SCHOOL OF BUSINESS

OPIM 5671 – Data Mining and Business Intelligence

Prof - Sudip Bhattacharjee

WALMART SALES FORECAST

GROUP-5

Team members:

Manas Joshi

Anand Shiv Sharma

Sri Varshini Chava

Anuja Sunil Kamble

Mitul Krishnaswamy Sivakumar

TABLE OF CONTENTS:

1. Summary:

Using time series forecasting techniques and SAS Studio 3.8 software, the research seeks to predict the weekly sales in Walmart stores. Weekly Sales data set consists of data from 45 stores and 81 departments with 16 columns and 420,212 rows. It makes predictions about upcoming weekly sales using historical data from weekly sales in 4 different Walmart Store Numbers 4,14,13,20 between February 5, 2010, and November 1, 2012 and 92 departments.

The aim of the project is to build a model which predicts sales of the stores. With this model, Walmart authorities can decide their future plans which are very important for arranging stocks, calculating revenue and deciding to make new investments or not.

The group created a number of time series models and selected the Arimax Models for Store 2 department 92, and compared them and selected the best one with the least error metrics.

2. Problem Statement

Intense competition and narrowing margins are causing rising anxiety for the US retail business. Several E-commerce businesses have increased their market share in this industry with the arrival of technology developments and consumer awareness. Traditional brick and mortar establishments are facing increasing competition from internet retailers that use various business models. The industry's overall SWOT analysis leads to the conclusion that it is essential to estimate future sales in order to balance the trade and take proactive measures to meet client demands. The following inputs can be controlled and managed with the use of sales forecasting.

1. Supply chain management and operational effectiveness
2. Inventory control - Just-in-time implementation
3. Marketing & Promotions – Forecasting approaches for client discounts and promotions
4. Financial Planning - Internal Controls and Budgeting

In order to generate strategic insights through in-depth data analysis, this project aims to analyze both the macro and micro elements of the fundamental difficulties presented in the data set.

3. Data Description

The dataset (train.csv) consists of 420212 rows and 16 columns, which has Walmart's weekly sales records at 45 stores and 92 departments from February 05, 2010 to November 01, 2012 (143 weeks).

The columns are as follows:

- **Time Variable**:
  - Date - the end date of the weekly cycle from Feb 2010 to Nov 2012.
- **Dependent Variable**:
  - Weekly Sales - Sales for a given department in the given store in dollars($).
- **Independent Variables (Used in Modeling)**:
  - **Is Holiday** - Whether the week is a holiday week where 1 - Holiday and 0 - Not Labor day, Thanksgiving and Christmas can be considered as Holidays.
  - **Mark Down1** - 1st round of Promotional Markdown in dollars ($).
  - **MarkDown2** - 2nd round of Promotional Markdown in dollars ($).
  - **Mark Down3** - 3rd round of Promotional Markdown in dollars ($).
  - **Mark Down4** - 4th round of Promotional Markdown in dollars ($).
  - **Mark Down5** - 5th round of Promotional Markdown in dollars ($).
- **Independent Variables (Not used in Modeling)**
  - **Dept** - The department number numbered from 1-99 in Walmart.
  - **Store** - Store numbered from 1-45 in Walmart.
  - **Size** - Sizes in Square Feet for each Store.
  - **Type** - Types of Stores labeled with A, B and C on the basis of sales.

○ **CPI -** Consumer Price Index.

○ **Unemployment** - Unemployment rate of the store in percentage.

○ **Temperature (Fahrenheit)** - Temperature in the region where the store is located.

○ **Fuel Price** - Cost of fuel in dollars ($) in the region where the store is located.

