Walmart Sales Prediction

USING TIME SERIES ANALYSIS

PRESENTER –SUYASH PANDEY

Problem Statement:

- Retail Giant Walmart is facing issues managing the sales and inventory of its stores in US.
- I will be analyzing their data for individual stores and some up with measures which are effecting the sales and remedies which can be offered to improve the performance of the stores.
- In this project I will be doing Exploratory , Predictive and prescriptive analysis for the Walmart group retail stores.

Problem Objective

1. You are provided with the weekly sales data for their various outlets. Use statistical analysis, EDA, outlier analysis, and handle the missing values to come up with various insights that can give them a clear perspective on the following:

a. If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?
b. If the weekly sales show a seasonal trend, when and what could be the reason?
c. Does temperature affect the weekly sales in any manner?
d. How is the Consumer Price index affecting the weekly sales of various stores?
e. Top performing stores according to the historical data.
f. The worst performing store, and how significant is the difference between the highest and lowest performing stores.

2. Use predictive modeling techniques to forecast the sales for each store for the next 12 weeks.

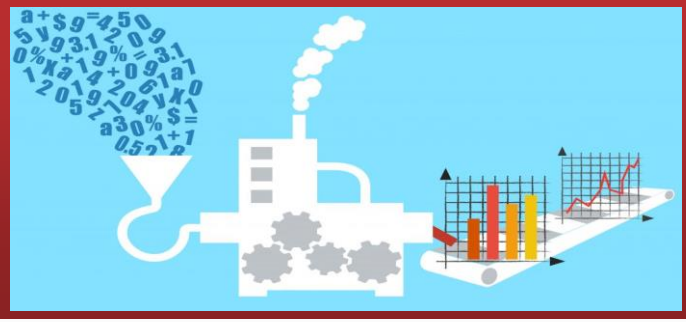
Data Description

- Store – Contains store Id of the respective store.
- Date – Week for the sales is done.
- Weekly sales – Total amount od sales for the week.
- Holiday Flag – Whether it is holiday weekend week or not
- Temperature – Temperature on the day of the sale
- Fuel price – Cost of fuel in the region of the store
- CPI – Consumer Price Index in US
- Unemployment – Unemployment Rate for US

```
walmart_df = pd.read_csv("Walmart DataSet.csv")
walmart_df
```

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
0	1	05-02-2010	1643690.90	0	42.31	2.572	211.096358	8.106
1	1	12-02-2010	1641957.44	1	38.51	2.548	211.242170	8.106
2	1	19-02-2010	1611968.17	0	39.93	2.514	211.289143	8.106
3	1	26-02-2010	1409727.59	0	46.63	2.561	211.319643	8.106
4	1	05-03-2010	1554806.68	0	46.50	2.625	211.350143	8.106

Data pre-processing steps



- Identification and removal of Missing Values – No Such values observed.
- Identification and removal of Missing Values – No Such values observed.
- Outliers – Performed both Z-Score and IQR method but not removed outliers as they seemed vital for models.
- EDA – Basic Visualizations and summary was done to understand the data better.

Algorithm Selection:

- FOR MY TIME SERIES PROJECT, I HAVE EMPLOYED TWO PROMINENT ALGORITHMS: ARIMA (AUTO REGRESSIVE
- INTEGRATED MOVING AVERAGE) AND SARIMAX (SEASONAL AUTO REGRESSIVE INTEGRATED MOVING
- AVERAGE WITH EXOGENOUS VARIABLES). THESE ALGORITHMS PROVIDE ROBUST AND EFFECTIVE METHODS FOR
- MODELING AND FORECASTING TIME SERIES DATA, ENABLING ME TO GAIN VALUABLE INSIGHTS AND MAKE INFORMED PREDICTIONS FOR MY PROJECT'S OBJECTIVES.

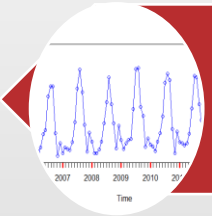
Motivation and Reasons For Choosing the Algorithm

- After implementing both the Algorithm it is found that SARIMAX is giving me more accuracy compared to ARIMA.

