CAPSTONE PROJECT- WALMART ANAYSIS AND SALES PREDICTION.

Table of Contents

PROBLEM STATEMENT

A retail store that has multiple outlets (45) across the country are facing issues in managing the inventory - to match the demand with respect to supply.

This mis-match of demand is creating hap-hazard in management and not letting company achieve the desired output. Non-linear supply, irregular demands is acting as bottleneck for the development of the store.

PROBLEM OBJECTIVE

The main objective of this project is to help the retail store with detailed EDA, predictions , statistical analysis so that the data can act as life-saviour for the firm.

We are provided with weekly sales data for various outlets, as a data scientist or Analyst we need to do various statistical analysis, EDA etc to come up with various insights that can help the retail store manage the mismanagement and help company in Business Oriented Decision and create a separate roadmap for separate stores.

We must perform every possible thing to get best insight of the data.

We also need to predict the next 12-week sales for the store so that they can manage the inventory accordingly.

DATA DESCRIPTION

The table contains 6435 rows and 8 columns.

The columns are: -

1. Store: - This column contains store number from (1-45) as there are 45 stores in the dataset.
2. Date: -This column contains dates on the weekly basis (Friday to Friday) starting from 5th February 2010 to 26th October 2012.
3. Weekly sales: -This column Contains the total sales for the said week.
4. Holiday flag: - This column tells if that entire week was a country holiday or not.
5. Temperature: - This column talks about the Average temperature for the entire week.

6. Fuel Price: -This column states the Average fuel price for the entire week.
7. CPI: - It tells us about the Consumer Price Index (Spending capacity of peoples)
8. Unemployment.: - This column tells Unemployment ratio (Number of people employed) for the entire week.

DATA INSIGHTS

- The data contains overall 6435 rows and 8 columns.
- Neither there are any null values in any columns, nor the dataset contains any duplicates.
- All the columns of the data are of int or float datatype except date that is of object datatype.
- There are overall 45 stores in the dataset.

```
#number of stores
data['Store'].nunique()
```

```
45
```

- For all store weekly sales of 143 weeks are given.

```
# number of weeks for which data is present.
data['Date'].nunique()
```

```
143
```

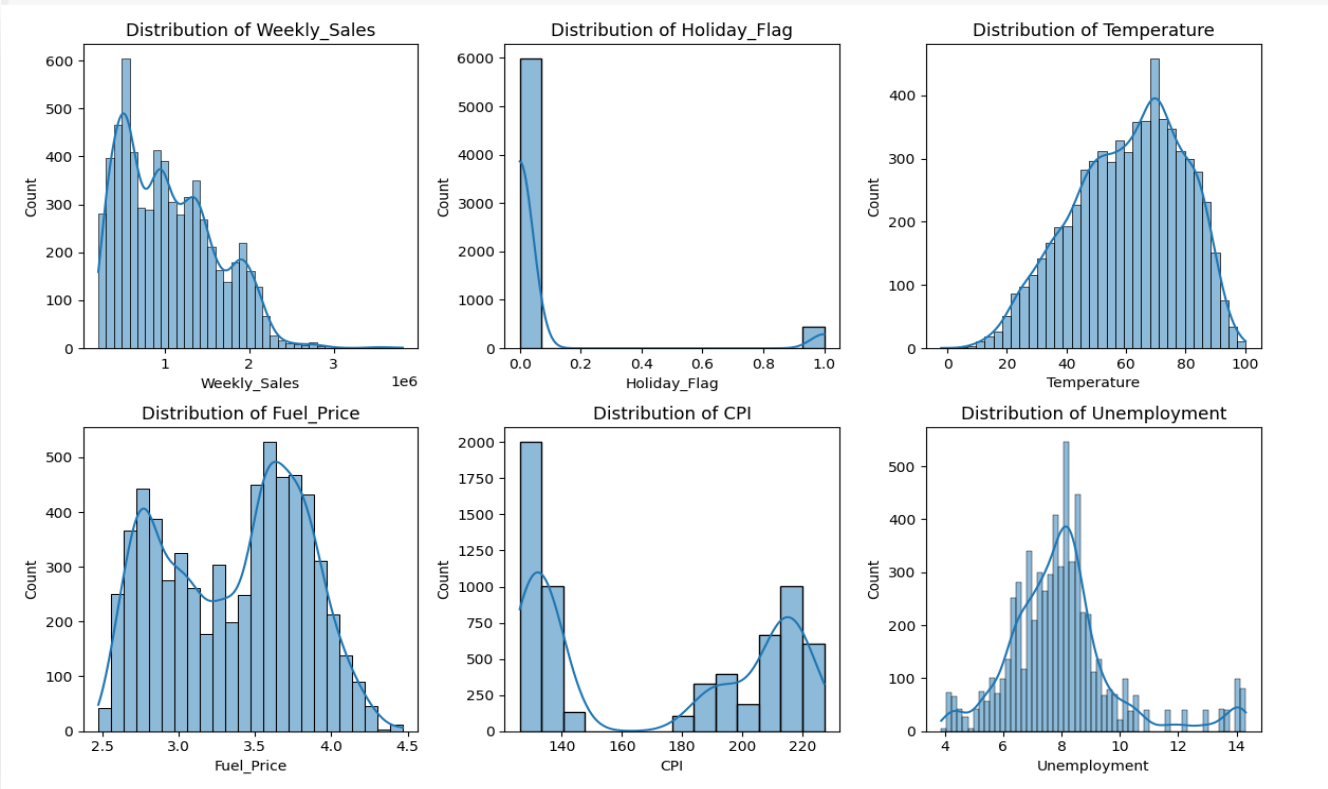
- We have weekly sales data of 33 months(approx.)

```
#number of months for which the data is present
data['Date'].apply(lambda x :x.split('-')[1]+'-'+x.split('-')[2]).nunique()
```

```
33
```

OBSERVATION FOR DISTRIBUTION OF VARIOUS NUMERICAL COLUMNS OF THE DATA

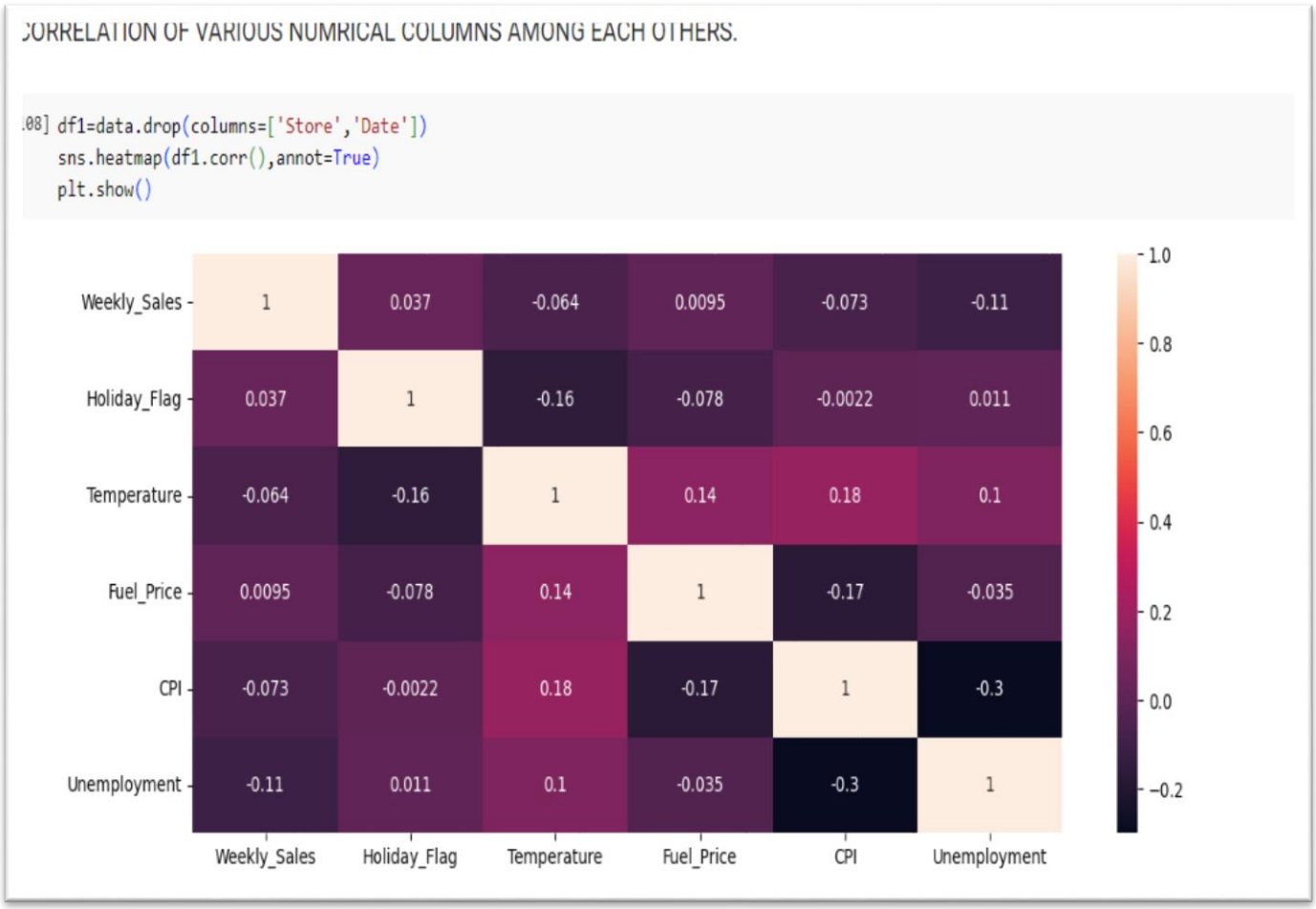
- Weekly sales is right distributed.
- Our data is not balanced with references to holiday or not , most of the data is for no holiday week and data for holiday week is very less as comparison to no holiday week.
- Temperature is left skewed and most the temperature varies from 60-75 degrees.
- Fuel price is also not normally distributed.
- CPI ranges from60-150 and 175-230 only.
- Unemployment is right skewed and mostly the unemployment index ranges from 7-9.3



