Predicting Walmart Sales, Exploratory Data Analysis, and Walmart Sales Dashboard

By

Smit Shah

A Project Report Submitted
in
Partial Fulfilment of the
Requirements for the Certification
of
Data Science and Artificial Intelligence
FROM

IIT- MADRAS

Contents

1.Problem Statement

A retail company with multiple outlet stores is having poor revenue returns from the stores with most of them facing bankruptcy. This project undertakes to review the sales records from the stores with a view to provide useful insights to the company and also to forecast sales outlook for the next 12-weeks.

2. Project Objective and Goal

Retail giants like Walmart consider this data as their biggest asset as this helps them predict future sales and customers and helps them lay out plans to generate profits and compete with other organizations. Walmart is an American multinational retail corporation that has almost 11,000 stores in over 27 countries, employing over 2.2 million associates (Wikipedia, n.d.)

Catering to their customers with the promise of ‘everyday low prices’, the range of products sold by Walmart draws its yearly revenue to almost 500 billion dollars thus making it extremely crucial for the company to utilize extensive techniques to forecast future sales and consequent profits. 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 study also 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. Walmart runs several promotional markdown sales throughout the year on days immediately following the prominent holidays in the United States; it becomes crucialfor the organization to determine the impact of these promotional offerings on weekly sales to drive resources towards such key strategic initiatives. 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 study can help the organization understand market conditions at various times of the year and allocate resources according to regional demand and profitability.

Additionally, the application of big data analytics will help analyze past data efficiently to generate insights and observations and help identify stores that might be at risk, help predict as well as increase future sales and profits and evaluate if the organization is on the right track.

The retail company with multiple outlets across the country are facing issues with inventory management. The task is to come up with useful insights using the provided data and make prediction models to forecast the sales the next twelve weeks.

3. Purpose Statement

The purpose of this study is to predict the weekly sales for Walmart based on available historical data (collected between 2010 to 2013) from 45 stores located in different regions around the country. The main deliverable is to predict the weekly sales for all such departments. The data contains the weekly sales for 45 stores, the amount of weekly sales, and whether the week is a holiday week or not. There is additional information in the dataset about the factors that might influence the sales of a particular week. Factors like Consumer Price Index (CPI), temperature, fuel price, promotional markdowns for the week, and unemployment rate have been recorded for each week to try and understand if there is a correlation between the sales of each week and their determinant factors.

4. About the Dataset

It contains historic weekly sales information about 45 Walmart stores across different regions in the country along with department-wide information for these stores. The main goal of this study is going to be to predict the department-wide weekly sales for each of these stores.

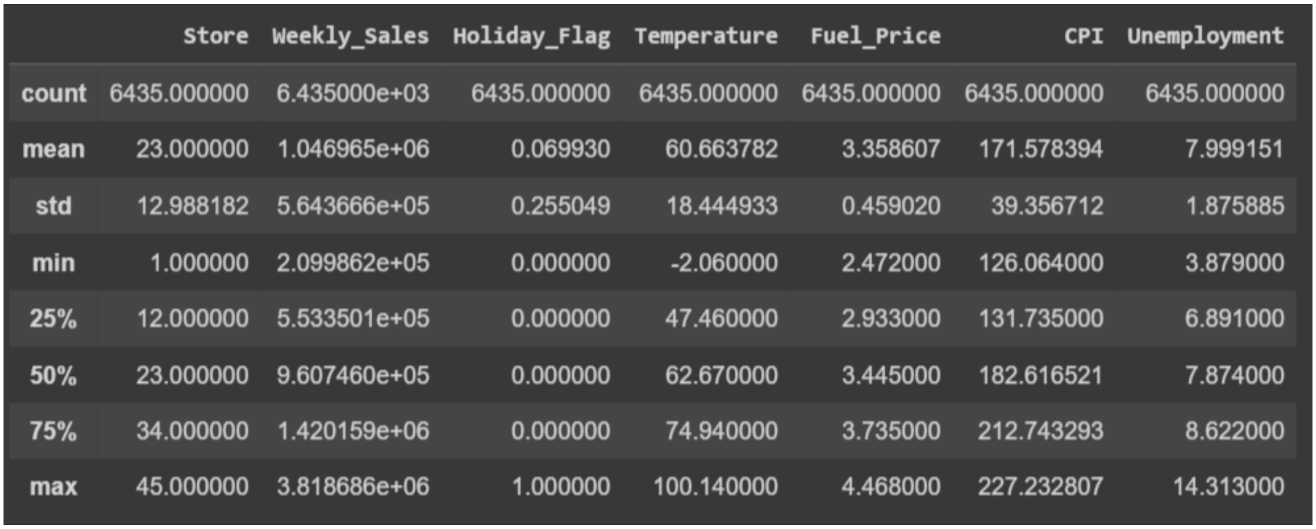
4.1 Dataset statistics

# Number of variables / features : 8

# Number of observations/ rows : 6435

# Variable types :

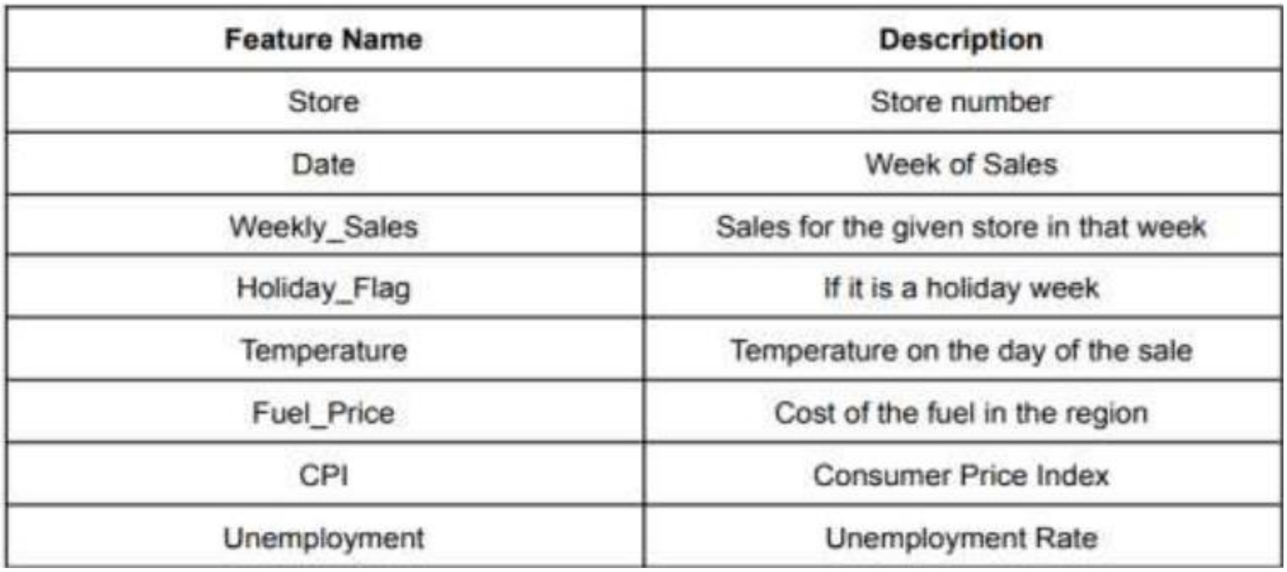
1. **Numeric Features : 6**
2. **Datetime Features : 1**
3. **Categorical Features : 1**



	Store	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
count	6435.000000	6.435000e+03	6435.000000	6435.000000	6435.000000	6435.000000	6435.000000
mean	23.000000	1.046965e+06	0.069930	60.663782	3.358607	171.578394	7.999151
std	12.988182	5.643666e+05	0.255049	18.444933	0.459020	39.356712	1.875885
min	1.000000	2.099862e+05	0.000000	-2.060000	2.472000	126.064000	3.879000
25%	12.000000	5.533501e+05	0.000000	47.460000	2.933000	131.735000	6.891000
50%	23.000000	9.607460e+05	0.000000	62.670000	3.445000	182.616521	7.874000
75%	34.000000	1.420159e+06	0.000000	74.940000	3.735000	212.743293	8.622000
max	45.000000	3.818686e+06	1.000000	100.140000	4.468000	227.232807	14.313000

4.2 Feature Description

The available dataset contains 6,435 records (rows) and eight features (columns) as shown in the Table 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

>>> **From the given dataset of the company, it is observed that the data consists of six thousand four hundred and thirty-five (6,435) records with eight features (captured weekly) as follows:**

#. Stores: There are 45 stores and each store has 143 recorded entries of:

a. Date of record (weekly),

b. Total sales record for the week,

c. Holiday flag for the week (1 or 0),

d. Temerature: average temperature recorded during the week,

e. Fuel Price: average fuel price for the week

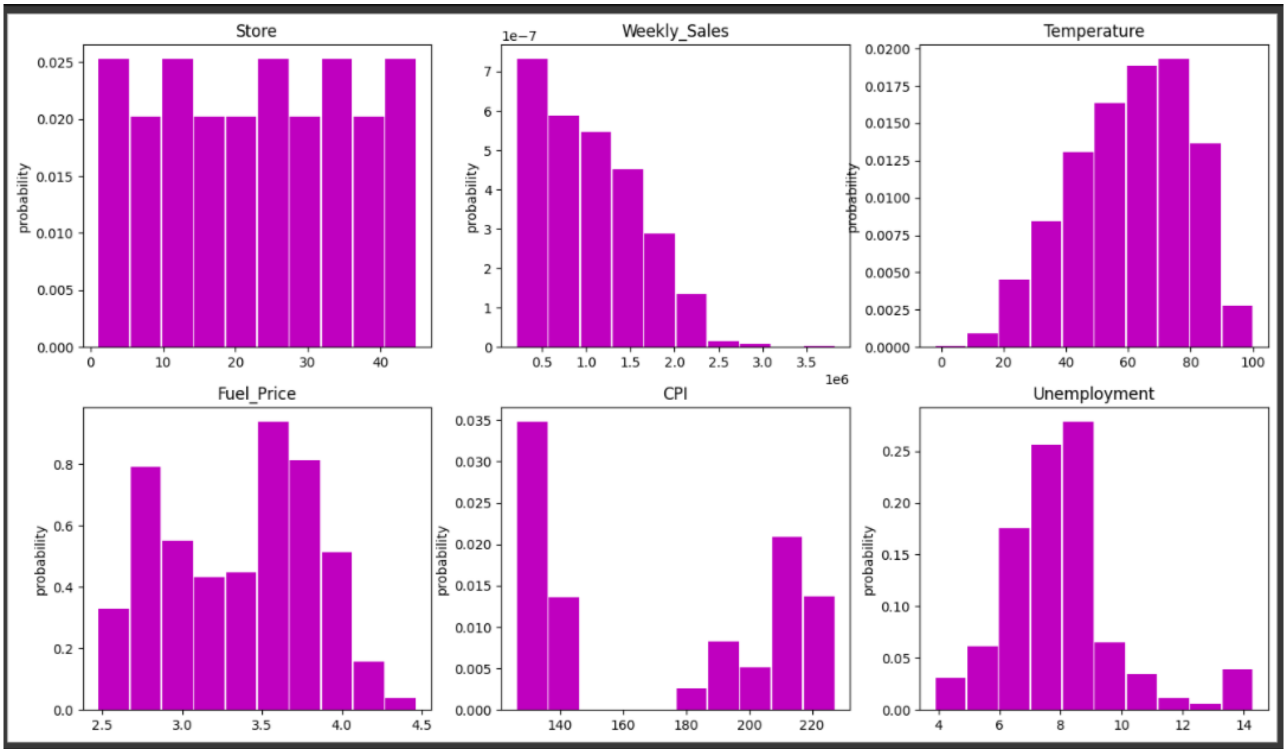
f. CPI: average Consumer Price Index for the week

g. Unemployment: rate of unemployment for the week of record

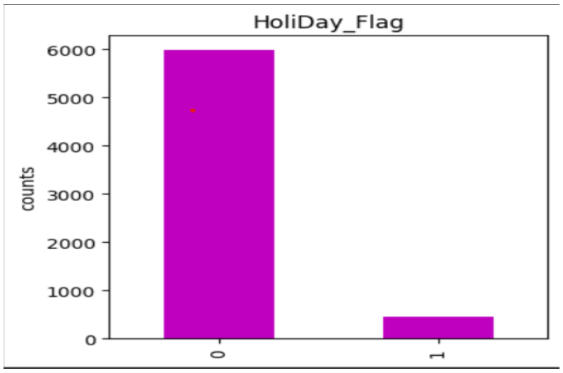
5. Data Pre-processing and Exploratory Data Analysis (EDA)

There are no missing or duplicate values in the dataset at all.

