CAPSTONE PROJECT-1

Predictive Modelling on Walmart Time Series Data of 45 Stores

Table of Contents

1. Problem Statement

A retail store that has multiple outlets(45 stores) across the country are facing issues in managing the inventory - to match the demand with respect to supply.To address these challenges and gain actionable insights into their sales operations by understanding customer demand leads to potential profits and operational efficiency.

2. Project Objective

- The Primary objective of this project is to analyze the impact of various factors like unemployment rate ,temprature ,holidays,fuel prices,and consumer price index on weekly sales
- Find the top performing and worst performing store based on historical sales data
- Predict the sales for the next 12 weeks

3. Data Description

```
<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
```

By using data1.info() it is found that the data contains 6435 rows and 8 columns .The columns are

Store: It is the store number from 1-45

Date: It is the date one which sales data is recorded weekly starting form 05-10-2020 to26-10-2010 for all the 45 stores

Weekly_Sales : It is the sales for the given store in that week

Holiday_Flag: It is a binary value representing if the week includes any holiday or not.If the given day is a holiday then it is represented by 1 else it is represented by 0

Temperature:Temprature in that particular week is taken at each store location

Fuel_Price: The cost of fuel in the region where store is located

CPI: It is a key economic indicator to measure inflation and changes in overall price levels of goods and services in an economy.

Unemployment: Unemployment rate in that particular region

4. Data Pre-processing Steps and Inspiration

1. Loading the required modules like numpy, pandas, matplotlib, seaborn.

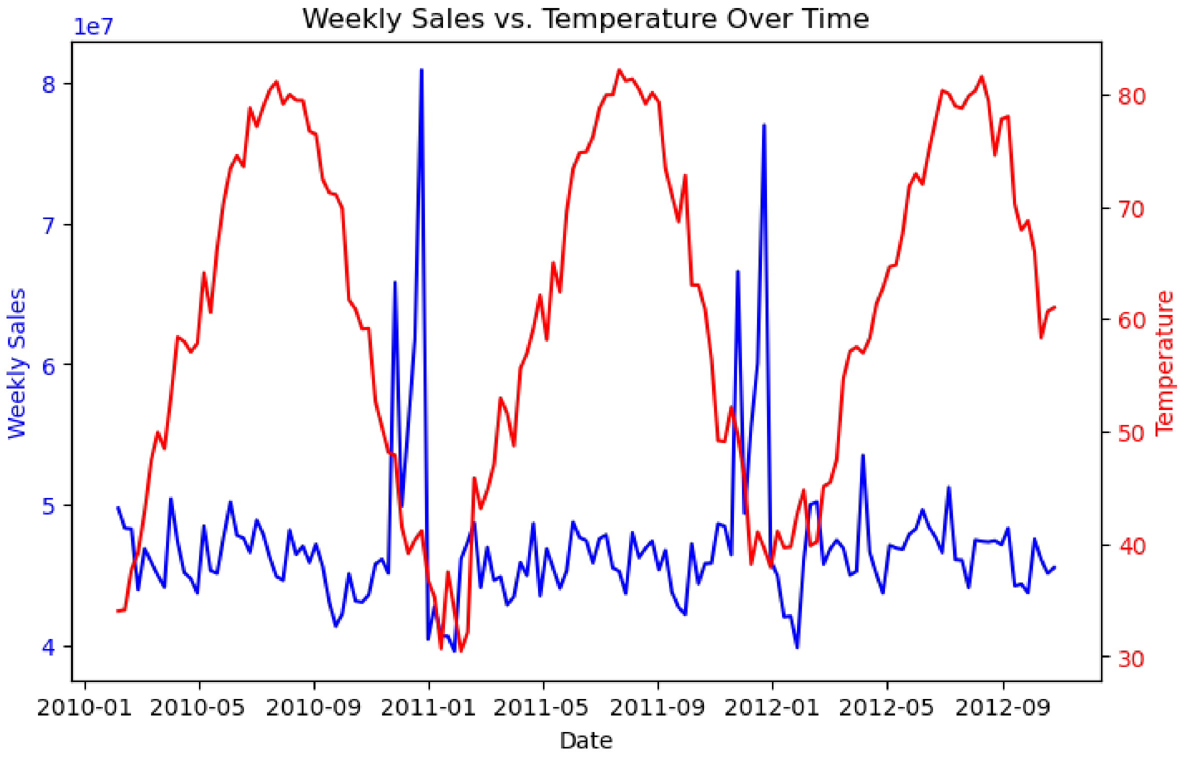
2.Converting date column to date time format