ANALYSIS THROUGH VISUALIZATIONS

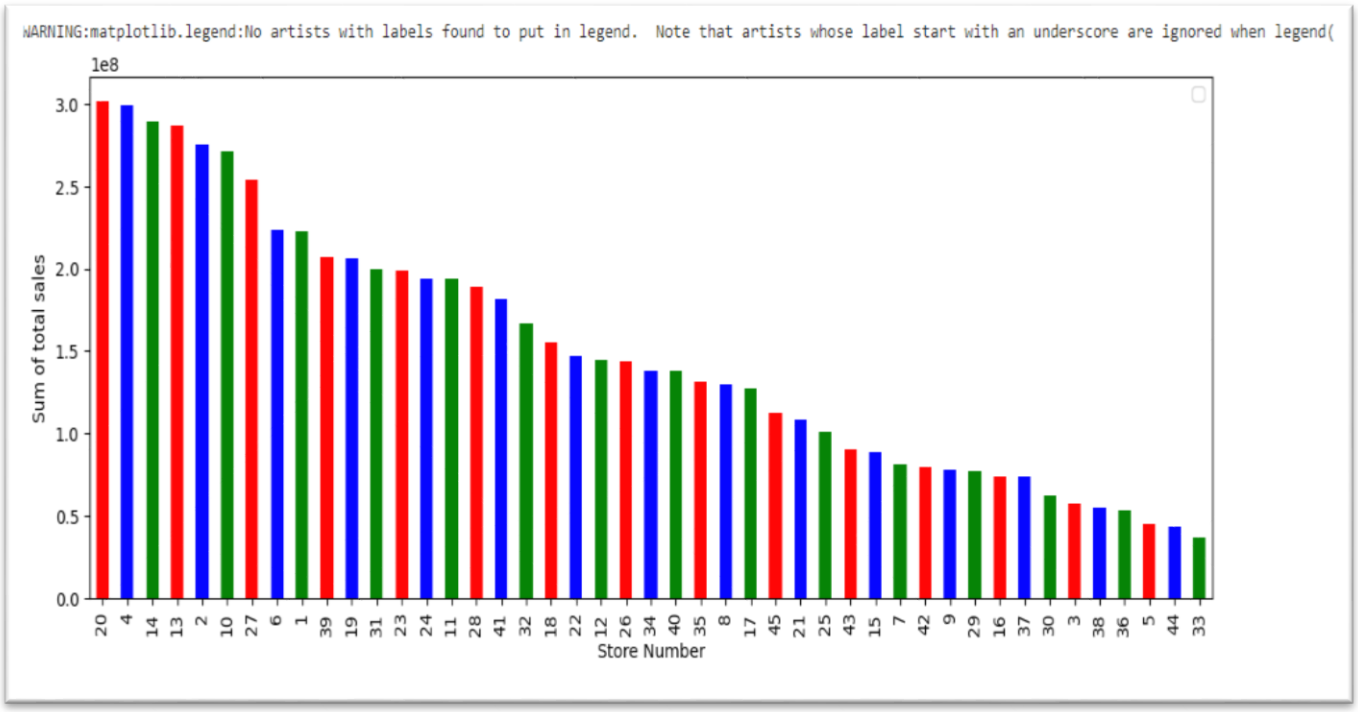
CORRELATION OF WEEKLY_SALES TO ALL OTHER NUMERICALS COLUMNS.

Weekly sales shows a very weak positive correlation with weekly sales and fuel prices, and shows a very weak negative correlation with Temperature, CPI and Unemployment.



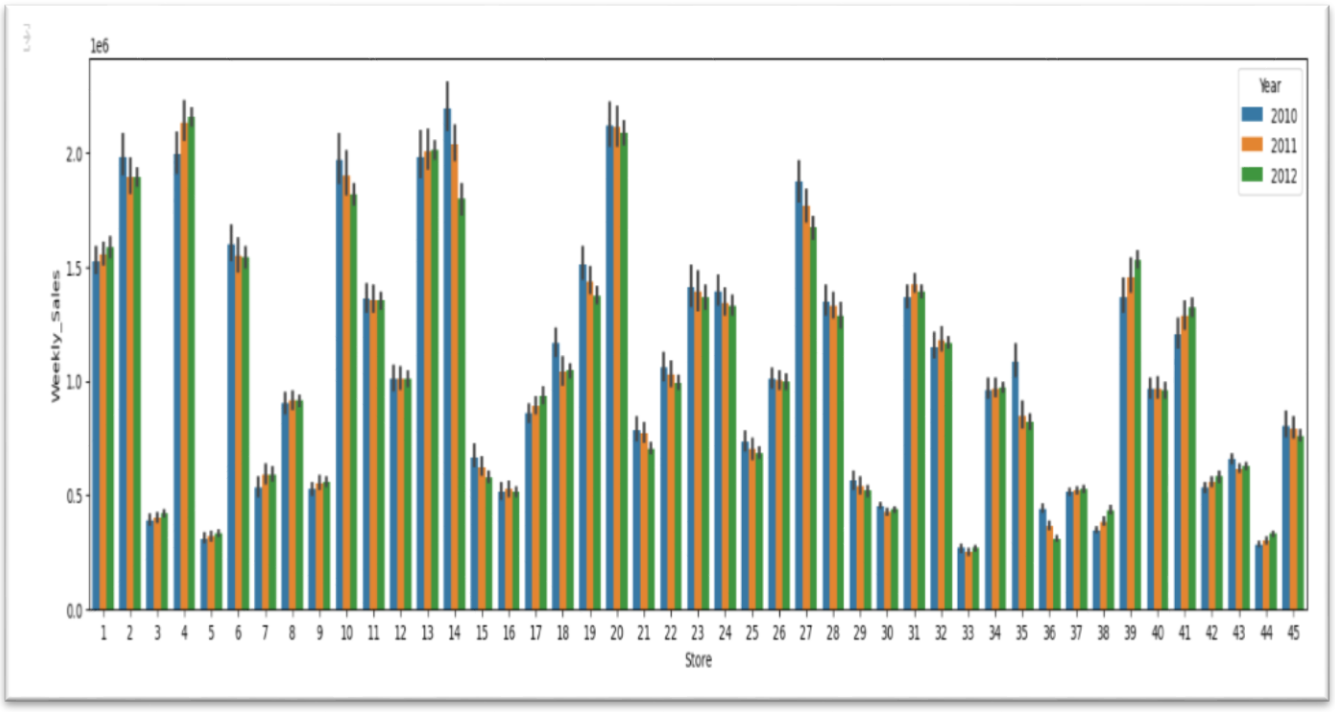
ENTIRE TOTAL OF SALES PER STORE

1. Maximum total sales is for store number 20, followed by 4 and 14.
2. Lowest sale is for the store number 33 followed by 44 and 5
3. Difference between the maximum overall sales and minimum overall sales is 264237570.50.