5.1 Histogram for Numerical Features



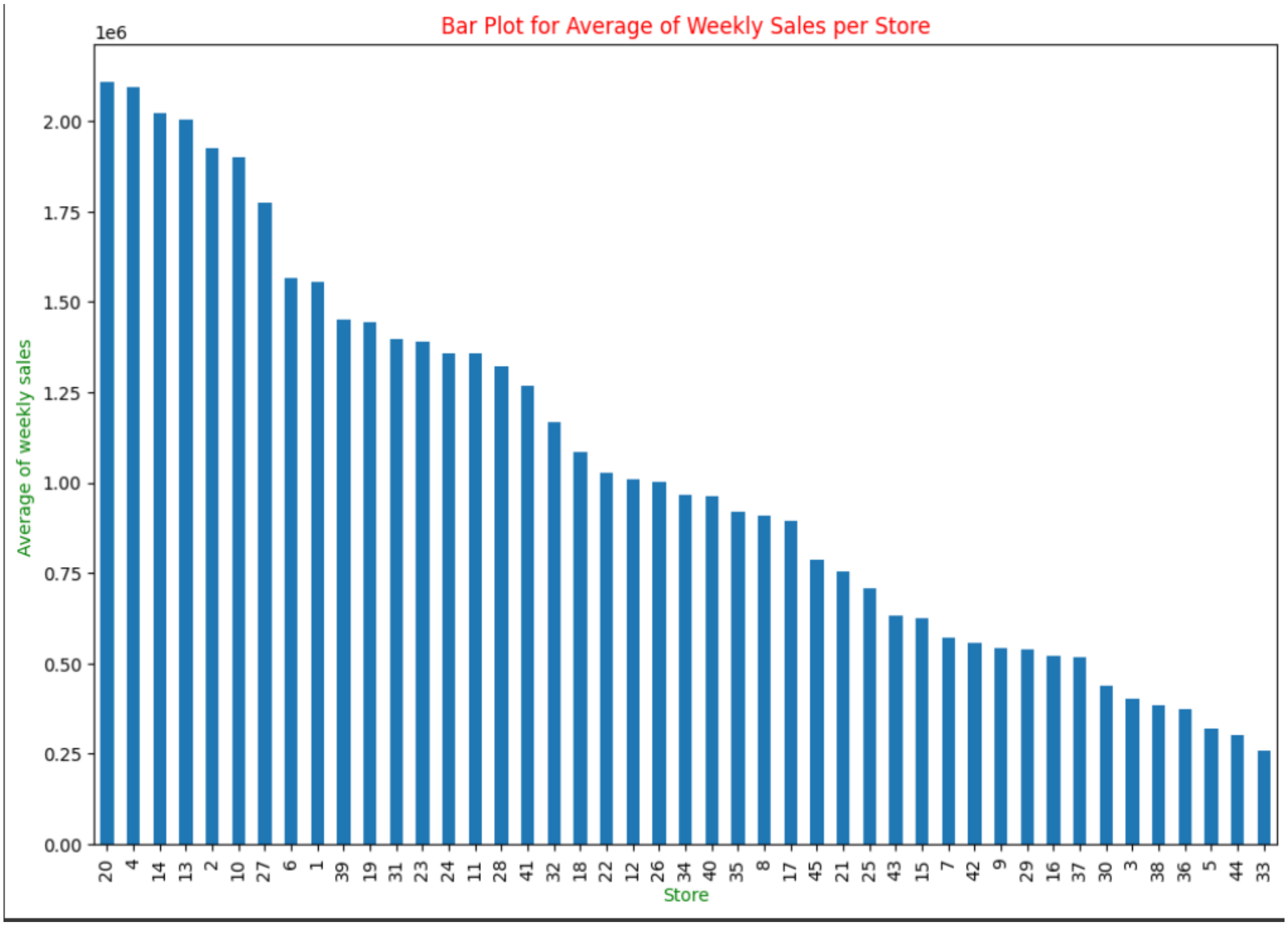
Count plot for column Holiday Flag



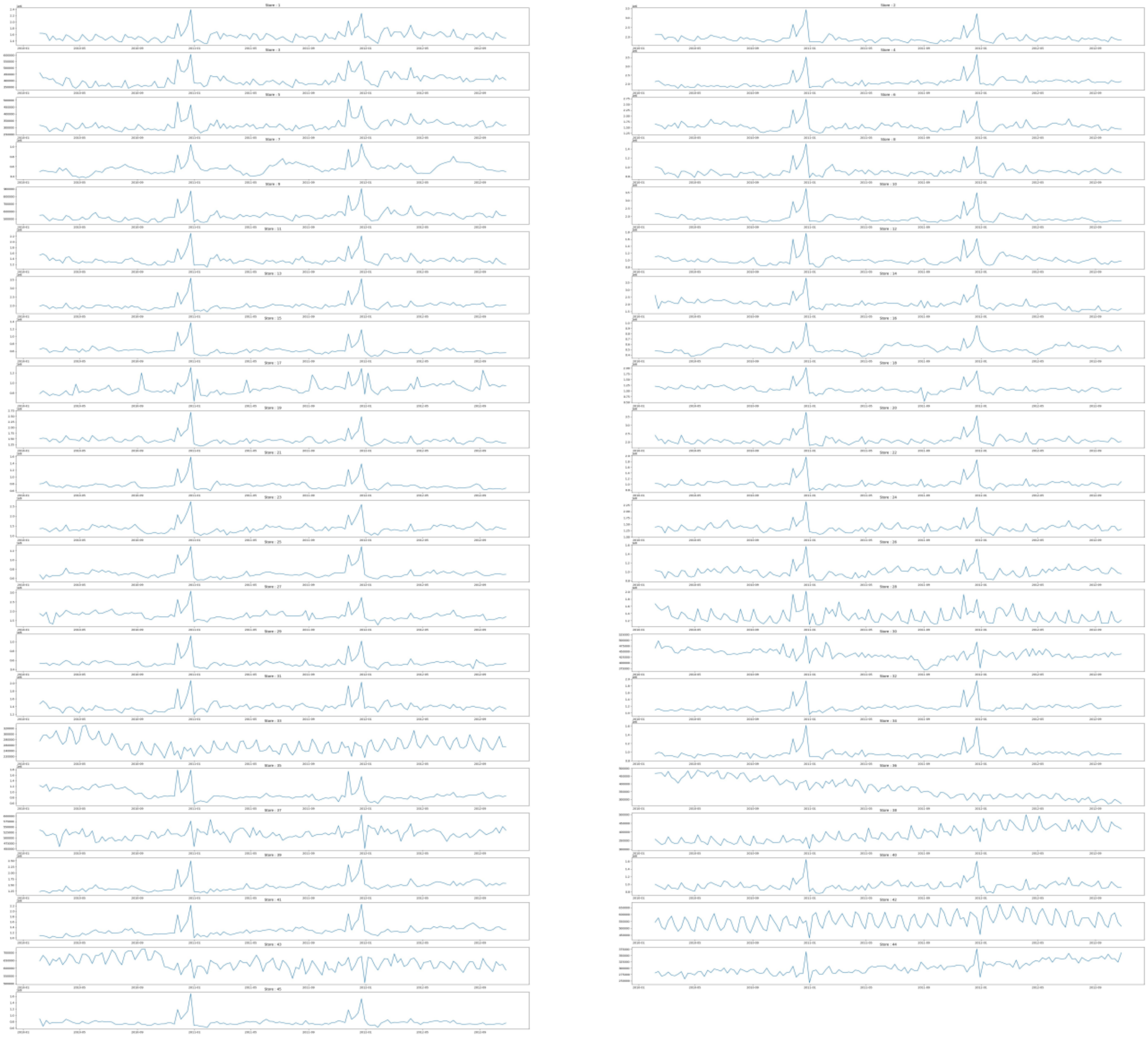
5.2 The Average of Weekly Sales per store

Here, the Sales are demonstrated in a descending order. In total there are 45 stores presented in the Walmart dataset and from the image below it can be concluded that store numbers 4, 14, and 20 have the highest average sales. We can infer that the total weekly sales in the stores are not uniform. In some stores the total sales are much higher and in some stores the total sales are effectively low.It should also be noted that there is a very high difference between the average sales for each of the stores; while some stores record huge sales, some others lack vastly in the area.

This could be dependent on factors like the kinds of products sold by the store,the geographic location, temperature, unemployment in the vicinity, etc.It is necessary to take steps to increase the sales for the stores where the sales are effectively low.