3. Checking if there are any null values(no null values in the data)

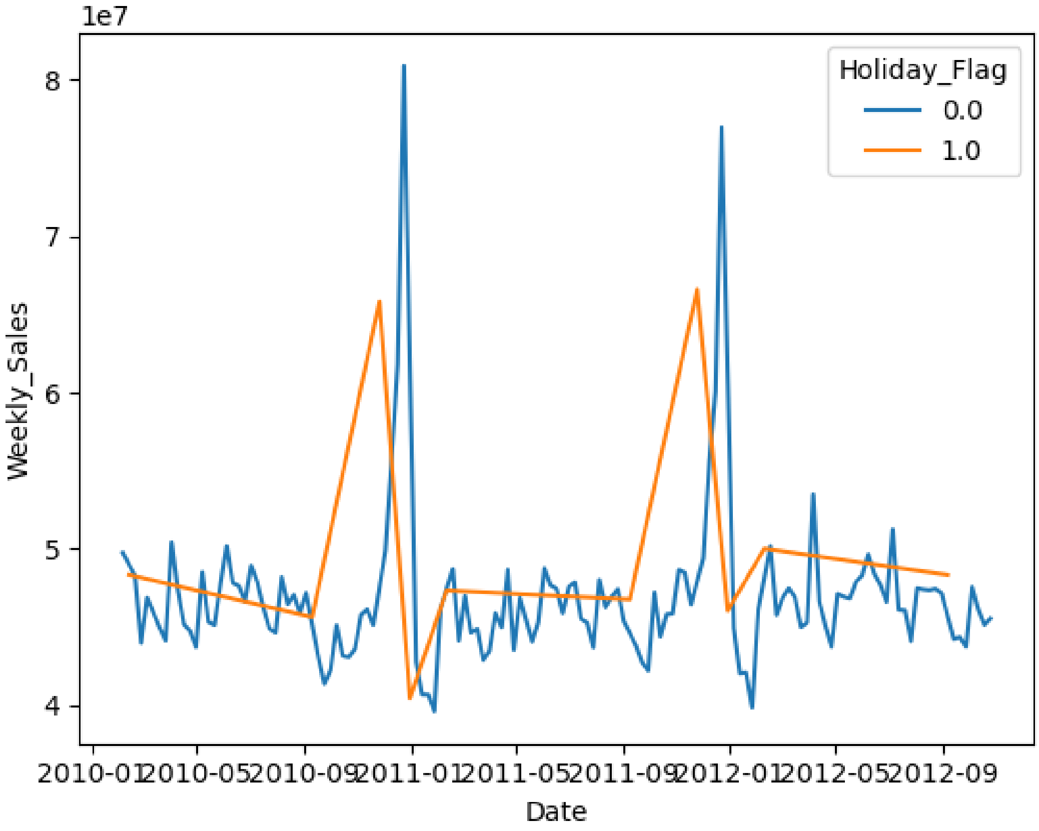
4.Checking the outliers in the data

5.Checking descriptive statstical values like mean median mode standard deviation ,percentiles , and minimum and maximum values of numerical features