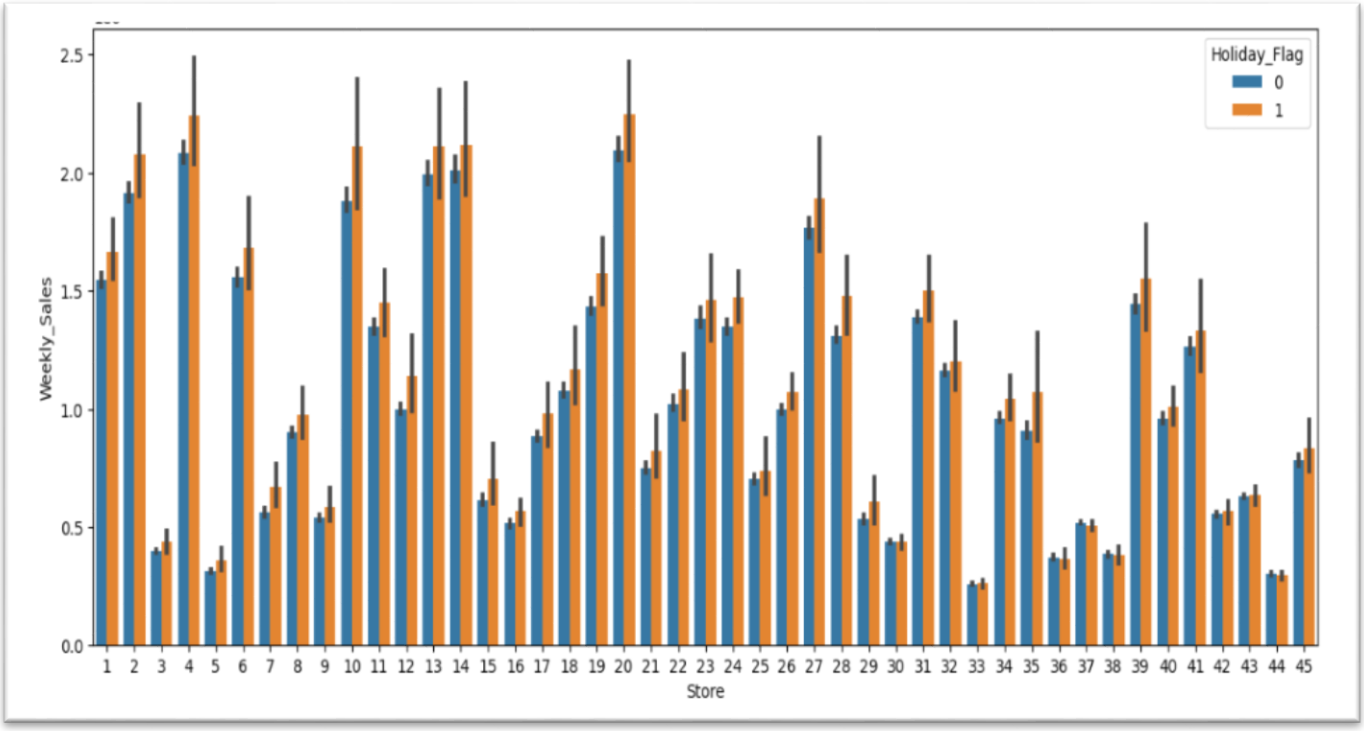
AVERAGE SALES PER STORE , YEAR WISE.

- Sales increased continuously for year (2010,2011,2012) For the store numbers: - 1,3,4,5,7,8,9,13,17,38,39,41,42,44.
- Sales decreased continuously for year (2010,2011,2012) for the store numbers: 2,6,10,14,15,19,20,21,22,23,24,25,26,27,28,29,35,36,45
- Weekly sales fluctuated in non-continuous manner or remain stable for the year(2010,2011,2012) for store numbers: - 11,12,16,18,30,31,32,33,34,43.



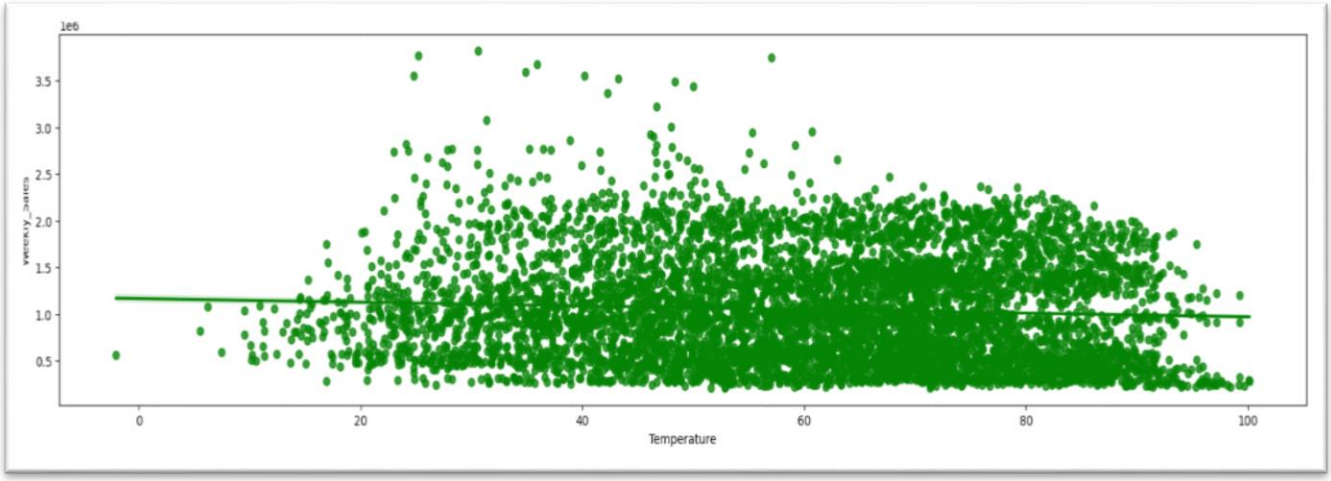
AVERAGE SALES OF STORES, WITH RESPECT TO HOLIDAY OR NOT

In all the stores, average sales on holiday week is more than that average sales on non-holiday weeks except for store number 36 ,37,38 and 45 where average holiday sales are slightly less than non-holiday sales and for store number 30 both the sales (holiday and non-holiday) are almost equal.



