Capstone Project – Project Walmart

Submitted by: Bibhu Kalita

Table of Contents

1. Problem Statement:

A retail store that has multiple outlets across the country are facing issues in managing the inventory - to match the demand with respect to supply. You are a data scientist, who has to come up with useful insights using the data and make prediction models to forecast the sales for 12 number of weeks.

2. Project Objective:

Estimating the store's sales for the upcoming 12 weeks is the project's goal. As size and time-related data are provided as features in the dataset, determine whether time- and space-based parameters have an impact on sales. Above all, how do holidays falling within a week increase store sales?

3. Data Description:

This dataset contains historical sales information for 45 Walmart locations broken down by week and retail location from 2010 to 2012. Every week, sales are impacted by a few events like Holiday, Temperature, Fuel Price, CPI, and Unemployment. Walmart wants to make precise sales and demand projections. The goal is to identify the variables influencing sales and assess how markdowns around holidays affect sales. Along with this, we need to predict the forecast for the next 12 weeks for each of the store.

In total, there are 6435 rows and 8 features in the dataset. (Figure 1)

```
#Checking the dtypes of all the columns
data.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   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
```

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

4. Data Pre-processing Steps and Inspiration:

It is crucial to have an in-depth understanding of the dataset that is used in this analysis to understand the models that would give the most accurate prediction. Several times there are underlying patterns or trends in the data that would not be identified as easily, hence the need for an extensive exploratory data analysis. This thorough examination is necessary to understand the underlying structure of the dataset and to draw conclusions or insight about the validity of our analysis.

4.1 Data Cleaning and Pre-processing:

4.1.1 Data Cleaning

a. On checking the datatype of columns, the 'Date' columns format was not in the correct data type. Therefore, I converted the 'Date' columns to correct 'DateTime' format.

"data['Date']=pd.to_datetime(data['Date'],dayfirst=True)"

b. I checked for the presence of the null values in the data. However, data set was clean.

```
: #checking null values
  data.isnull().sum()

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

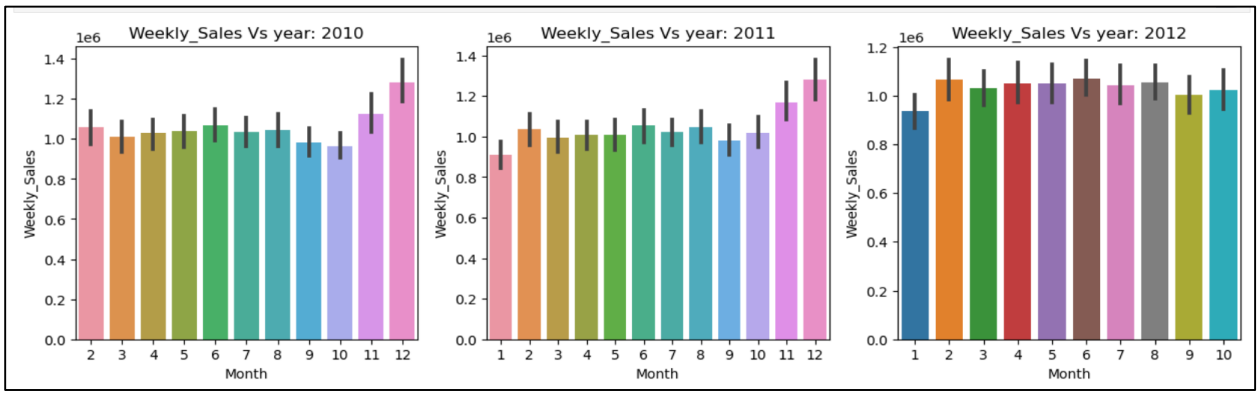
c. I have broken down the 'Date' columns into 'Week', 'Year' and 'Month' to analyze the data more deeply on various features

