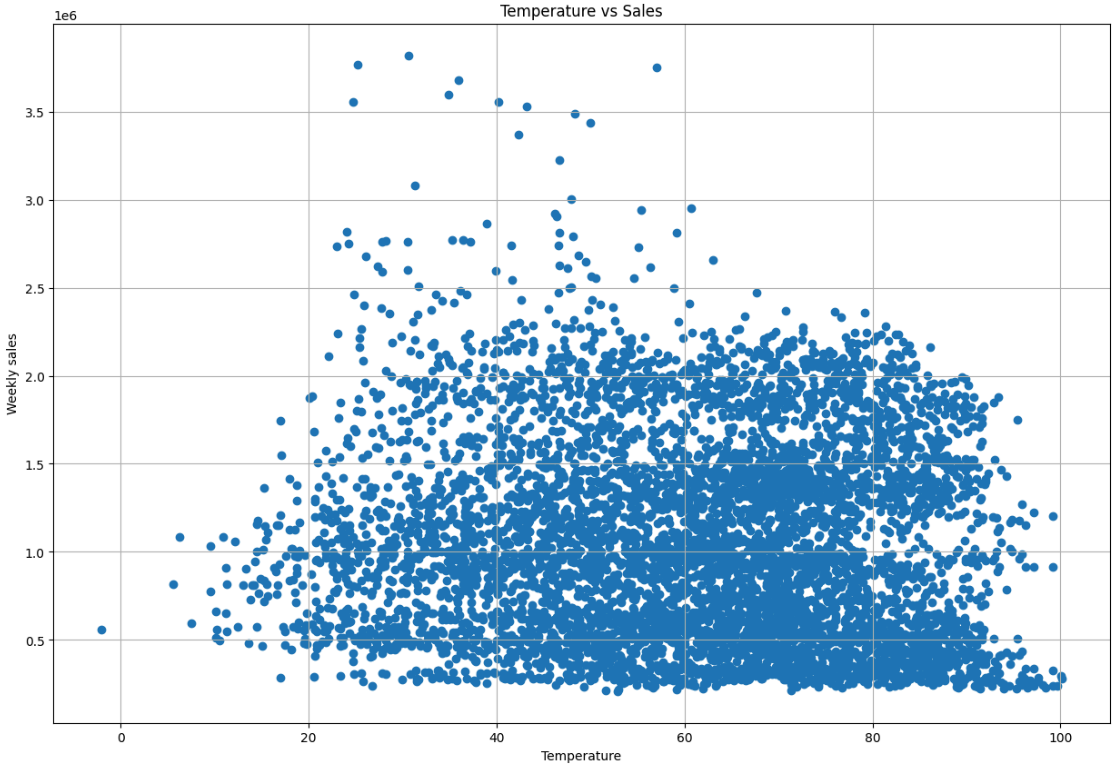
In general, unemployment rate between 6 and 9 has the maximum concentration of sales which is depicted in the below graph.However total sales are higher in the regions with an unemployment rate greater than 9 compared to regions with unemployment rate less than 6, this is probably because of the higher number of stores in that region. Based on the p-Value the correlation between unemployment and sales came out to be statistically significant.



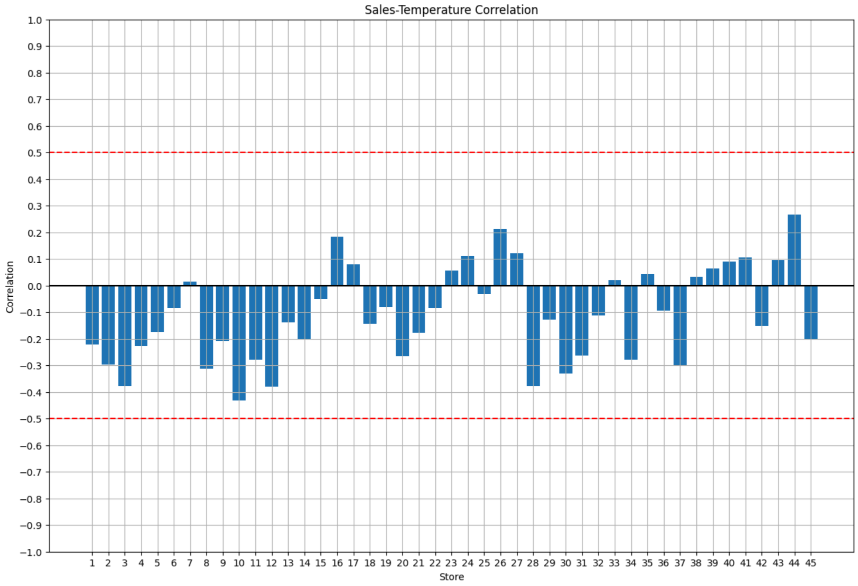
Further, upon examining the correlation between unemployment and weekly sales for each store, Store 36 shows a strong positive correlation of sales and unemployment rate suggesting an increase in sales with increase in unemployment rate. A similar trend is observed for the store 35 but with a lesser impact. Whereas store 38, 44 have a strong negative correlation of sales with unemployment rate suggesting an increase in sales with decrease in unemployment rate.

Weekly sales vs Temperature:

In general, maximum sales are between 40 and 80 degrees whereas for extreme temperatures the sales seem to reduce, suggesting proper sales in a pleasant weather which is indicated by the below plot. Based on the p-Value the correlation between temperature and sales came out to be statistically significant.