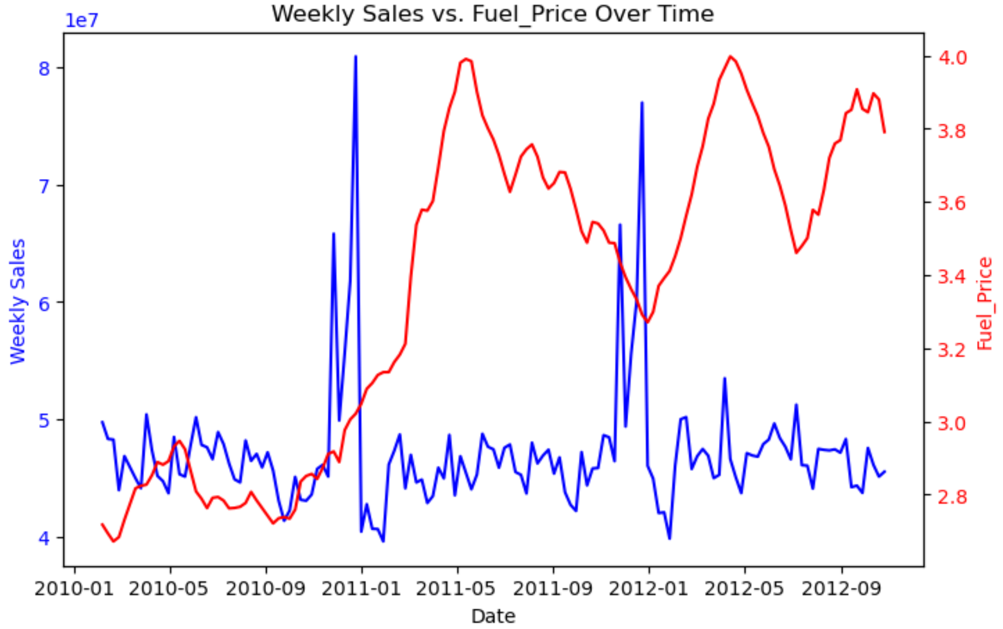
6.Checking how temprature is affecting the sales



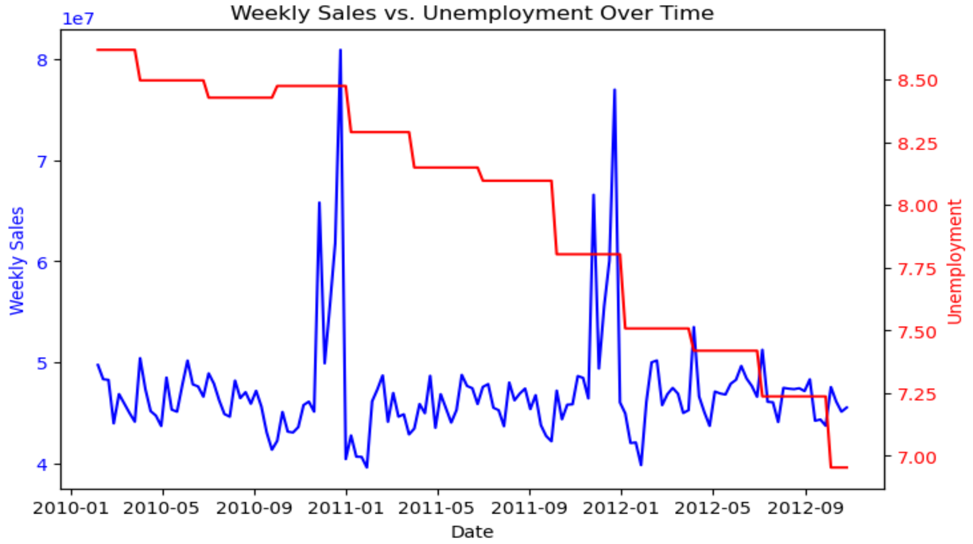
7.Checking how sales are being varied during holidays