4. Exploratory Data Analysis

- The data was not preprocessed and contained outliers, missing values, and repetitive columns.
- We merged our data set on the basis of Date and Store columns from three csv files from Kaggle.
- After merging the dataset, we removed the repetitive column **Isholiday_y** and renamed **Isholiday_x** to **Isholiday**.
- There were many missing values in MarkDown1 to MarkDown 5 columns. We filled all of the na or null values with zeros.
- 0.3% of the data has less than or equal to zero sales, those outliers were removed for proper functioning of the data.
- We removed the columns which are not correlated with the output sales variable.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 420212 entries, 0 to 421569
Data columns (total 16 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   Store         420212 non-null  int64
 1   Dept          420212 non-null  int64
 2   Date          420212 non-null  object
 3   Weekly_Sales  420212 non-null  float64
 4   IsHoliday     420212 non-null  int64
 5   Temperature   420212 non-null  float64
 6   Fuel_Price    420212 non-null  float64
 7   MarkDown1     420212 non-null  float64
 8   MarkDown2     420212 non-null  float64
 9   MarkDown3     420212 non-null  float64
 10  MarkDown4     420212 non-null  float64
 11  MarkDown5     420212 non-null  float64
 12  CPI           420212 non-null  float64
 13  Unemployment  420212 non-null  float64
 14  Type          420212 non-null  object
 15  Size          420212 non-null  int64
dtypes: float64(10), int64(4), object(2)
memory usage: 54.5+ MB
```

Fig 1: Data Distribution

- Below table shows the basic statistics of all of the columns

	count	mean	std	min	25%	50%	75%	max
Store	420212.0	22.195611	12.787236	1.000	11.000000	22.000000	33.000000	45.000000
Dept	420212.0	44.241309	30.508819	1.000	18.000000	37.000000	74.000000	99.000000
Weekly_Sales	420212.0	16033.114591	22729.492116	0.010	2120.130000	7661.700000	20271.265000	693099.360000
IsHoliday	420212.0	0.070345	0.255729	0.000	0.000000	0.000000	0.000000	1.000000