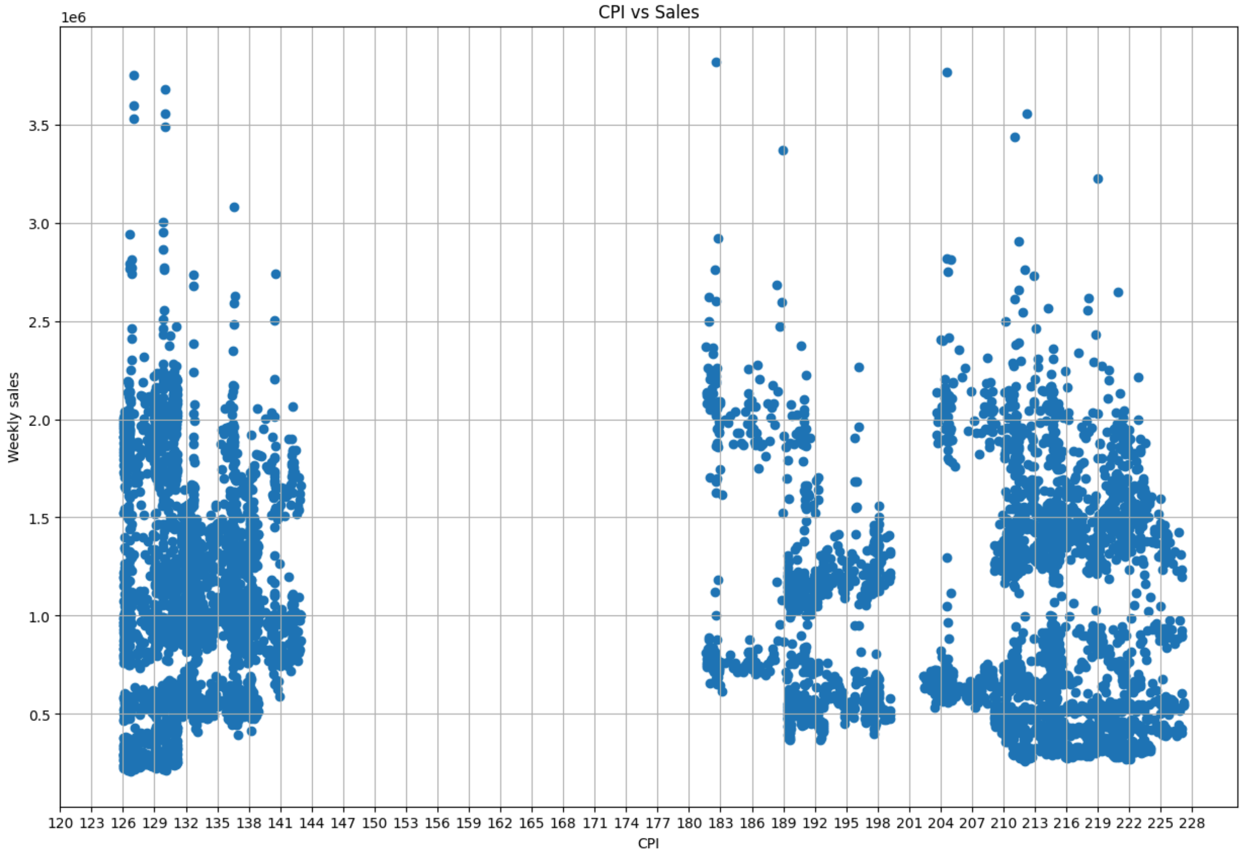


Upon investigating correlation store wise, sales of most stores are negatively affected by the rise in temperature but the effect is not that pronounced as the sales do not change much.