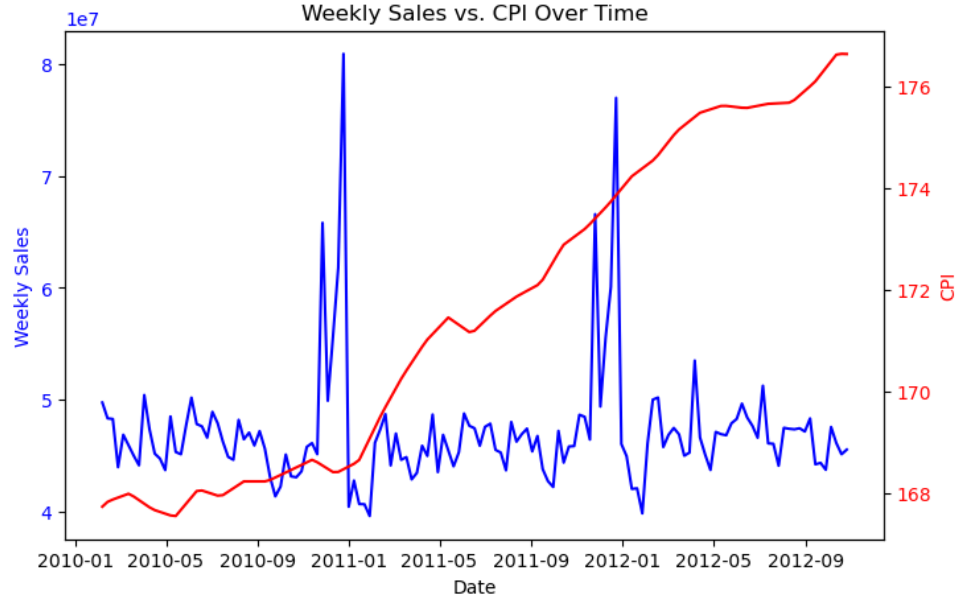


8.Checking how fuel prices are affecting the sales



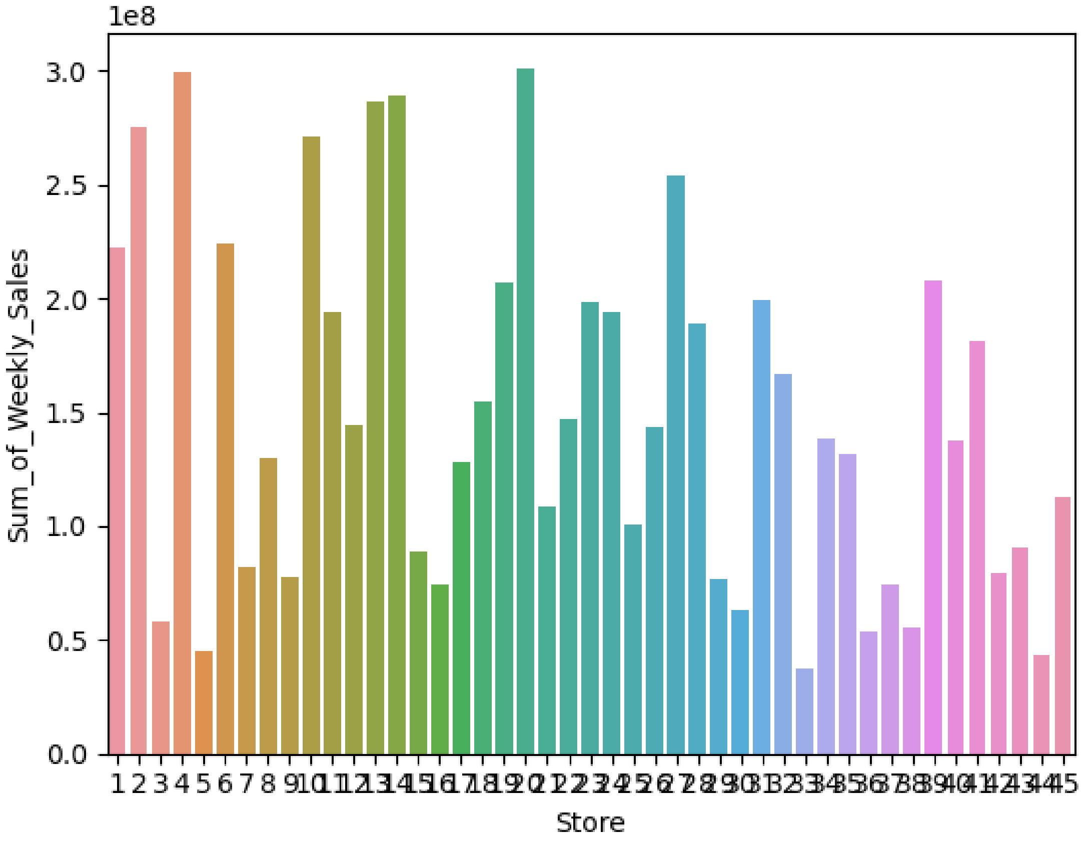
9.Checking how CPI and unemployment in a particular area affects sales