5.3 Weekly Sales for each store over a time period.

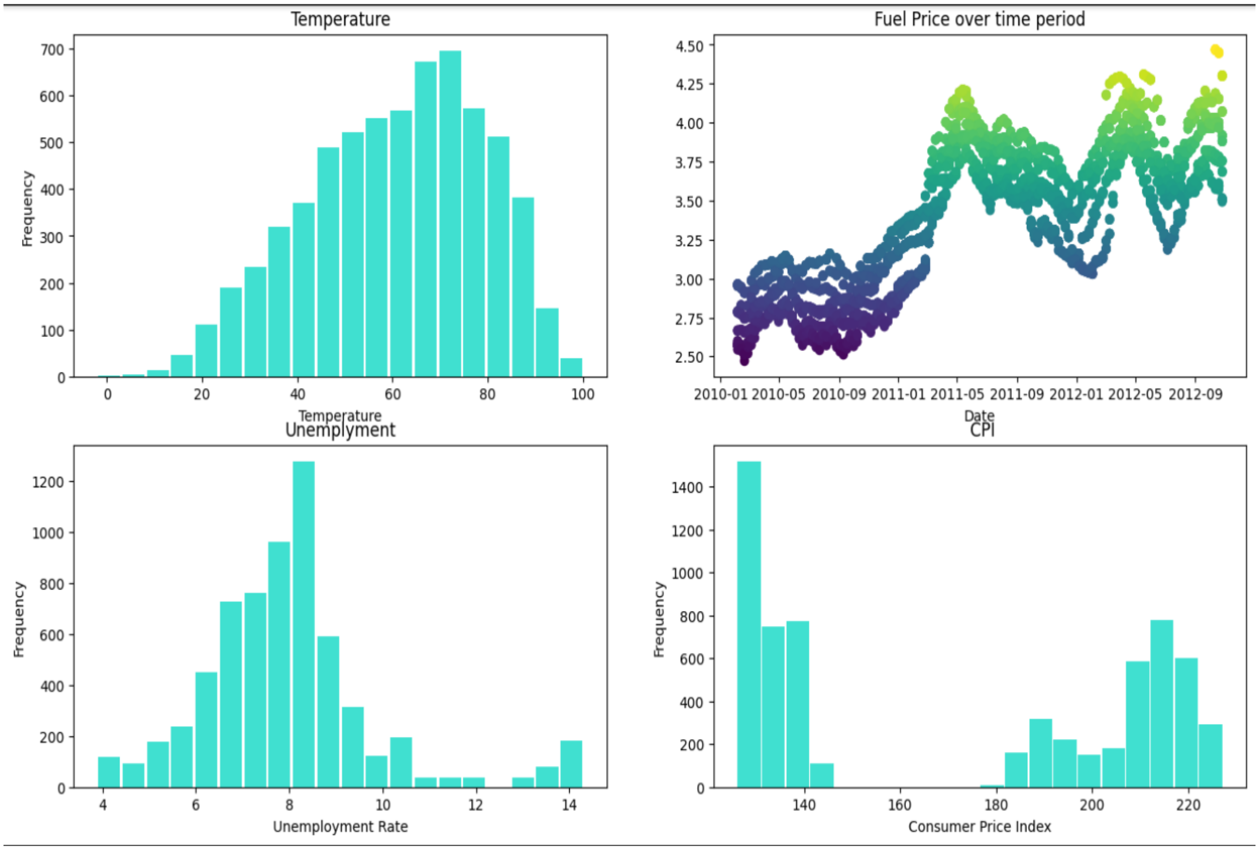


Here, in the weekly sales:

In majority of the stores the weekly sales of month of Jan in each year is more than the other months of the year. As the pattern is repeating each year this indicates the sales follows the seasonal pattern in the months of Dec -Jan.

6. Data Visualization of Exogenous Features

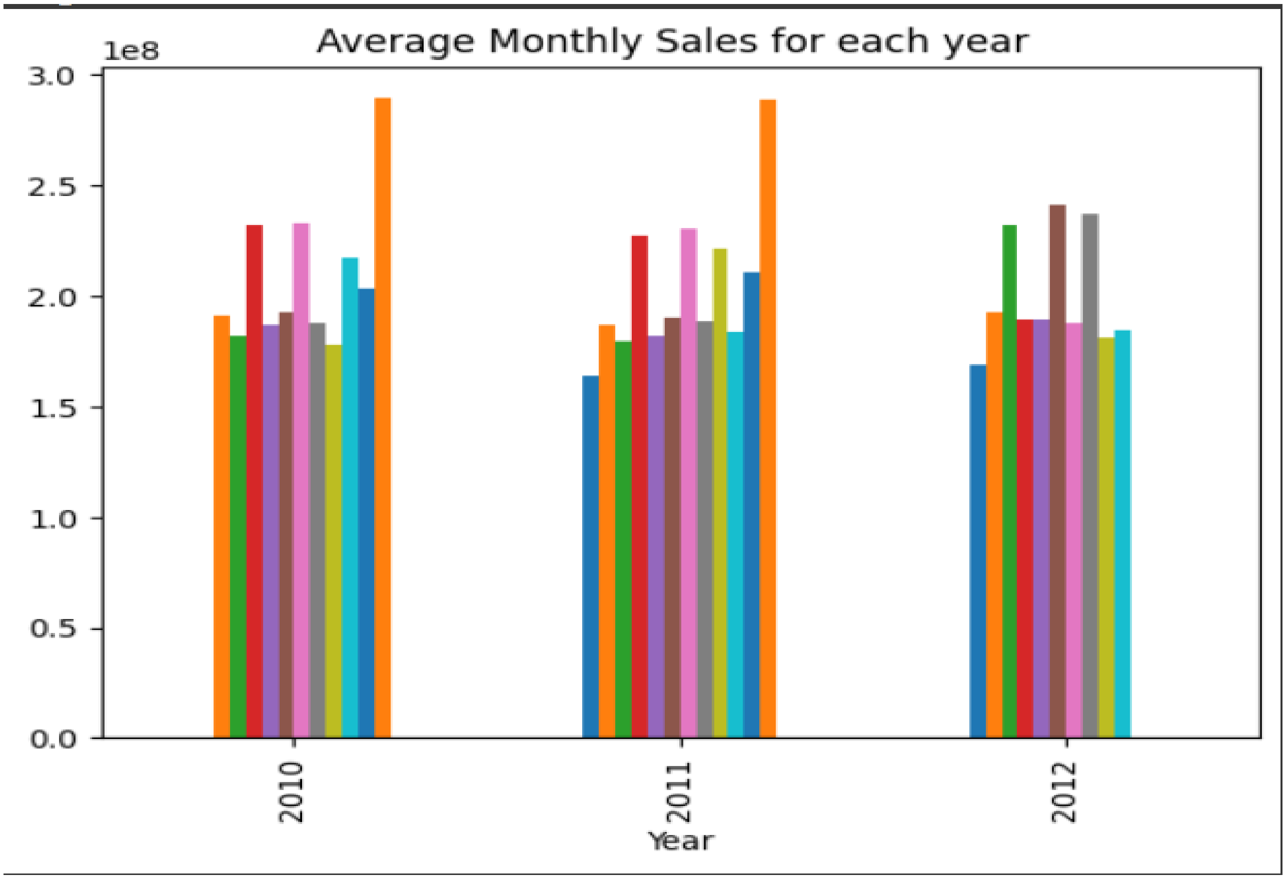
Temperature and unemployment rate fairly normally distributed. Fuel price has increased over time period. We might need to find some insights of the data over a weekly-sales.



6.1 Identifying Monthly Sales for Each Year

With the holiday information provided in the original dataset, it is known that the major holidays fall at the end of the year.

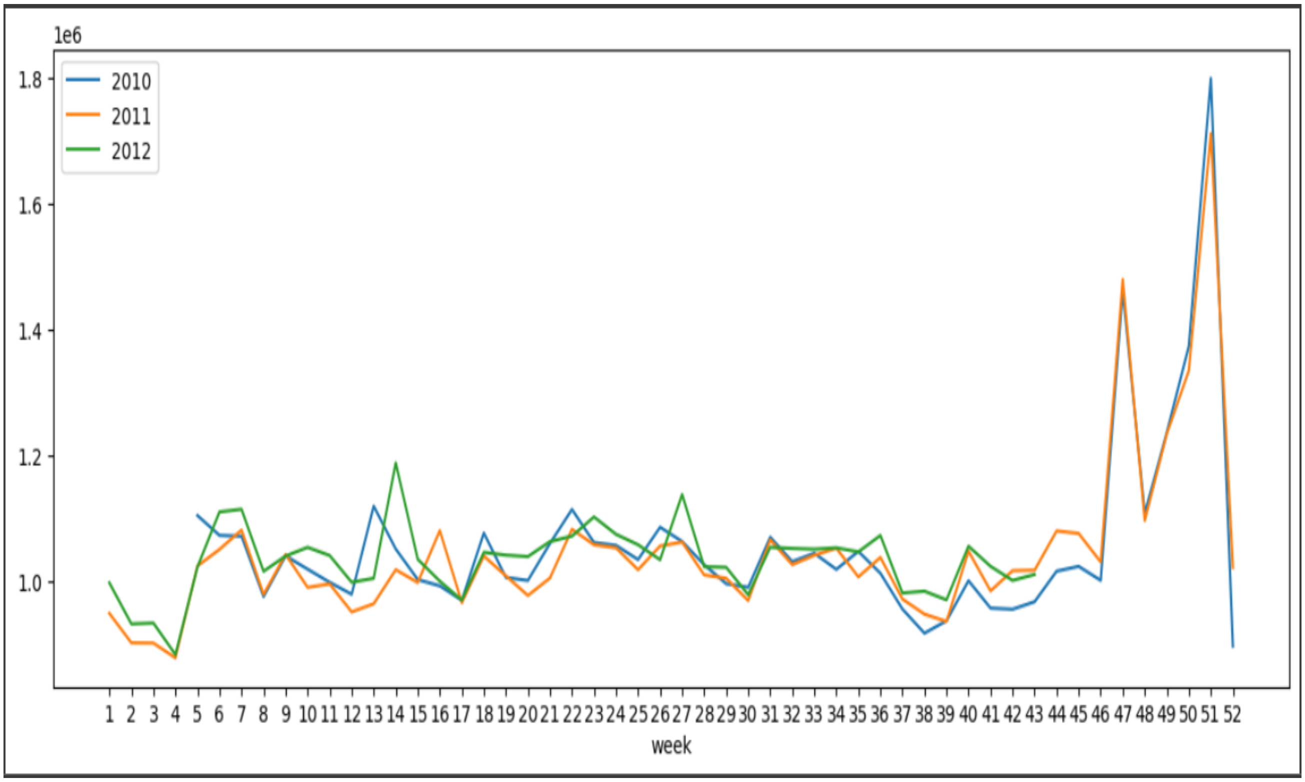
The graph below clearly depicts that the months of december recorded the highest average sales for 2010 and 2011. The dataset provided by Walmart contained no weekly sales information for the last two months of the year 2012, hence no conclusion can be drawn for that year.



6.2 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 during holiday weeks each year

There is an evident hike in sales in weeks 47 and 51 that correspond to 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 tells that there is a distinguished pattern of decline immediately following Christmas and New Year's.