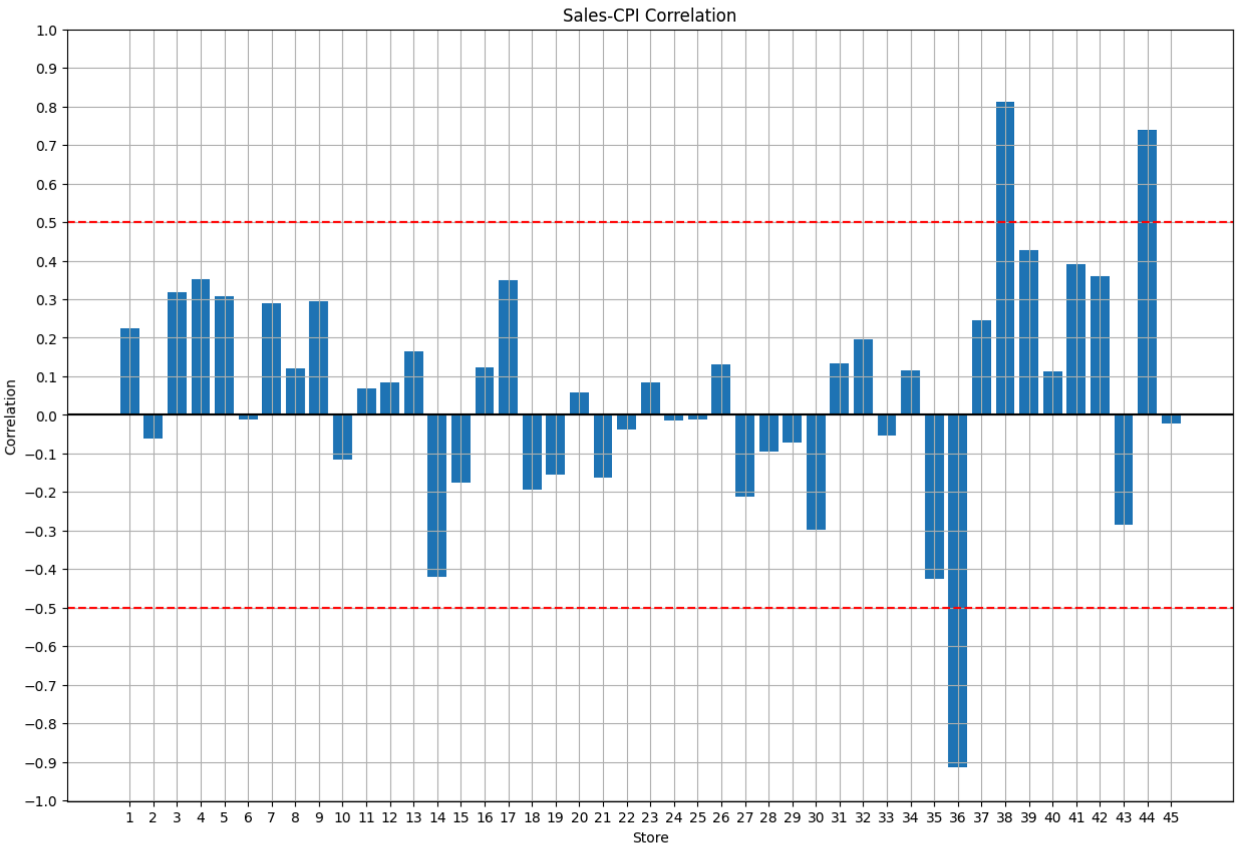


Weekly sales vs Consumer Price Index(CPI):

There are mainly three clusters of CPI with respect to sales, CPI <= 143 makes cluster 1, 181 <= CPI <= 200 makes cluster 2, CPI >= 202 makes cluster 3. Maximum concentration of sales are on cluster 1 & 3 indicating higher sales in low and high CPI regions. Based on the p-Value the correlation between temperature and sales came out to be statistically significant.

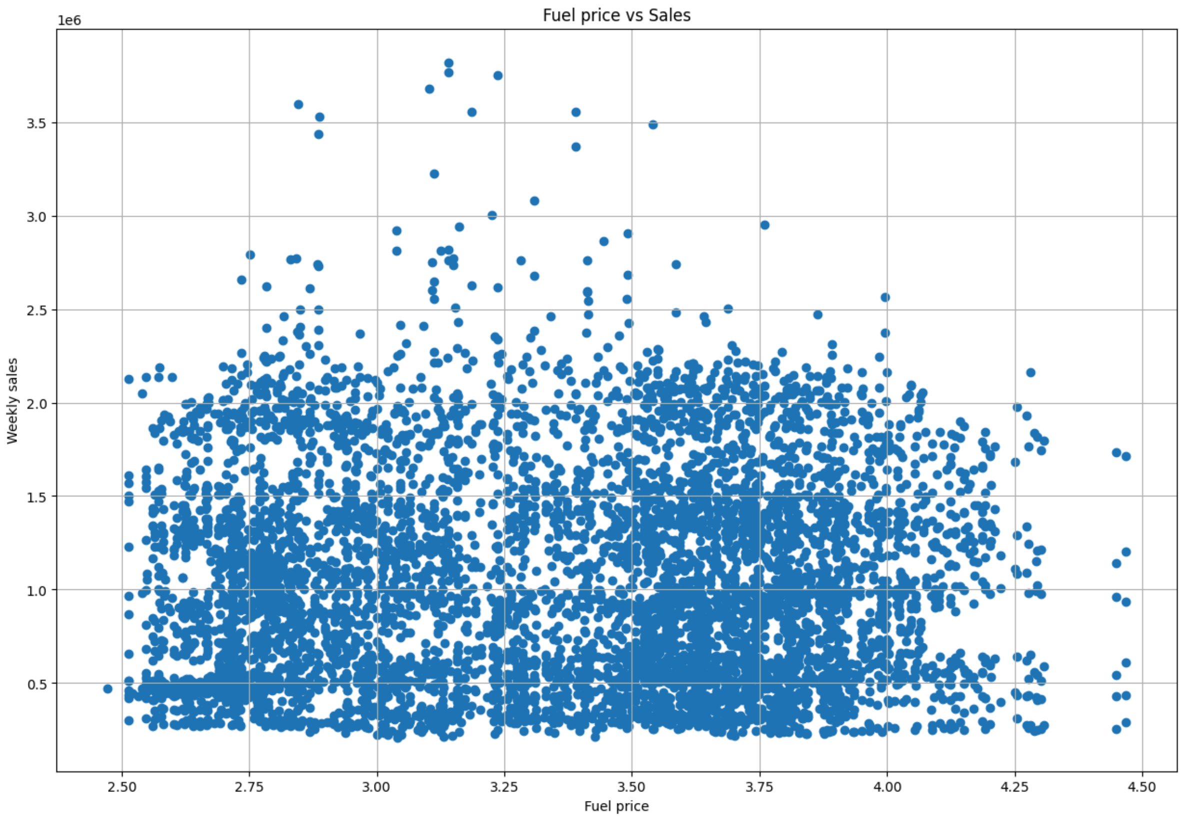


Upon investigating correlation store wise, sales of store 36 have decreased with the increase in the CPI . A similar trend is observed for stores 14, 35, 30, 43 but the impact is not that pronounced. Whereas store 38 and 44 has shown an increase in sales with the increase in CPI, followed by a lot of other stores showing the same trend but less pronounced. Rest stores are moderately affected by the CPI with a higher percentage of them having an increase in sales with increase in CPI.