- So I moved ahead with SARIMAX algorithm.

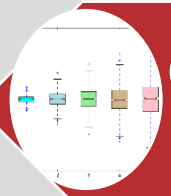
Assumptions:

Data Quality – Data quality is assumed to be accurate and the figures provided are true the best of knowledge



Stationarity – Data is taken as stationary which is a general assumption for time series analysis

Seasonality – Keeping in mind that it is a retails store sales data , seasonality is a unavoidable factor and while preparing model and forecasting the seasonality is expected to continue in the future.



Outliers – It is assumed that outliers in the data are not significantly affecting the model performance , outliers were identified but not removed.

Model Evaluation and Techniques:

Model Evaluation :

Selection – There are many models available but due to the need of this data I choose to go with ARIMA and SARIMA because it's a retail stores data which deals with factors like seasonality and other influencing parameters.

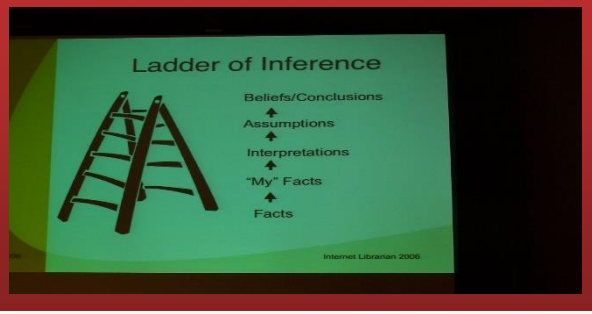
Training – The dataset for store was divided into ratio of 80:20 for training and testing.

Performance metrics – To measure the accuracy I used RMSE metrics.

Techniques :

Data Preprocessing – Grouping the data store wise to perform EDA.
Time series decomposition – To analyse the seasonality and trends.
Forecasting horizon – Forecasted weekly sales for coming 12 weeks.

Inferences from the Same



Effect of Unemployment – As concluded from the model analysis, Unemployment overall is not a much contributing factor for overall sales of Walmart but some stores have been hit by unemployment the highest effected store is Store number 38