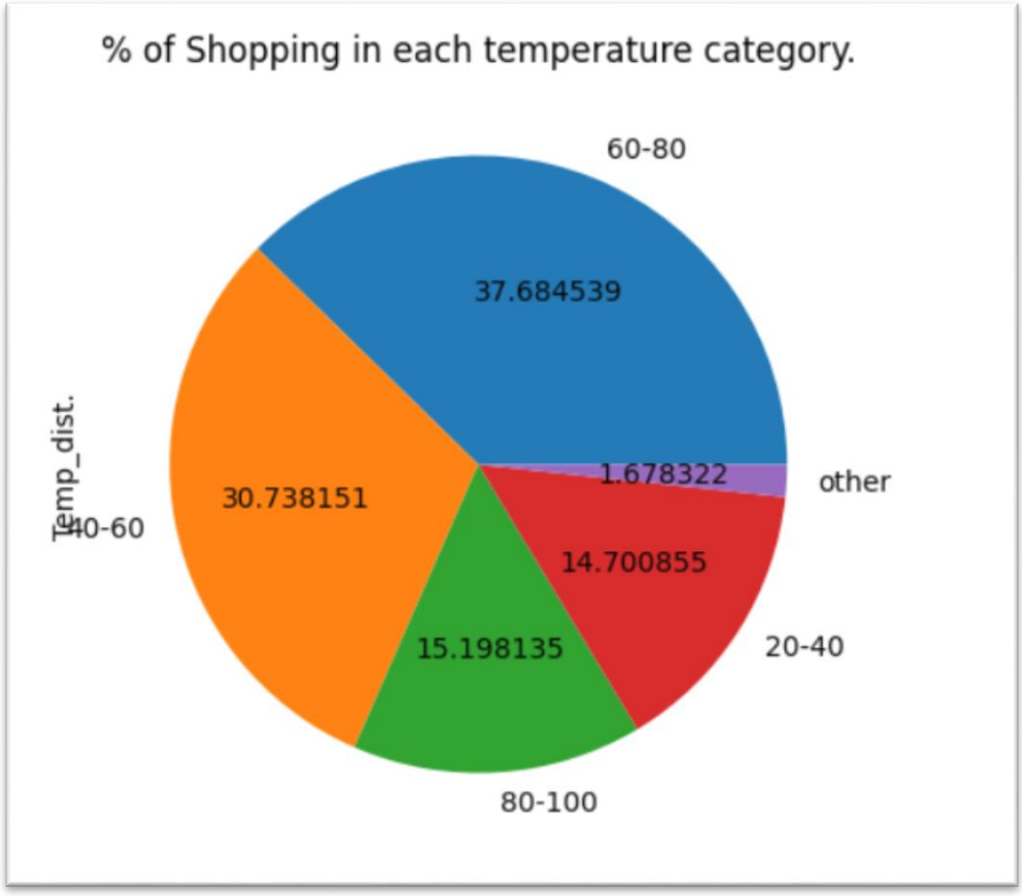
EFFECT OF TEMPERATURE ON WEEKLY SALES

With increase in temperature there is no drastic decrease, however there is little decline in sales with increase in temperature.



FOOTFALL(SHOPPING) FOR EACH GROUP AFTER DIVIDING TEMPERATURE INTO GROUP OF 20-20

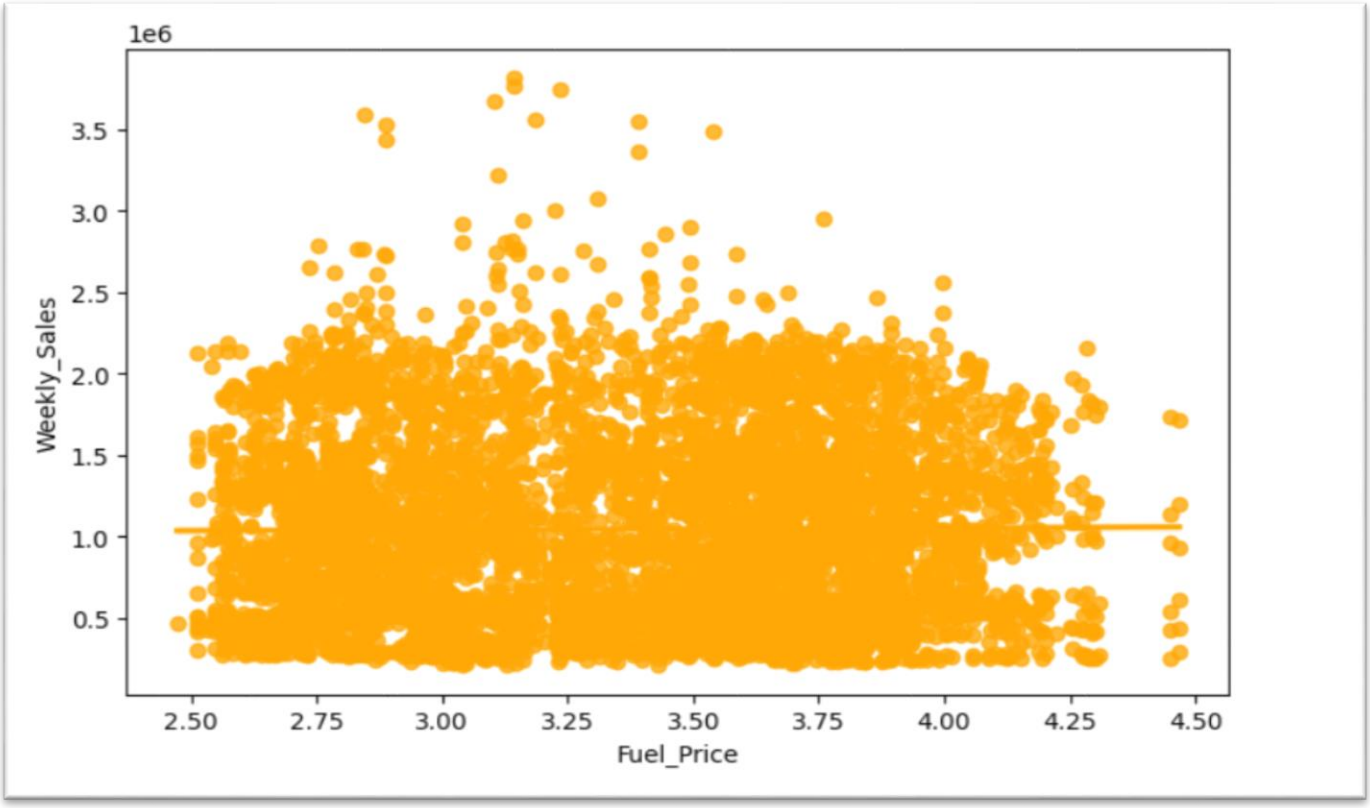
Maximum shopping was done when the temperature was between 60-80(37.68%), followed by when temperature was in range 40-60(30.73%).



EFFECT OF FUEL PRICES ON WEEKLY SALES.

With increase in fuel prices, it is generally observed that there is decrease in sales, as prices of commodities increases. However here contrary is happening, with the increase in fuel prices there is a little increase in weekly sales.

There is no such drastic effect of fuel prices on weekly sales however with the increase in fuel prices, weekly sales increases a little bit.



EFFECT OF FUEL PRICE ON WEEKLY SALES FOR EACH STORE.

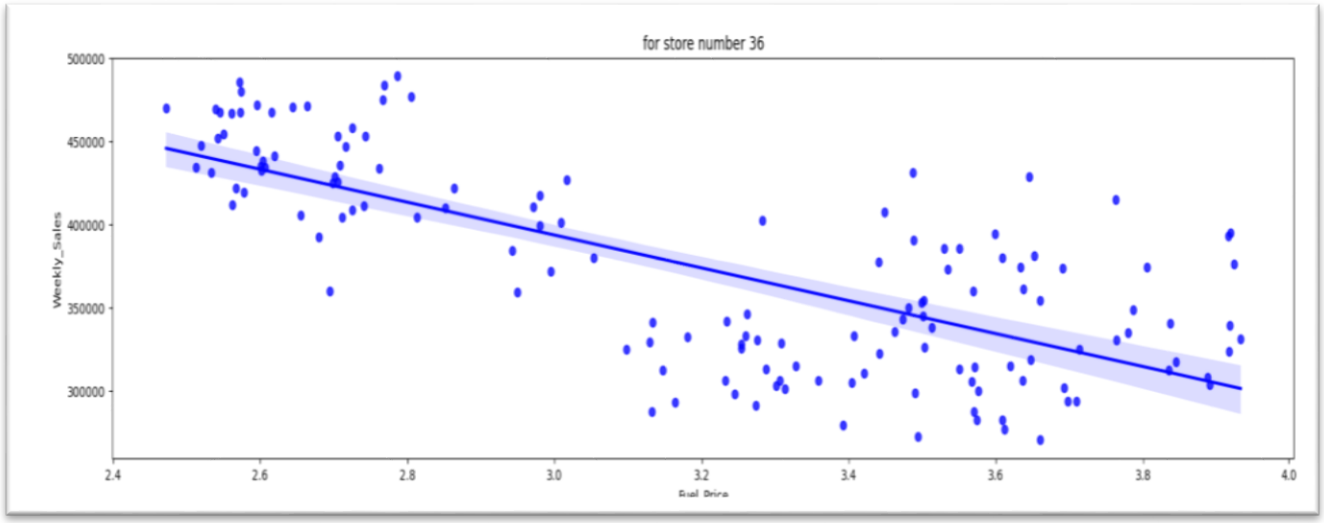
With increase in fuel prices weekly sales is increasing with a minor extent.