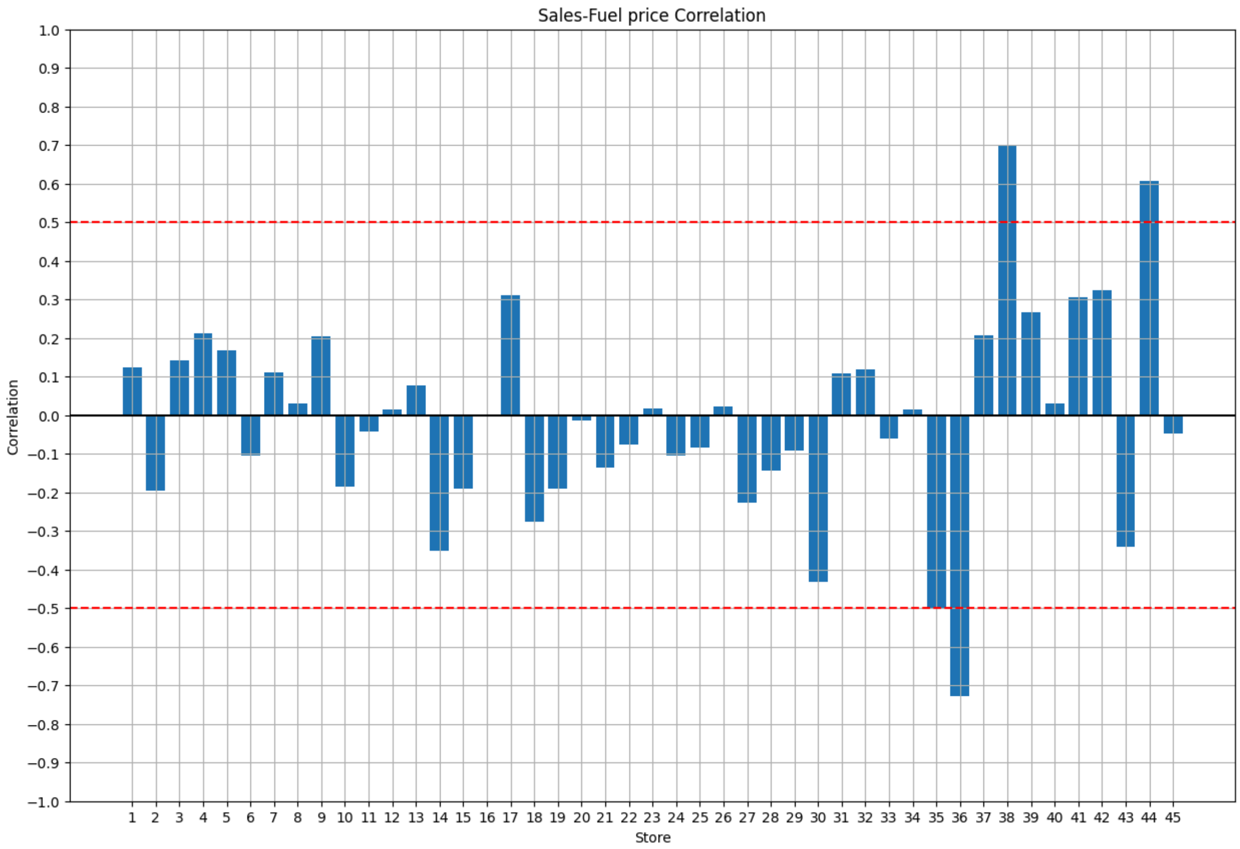


Weekly sales vs Fuel price:

In general, the amount of sales seems to decrease at the higher end of Fuel price which signifies that higher fuel prices causes inconvenience for people to commute to stores easily. Based on the p-Value the correlation between temperature and sales came out to be statistically insignificant.