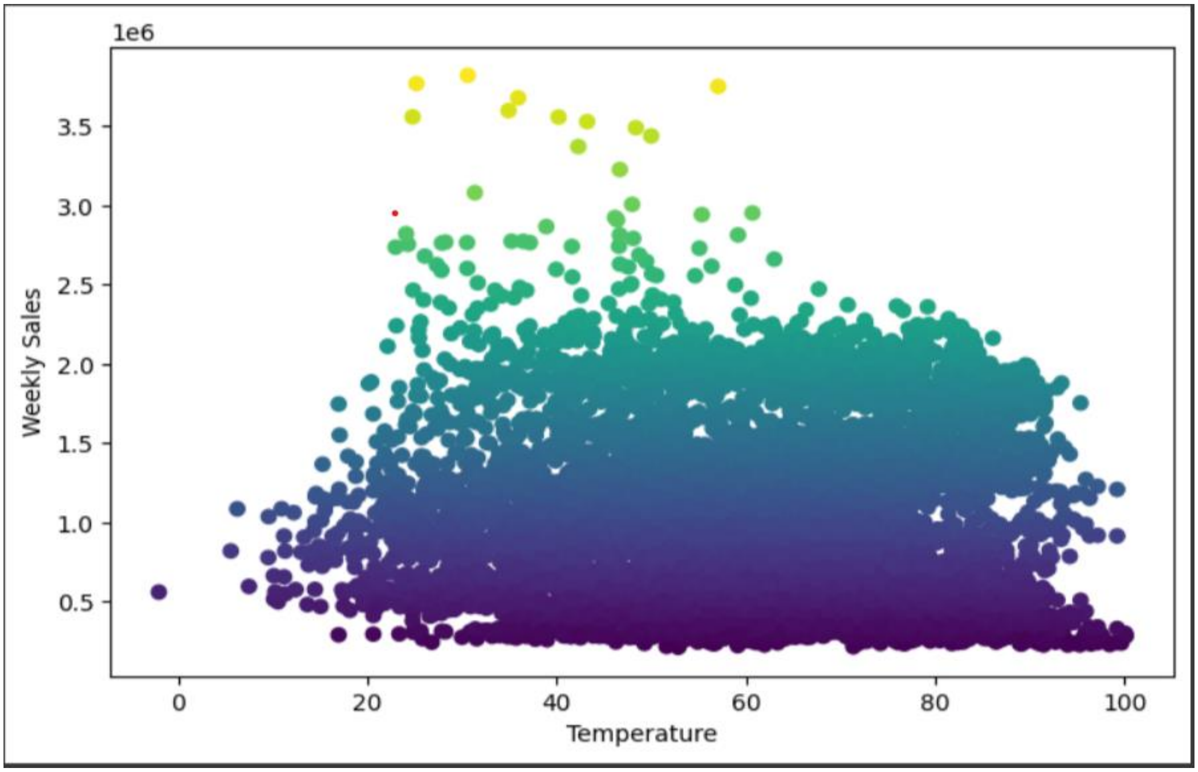


After studying the overall sales summaries of different components of the Walmart dataset, this report will now throw light upon the effect of different factors (such as holidays, markdowns, CPIs, unemployment, etc.) on the weekly sales. It has always been an integral part of this study to understand the effect that these factors have on Walmart's sales performance. I will also create several visualizations that shed light on the difference in Walmart store sales on holidays versus non-holiday days, the impact on weekly sales, and finally create a correlation matrix to examine the correlation between the many factors included in the study

6.3 Impact of Temperature on Sales

It has widely been known in the retail sector that weather has a profound effect on sales. While warmer weather promotes sales, cold/harsh or extremely hot weather is generally not a great encouragement for shoppers to get outdoors and spend money. Generally speaking, temperatures between 40 to 70 degrees Fahrenheit are considered as favorable for humans to live in considering they are not as hot or cold.

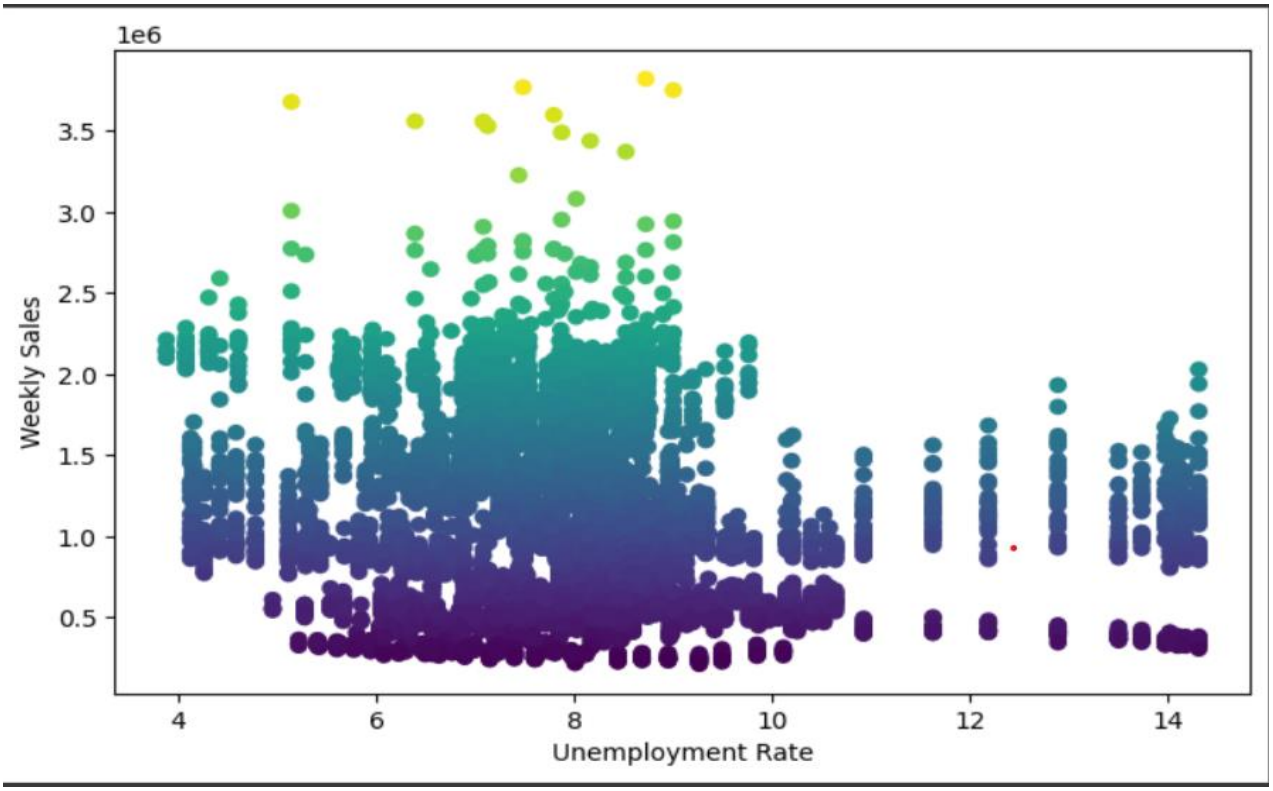
As seen below, the highest sales occur for most store types between the range of 25 to 60 degrees Fahrenheit, thus proving the idea that pleasant weather encourages higher sales. Sales are relatively lower for very low and very high temperatures but seem to be adequately high for favorable climate conditions.



6.4 Impact of Unemployment on Sales

Spending sharply drops on the onset of unemployment; a higher unemployment index would generally result in a dip in sales as individuals tend to decrease overall spending. In our dataset, unemployment is presented through an index of the unemployment rate during that week in the region of the store. From our scatter plot, it is easier to gather the following information:

For the given store sales, there seems to be a visible decrease in sales when the unemployment index is higher than 9

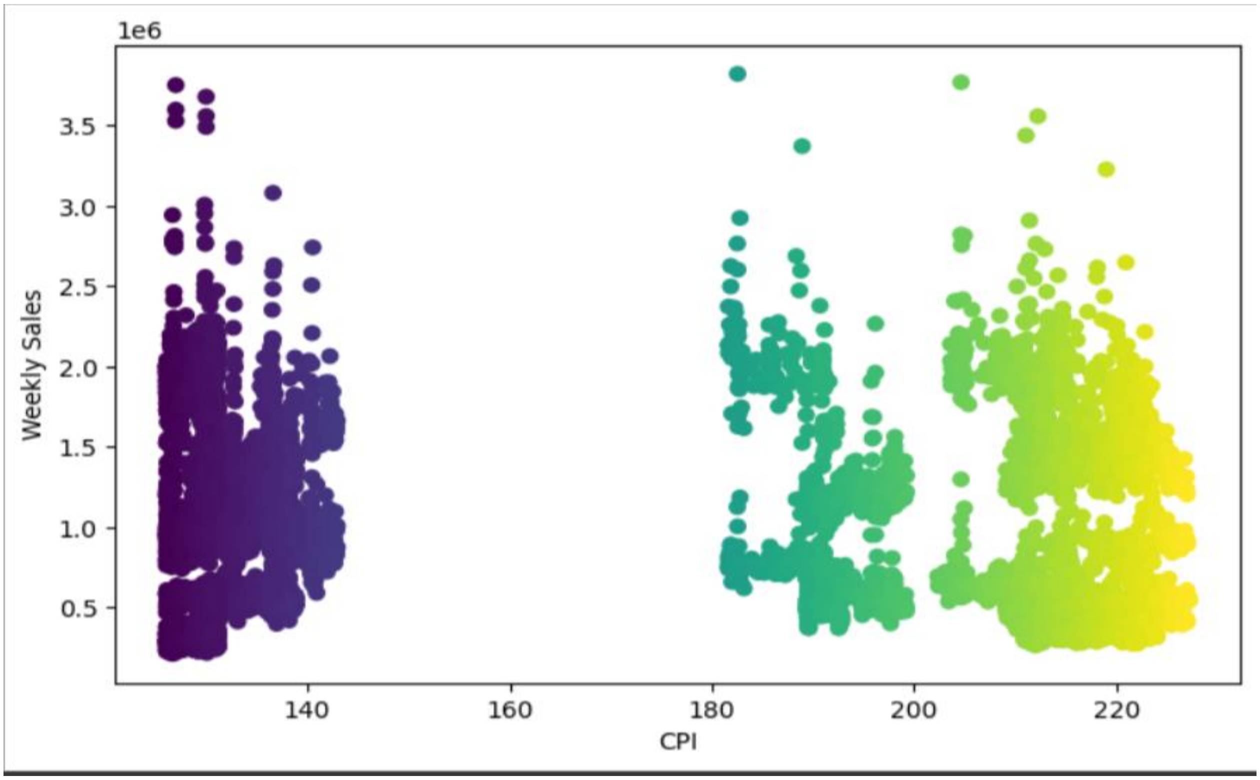


6.5 Impact of CPI on Sales

CPI (Consumer Price Index) is defined as the measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services.

In layman's terms, CPI is a measure that assesses the price changes that are associated with the cost of living for each individual. CPI is a great measure for the government when studying inflation (i.e. an increase in the prices of a representative basket of goods consumed) and is often used to evaluate and adjust government assistance needs based on income levels and provide wage adjustments with respect to changing cost of living expenses. A higher CPI generally means that the price of goods has increased and that an individual needs to spend more money to maintain the same standard of living.