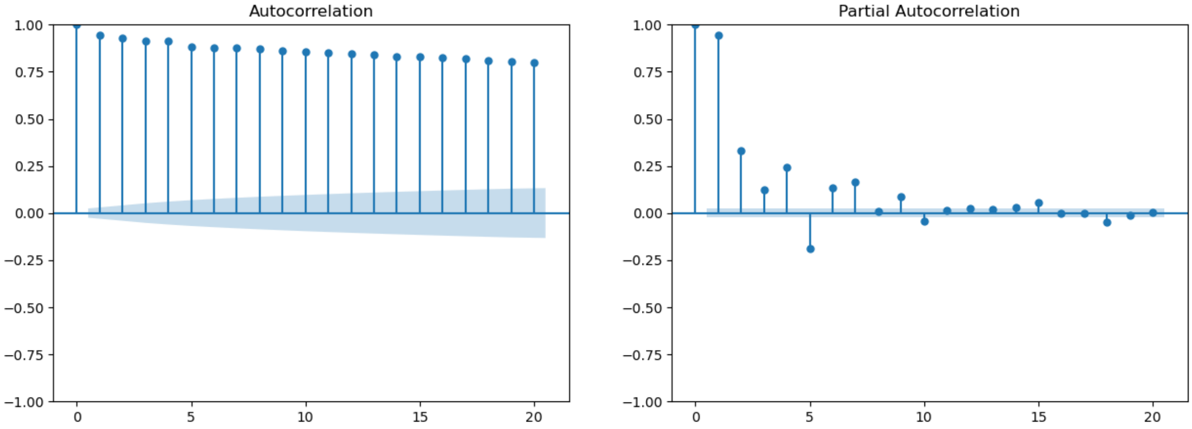
10.Finding the best performing and worst performing stored based on weekly sales and the difference in their sales

Store   Sum_of_Weekly_Sales
20   3.013978e+08 maximum
Store   Sum_of_Weekly_Sales
33   37160221.96 minimum