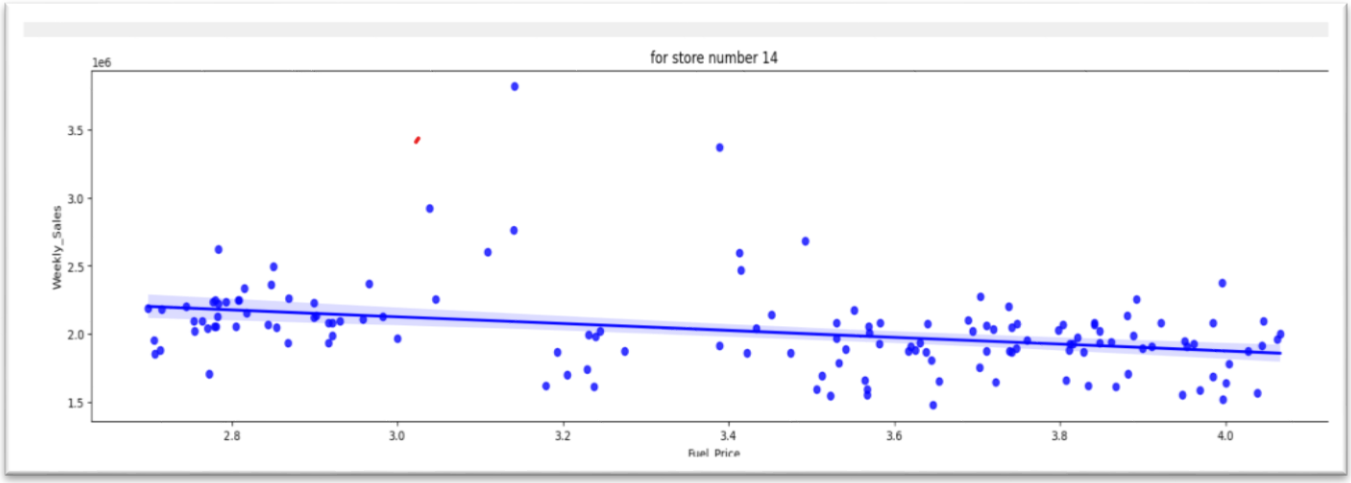
However, there are some stores which is not following the trend that Is being followed by the overall dataset.

Store with number 2,6,5,18,19,20,21,24,25,28,43 shows a trend that is opposite to the general trend of the entire dataset. For these stores the weekly sales in decreasing with increase in fuel prices but with minor effect.

However Store with store number 14,30,35,36 shows drastic decrease in weekly sales with increase in fuel prices.

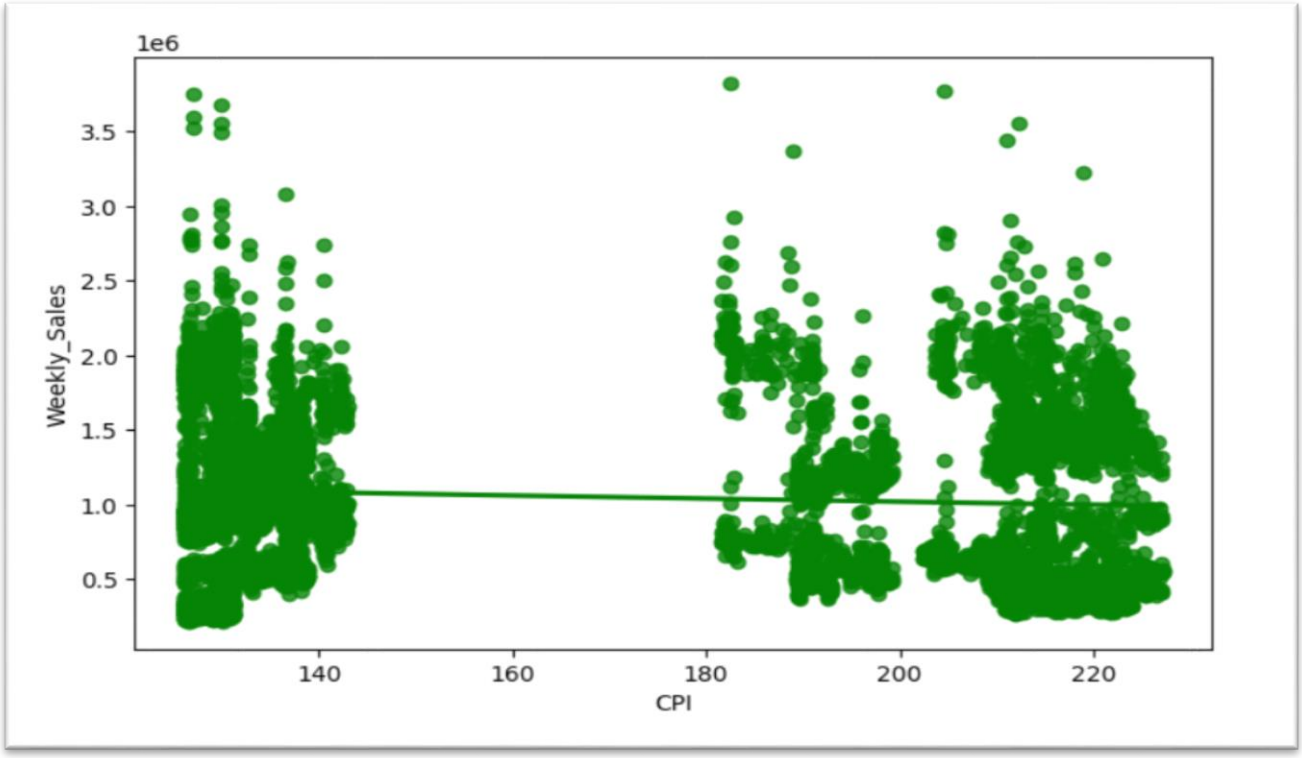
Below are some graphs for references.



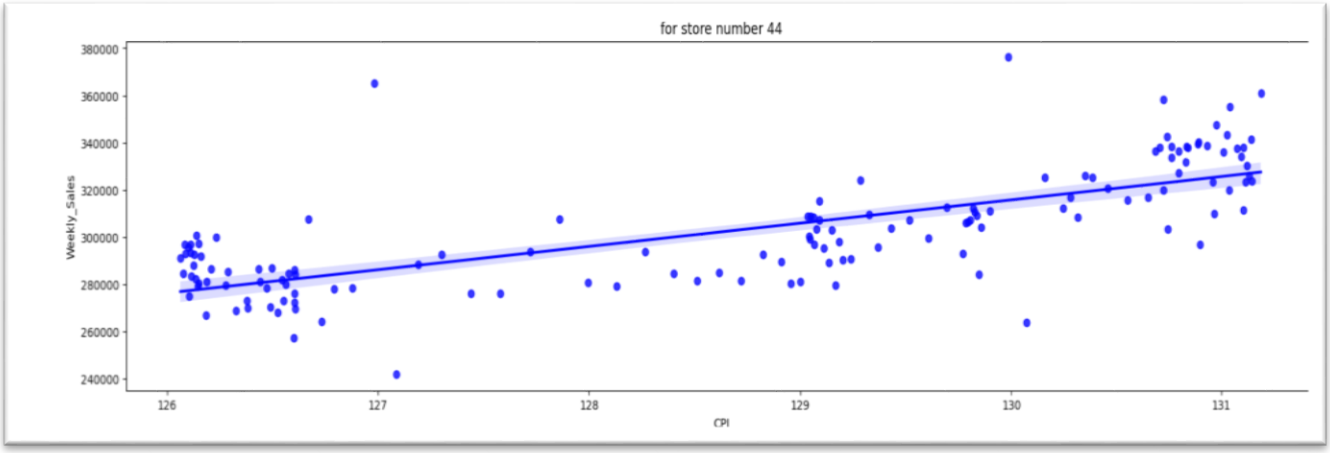


EFFECT OF CPI ON OVERALL WEEKLY SALES.