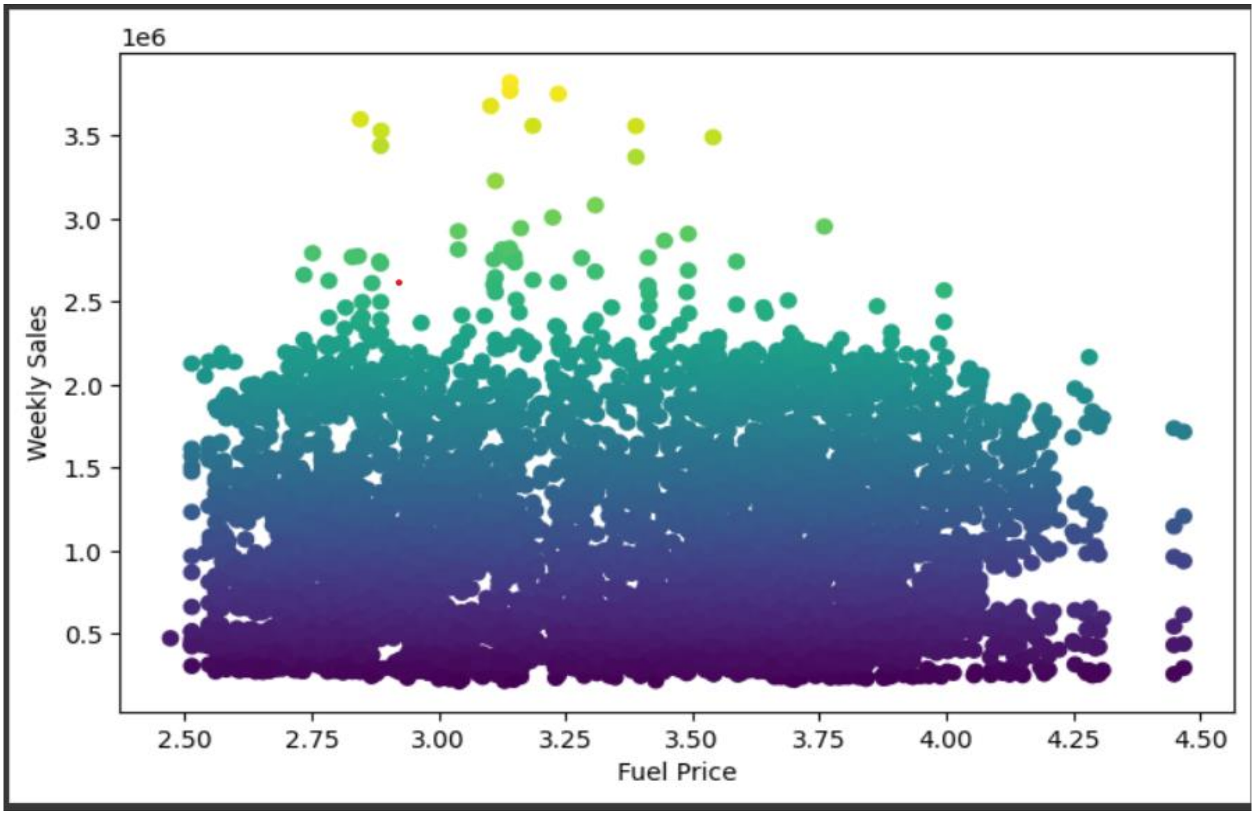
In our scatter plot above, we can identify three different clusters around different ranges of CPI; while there seems to be no visible relationship between the change in CPI and weekly sales for Walmart stores (sales still occur at high CPI rates)



6.6 Impact of Fuel Price on Sales

The economist assumes that even a slight increase in fuel prices significantly adds up to the annual expenditure and thus discourages consumers from actively buying their required goods and services.

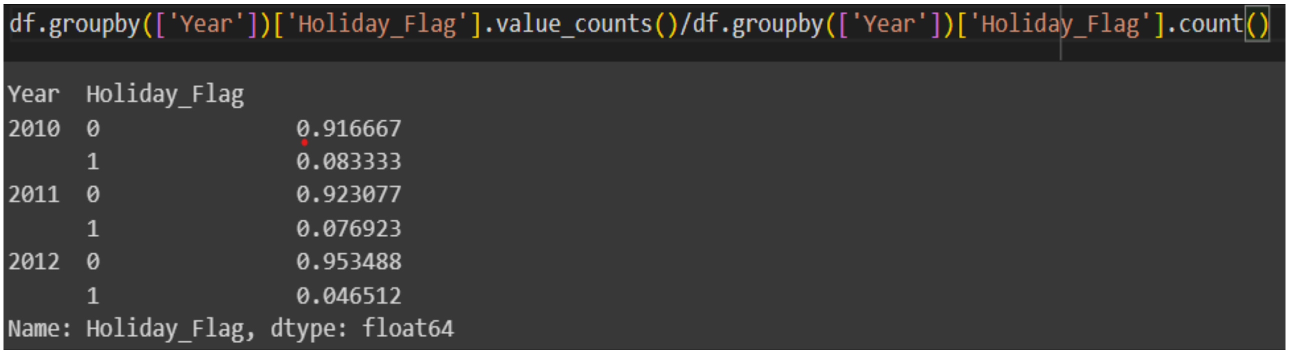
This can also slightly be observed in the visualization below; while there seems to be a decrease in sales when fuel price is higher than 4.00 dollars, sales are higher when fuel price ranges between 2.75 to 3.75 dollars.



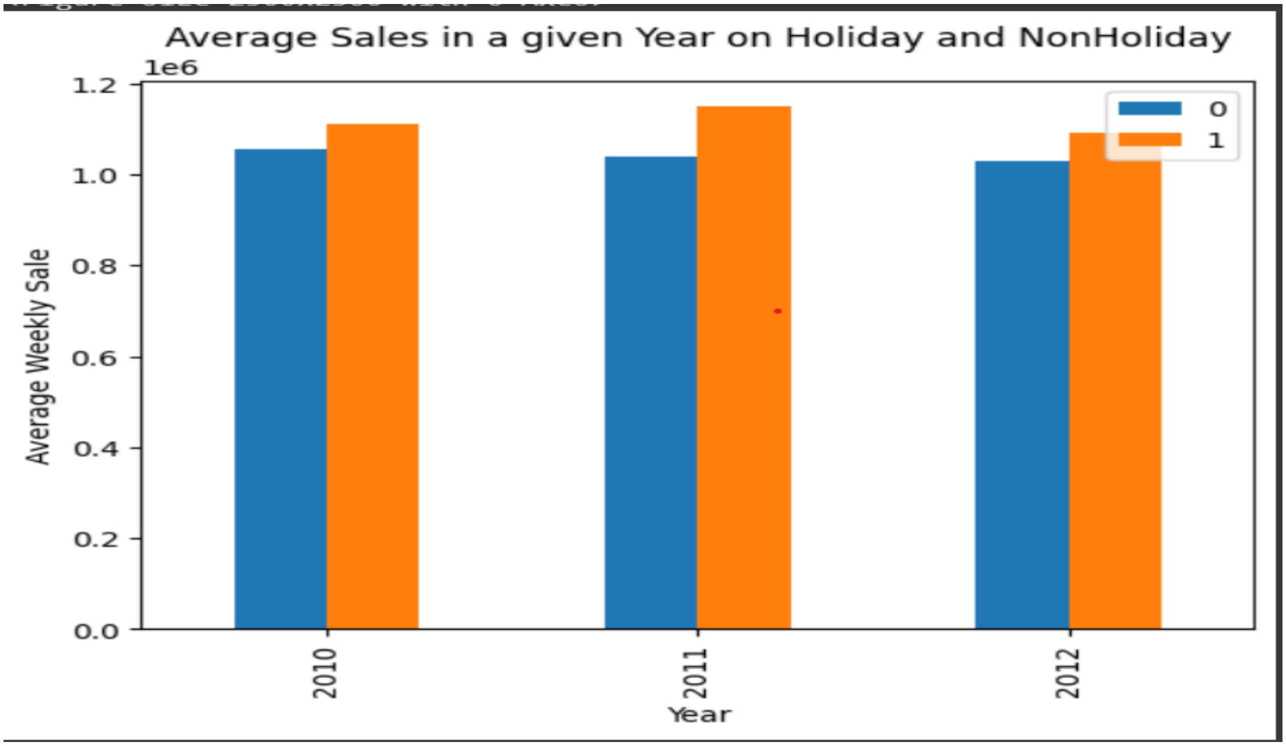
6.7 Holiday VS Non-Holiday Sales

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 7 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).



```
df.groupby(['Year'])['Holiday_Flag'].value_counts()/df.groupby(['Year'])['Holiday_Flag'].count()

Year  Holiday_Flag
2010  0               0.916667
      1               0.083333
2011  0               0.923077
      1               0.076923
2012  0               0.953488
      1               0.046512
Name: Holiday_Flag, dtype: float64
```



7. Correlation Matrix

A correlation matrix describes the correlation between the various variables of a dataset. Each variable in the table is correlated to each of the other variables in the table and helps in understanding which variables are more closely related to each other.

With the numerous variables available through this dataset, it became imperative to study correlations between some of them. By default, this matrix also calculates correlation through Pearson’s Correlation Coefficient (Johnson, 2021) that calculates the linear relationship between two variables, within a range of -1 to +1. The closer the correlation to |1|, the higher the linear relationship between the variables and vice versa.

Pearson’s Correlation Coefficient (Johnson, 2021) is calculated as:

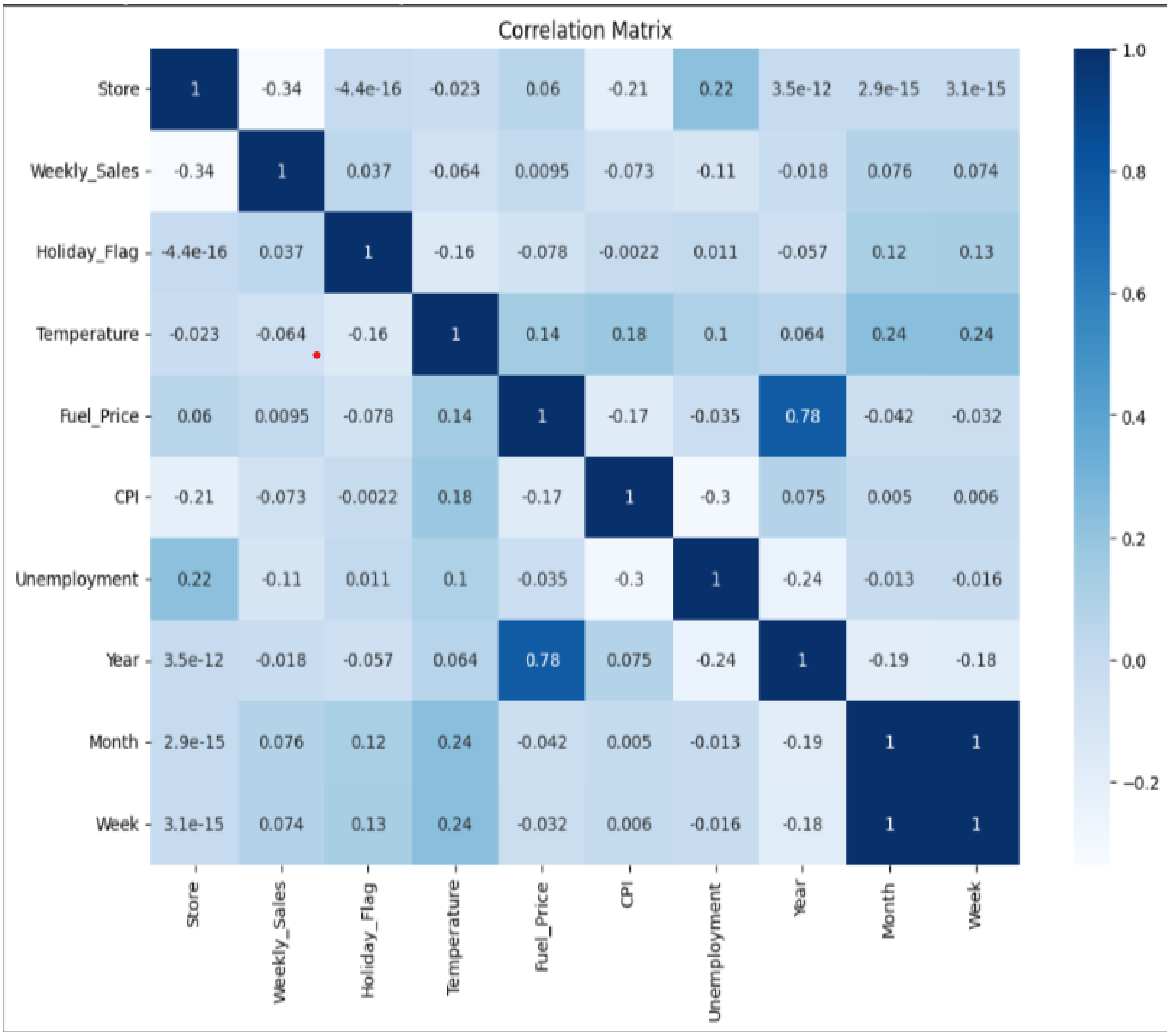
$r = \mathrm{Cov}(x, y) / \sigma x \sigma y$ ......Eq.

with

- $\mathrm{Cov}(x, y) = \sum(x-x)(y-\bar{y}) / n-1$
- $\sigma x = \sum(x-\bar{x})2$
- $\sigma y = \sum(y-\bar{y})2$

The heatmap/correlation matrix in figure, created using the seaborn library in Python gives the following information:

- There seems to be a neglected correlation between weekly sales and temperature, unemployment, CPI, and fuel price. This could suggest that sales are not impacted by changes in these factors.