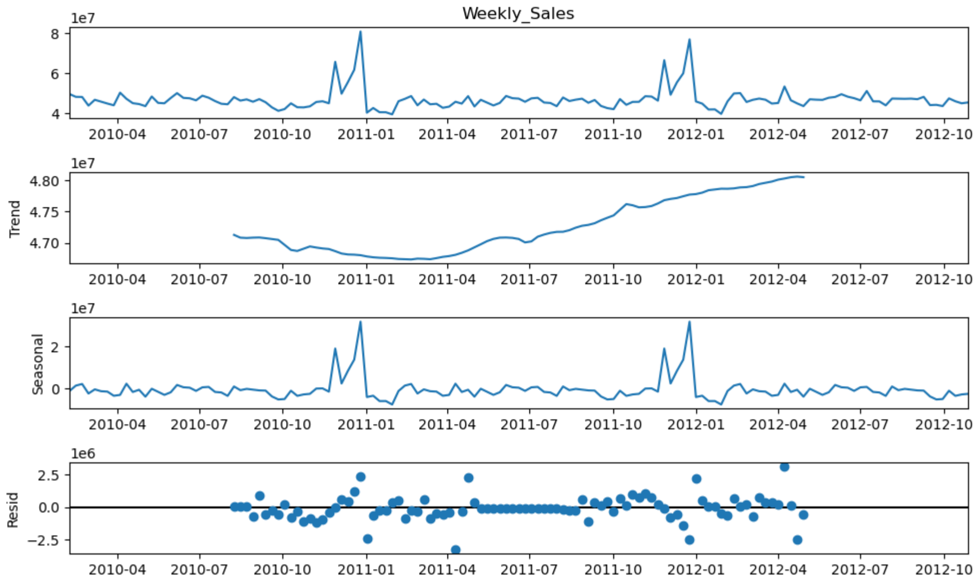


5. Choosing the Algorithm for the Project

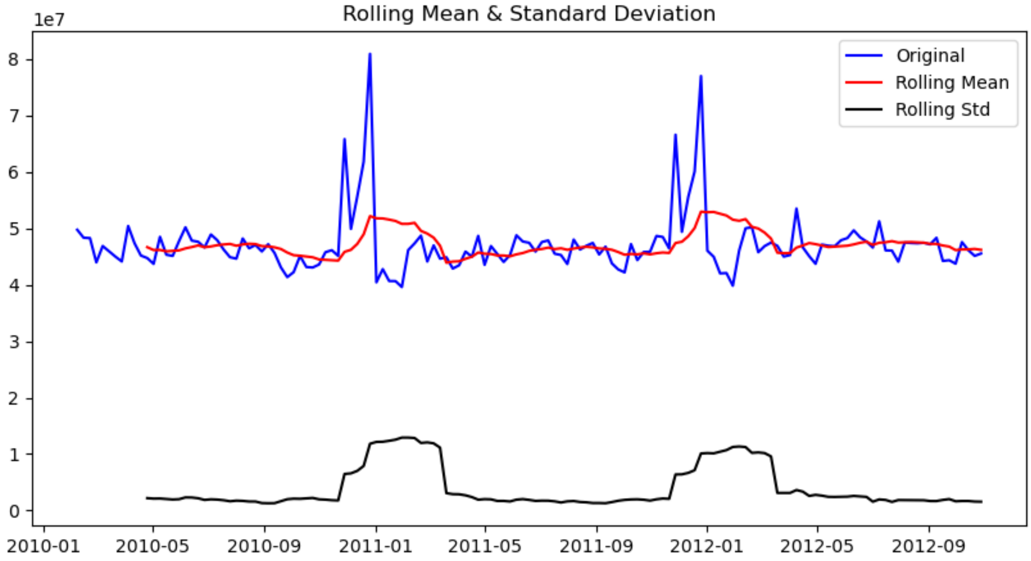
1.Here initially I used ACF and PACF functions to understand the autocorrelation between timeseries and also the lagged values