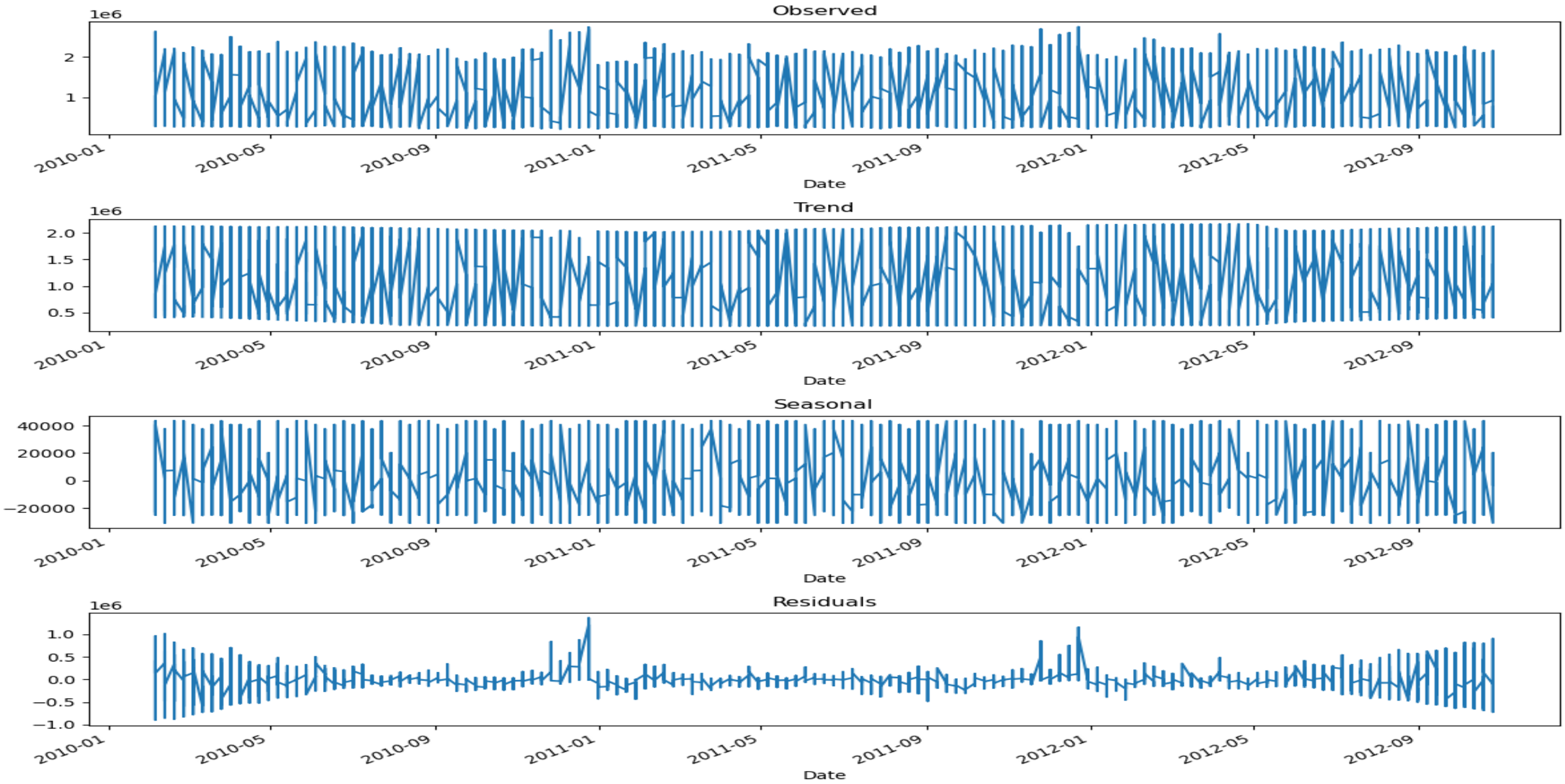
Seasonality – Since it is retail store data seasonality is a major factor as sales are been affected by seasons such as festival times highest number of sales can be seen in Q3 sales as it has the Christmas new year time which is the most vital time for sales

Temperature Effect – Clearly visible from the model the temperature is a factor effecting the sales whenever the temperature is too high or low the sales is reduced.

Holiday – When compared between holiday week vs non – holiday week it clearly visible that sales are increased in holiday week by 8% compared to non – holiday week with avg. sales for holiday week being – $2079267 against non – holiday sales being - $1914208

CPI – While evaluating the model it is observed that CPI does not prove to be a much affecting factor to the sales as the corelation of CPI with sales is -0.06