	Store	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	Year	Month	Week
Store	1	-0.34	-4.4e-16	-0.023	0.06	-0.21	0.22	3.5e-12	2.9e-15	3.1e-15
Weekly_Sales	-0.34	1	0.037	-0.064	0.0095	-0.073	-0.11	-0.018	0.076	0.074
Holiday_Flag	-4.4e-16	0.037	1	-0.16	-0.078	-0.0022	0.011	-0.057	0.12	0.13
Temperature	-0.023	-0.064	-0.16	1	0.14	0.18	0.1	0.064	0.24	0.24
Fuel_Price	0.06	0.0095	-0.078	0.14	1	-0.17	-0.035	0.78	-0.042	-0.032
CPI	-0.21	-0.073	-0.0022	0.18	-0.17	1	-0.3	0.075	0.005	0.006
Unemployment	0.22	-0.11	0.011	0.1	-0.035	-0.3	1	-0.24	-0.013	-0.016
Year	3.5e-12	-0.018	-0.057	0.064	0.78	0.075	-0.24	1	-0.19	-0.18
Month	2.9e-15	0.076	0.12	0.24	-0.042	0.005	-0.013	-0.19	1	1
Week	3.1e-15	0.074	0.13	0.24	-0.032	0.006	-0.016	-0.18	1	1

(Correlation Matrix)

8. Model Selection and Implementation

Trying to find and implement the most effective model is the biggest challenge of this study. Selecting a model will depend solely on the kind of data available and the analysis that has to be performed on the data (UNSW, 2020)

The main purpose of the project to forecast the sales of the stores for a given dataset using Time Series models ARIMA, SARIMAX, etc.

8.1 Brief on Time Series Analysis

Time series analysis is a statistical technique used to analyse and interpret data points collected over time. It involves studying the pattern and behaviour of data, identifying trends, seasonality, and any other underlying patterns that may exist within the dataset.

In a time series, the data points are ordered chronologically, typically with a constant time interval between observations. Time series data can be found in various fields such as economics, finance, meteorology, and many others.

Time series analysis is a statistical technique used to analyze and interpret data points collected over time. It involves studying the pattern and behavior of data, identifying trends, seasonality, and any other underlying patterns that may exist within the dataset.

In a time series, the data points are ordered chronologically, typically with a constant time interval between observations. Time series data can be found in various fields such as economics, finance, meteorology, and many others.

Some key concepts in time series analysis include:

1. Trend: The long-term movement or direction of the data. It indicates whether the data is increasing, decreasing, or remaining relatively constant over time.

2. Seasonality: Patterns that repeat at regular intervals within a time series. For example, sales of ice cream might show a seasonal pattern with higher sales during the summer months.

3. Cyclical patterns: Longer-term patterns that are not necessarily of fixed duration. These patterns often occur in business cycles or economic cycles

and can span multiple years.

4. Stationarity: A time series is said to be stationary if its statistical properties, such as mean and variance, do not change over time. Stationary time series are easier to analyse and model.
5. Autocorrelation: The correlation between observations at different time points within a time series. Autocorrelation helps identify dependencies and patterns in the data.
6. Forecasting: Predicting future values of a time series based on its historical patterns and trends. Forecasting models can be used to estimate future behaviour and make informed decisions.

   Time series analysis techniques include data visualization, decomposition, smoothing, autocorrelation analysis, and various statistical models such as autoregressive integrated moving average (ARIMA), exponential smoothing, and seasonal decomposition of time series.

These techniques allow analysts to gain insights, make predictions, and identify anomalies or outliers in the data, enabling informed decision-making and effective planning in various domains.

To predict the weekly sales for different Walmart stores the following two time series models have been used:

1. ARIMA
2. SARIMAX

Each of these methods have been discussed briefly in the upcoming report.

8.2 ARIMA (Auto Regressive Integrated Moving Average)

ARIMA (Autoregressive Integrated Moving Average) is a popular and widely used time series forecasting model. It is a combination of three components: autoregression (AR), differencing (I), and moving average (MA).

1. Autoregression (AR): The autoregressive component of ARIMA predicts the future values of a time series based on its own past values. It assumes that the current value of the series is linearly related to its previous values. The "p" parameter in ARIMA represents the order of the autoregressive component, indicating how many past values are used for prediction.

2. Differencing (I): The differencing component of ARIMA is used to make the time series stationary. Stationarity is a property where the statistical properties of a time series, such as mean and variance, do not change over time. If a time series exhibits trends or seasonality, differencing can be applied to eliminate them. The "d" parameter in ARIMA represents the order of differencing required.
3. Moving Average (MA): The moving average component of ARIMA models the error or residual terms of the time series. It assumes that the current value of the series depends on the average of the past error terms. The "q" parameter in ARIMA represents the order of the moving average component, indicating the number of lagged forecast errors used for prediction.

The ARIMA model combines these three components to capture the autoregressive, differencing, and moving average properties of a time series. It is typically denoted as ARIMA(p, d, q), where "p" represents the order of the autoregressive component, "d" represents the order of differencing, and "q" represents the order of the moving average component.

The parameters p, d, and q are determined based on the characteristics of the time series, such as its trend, seasonality, and autocorrelation.

However, they may not be suitable for all types of time series, and other models like SARIMA (Seasonal ARIMA) or more advanced techniques may be needed to handle complex patterns and seasonality.

SARIMA stands for Seasonal Autoregressive Integrated Moving Average, and it is a popular time series forecasting model that combines autoregressive (AR), differencing (I), moving average (MA), and seasonal components. The SARIMA model is an extension of the ARIMA (AutoRegressive Integrated Moving Average) model, which includes a seasonal component to handle time series data with seasonal patterns.

8.3 SARIMA (Seasonal Auto Regressive Integrated Moving Average)

The components of SARIMA are as follows:

1. Autoregressive (AR) component: The AR component captures the relationship between the current value and its previous values. It models the dependence of the current value on its own lagged values.
2. Integrated (I) component: The I component represents the differencing of the time series data to make it stationary. Stationarity means that the statistical properties of the series do not depend on time, making it easier to model.
3. Moving Average (MA) component: The MA component models the relationship between the current value and the residual errors from past predictions. It helps to capture the influence of past errors on the current value.
4. Seasonal component: The seasonal component is an extension of the ARIMA model to handle data with seasonal patterns. It captures the seasonal patterns by including seasonal lags (lags of the series at previous seasonal periods) in the model.

The SARIMA model is defined by three sets of parameters:

- p: The order of the autoregressive component.
- d: The degree of differencing required to make the series stationary.
- q: The order of the moving average component.

Additionally, SARIMA models include seasonal parameters:

- P: The seasonal order of the autoregressive component.
- D: The degree of seasonal differencing required to make the seasonal component stationary.
- Q: The seasonal order of the moving average component.
- s: The length of the seasonal period (e.g., s=12 for monthly data with a yearly seasonality).

By properly selecting these parameters, the SARIMA model can effectively capture the underlying patterns and variations in the time series data, making it a powerful tool for time series forecasting. SARIMA models can be fitted and adjusted using historical data and then used to make predictions for future time points.

8.4 PACF (Partial Autocorrelation Function)

The Partial Autocorrelation Function (PACF) measures the correlation between a time series and its lagged values, while removing the effects of the intermediate lags. In other words, it captures the direct relationship between the current value and its past values without considering the influence of the other lags.

The PACF plot is a graphical representation of the PACF values. It displays the partial correlation coefficients for each lag, with the lag on the x-axis and the partial correlation coefficient on the y-axis. The PACF plot is useful in determining the order of the autoregressive (AR) component in a time series model.

Interpreting the PACF plot can provide insights into the appropriate order of the AR component. Here are a few general guidelines:

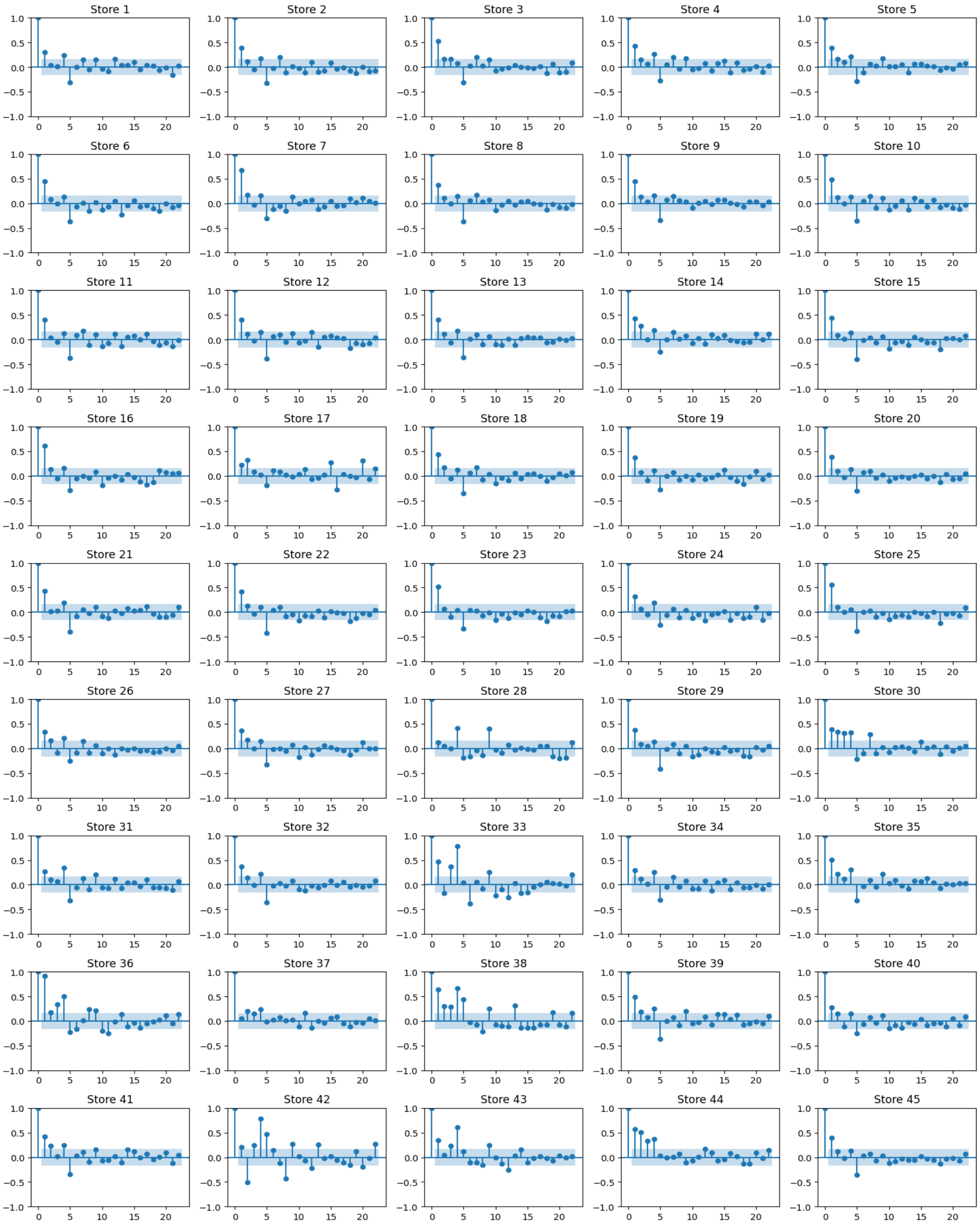
1. Significant spikes at certain lags: If the PACF plot shows a significant spike at a specific lag (lag k) and the values for other lags are not significant, it suggests that an autoregressive model of order k (AR(k)) might be appropriate. The spike indicates a direct relationship between the current value and the value at lag k, while the insignificant values for other lags indicate no significant direct relationship.