```
9]: # adding 'Year', 'Month' and 'Week' column
    data['Year'] =data['Date'].dt.year
    data['Month'] =data['Date'].dt.month
    data['Week'] =data['Date'].dt.strftime('%U')
```

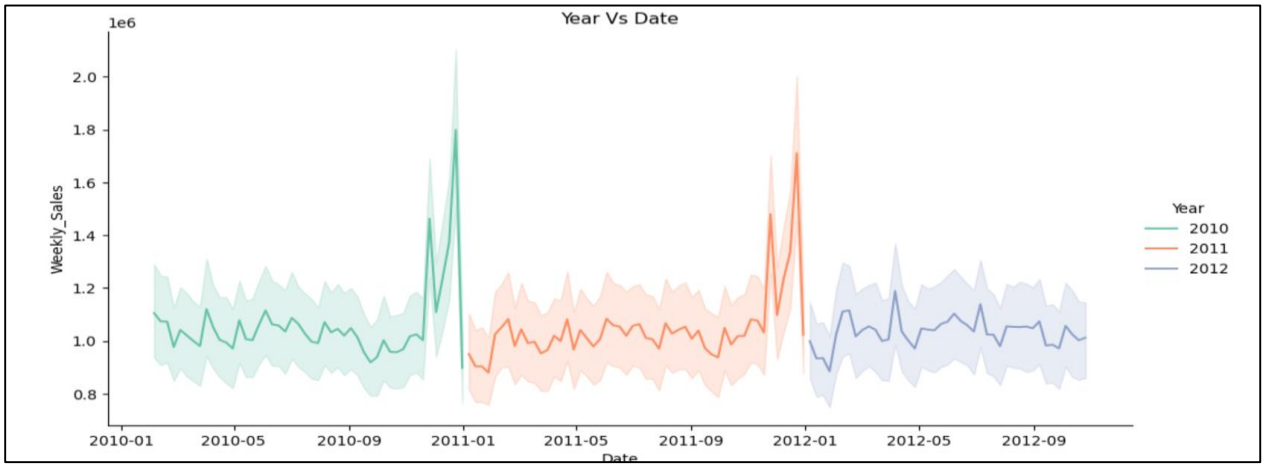
d. I change the data type of 'Store' and 'Holiday_flag' columns to 'object' data type from int type.

4.1.2 Identifying Monthly Sales for Each Year:

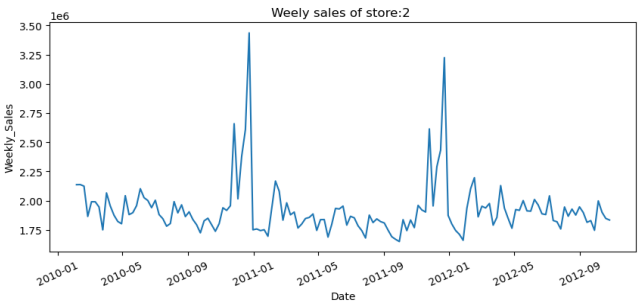
The graph below clearly depicts that the months of November and December recorded the highest average sales for 2010 and 2011. The dataset provided byWalmart contained no weekly sales information for the last two months of the year 2012, hence no conclusion can be drawn for that year. This graph also shows that the month of January tends to have the lowest average sales in the whole year.

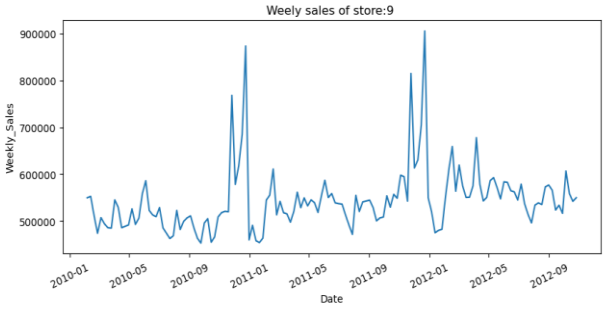


We can also observe that the sales pattern in each year are almost same which is clearly visible in the below graph.



On checking the sales pattern in each store are also same. I am showing below two graph for reference.

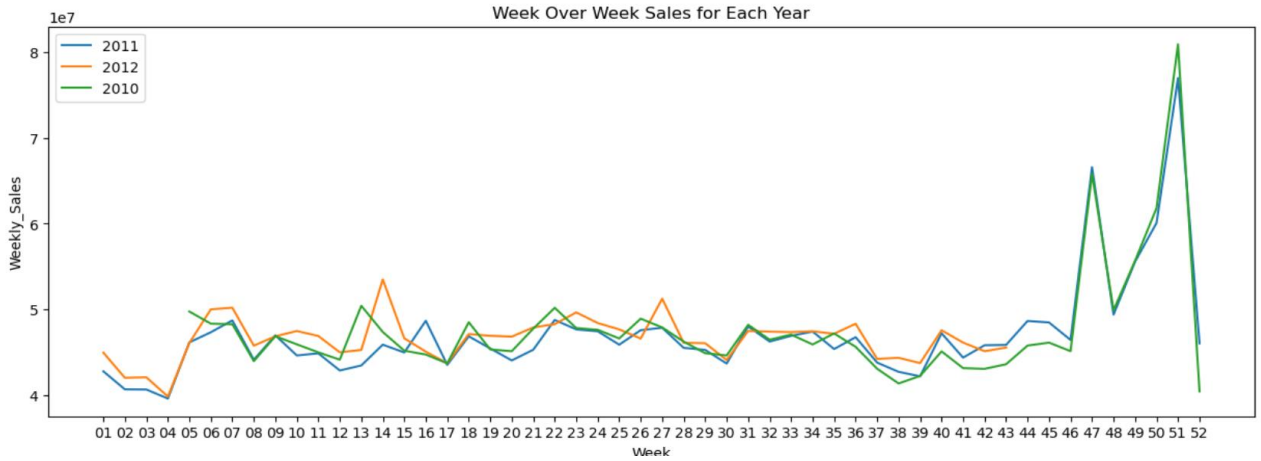




4.1.3 Identifying Week Over Week Sales for Each Year:

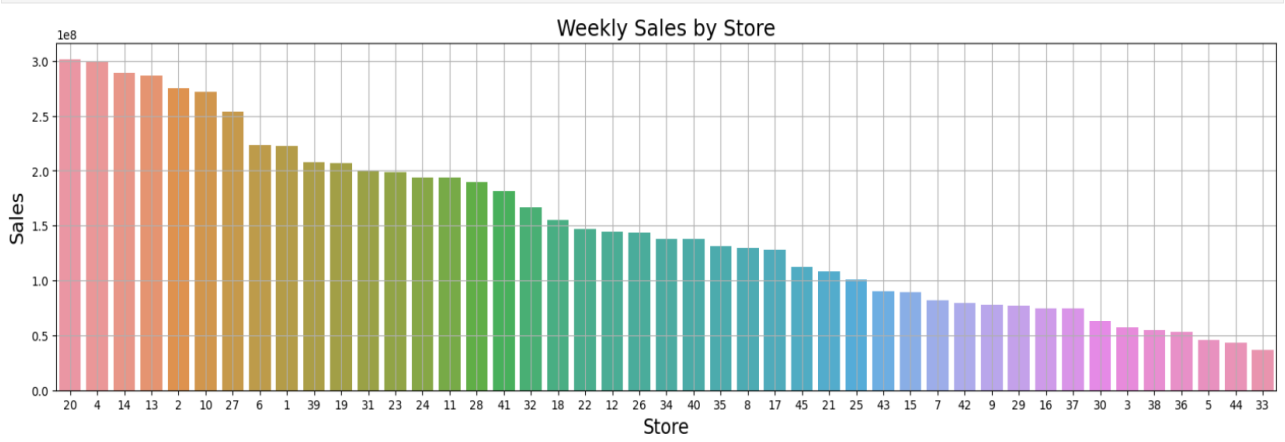
The week over week overview again helps us in understanding if there is an increase in sales in particular weeks each year

There is an evident hike in sales in weeks 47 and 51 that correspond to the holiday season (Thanksgiving and Christmas respectively) proving again that sales rise during the holiday season. Due to the insufficiency of data for the year 2012, these conclusions have only been made based on the data available from 2010 and 2011. This graph also shows that there is a distinguished pattern of decline immediately following holidays.



4.1.4 Identifying Specific Stores and Departments with Highest Sales

After looking at specific stores with the highest sales, it is also imperative to analyse whether these specific stores with the highest sales belong to a specific store type.

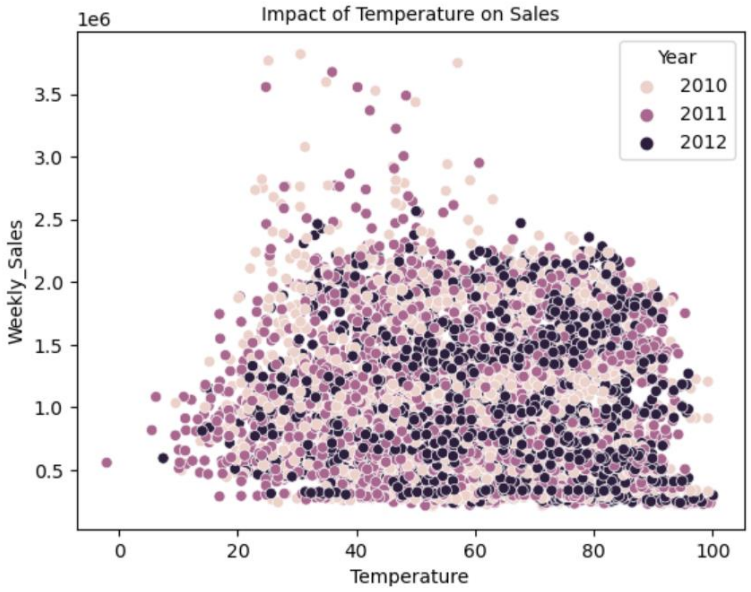


As observed in the figure, store 20, 4, 14, 13 and 2 are certainly the top five revenue-generating stores whereas stores 38, 36, 5, 44, and 33 are certainly the bottom five. As there is no information in this dataset about the type of these stores, it is unclear to establish any conclusion on the revenue of the above-mentioned stores.

4.1.5 Impact of Temperature on Sales

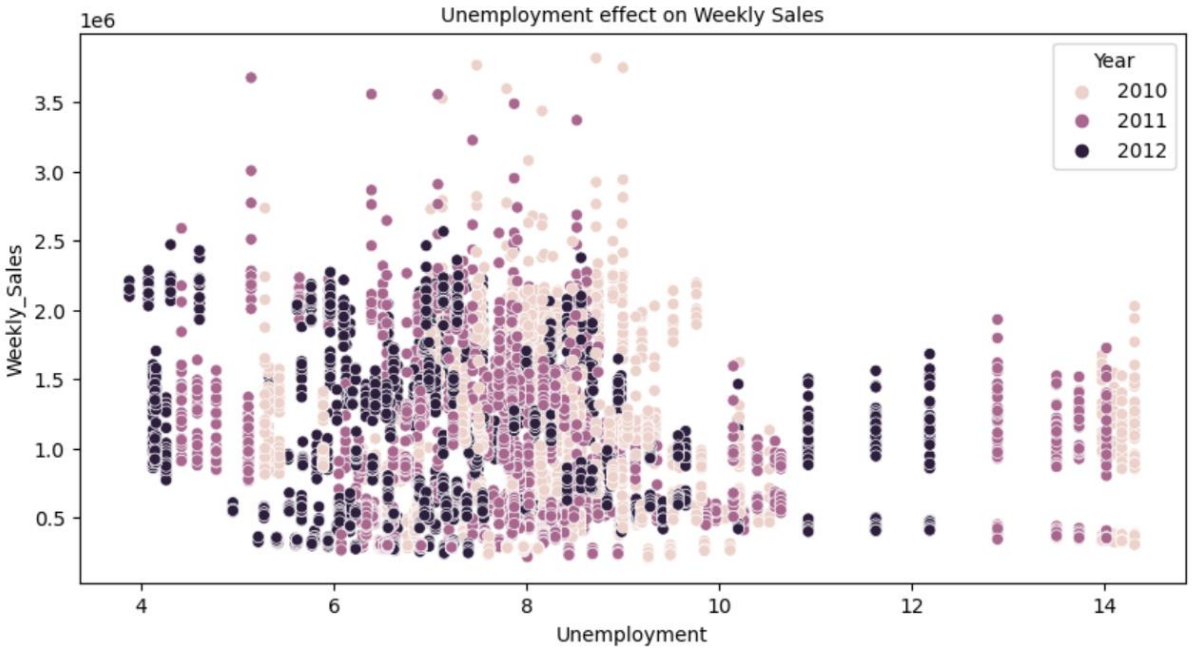
The industry of retail is well aware that weather has a significant impact on sales. While warmer weather encourages sales, excessively hot or cold weather typically discourages people from going outside to make purchases. Since they are neither as hot or cold, temperatures between 30 and 70 degrees Fahrenheit are generally regarded as suitable for human habitation.