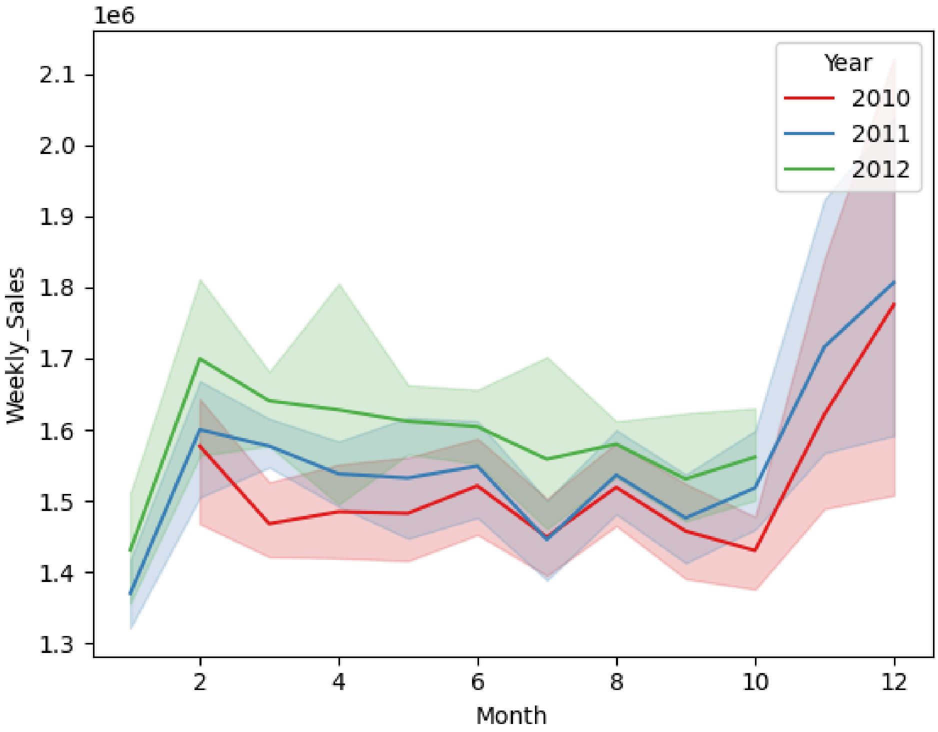


However, upon investigating correlation store wise, sales of store 36 and 35 seems to be most affected negatively with the increase in fuel prices. A similar trend is observed for stores 30, 14, 43 but the impact is not that pronounced. Whereas store 38 and 44 seems to be most affected positively with the increase in fuel prices, followed by stores 40, 41 and 17 showing the same trend but less pronounced. Rest stores are moderately affected by the fuel price with a higher percentage of them having a negative effect with increase in Fuel price.



Seasonal trend in sales:

Upon examining the overall trend in sales for all stores combined, there is a seasonal increase in sales during the last two months of every year.



Further analysis of sales trend store wise concreted the fact that over 90% of stores have a seasonal increase in sales during the last two months of 2010 and 2011 and the reason can be accounted for by the holidays and festivals during those months.

Storewise prediction of future sales

Three algorithms were implemented for the predictions of sales.

- The standard ARIMA
- The SARIMA (Seasonal ARIMA)
- The Prophet model

Predictions from all three models were compared store wise to check the best fit for a particular store.

The **ARIMA model** works well with time series data to predict the future trend and how the data behave by examining the differences between values in the series instead of through actual values. ARIMA model integrates three components:

- **Autoregression (AR):** The AR part refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
- **Integrated (I):** The I part represents the differencing of raw observations to allow the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).
- **Moving average (MA):** The MA part incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

The **SARIMA model** is an extension of the ARIMA model which along with all ARIMA components also includes the Seasonal component of the data to predict the future values. The sales data upon analysis has shown to have a seasonal pattern which makes SARIMA a good model for prediction.

The **Prophet model** is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well. As the data provided also has some outliers, prophet model is a good contender for forecasting.

Parameters evaluation:
The parameters for ARIMA and SARIMA models were evaluated by following methods:

- **Value of d:** Stationary check were done for the data by statistical and visual methods.
  - **Statistical method:** Both ADFuller and KPSS tests were implemented on the time series and depending on respective p-values, the stationary of data was determined.
  - **Visual method:** The rolling mean of time series was plotted to cross check the stationary of data.
- **Value of p and q:** The AutoCorrelation and Partial AutoCorrelation plots for the time series were examined to get an idea on the AutoRegressive(p) and Moving average(q) part of the models. But to find the exact values of p and q, the AIC scores were used and the values with least AIC scores were selected.

- **Value of m:** The value of seasonality(m) was set to 52 as the time series was a weekly data.

The parameters for the Prophet model were auto detected by the function but the interval width was set to 0.95.

Prediction evaluation:

The time series was divided into train and test samples. All the models were trained using the train sample and predictions were done on the test sample.

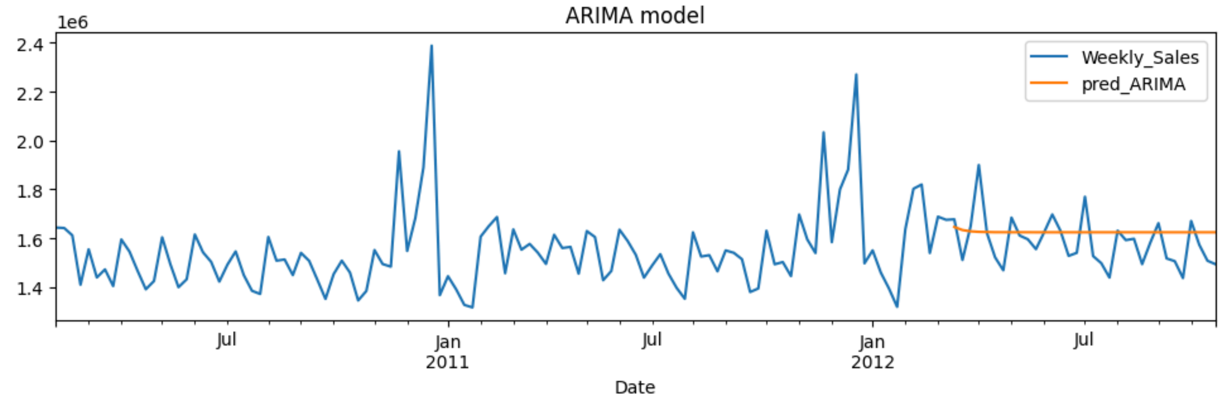
All the predictions of three models were evaluated by plotting the actual test sample with the ones predicted by each model and also by calculating the RMSE values for each model.