CPI is between 120-145 or from 180-240, and no customer lies in this range140-180. CPI and weekly sales have a very week negative co-relation ,with increase in CPI , weekly sales is decreasing but the impact is very less.

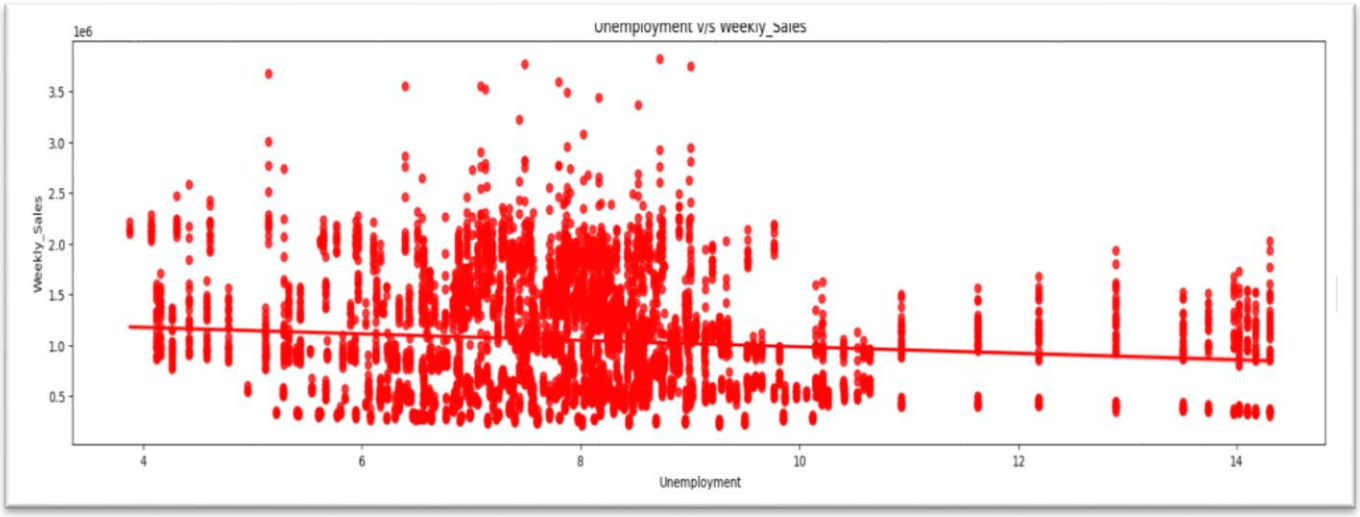


However, some stores shows opposite trend from that off overall weekly sales like store number 38,44 etc.

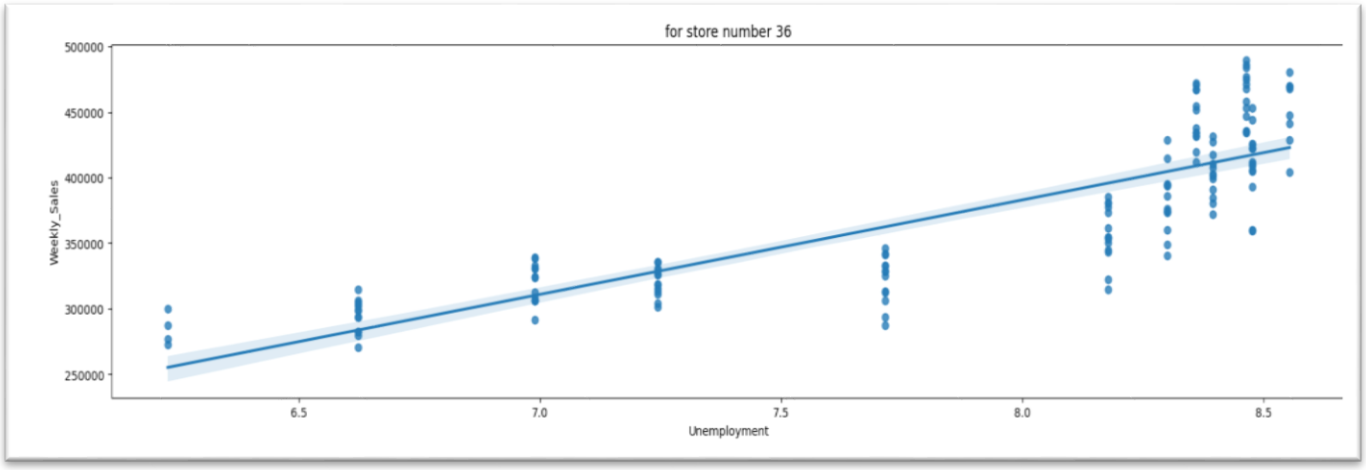


EFFECT OF UNEMPLOYMENT ON OVERALL WEEKLY SALES

Weekly sales and unemployment also show week negative co-relation, With increase in unemployment rate the weekly sales is decreasing but by very little margin.

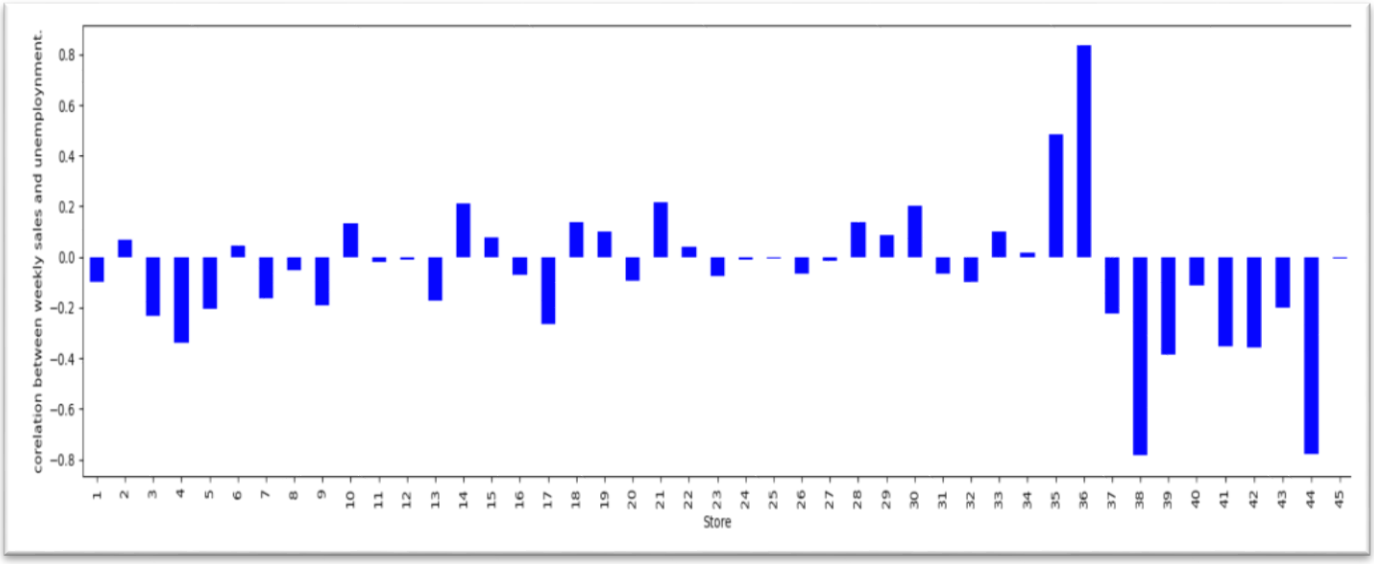


However certain stores like store number: -10,14,18,21,30 shows increase in weekly sales with increase in unemployment, and store number 35 and 36 , shows drastic increase in weekly sales with increase in unemployment.



WORST EFFCTED STORE BY UNEMPLOYNMENT RATIO.