Seasonal Decomposition - Visualised



Future Possibilities

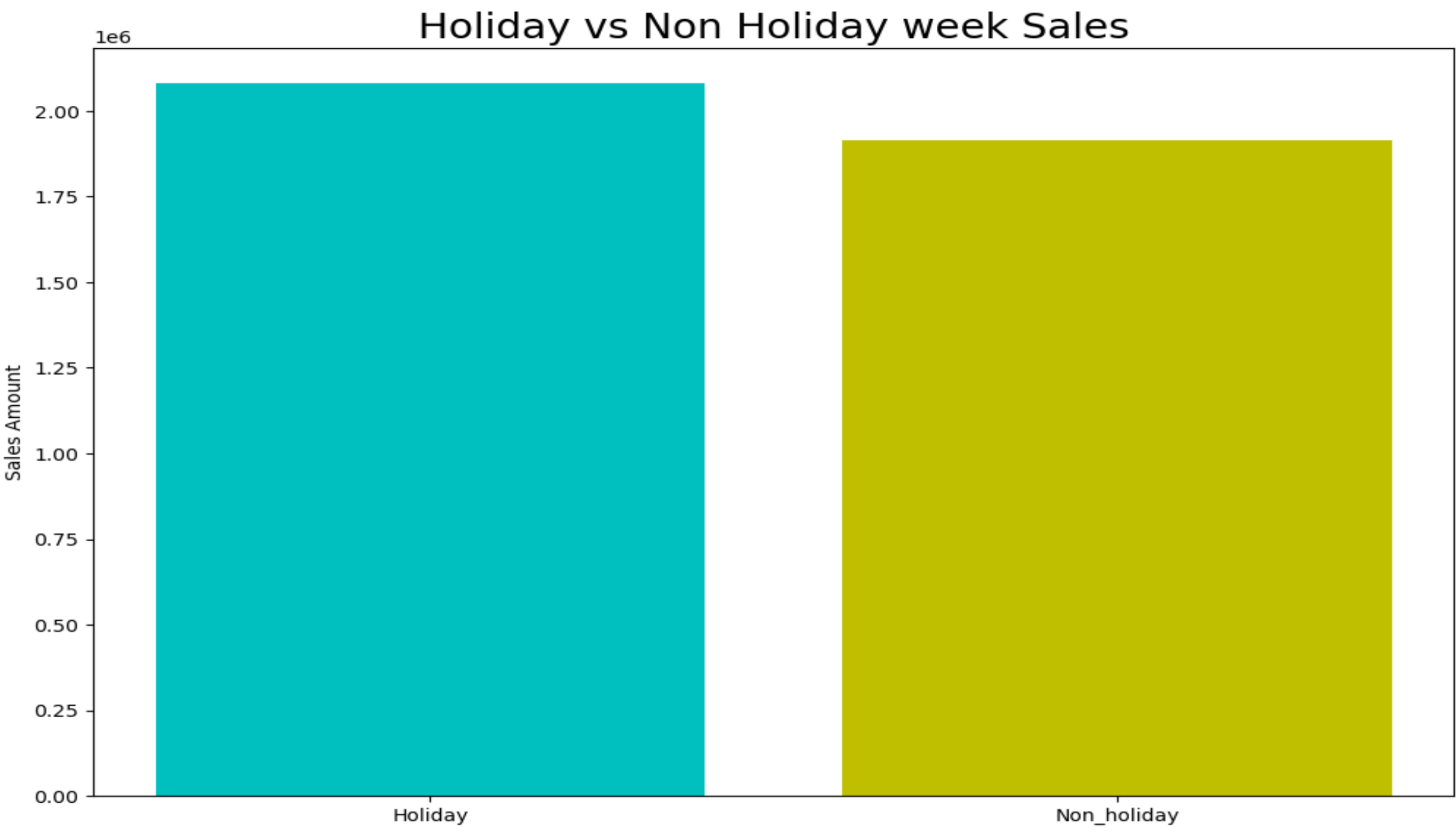
Based in the findings from the analysis below possibilities can be inferred :

Holiday Planning – Since holiday weeks are the ones with more engagement the store can come up with a more vibrant scheme for sales like discounts or entertainment activities which attracts children to visit store.

Fuel Price – Plan to manage inventory when there is a possibility of fuel price hike in the region.

Festival planning – Since the highest sales are coming in the festival week of Q3 special focus and pre planning should be done to make the max of the opportunity. Special campaign for gifts should be launched.

Store wise strategies – Worst performing stores mentioned in the model should be focused primarily and Corrective measure should be taken to make them better in performance.


Holiday vs Non Holiday week Sales
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Sales Amount
Holiday
Non_holiday

Conclusion

Summary – In this project I worked on the data to find the key factors affecting the sales.

Accuracy – Both the ARIMA and SARIMAX were deployed and SARIMAX proved to be more efficient after metrics evaluation by RMSE.

Seasonality and trends – Presence of seasonality and trend is expected to continue in the future also.

References

- Intellipaat Hands-on Lecture on Time series by Akanksha.
- YouTube channel Unfold data science - https://youtube.com/playlist?list=PLmPJQXJiMoUVr07-VnwDiki89DqyuSS21&si=yfxLI3ffzOpqFYgu
- Blog for ACF and PACF - https://towardsdatascience.com/interpreting-acf-and-pacf-plots-for-time-series-forecasting-af0d6db4061c

Thank You 

SUYASH PANDEY