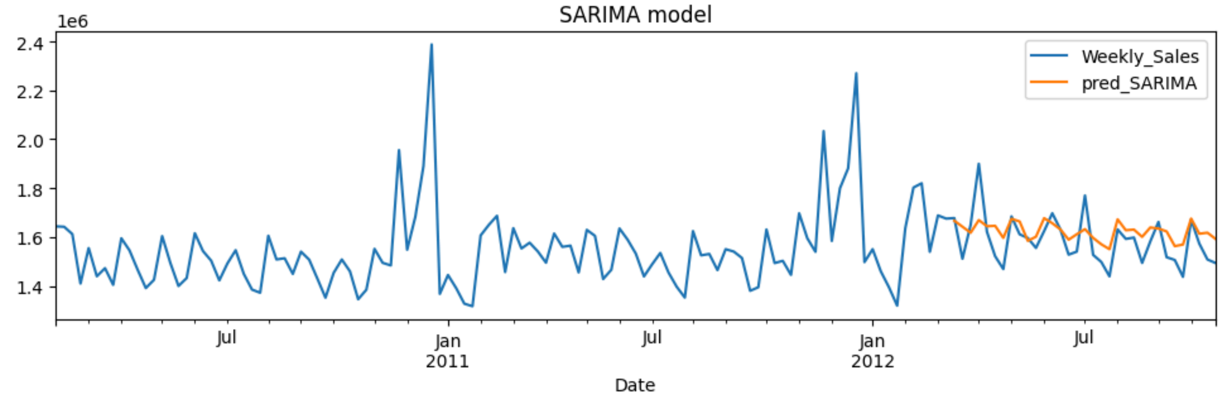
Prediction

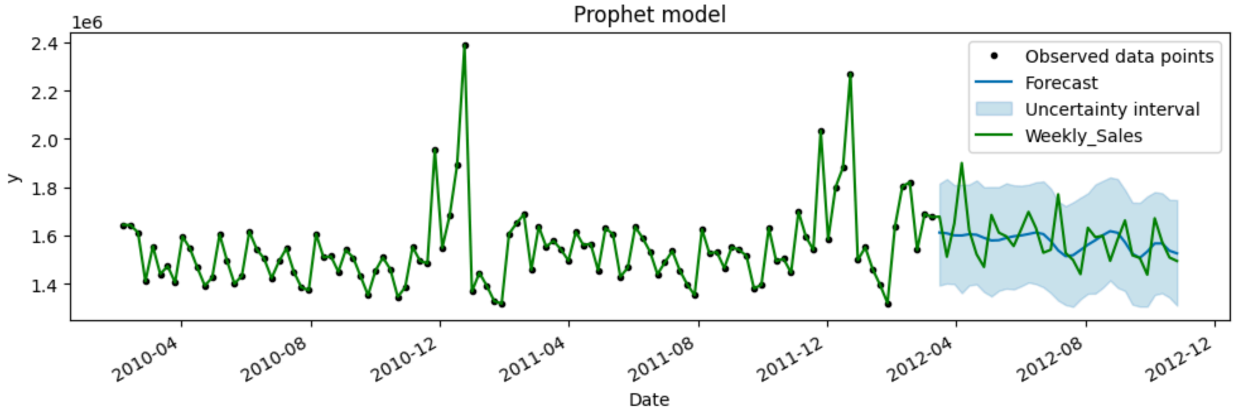
STORE - 1:

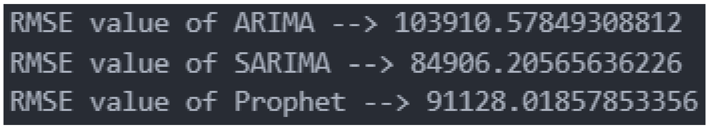
An evaluation for Store 1 was done using the three models and the best model for prediction of sales was the SARIMA model. Values of “d” by inspecting stationary of data was found to be 1.

Here is a comparison of predictions from three models along with their RMSE scores:



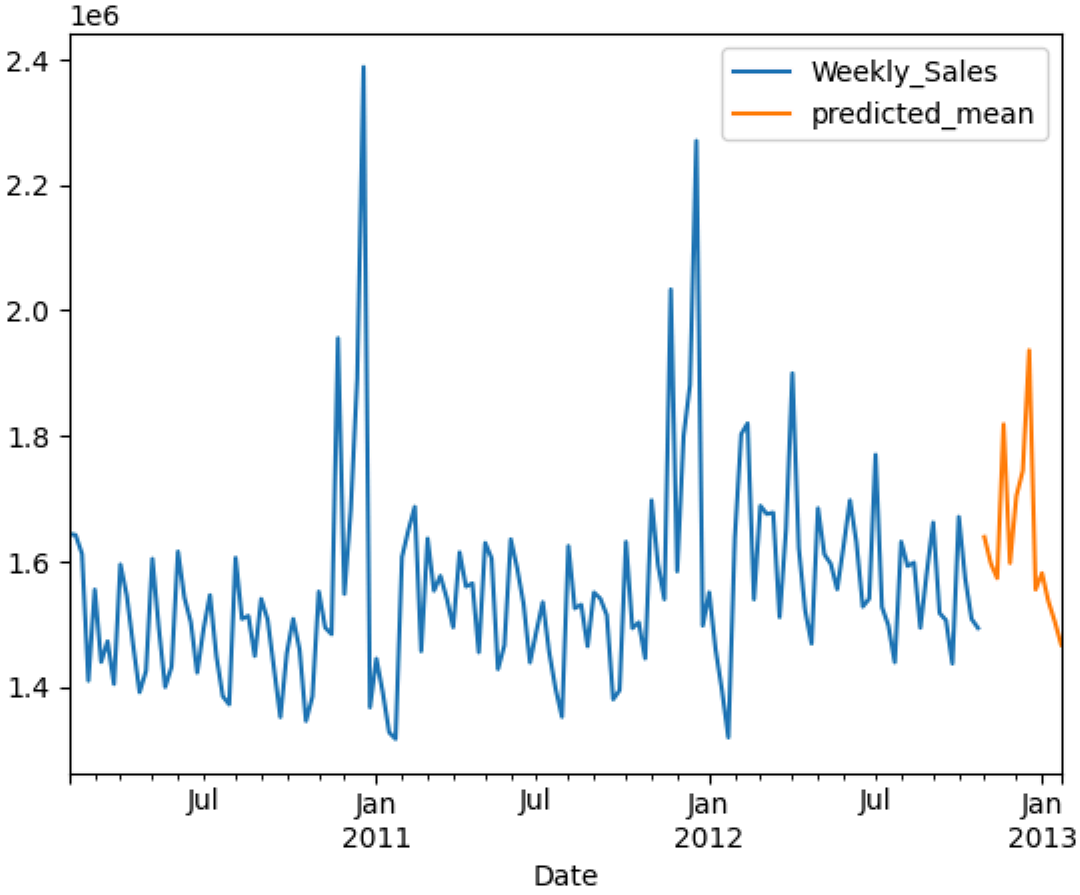






```
RMSE value of ARIMA --> 103910.57849308812
RMSE value of SARIMA --> 84906.20565636226
RMSE value of Prophet --> 91128.01857853356
```