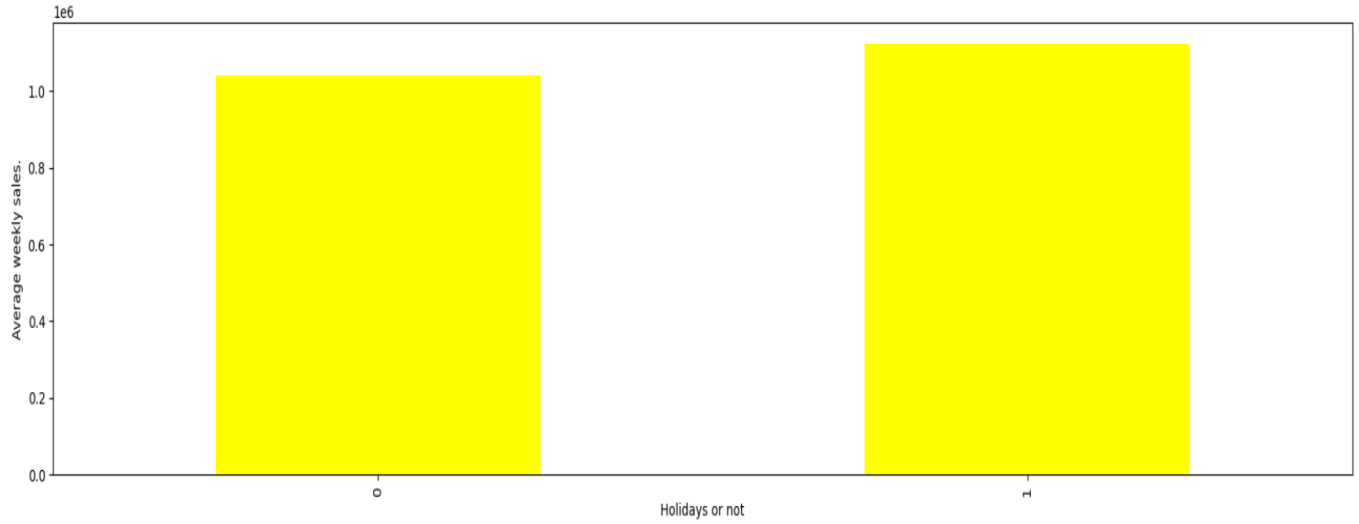
Weekly sales of Store number 38 was worst affected by the unemployment followed by weekly sales of store number 44 and 39. However, sales of 36 and 35 shows a great increase in weekly sales with increase in unemployment.



WHAT HAD BETTER SALES WORKING DAYS OR HOLIAYS.

After calculating average weekly sales on working days and holidays it was found that holidays have more sales than working days.

Average weekly sales on holidays exceeded average weekly sales on weekdays by 7.8%



SEASONAL DECOMPOSITION FOR SALES OF ENTIRE DATA.

On seasonal decomposition it was observed that , data follows an increasing trend but not linearly increasing.

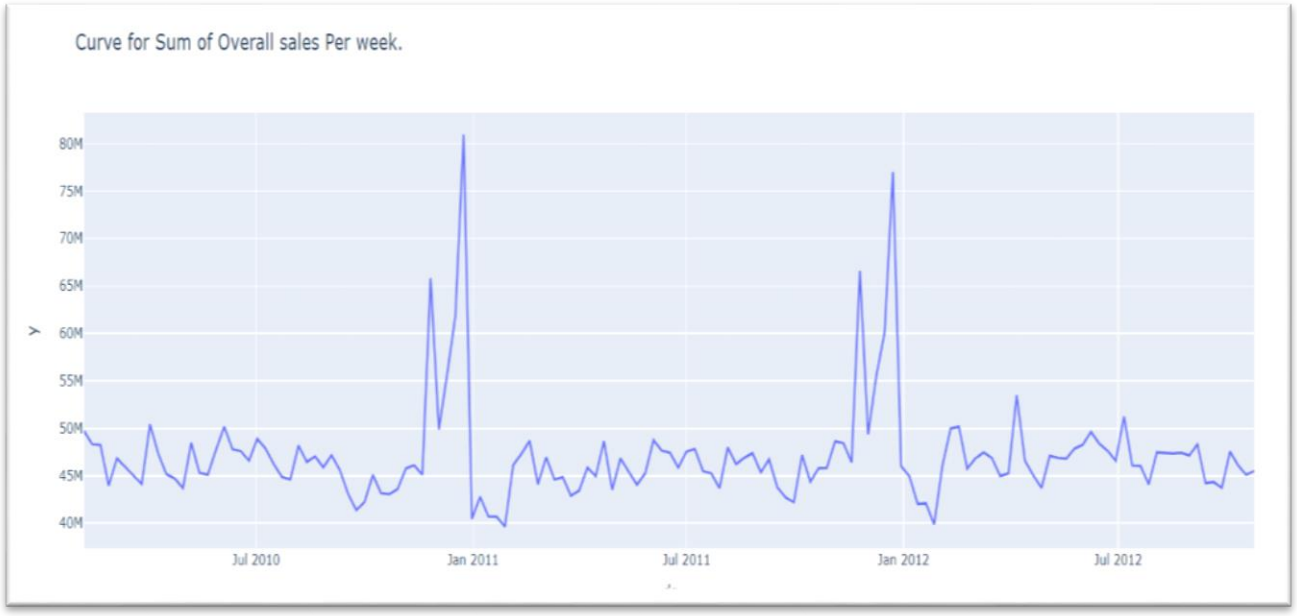
There are noises in data specially in the months of December and January in year 2011 and 2012.

Data also follows a seasonal pattern :-

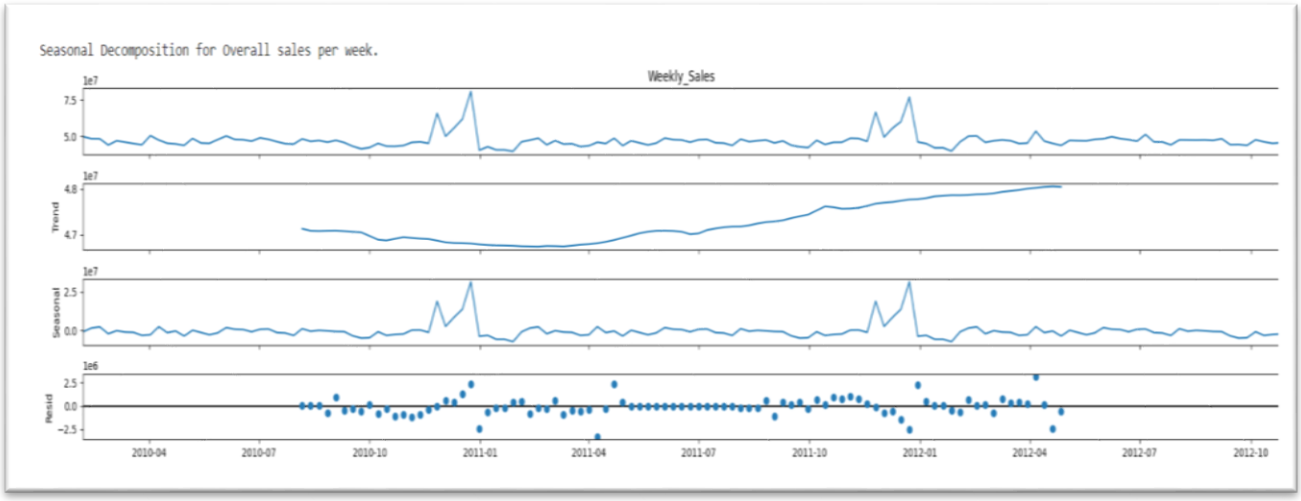
from the month of February to month of October, data shows lot of noises but data in this region don't show any clear increase or decrease, however from November starting to December mid , weekly sales get momentum and touches the peak and increases to maximum and then the sales decreases after December 24 with great margin.

REASONS POSSIBLE FOR SEASONAL TREND.