As can be seen here, most store kinds have their peak sales between 25 and 80 degrees Fahrenheit, supporting the theory that nice weather boosts sales. Sales appear to be sufficiently high under favourable weather circumstances, but are comparatively lower in extremely low and extremely hot temperatures.



4.1.6 Impact of Unemployment on Sales

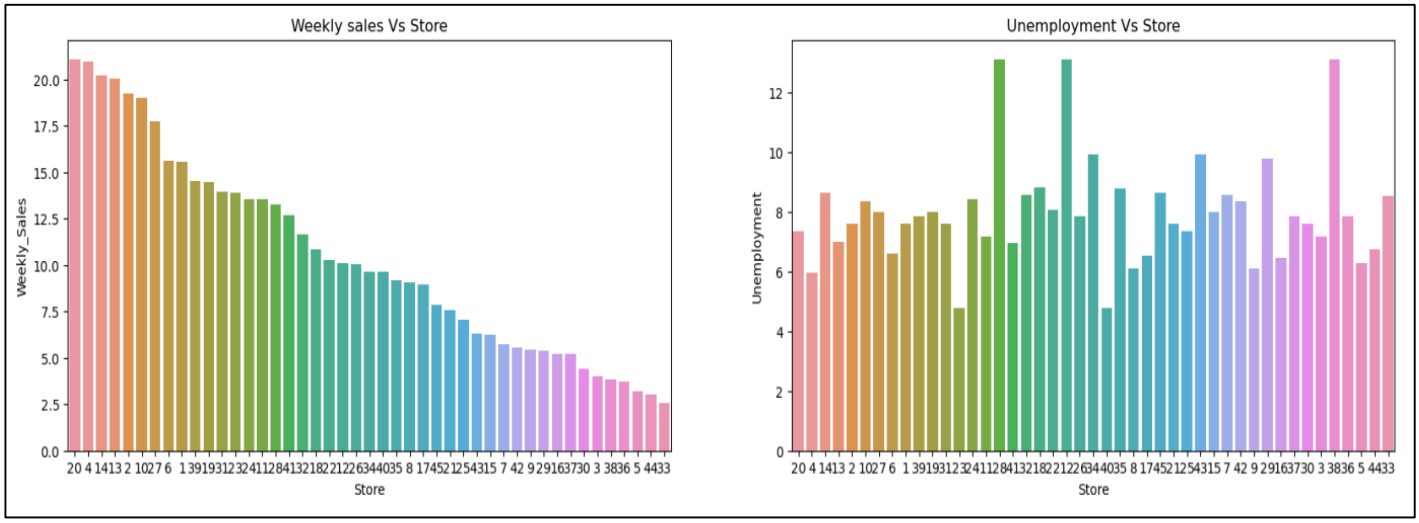
The start of unemployment causes expenditure to drop precipitously; a higher unemployment index would typically cause sales to decline as people tend to spend less money overall. The unemployment rate for that particular week in the store's vicinity is represented as an index in our dataset. The following data is more easily gathered from our scatter plot:



- For the given store types, there seems to be a visible decrease in sales when the unemployment index is higher than 10

- Even when the unemployment index is higher than 11, there is no significant change in the average sales in stores when compared to the overall sales each year.

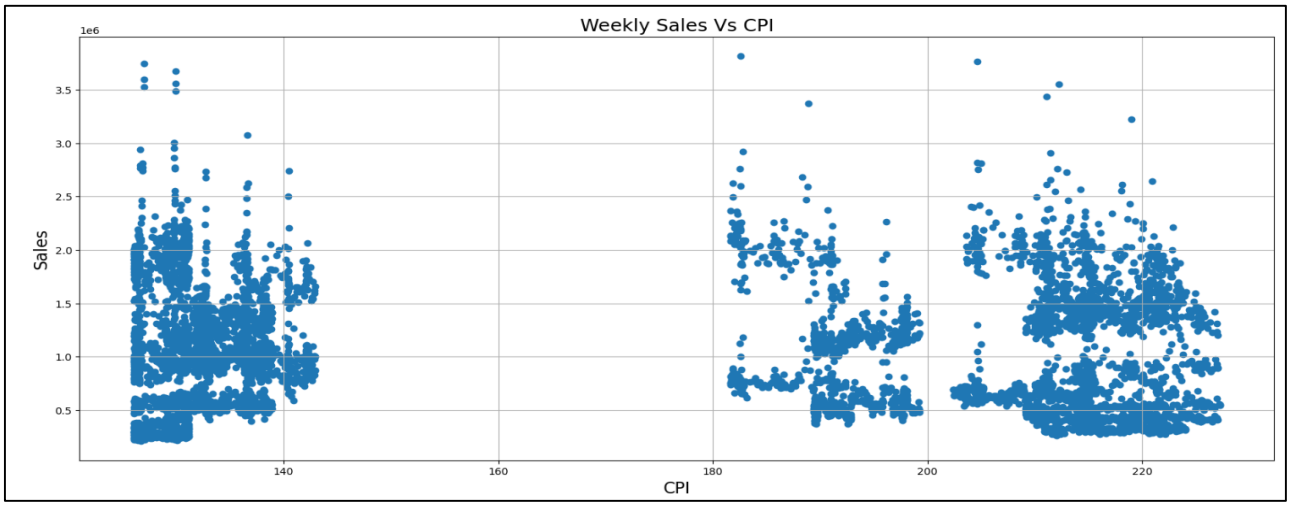
- The highest recorded sales in stores around the unemployment index of 6 to 9; this gives ambiguous ideas about the impact of unemployment on sales for each of the stores which is visible through the below bar graph.



4.1.7 Impact of CPI on Sales

Simply put, the Consumer Price Index (CPI) is a metric that evaluates price fluctuations linked to an individual's cost of living. Generally speaking, a higher CPI indicates that goods have become more expensive and that an individual must spend more money to maintain the same standard of living.

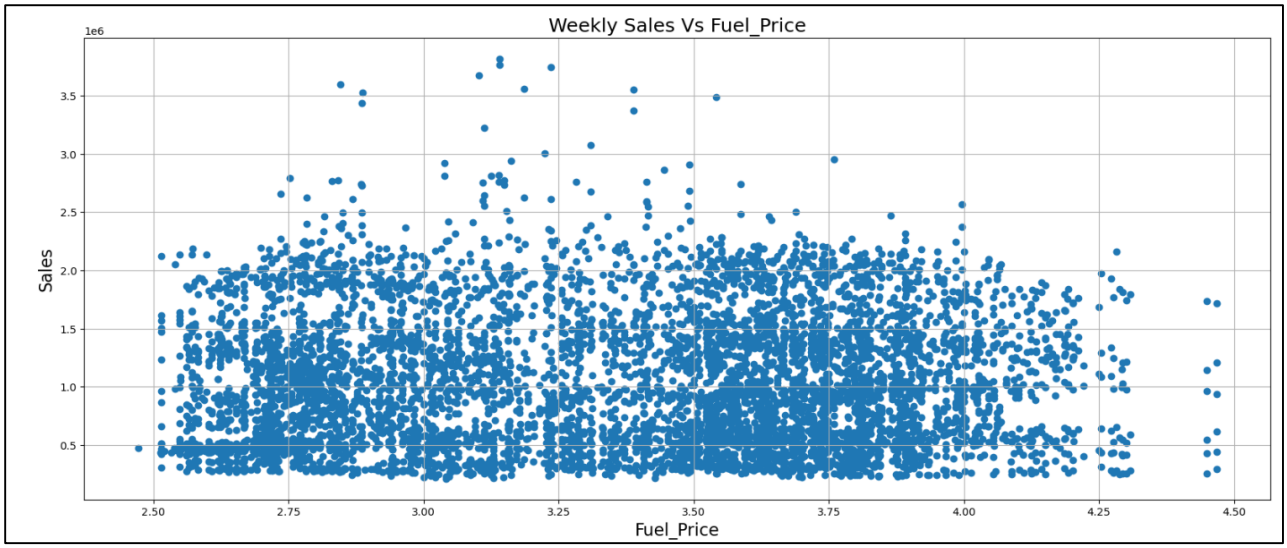
Our scatter plot illustrates this phenomenon, showing three distinct clusters around varying CPI ranges, while there appears to be no discernible correlation between weekly sales for Walmart stores and CPI changes (sales continue to occur at high CPI rates).



4.1.8 Impact of Fuel Price on Sales

The economist assumes that even a little increase in fuel prices drives up annual expenses considerably and deters customers from actively purchasing the goods and services they need.

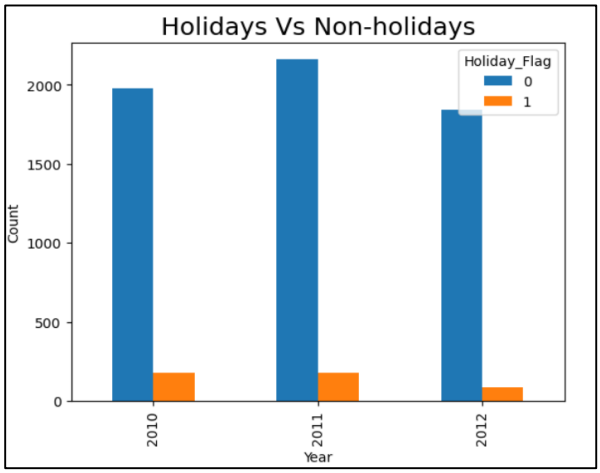
This is shown in the graphic below as well; sales appear to somewhat decline when fuel prices exceed $4.25, but they increase when fuel prices fall between $2.75 and $3.75. The idea that cheaper fuel prices lead to more sales is supported by various findings, even though there isn't a clear trend to support this.



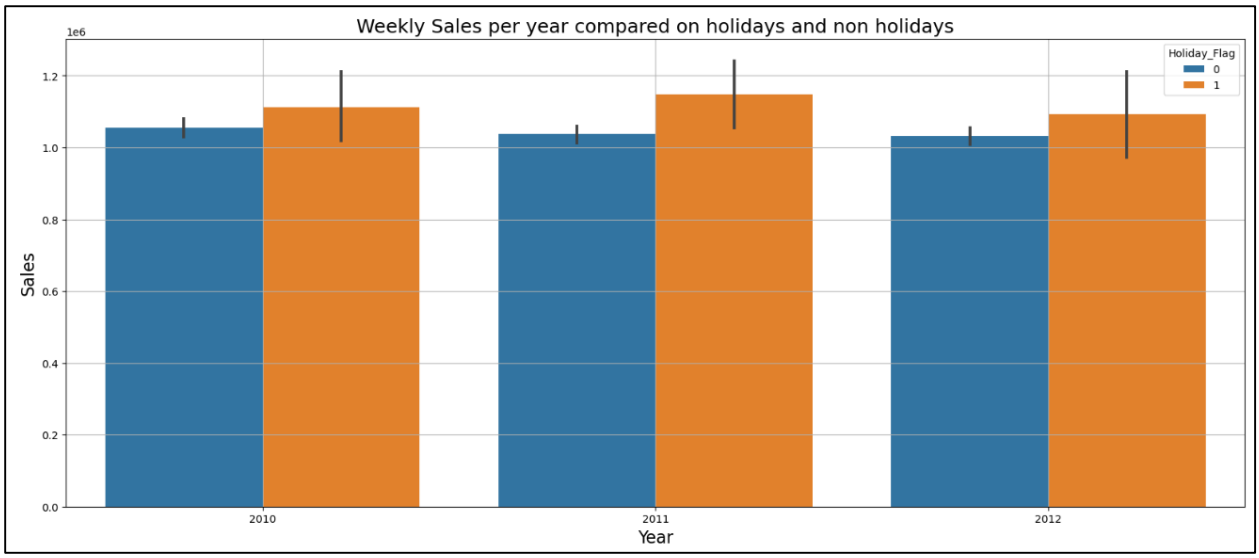
4.1.9 Holiday VS Non-Holiday Sales

The dataset provided contains data about weekly Walmart sales over various periods of time in a year, this includes data about the sales that occur during holiday periods. It was crucial to compare the difference between sales during holidays and normal weeks to understand if the holiday season gathers higher sales.