So the SARIMA model was implemented for prediction of sales for next 12 weeks of Store 1.



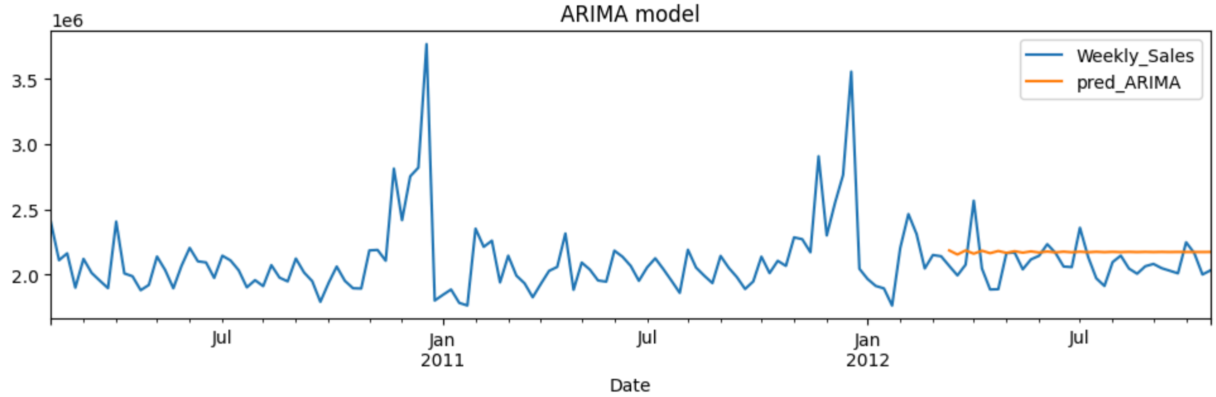
The forecast sales have shown an increase in sales at the year end, which is in accordance with the observed seasonality. So store 1 should stock up its inventory to match the demand. However the predicted sales won't be that high as previous years so a moderate stocking of inventory can be done compared to previous years.

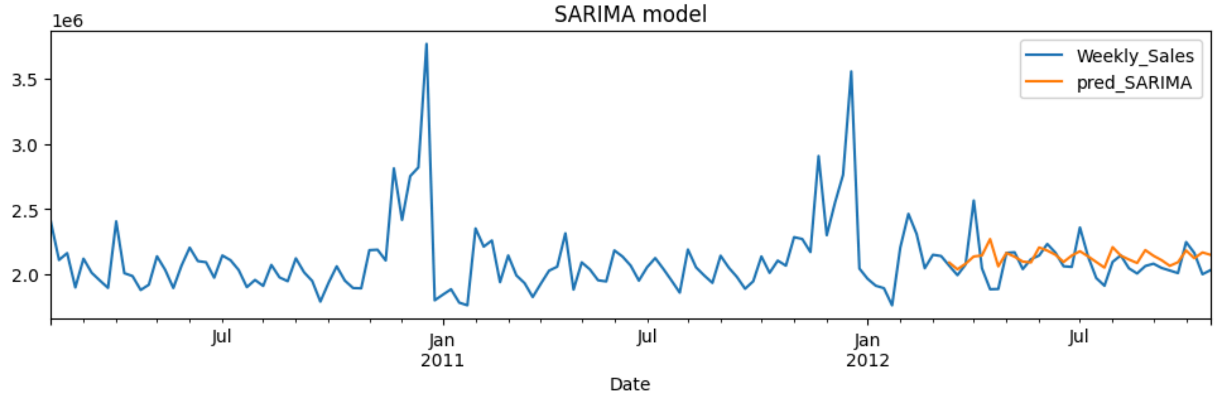
STORE - 20:

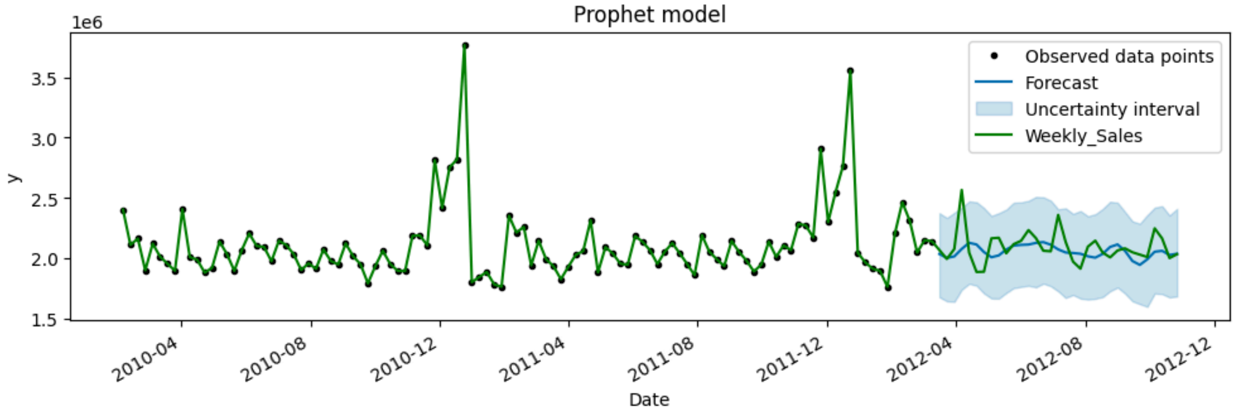
The next store picked up for analysis was Store 20 which has the highest total sales in the recorded period of time.

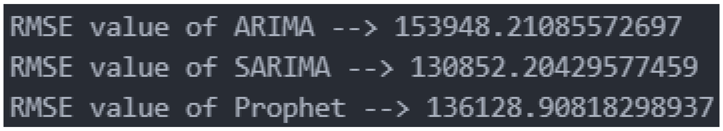
Similar to before, three models were evaluated to pick the best fitting one and in this case also the SARIMA model gave the best predictions. Values of “d” by inspecting stationary of data was found to be 1.

Here is a comparison of predictions from three models along with their RMSE scores:



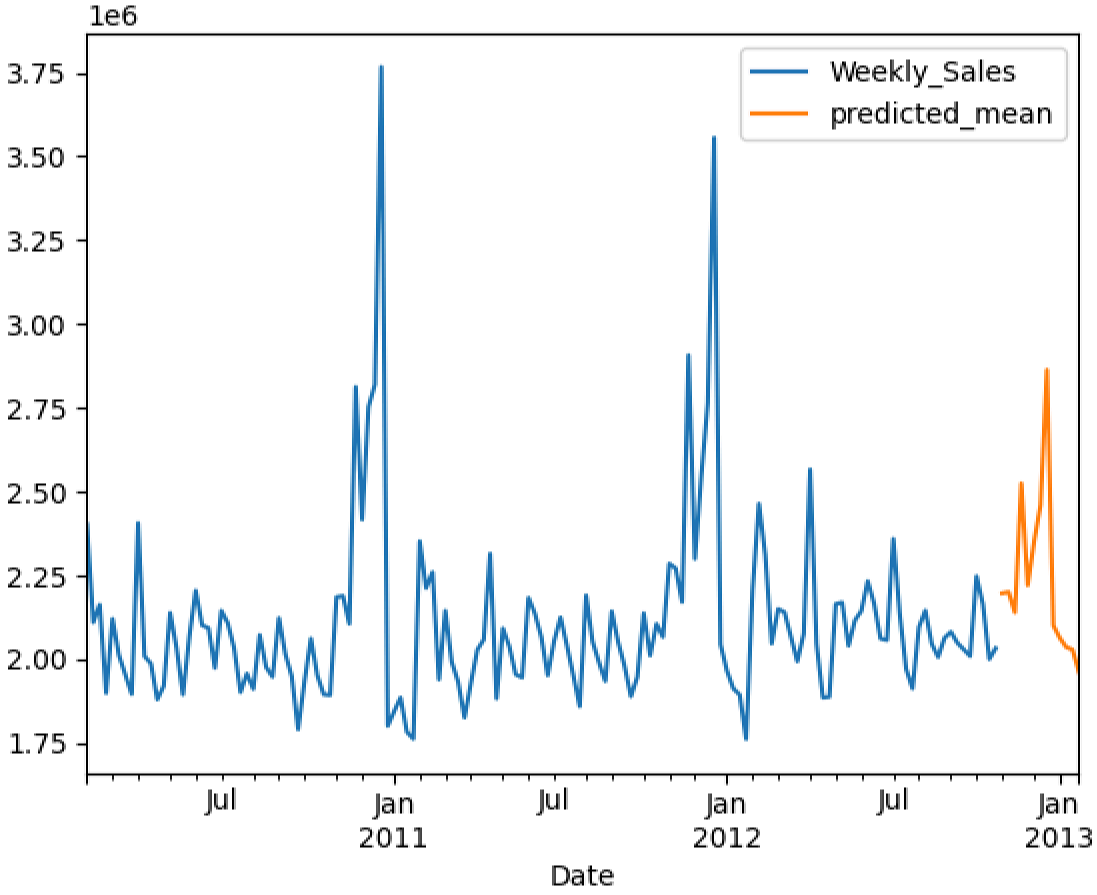






```
RMSE value of ARIMA --> 153948.21085572697
RMSE value of SARIMA --> 130852.20429577459
RMSE value of Prophet --> 136128.90818298937
```

So the SARIMA model was implemented for prediction of sales for next 12 weeks of Store 20.



The forecast sales have shown an increase in sales at the year end, which is in accordance with the observed seasonality. So store 20 should stock up its inventory to match the demand. However the predicted sales won't be that high as previous years so a moderate stocking of inventory can be done compared to previous years.

Future scope

The project can be expanded by including factors like holidays and temperatures while predicting so that more accurate sales are available for future, giving an insight on the instantaneous requirement of inventories and a focus on how to improve the low performing stores.

Also on a larger scale, the location for opening new stores can also be optimised by considering factors like Unemployment, Temperature, CPI and Fuel price that affect sales.