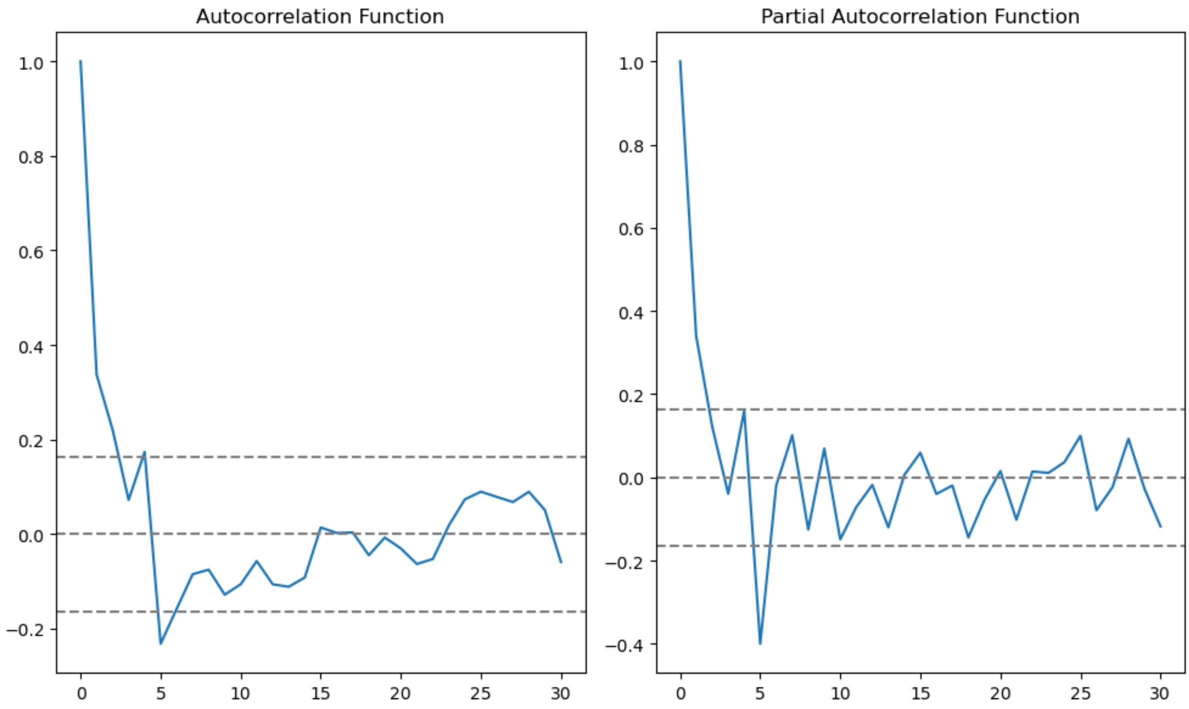
2.The Seasonal decomposition is performed to understand the trend and seasonality



Rolling statistics is calculated to identfy the rolling mean and standard deviation to understand tred and seasonality



ACF and PACF analysis is done to generate ACF and PACF plots that visua;ize autocorrelation and partialautocorrelation which is used to selct p,d,q for ARIMA model



ARIMA model is used to fit the data based on AIC by using auto_arima that helps in finding p,d,q valuse that best fits the data

The data set is splitted into training data '2010-02-5':'2012-02-17' and testing data the remaining part

The data is trained using ARIMA and forecasts where generated

6. Motivation and Reasons For Choosing the Algorithm

ARIMA model is used here because

1.It can capture temporal dependencies in short and long term patterns

2.It can mode auto regressive behaviour ,differencing and moving average components

3.It works well with stationary data and as well as non stationary data by using the differencing to make the data stationary

4.auto_arima function automate the selection of p,d,q

7. Assumptions

The time series analysis should always be performed on stationary data that means the statistical properties do not change over time

The auto_arima function selects the model parameters and seasonal parameters based on AIC

The seasonal decomposition is made by assuming that the data contains seasonality and trend components

Data points are after equaltime periods

8. Model Evaluation and Techniques

The model evaluation techniques used in this project

1. Mean Absolute Error which measures the average absolute difference between the actual and forecasted values
2. Mean Squared Error which calculates the average of the squared differences between the actual and forecasted valeus
3. Root Mean Squared Error which provides average magnitude of errors in the same units as the data

```
Absolute Error (MAE): 3023338.58
Mean Squared Error (MSE): 12734968035761.96
Root Mean Squared Error (RMSE): 3568608.7044
```

9. Inferences from the Same

Observation: As Temperature increase in summer sales are not as high as in winter which indicates that people go out less frequently in summer due to high temperatures

Observation: sales show greater increase (spikes) during holidays and non holiday periods in december and drops during the period of january

Observation: No direct correlation between any feature with the other based on pearson correlation values

Observation: unemployement doesn't affect sales. however it decreases considerably over time

Observation: CPI has no direct correlation with sales overtime.