For this comparison, I first counted the number of holidays in a year and compared sales during the holiday dates versus the normal days. While the holiday dates only accounted for almost 8 to 5 percent of the days in the year, they still have higher weekly sales than the rest of the year combined (as seen in the image below).

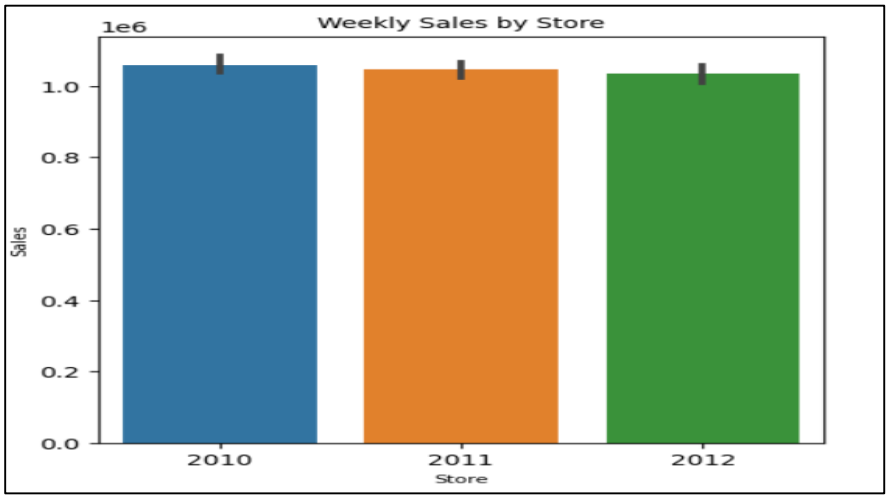


Year	0	1	%Holiday
2010	1980	180	8.33
2011	2160	180	7.69
2012	1845	90	4.65



4.1.10 Annual Sales over the year

The three-year trend in annual sales indicates a slow decline. Nevertheless, there is not enough data in the dataset to investigate the probable cause of the declining trend.



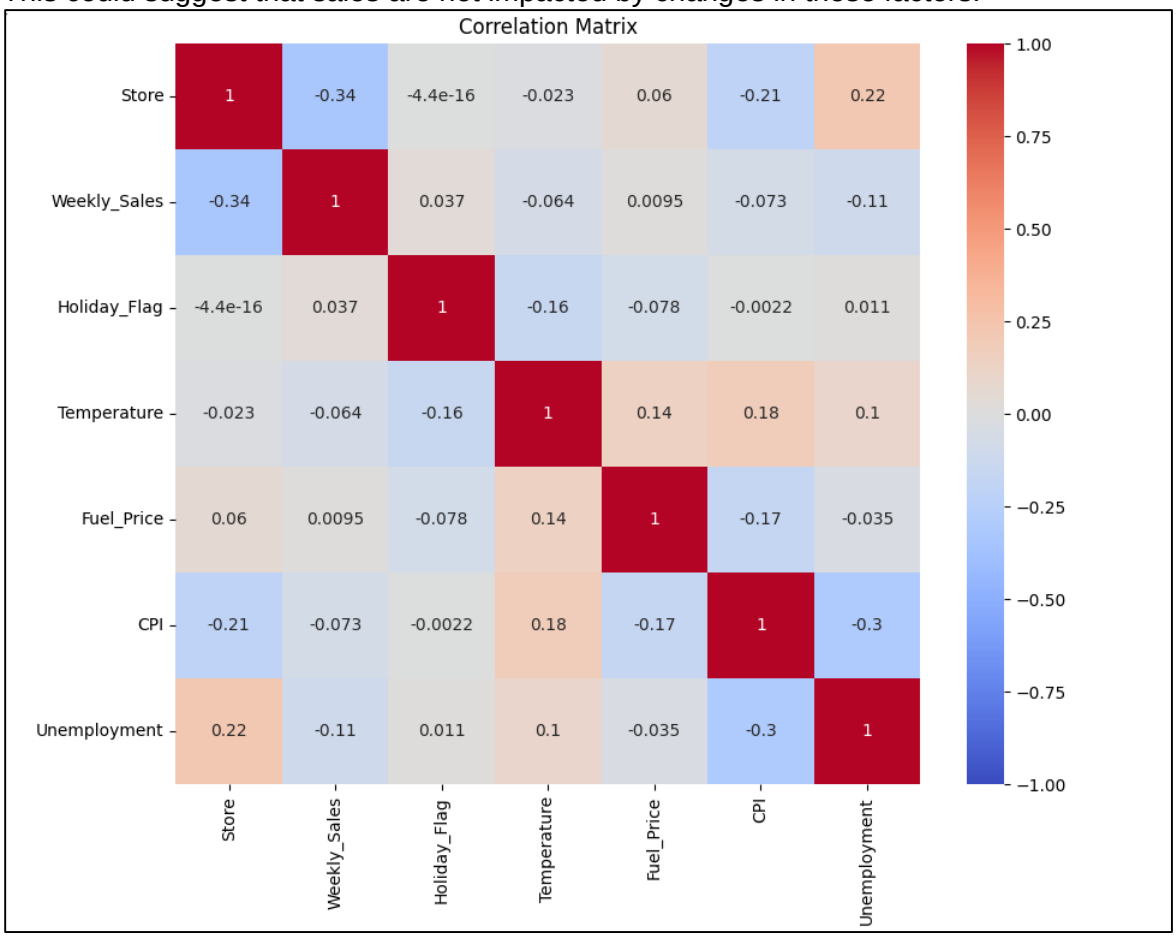
4.1.11 Correlation Matrix

A correlation matrix describes the correlation between the various variables of a dataset.

The heat map/correlation matrix below gives the following information:

- There is a slight correlation between weekly sales Holidays, and Fuel_price.
- There seems to be a negative correlation between weekly sales and temperature, unemployment, and CPI.

This could suggest that sales are not impacted by changes in these factors.



4.1.12 Pre – Processing

Before creating the appropriate model for future weekly sales, we need to pre-process the dataset. Therefore, for prediction model building we perform the below steps:

1. drop the ‘Data’ columns
2. Encode the object data using LabelEncoder()
3. Split the dataset into independent and dependent variables
4. Split the independent and dependent variables into test and training dataset for model building

5 Choosing the Algorithm for the Project:

Several models have been studied as part of this study that were selected based on different aspects of our dataset; the main purpose of creating such models is to predict the weekly sales for different Walmart stores and departments, hence, based on the nature of models that should be created, the following four machine learning models have been used:

- Linear Regression
- Decision Tree
- Random Forest
- KNN

For the weekly forecasting below models are used

- ARIMA
- SARIMAX

Each of these methods have been discussed briefly in the upcoming report. For each of the models, why they were chosen, their implementation, and their success rate have been included.

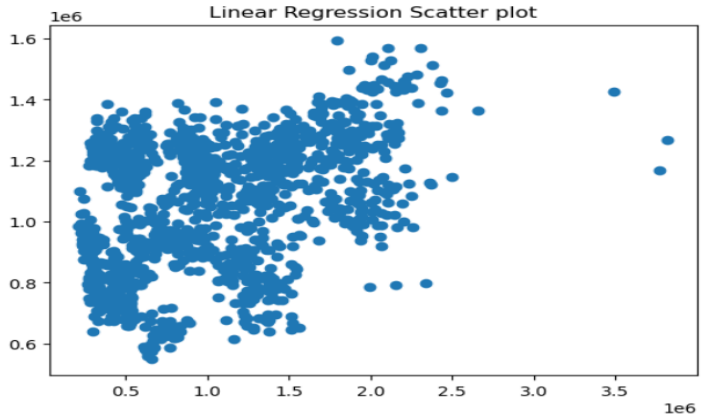
5.1 Linear Regression

Starting with the most basic and straightforward model for this analysis, linear regression aims at finding relationships between two linear variables, a predictor and a target. For this study, multiple linear regression helps in predicting the future value of a numeric variable based on past data.

The 'scikit-learn' library along with the 'Linear Regression' function in Python has been used to create the linear regression model. A r2_score function has been created that provides a measure of success for the model applied. In conclusion, the higher the r2_score, the more efficient the model. Moreover, the scatter plot between the predicted y value and actual y value has been used to check the model behaviour.