2. Gradual decay: If the PACF plot exhibits a gradual decay in the correlation coefficients, it suggests that an autoregressive model may be suitable, but the order is unclear. In such cases, further analysis or model selection techniques can help determine the appropriate order.

3. No significant values: If all the PACF values are within the confidence bounds (typically shown as dashed lines), it indicates no significant correlation beyond lag 0. This suggests that an autoregressive model might not be suitable, and other components, such as moving average (MA) or

combination of AR and MA, should be considered.

Remember that interpreting the PACF plot is not always straightforward, and other factors, such as the behaviour of the ACF (Autocorrelation Function) plot and the characteristics of the time series, should also be considered.



(PACF plots of Walmart Stores)

8.5 ACF (Autocorrelation Function)

The Autocorrelation Function (ACF) plot is a graphical representation of the autocorrelation coefficients of a time series with its lagged values. The ACF measures the correlation between a time series and its past values at different lags.

The ACF plot helps in understanding the overall correlation structure of the time series. It displays the autocorrelation coefficients on the y-axis, while the lags are shown on the x-axis. The ACF values range between -1 and 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

Interpreting the ACF plot can provide insights into the presence of autocorrelation in the time series. Here are a few general guidelines:

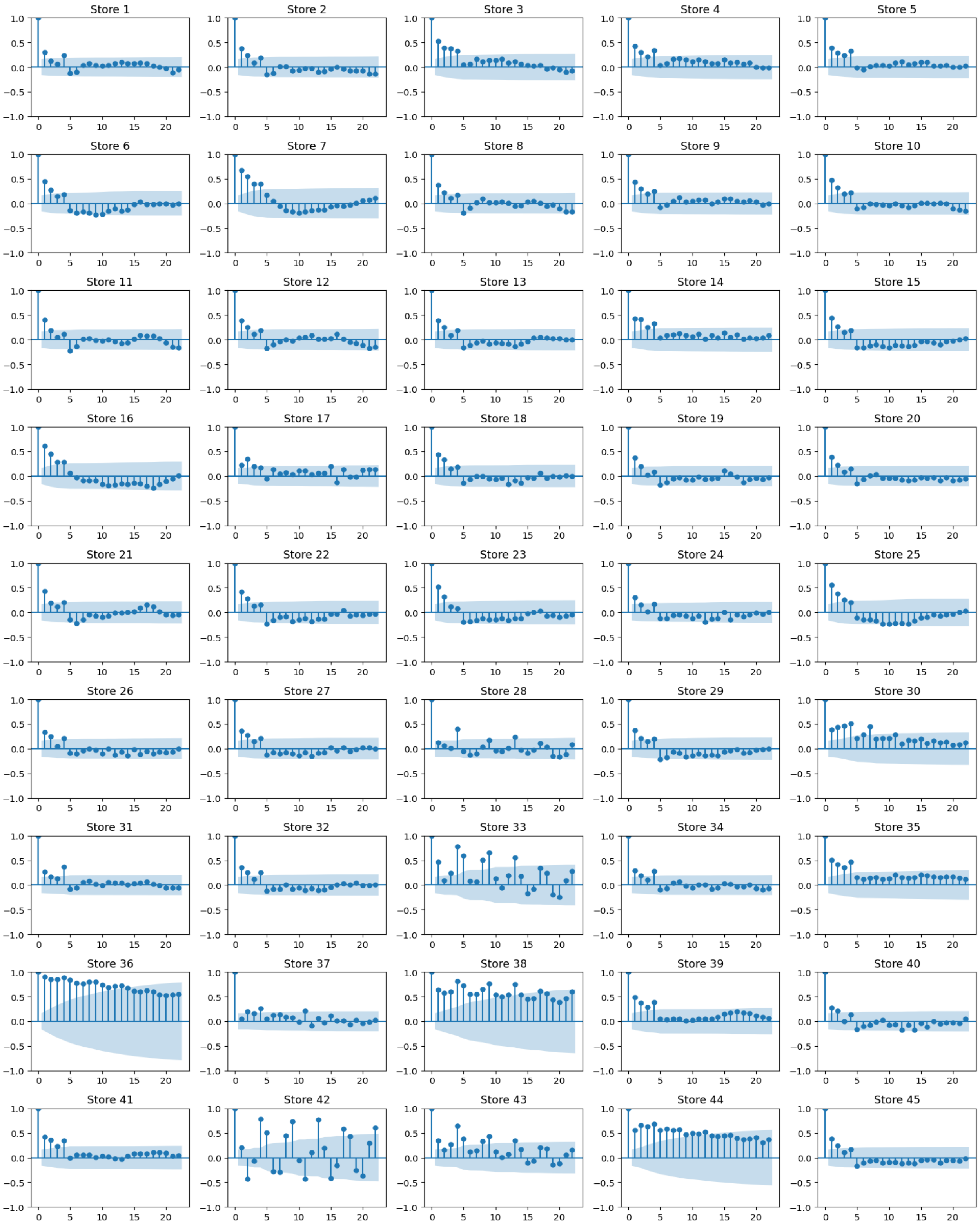
1. Decay in autocorrelation: If the ACF plot shows a gradual decay in the autocorrelation coefficients as the lag increases, it suggests that the current value of the time series is correlated with its past values up to a certain lag. The gradual decay indicates a pattern of dependence, where the correlation weakens as the lag increases.

2. Significant spikes at certain lags: If the ACF plot exhibits significant spikes at certain lags (lags with correlation coefficients outside the confidence bounds), it suggests the presence of autocorrelation at those lags. These spikes indicate a strong correlation between the current value and the value at those specific lags.

3. No significant values: If all the ACF values are within the confidence bounds (typically shown as dashed lines), it indicates no significant correlation beyond lag 0. This suggests that the time series might not exhibit autocorrelation, or the autocorrelation is weak and not statistically significant.

It's important to note that the ACF plot alone may not provide a complete understanding of the time series dynamics. Other factors, such as the Partial Autocorrelation Function (PACF) plot, the behaviour of the time series itself, and the underlying patterns, should be considered for a more comprehensive analysis.

The ACF plot, along with the PACF plot, is commonly used in time series analysis to determine the appropriate order of autoregressive (AR) and moving average (MA) components in models like ARIMA (Autoregressive Integrated Moving Average). These plots are valuable tools for identifying and modelling the autocorrelation structure of a time series.

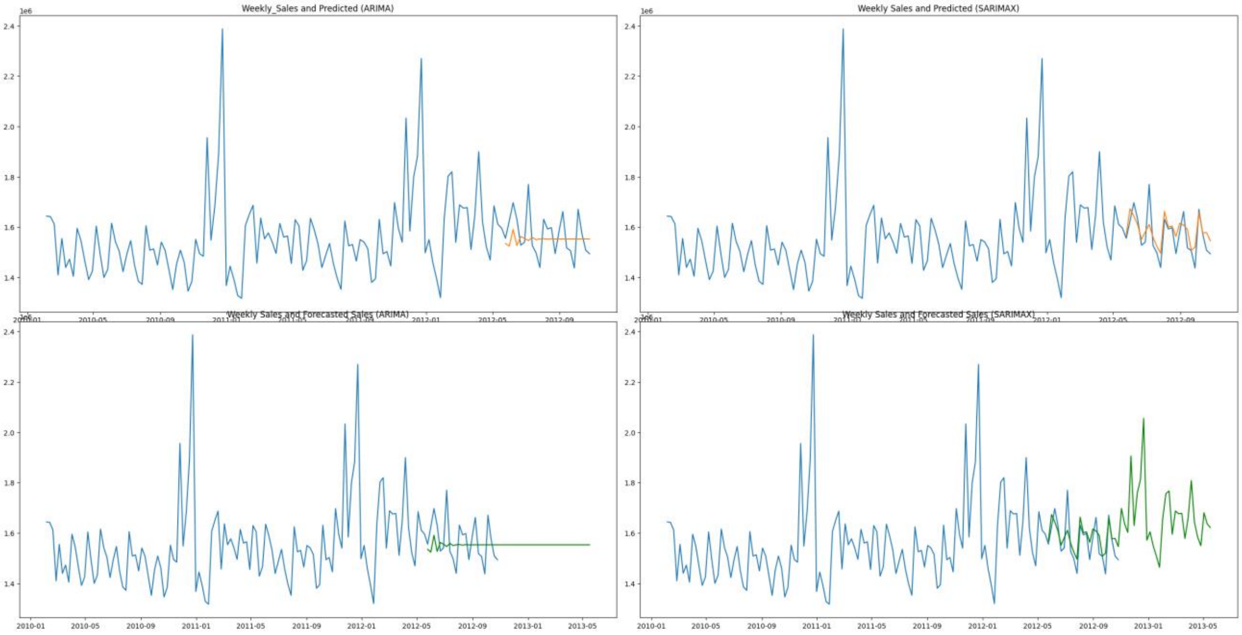


(ACF plots of Walmart Stores)

8.6 Model Comparison between ARIMA and SARIMA on Walmart Dataset

Store 1:

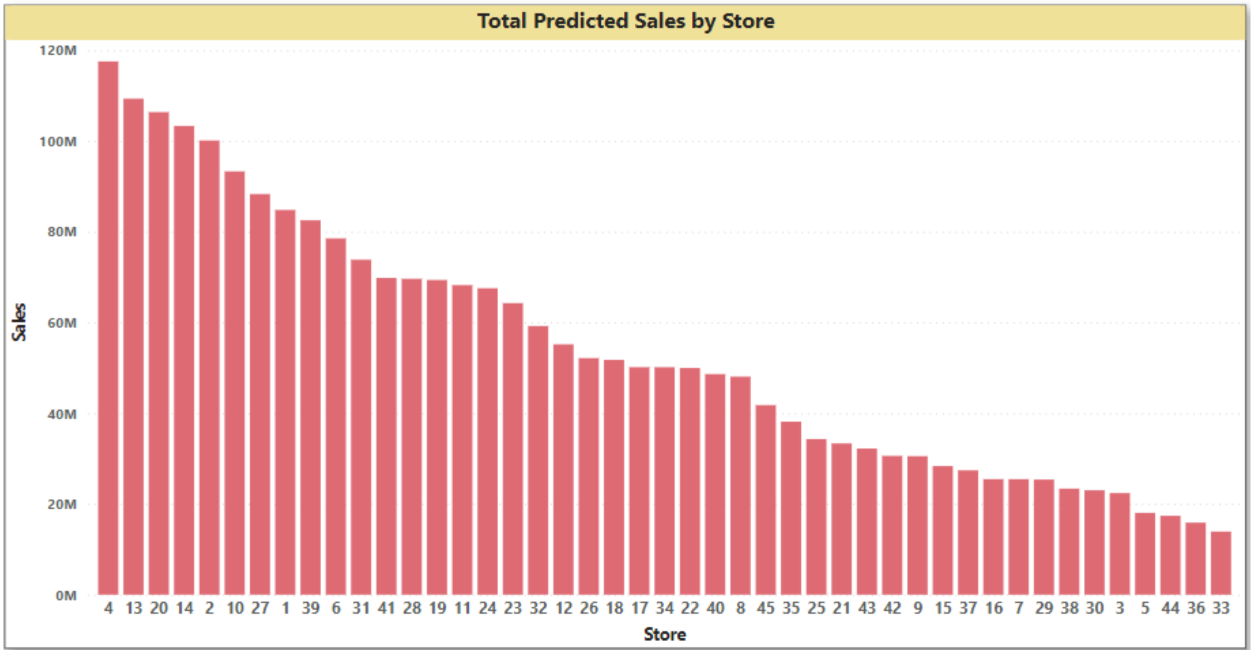
```
rmse for true values (test) and predicted (ARIMA) for store 1 :
84251.24469271746

rmse for true values (test) and predicted (SARIMAX) for store1 :
54465.314021180224
```



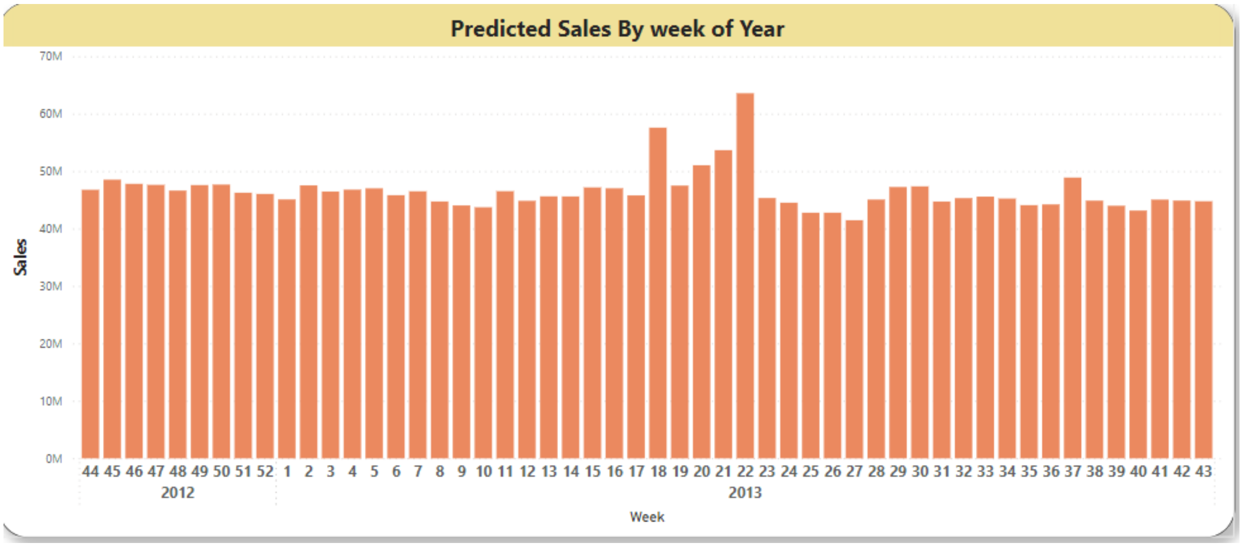
This is an example of store 1 only. The forecasting for all the other stores has already been captured in the code file.

From the above visualization it is pretty much clear that SARMA is performing better than ARIMA.

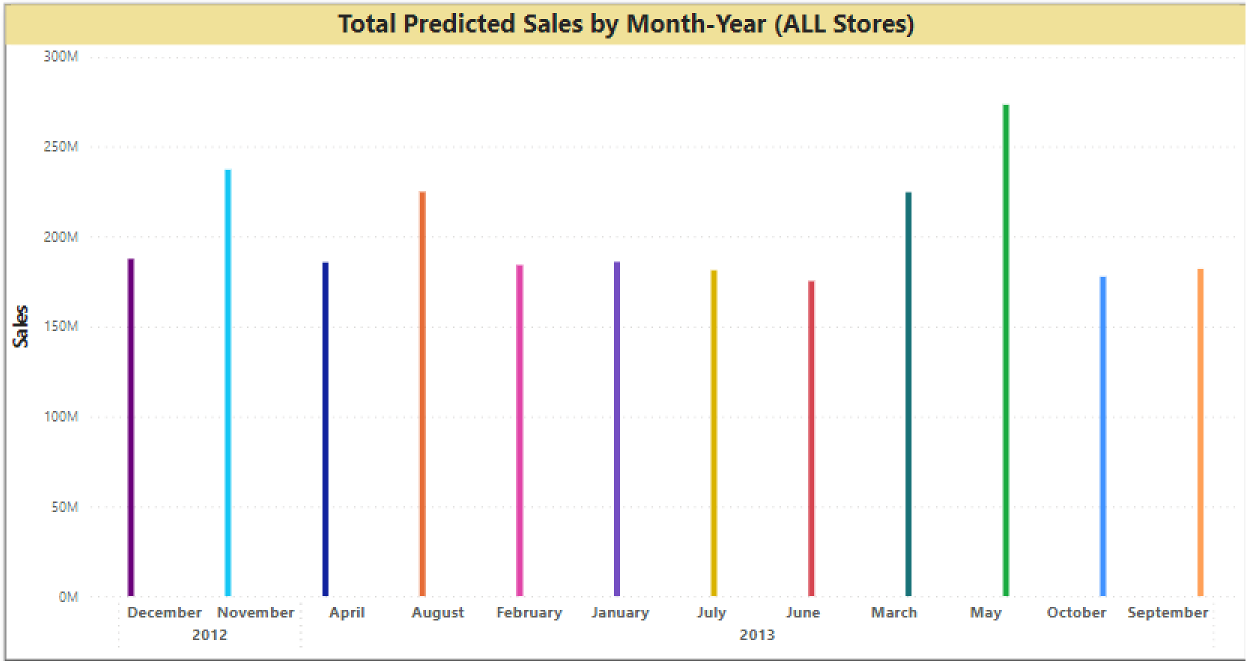
9.0 Power BI Visuals



Total Predicted Sales for Each Store



Total Predicted Sales by Week of Year



Total Predicted Sales by Month of Year

Based on the visualizations created based on the predicted sales, the following observations can be made:

- Sales still seem to be the highest during the holiday season (in the months of November and December)
- Stores 4, 13, and 20 are the three stores with the highest sales; similar to this, other than store 14, stores 4 and 20 still have the highest predicted sales

These predictions can be found in the ‘forecast.csv’.

10. Conclusion

10.1 Overall Results

The main purpose of this study was 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 study also aims to understand whether sales are relatively higher during holidays like Christmas and Thanks giving than normal days so that stores can work on creating promotional offers that increase sales and generate higher revenue.

Pertaining to the specific factors provided in the study (temperature, unemployment, 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. By 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. This was necessary because of the presence of a high number of predictor variables in the dataset. While the interaction effects were tested on a combination of significant variables, a statistically significant relationship was only observed between the independent variables of temperature, CPI and unemployment, and weekly sales (predictor variable). However, this is not definite because of the limitation of training data. 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 gradient boosting 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, store size, department, etc.

Finally, the tuned SARIMA model, with the lowest RMSE score, is the main model used to create the final predictions for this study.

10.2 Limitations

A huge constraint of this study is the lack of sales history data available for analysis. The data for the analysis only comes from a limited number of Walmart stores between the years 2010 and 2013. Because of this limited past history data, models cannot be trained as efficiently to give accurate results and predictions. Because of this lack of availability, it is harder to train and tune models as an over-constrained model might reduce the accuracy of the model. An appropriate amount of training data is required to efficiently train the model and draw useful insights. Additionally, the models created have been developed based on certain preset assumptions and business conditions; it is harder to predict the effects of certain economic, political, or social policies on the sales recorded by the organization. Also, it is tough to predict how the consumer buying behaviour changes over the years or how the policies laid down by the management might affect the company’s revenue; these factors can have a direct impact on Walmart sales and it is necessary to constantly study the market trends and compare them with existing performance to create better policies and techniques for increased profits.

11. Future Work

With growing technology and increasing consumer demand, Walmart can shift its focus on the e-commerce aspects of the business. Taking inspiration from Amazon’s business model, Walmart can grow its online retail business massively and gather huge profits. With already established stores and warehouses, it is easier for the organization to create a nationwide reach, limiting the presence of physical stores and helping their consumers save on fuel costs by delivering goods at their doorstep. It also makes it a lot easier to identify consumer buying patterns. An important aspect of this study is also to try and understand customer buying behaviour based on regional and departmental sales. This customer segmentation can help the organization in creating and communicating targeted messages for customers belonging to a particular region, establishing better customer relationships, focusing 57 on profitable regions, and identifying ways to improve products and services in specific regions or for specific customers.

12. References

1. "Time Series Analysis and Its Applications: With R Examples" by Robert H. Shumway and David S. Stoffer. This book provides a comprehensive introduction to time series analysis, including ARIMA and seasonal models, with practical examples in the R programming language.
2. "Forecasting: Principles and Practice" by Rob J Hyndman and George Athanasopoulos. This book covers various time series forecasting techniques, including ARIMA and seasonal forecasting methods. It also includes hands-on examples using the R forecast package.
3. "Introductory Time Series with R" by Paul S.P. Cowpertwait and Andrew V. Metcalfe. This book is an accessible introduction to time series analysis with a focus on practical applications using R. It covers ARIMA and seasonal models as well.
4. "Time Series Analysis" by James D. Hamilton. A classic textbook on time series analysis, which covers the theoretical foundations of ARIMA and other time series models. It includes examples and applications from various fields.
5. "Analysis of Financial Time Series" by Ruey S. Tsay. This book is focused on the analysis of financial time series data. It covers ARIMA and GARCH models, among other techniques, and provides real-world applications in finance.
6. "Econometric Analysis" by William H. Greene. While not solely dedicated to time series analysis, this comprehensive econometrics textbook covers ARIMA models along with other essential econometric methods.