Top performig store -20

Weak performing store-33

The data has seasonality and trend components

Based on the time series it is observed that the sales have greater spikes during december and falls highly during january

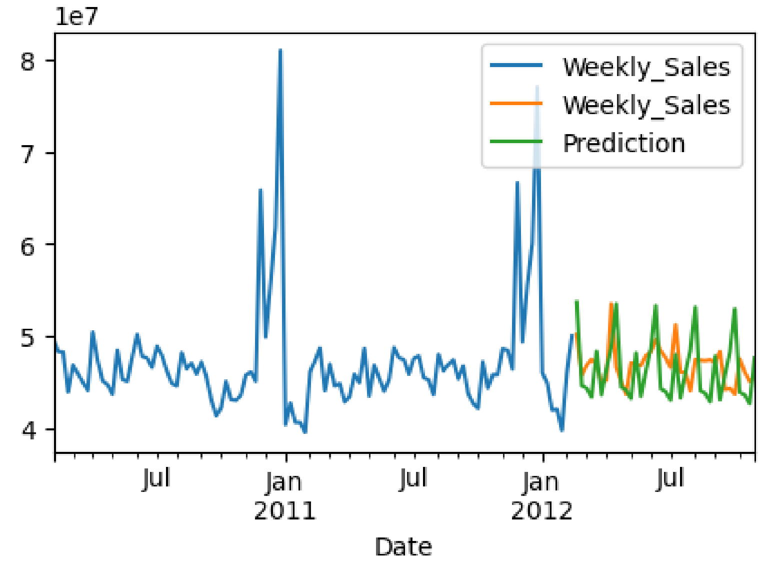
10. Future Possibilities of the Project

Devoloping a real time sales forecasting dashboard application for better sales management

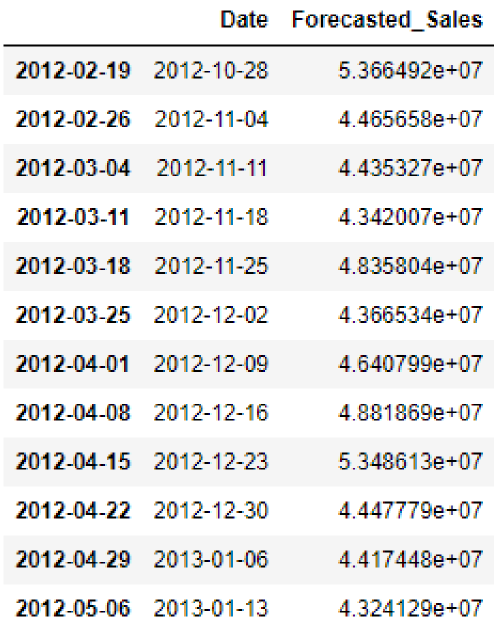
Incorporate other factors to have a more accurate forecast

11. Conclusion

The below graph shows the forecasted sales and forecasted.csv file has the predicted sales of next 12 weeks



It shows that the sales are increasing in the month of december aand dropping in january



	Date	Forecasted_Sales
2012-02-19	2012-10-28	5.366492e+07
2012-02-26	2012-11-04	4.465658e+07
2012-03-04	2012-11-11	4.435327e+07
2012-03-11	2012-11-18	4.342007e+07
2012-03-18	2012-11-25	4.835804e+07
2012-03-25	2012-12-02	4.366534e+07
2012-04-01	2012-12-09	4.640799e+07
2012-04-08	2012-12-16	4.881869e+07
2012-04-15	2012-12-23	5.348613e+07
2012-04-22	2012-12-30	4.447779e+07
2012-04-29	2013-01-06	4.417448e+07
2012-05-06	2013-01-13	4.324129e+07

12. References

- https://www.rit.edu/ischoolprojects/sites/rit.edu.ischoolprojects/files/document_library/Rashmi_Jeswani_Capstone.pdf
- https://www.statsmodels.org/stable/tsa.html
- https://scikitlearn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html
- https://matplotlib.org/stable/plot_types/basic/index.html
- https://pandas.pydata.org/docs/user_guide/index.html
- https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/
- https://seaborn.pydata.org/tutorial.html