```
r2_score_lr=r2_score(y_test, y_pred)
plt.scatter(y_test, y_pred)
plt.title('Linear Regression Scatter plot')
print("R2 Score: ", r2_score(y_test, y_pred))

# print("MSE Score: ", mean_squared_error(y_test, y_pred))
# print("RMSE : ", np.sqrt(mean_squared_error(y_test, y_pred)))
```

```
R2 Score:  0.1555316049960288
```



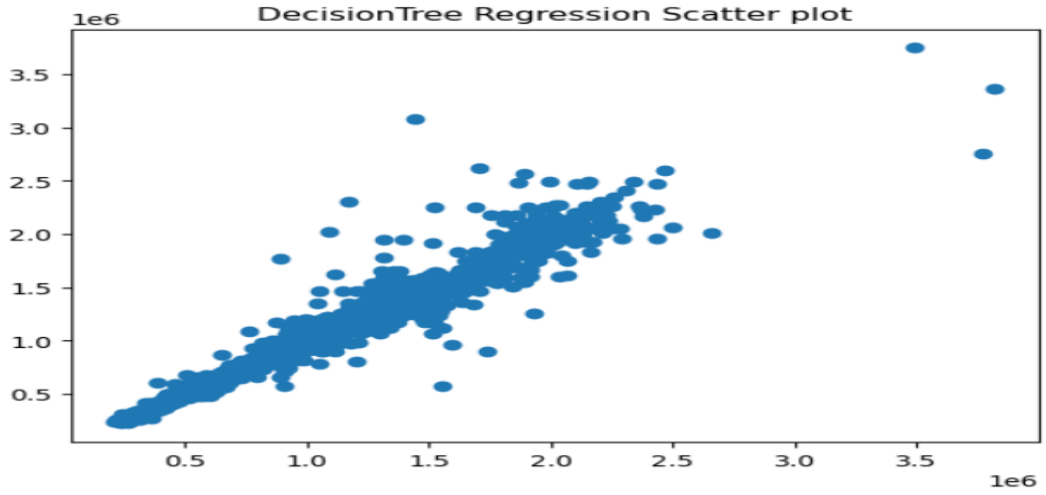
5.2 Decision Tree

This model uses a decision tree to classifies the dataset into smaller subsets, and to define a conclusion about a target value. The tree consists of leaves, where the intermediate ones are the decision nodes and the ones from the extremes are the final outcomes.

The 'scikit-learn' library along with the 'DecisionTreeRegressor' function in Python has been used to create the DecisionTree regression model. A r2_score function has been created that provides a measure of success for the model applied. . Moreover, the scatter plot between the predicted y value and actual y value has been used to check the model behaviour

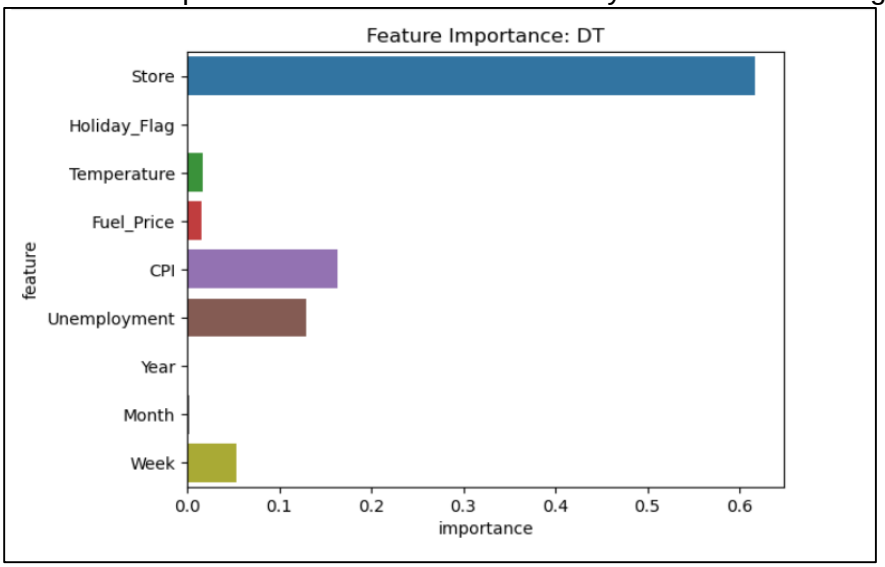
```
r2_score_dt=r2_score(y_test, y_pred1)
plt.scatter(y_test, y_pred1)
plt.title('DecisionTree Regression Scatter plot')

print("R2 Score: ", r2_score(y_test, y_pred1))
# print("MSE Score: ", mean_squared_error(y_test, y_pred1))
# print("RMSE : ", np.sqrt(mean_squared_error(y_test, y_pred1)))
```