Temperature	420212.0	60.090599	18.447857	-2.060	46.680000	62.090000	74.280000	100.140000
Fuel_Price	420212.0	3.360890	0.458519	2.472	2.933000	3.452000	3.738000	4.468000
MarkDown1	420212.0	2590.323565	6053.415601	0.000	0.000000	0.000000	2809.050000	88646.760000
MarkDown2	420212.0	878.905242	5076.928566	-265.760	0.000000	0.000000	2.400000	104519.540000
MarkDown3	420212.0	468.845949	5534.069859	-29.100	0.000000	0.000000	4.540000	141630.610000
MarkDown4	420212.0	1083.534361	3896.068938	0.000	0.000000	0.000000	425.290000	67474.850000
MarkDown5	420212.0	1662.805002	4206.209357	0.000	0.000000	0.000000	2168.040000	108519.280000
CPI	420212.0	171.212496	39.162445	126.064	132.022667	182.350989	212.445487	227.232807
Unemployment	420212.0	7.960000	1.863879	3.879	6.891000	7.866000	8.567000	14.313000
Size	420212.0	136749.732787	60993.084568	34875.000	93638.000000	140167.000000	202505.000000	219622.000000

Fig 2 : Data Statistics

- There are 45 stores and each store has 81 departments. As every department in a store has its own time series. We tried to run our model on specific stores and specific departments. For our hypothesis, we chose departments and stores with the highest average sales. As a result we chose stores [4,13,14,20] and department 92 for forecasting purposes.

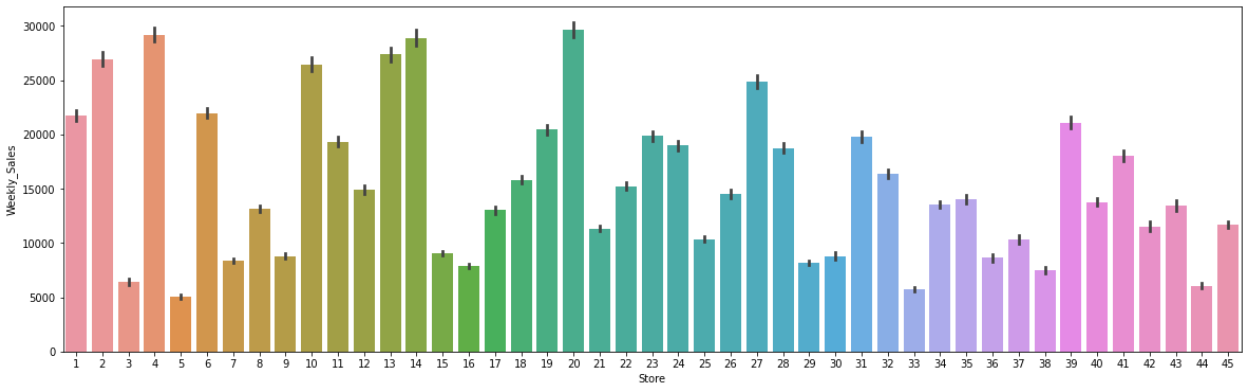

Fig 3: Relationship between Average Weekly Sales and Stores

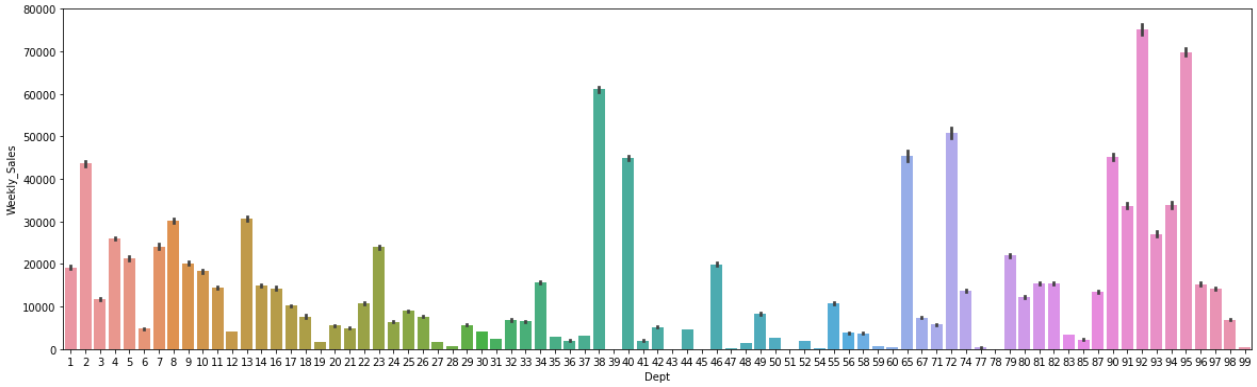

Fig 4: Relationship between Average Weekly Sales and Departments

- We also checked all the variables and their correlation with sales. Out of all variables we found **MarkDown3** and **IsHoliday** are highly correlated with **weekly sales.**

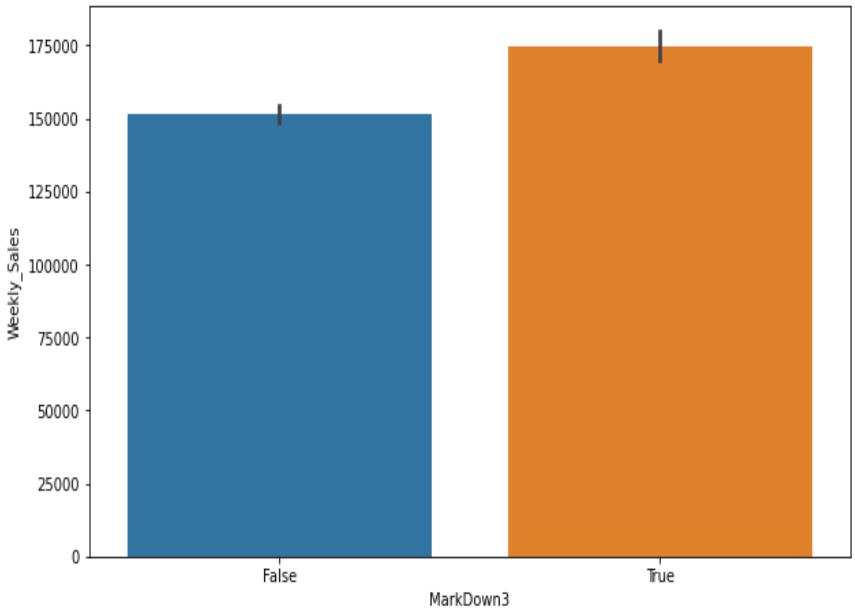

Fig 5: Impact of MarkDown3 on Weekly_Sales

- Below graph represents the monthly sales for three years.

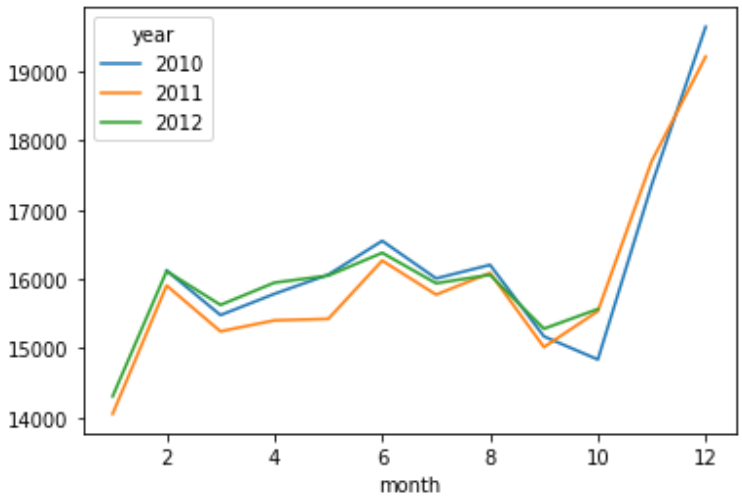

Fig: 6: Monthly sales of last three years

5. Times Series Exploration