- Seasonality could be due to festivals, it is observed that in month of December there is increase in sale, this could be due to Christmas.
- Seasonality in sales could be due to seasonal merchandise, Walmart being a retail store sells merchandise as per season.
- Different cultures and seasons have their own tradition, festivals and culture , this may be reason for seasonality.
- Retailers often releases new products or product line at certain times of the year. If this continues on year to year basis can be the reason for seasonality.
- Temperature and weather can also leads to seasonality.



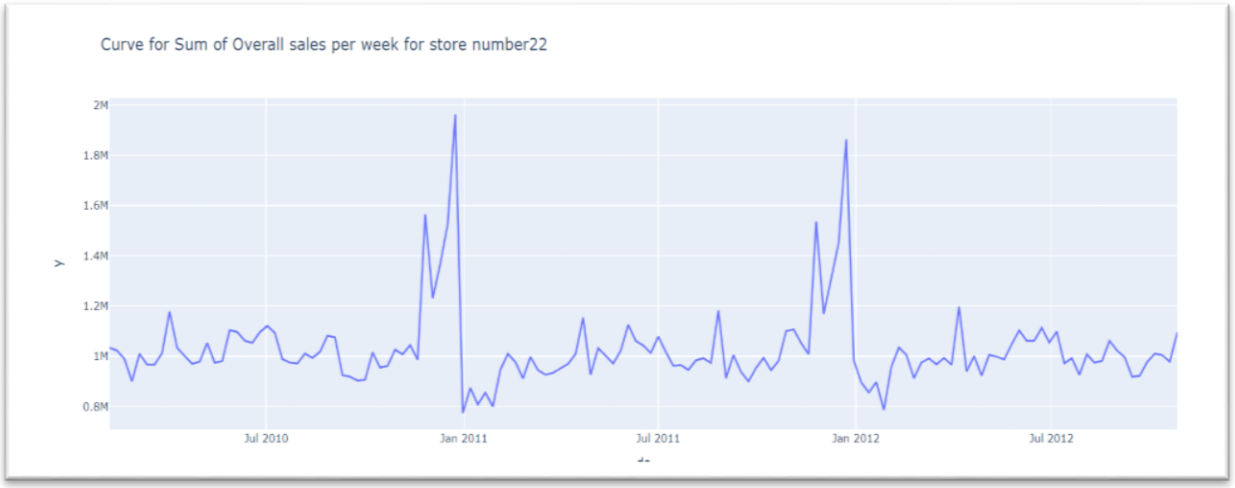
Seasonal decomposition for the overall weekly sales.



SEASONAL DECOMPOSE OF WEEKLY SALES FOR EACH STORE.

Store number 3,5,26,27,12,35 Shows slight variations from the overall sales distribution for the entire table.

Store number 16,17,18,28,30,33,36,27,38 ,39,42,43,44 behaved abnormally as compared to overall sales distribution and showed irregular patterns.



DATA PREPROCESSING STEPS AND INSPIRATION.

For making data fit for training the model so that I get best accuracy for the future predictions , various pre-processing steps were taken in the following order.

- Separating date and weekly sales(target column) from entire dataset, This was done because our time series models delas with date and target column only , Time series model generally learn trend , seasonality from the data and then makes predictions on the basis of learning.

- Renaming the date columns to ‘ds’ and weekly sales column to ‘y’, I did this transformation because , after trying different models , I found Fb prophet giving the best accuracy. FB prophet have a format of taking date input by the column name ‘ds’ and sales input by the column name y.

- Changed ‘ds’ column datatype to ‘datetime’ using ‘to_datetime’ function of pandas, this was done to make our date column datatype date so that considering these dates patterns can be learnt. By default the datetime column is of object datatype.

MODEL CHOOSED, MOTIVATION AND REASON FOR CHOOSING THE MODEL.

After trying various models and various hyper-parameters of various model. Fb prophet came out to be champion model.

- Our data was having seasonality, trends and noises, these things can be best captured and hyper-tuned by fb prophet.

- Fb prophet also don’t need data to be stationary, it don’t take date to be stationary and consider only seasonality and trends.

- Moreover, our data also had holiday as a column. Fb prophet provide holiday hyper parameter tuning to handle target values affected by holidays.

- Prophet models both short-term fluctuations and long-term trends in your data, making it versatile for capturing various patterns.
- Prophet is designed to be robust to outliers, so it can provide reasonably accurate forecasts even when your data contains sporadic extreme values.

PREPROCESSING STEPS AND MODEL CREATION.

```
[71] def predict_sales():
       a= int(input("Enter the sotre number[1,45] for which you want to predict the sales."))
       df1= data[data['Store']==a][["Date",'Weekly_Sales']]

       #Renaming the columns.
       df1.columns=['ds','y']

       #Changing the datatype of ds column to datetime.
       df1['ds']= pd.to_datetime(df1['ds'],dayfirst=True)

       #Calling model
       m = Prophet(
          seasonality_prior_scale=50,
          holidays_prior_scale=10.0,
          changepoint_prior_scale=0.10,
          yearly_seasonality=25,
          growth='linear',
          interval_width=0.95 )

       m.add_country_holidays(country_name='US')
       m.fit(df1)

       #Creating future dataframe
       future=m.make_future_dataframe(periods=12,freq='W')

       #Predicting.
       forecast= m.predict(future)
       f1=forecast[['ds','yhat_lower','yhat_upper','yhat']]
```

MODEL EVALUATION AND TECHNIQUES

- Prophet provides several default settings, it also allows for customization, enabling you to fine-tune parameters to better fit your specific dataset.
  Model evaluation includes accessing the performance of our model after training it.
  I used various techniques for model evaluation in this project.

Plotted line and scatter graph using plotly library, that shows the true values, predicted values and forecasted values. And for each data point while moving cursor over the point, It will show the true values and predicted values for the same point.

Used R2_score(R squared score)- it measures the proportion of the variance in the dependent variable that is explained by the model. It assesses how well the model fits the data.

Mean Absolute Error (MAE)- It is the absolute differences between actual and predicted values. It measures the average magnitude of errors.

Mean Absolute Percentage Error (MAPE)- MAPE is the average of the absolute percentage differences between the actual and predicted values. It expresses errors as a percentage of the actual values.

My values for:-

1. R2_score ranged from (0.92,0.95) for different stores.
2. MAE ranged from (22000, 32000) for different stores.
3. MAPE ranged from (2%-4%) for different stores.

SCOPE OF THE PROJECT

Inventory Management:- This project can help companies to manage their inventory , cost , demand and supply. By tuning the model to highest efficiency companies can Rely on the model predictions with some margin of errors.

Promotion Optimization:- We can develop a model for optimizing promotional campaigns. Determine the most effective store , timing, duration, workforce etc for the store. This can be very helpful during the peak seasons or festive seasons.

Collaboration with vendors and Partners:- The insights can be used to collaborate business partners , vendors to share insights and optimize the supply chain and retail operations further.

Business Oriented Decisions:- The analysis and predictions can help senior management to prepare strategies , tactics and form organisational structures that will result in upliftment of the company.

Supply and Demand Management:- Having idea in advance regarding the demand , core management can take a very precise decision on how much demand is how the demand is to be fulfilled and how the material is to be supplied to various stores to fulfil the demand

Early opportunity and early threat detection:- By feeding this model with new data , and improving the accuracy of the model every time. We can predict the near future and with the help of statistical tools we can figure out any opportunity and curb any threat that will arise in near future.

CONCLUSION

Following conclusion can be drawn from the various graphs, charts and model predictions.

- All the 6 columns are not normally distributed.
- All the columns neither have any strong nor any week relationship with the output variable (Weekly Sales)
- Top 3 performing stores are 20,4,14
- Worst 3 performing stores are 33,44,5
- Average sales on holiday week is more than on non-holiday week by 7.8%
- temperature and weekly sales show very weak negative co-relation.
- Fuel Prices and weekly sales shows very weak positive co-relation.
- CPI and weekly sales show very weak negative corelation.
- unemployment and weekly sales also show very weak negative corelation.
- Champion model for the prediction was fb prophet.

REFRENCES

- Fb prophet webpage (https://facebook.github.io/prophet/)
- Plotly webpage (https://plotly.com/python/plotly-express/)
- NumPy (https://numpy.org/)
- Pandas (https://pandas.pydata.org/)