```
R2 Score:  0.9356179807077832
```



One important aspect of feature importance is that with the creation of the decision trees, it is easier to capture the importance of each attribute.

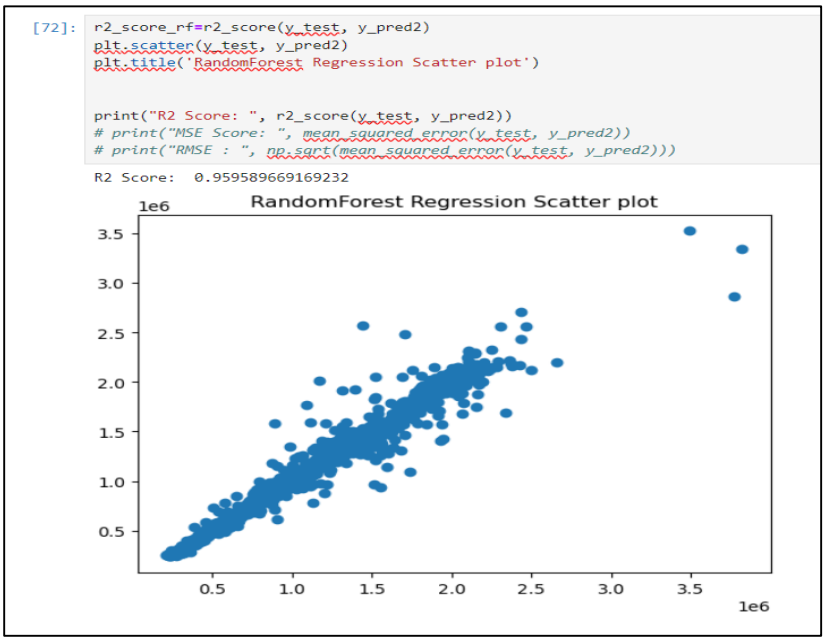
Harmonious to the findings in the EDA, the 'Store, 'CPI' and 'Unemployment' attributes are the three most important features that affect weekly sales for the retail giant.



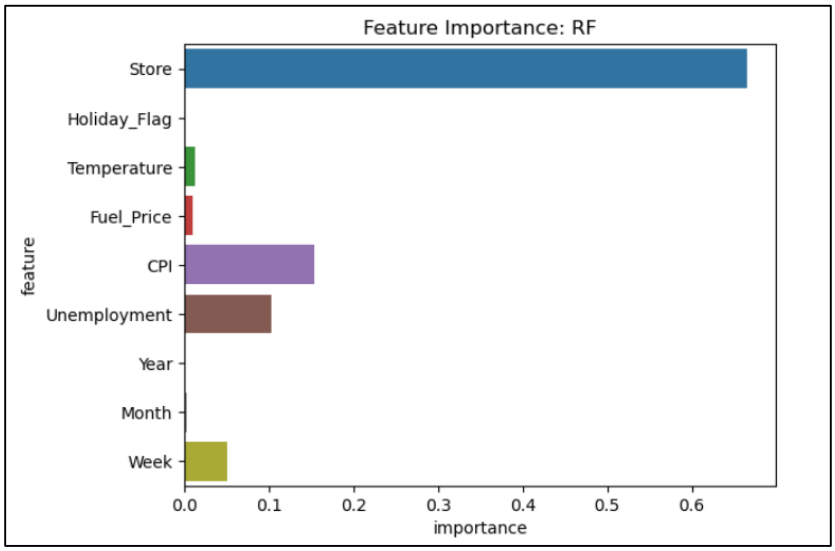
5.3 Random Forest

Random forest usually has good accuracy compared to other linear models and scales well with new features or samples. This regression model can handle missing data and outliers which makes it time-saving and easy to use.

Using the ‘RandomForestRegressor’ a basic model was created, with some initial parameters, and the r2_score was calculated for the test and predicted sets. . Moreover, the scatter plot between the predicted y value and actual y value has been used to check the model behaviour.



Just like the Decision Tree model, the feature importance plot below highlights the most important attributes in the random forest model. According to the feature importance barplot below, the ‘Store, ‘CPI’ and ‘Unemployment’ impact weekly sales the most, which corresponds to the correlation matrix findings in the EDA.



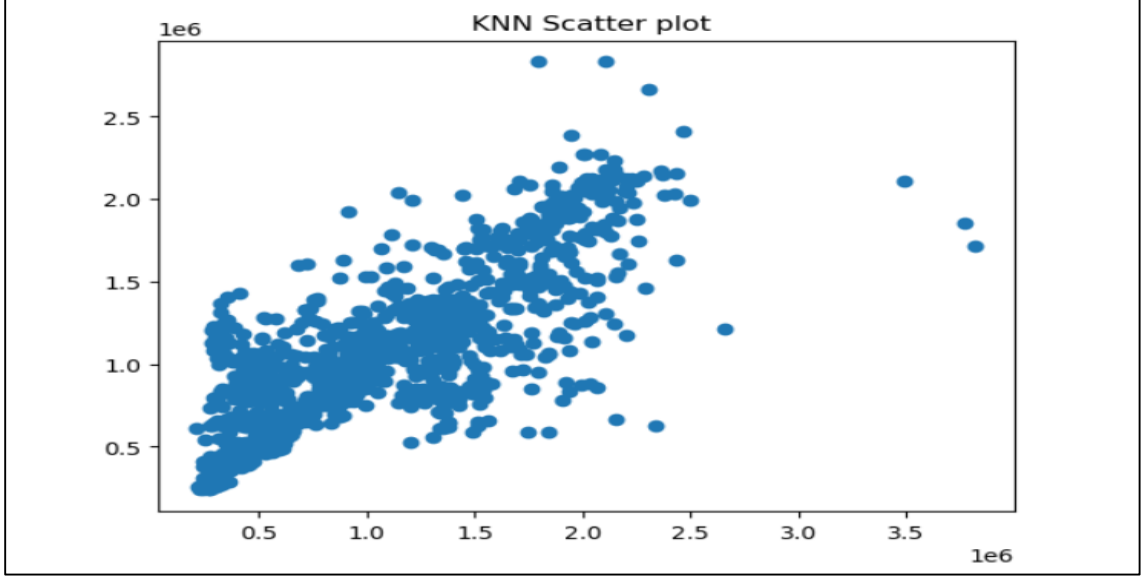
5.4 KNN

This algorithm takes in consideration the k closest points (neighbours) around the target and uses them to learn how to classify the desired point. This model was chosen, because it's simple to implement, no assumption about the data is necessary and the non-parametric nature of KNN gives an advantage in certain settings where the data may be highly unusual.

. An r2_score function has been created that provides a measure of success for the model applied. Moreover, the scatter plot between the predicted y value and actual y value has been used to check the model behaviour

```
[81]: r2_score_knn=r2_score(y_test, y_pred3)
      plt.scatter(y_test, y_pred3)
      plt.title(' KNN Scatter plot')

      print("R2 Score: ", r2_score(y_test, y_pred3))
      # print("MSE Score: ", mean_squared_error(y_test, y_pred3))
      # print("RMSE : ", np.sqrt(mean_squared_error(y_test, y_pred3)))
```

```
R2 Score:  0.6080301793025182
```



5.5 Time series analysis Model for forecasting

Techniques for examining time series data in order to derive significant statistics and other features of the data are referred to as time series analysis. A model is used in time series forecasting to project future values based on observed values from the past.

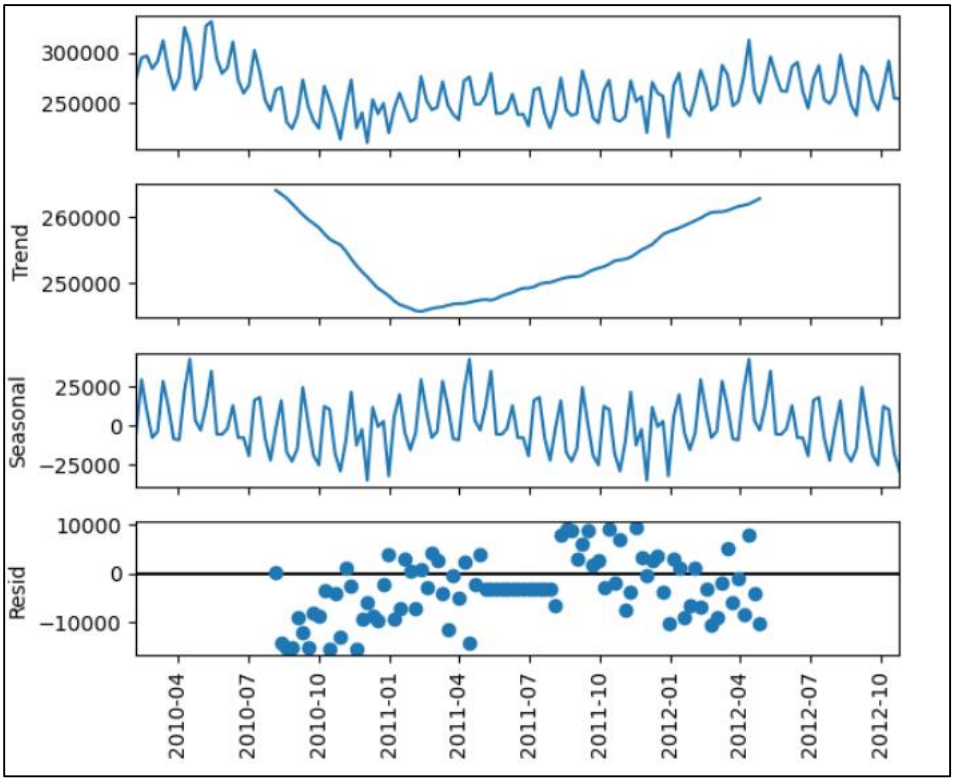
Time series are frequently used in this topic to analyze non-stationary data, including as stock prices, retail sales, weather, and the economy. Various methods for predicting retail sales time series will be exhibited. Let's get started!

As we need to forecast for each store, we select any of the store's weekly sales data randomly for our analysis. Before passing the data to the model, we did the below pre-processing steps

a. Indexing with Time Series Data

b. Visualizing Furniture Sales Time Series Data:

Some distinguishable patterns appear when we plot the data. The time-series has seasonality pattern. We can also visualize our data using a method called time-series decomposition that allows us to decompose our time series into three distinct components: trend, seasonality, and noise.



c. The plot above clearly shows that the sales of furniture is unstable, along with its obvious seasonality. The confirm our assumption, we did an adfuller test for stationarity.

d. to make the data stationary, we perform log transformation followed by rolling mean to make the dataset stationary.

5.5.1 Time series model ARIMA

We are going to apply one of the most commonly used method for time-series forecasting, known as ARIMA, which stands for Autoregressive Integrated Moving Average.

ARIMA models are denoted with the notation ARIMA(p, d, q). These three parameters account for seasonality, trend, and noise in data. Our first goal here was to use a “grid search” to find the optimal set of parameters that yields the best performance for our model. We found the optimal value as optimal_p,optimal_d and optimal_q. Based on these orders, we build our ARIMA to predict the forecast.

5.5.2 Time series model SARIMAX

The SARIMAX **(Seasonal Autoregressive Integrated Moving Average + exogenous variables)** model stands out as a powerful tool for modelling and forecasting both trends and seasonal variations in temporal data while incorporating exogenous variables into the analysis to improve prediction accuracy.

We created the SARIMAX model using the same optimal values of p,d,q and keeping seasonal order as 52.

6. Motivation and Reasons for Choosing the Algorithm

6.1 Selection of the Most Accurate Prediction Model

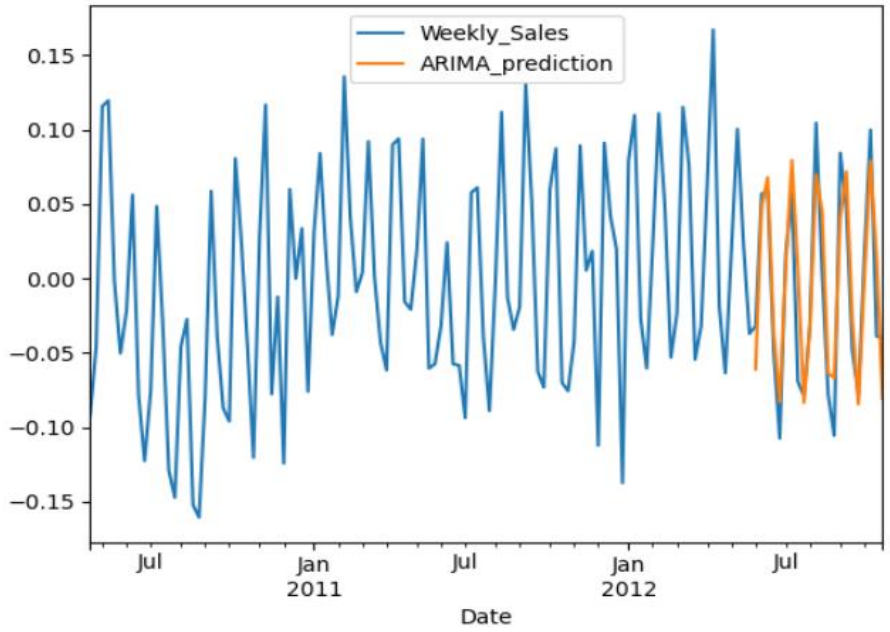
The model with the highest r2_score is the most efficient model for our predictions. The r2_score determines the proportion of variance in the dependent variable that can be explained by the independent variable.

Based on below Table, the highest r2_score is provided by the **Random Forest Machine model.**

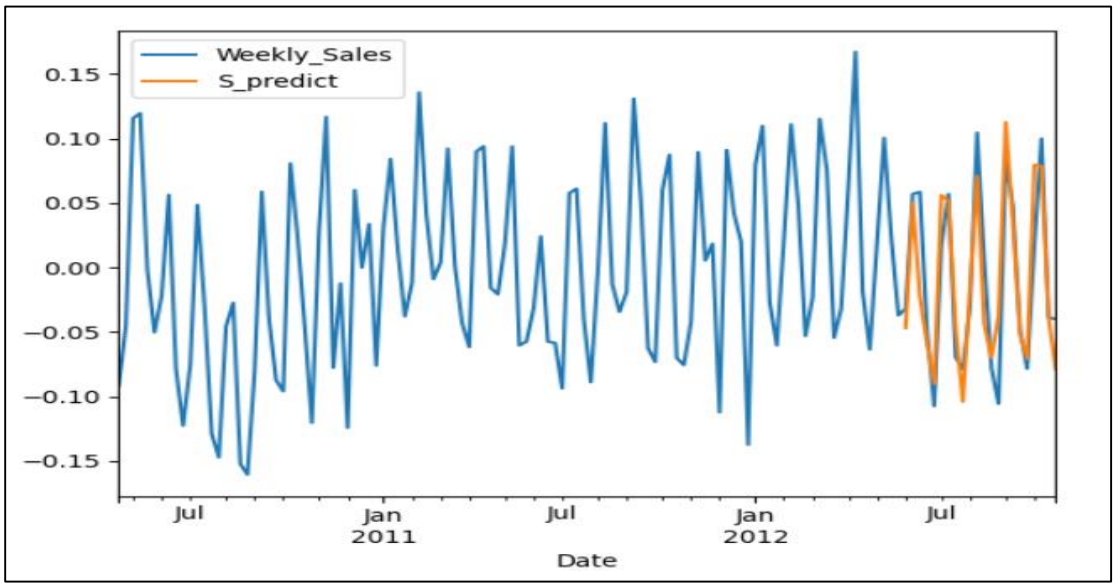
Model	R2 Score
RandomForestRegressor	0.959590
DecisionTreeRegressor	0.935618
KNeighborsRegressor	0.608030
LinearRegression	0.155532

6.2 Selection of the Most Accurate Forecast Model

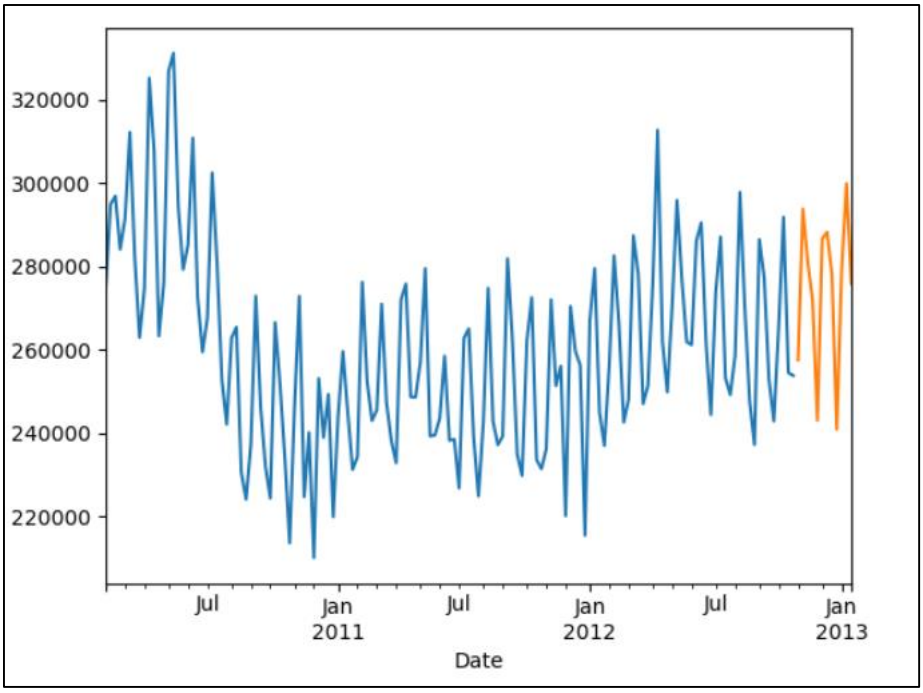
The below plot revels that the observed values from ARIMA compared to the rolling forecast predictions. Overall, our forecasts align with the true values very well.



However, if we check the prediction output of SARIMAX model is even more closer to the actual value.



Therefore, we will select SARIMAX model for our forecast prediction for next 12 weeks of the selected store.



7. Assumptions

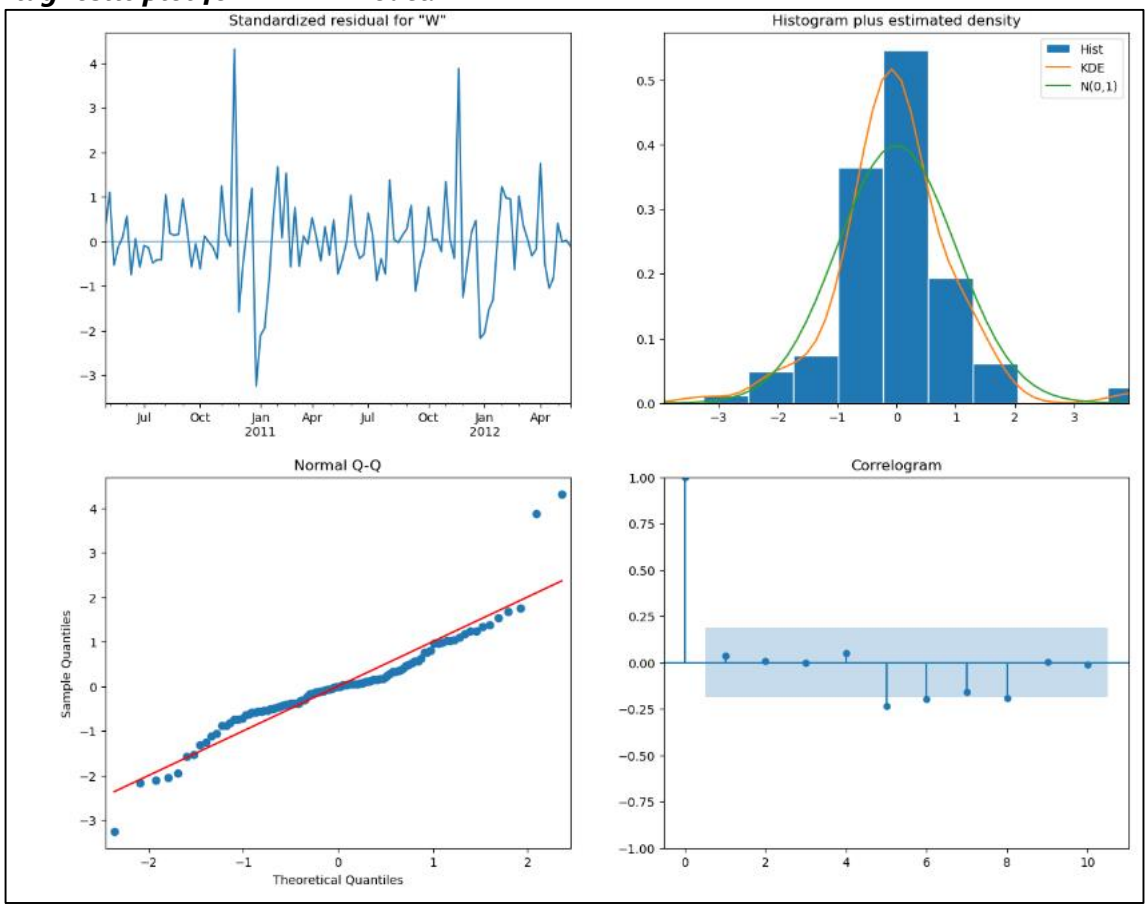
It is not possible to accurately forecast the sales for each store for the next 12 weeks using machine learning without additional information. Machine learning algorithms require data to be able to make predictions. This data could include historical sales data, customer demographics, store location, and other factors. Without this data, it is not possible to accurately forecast sales for each store for the next 12 weeks.

8. Model Evaluation and Techniques

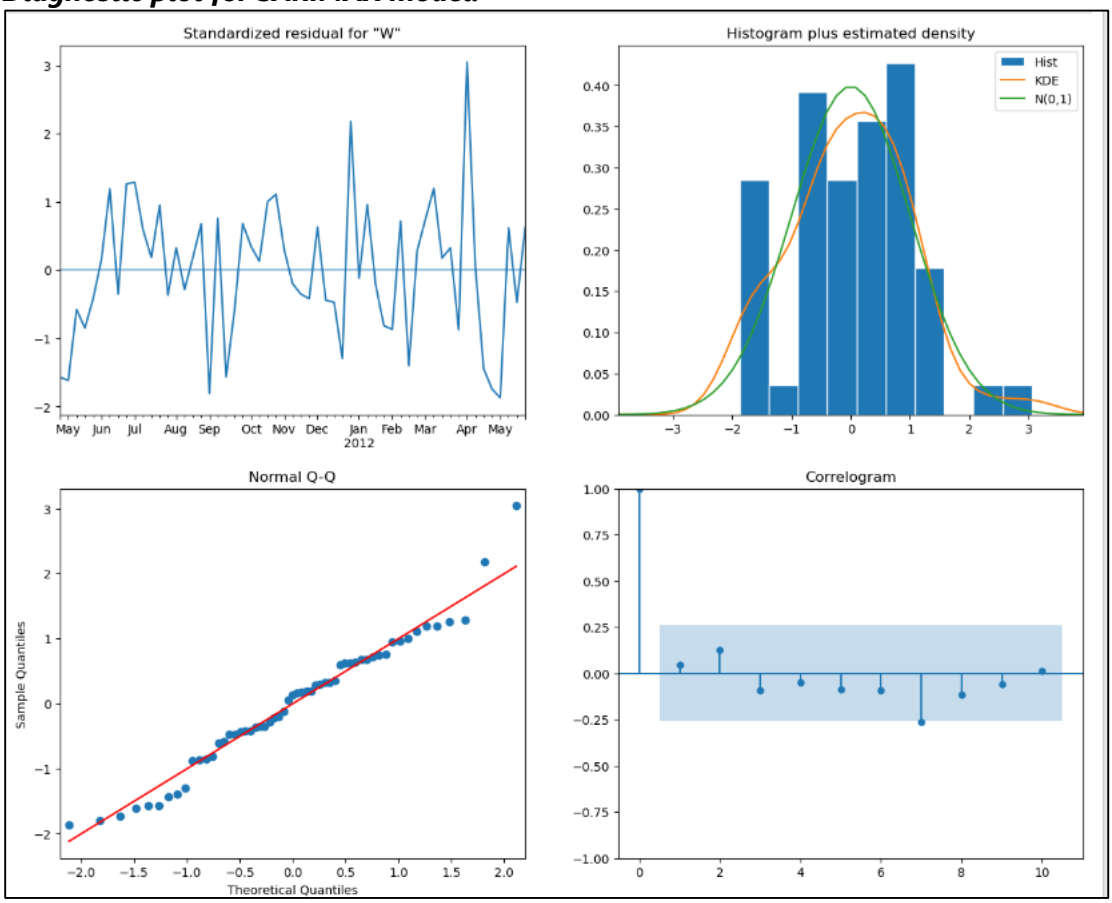
The most accurate way to forecast sales for each store using machine learning is to use a time series forecasting model. This type of model takes into account the historical sales data for each store and uses it to predict future sales. The model can be trained using a variety of techniques, such as ARIMA, SARIMAX, Prophet etc. Once the model is trained, it can be used to make predictions about future sales for each store. Additionally, the model can be evaluated using a variety of metrics like The Standardized residual, Histogram plus KDE estimate, Normal q-q, and the correlogram plot_diagnostics function.

If we compare the plot of both ARIMA and SARIMAX, we can observe that SARIMAX is a good fit for forecasting.

Diagnostic plot for ARIMA model:



Diagnostic plot for SARIMAX model:



In the SARIMAX diagnostics plots,

- The KDE curve should be very similar to the normal distribution.
- Most of the data points lie closer to the straight line (Normal Q-Q plot).
- In Correlogram (ACF plot) 95% of correlations for lag greater than zero should not be significant. The grey area is the confidence band, and if values fall outside of this then they are statistically significant. In SARIMAX model, there are a few values outside of this area. Therefore, SARIMAX model is performing much better than ARIMA model.

9. Inferences from the Same and Conclusion

The main purpose of this study was to predict Walmart's sales based on the available historical data and identify whether factors like temperature, unemployment, fuel prices, etc affect the weekly sales of particular stores under study. This study also aims to understand whether sales are relatively higher during holidays than on normal days so that stores can work on creating promotional offers that increase sales and generate higher revenue.

As observed through the exploratory data analysis, stores and holidays have a direct relationship with high Walmart sales. About the specific factors provided in the study (temperature, employment, CPI, and fuel price), it was observed that sales do tend to go up slightly during favourable climate conditions as well as when the prices of fuel are adequate. However, it is difficult to make a strong claim about this assumption considering the limited scope of the training dataset provided as part of this study. According to the observations in the exploratory data analysis, sales also tend to be relatively higher when the unemployment level is lower. Additionally, with the dataset

provided for this study, there does not seem to be a relationship between sales and the CPI index. Again, it is hard to make a substantial claim about these findings without the presence of a larger training dataset with additional information available.

Interaction effects were studied as part of the linear regression model to identify if a combination of different factors could influence the weekly sales for Walmart. Relationships between independent and target variables were tried to be identified through EDA components like the correlation matrix and scatter plots, feature importance plots created as part of the random forest and decision tree models as well as the interaction effects. It was discovered that, although there were no significant relationships between weekly sales and factors like temperature, fuel price, etc. in the correlation matrix, some significant relationships were observed between weekly sales and stores, CPI and unemployment in the feature importance plots created as part of the decision tree and random forest models. Finally, the random forest model, with the highest r2_score, is the main model used to create the final sales prediction for this study.

Another aspect that is worth exploring with this study is identifying trends in sales for each of the stores and predicting future trends based on the available sales data. Time series forecasting has been utilized (ARIMA and SARIMAX modelling) to predict future sales for each of the stores for 12 weeks. SARIMAX model gave us more accurate results which was supported by diagnostic plots and was used for the next 12 weeks' sales forecast.

11. Future Possibilities of the Project

Walmart may decide to concentrate more on the e-commerce side of the company as a result of developing technology and rising consumer demand. With already-existing stores, the company can more easily expand across the country, reducing the need for physical storefronts and enabling customers to save money on gasoline by having their purchases delivered right to their homes. Additionally, it makes identifying customer purchasing habits much simpler. This study's attempt to comprehend consumer purchasing behaviour based on departmental and regional sales is also quite relevant. The company can focus on profitable locations, build stronger customer relationships, create and distribute tailored messages for customers in those areas, and find methods to enhance products and services in those areas with the use of customer segmentation.

12. References

a. https://towardsdatascience.com/
b. https://github.com/
c. https://www.kaggle.com/
d. https://www.researchgate.net/publication
e. https://www.rit.edu/research