In order to conduct our modeling, we limited our final dataset to the combination of store_4 and dept_92, as this particular store and department exhibited the highest weekly sales. Subsequently, we proceeded with data processing and initiated an exploratory analysis of the resulting time series. This particular time series spans from 2010 to October 2012, encompassing a total of 143 weeks. The summary statistics reveal that the mean weekly sales value over this time period amounts to 159,365

Dept=92

Variable Information	
Name	Weekly_Sales
Label	Weekly_Sales
First	Fri, 5 Feb 2010
Last	Fri, 26 Oct 2012
Number of Observations Read	143

Time Series Descriptive Statistics	
Variable	Weekly_Sales
Number of Observations	143
Number of Observations Used	143
Number of Missing Observations	0
Minimum	122263.2
Median	158924
Maximum	239759.3
Mean	159365.1
Standard Deviation	19283.27

Fig 7: Time series Descriptive Statistics

Correlation Plot of The Dependent Variable

Based on our analysis of the correlation plots, we can make several inferences about the dependent variable, weekly_sales. Specifically, we can conclude that it exhibits an

1. autoregressive structure,

2. a moving average component, and

3. is not characterized by white noise.

These observations are supported by the presence of a discernible downward trend in the ACF plot, as well as spikes in the PACF and IACF plots. Furthermore, the downward trend in the PACF plot and spikes in the ACF plot further corroborate the existence of both auto regressive and moving average components.

Given the aforementioned observations, we identified potential values for the parameter p based on the spikes in the PACF plot at lag 4, 8, and 9. Additionally, we identified potential values for the parameter q based on the spikes in the ACF plot at lag 5 and 9.

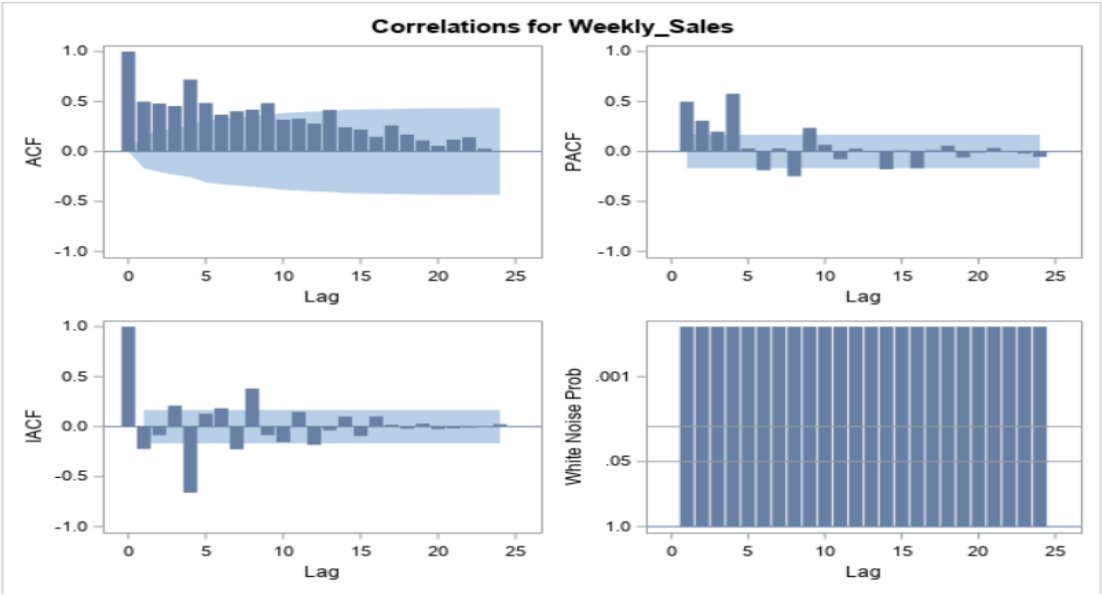

Fig 8: Correlation plot of weekly_sales

Decomposition Analysis on weekly _sales

The results of the decomposition analysis(**Fig 9**) indicate that the weekly_sales variable exhibits both trend and seasonal components. Specifically, the time series possesses a stable seasonal component with peak values occurring between the months of October and December, which correspond to major holidays such as Thanksgiving, Christmas, and New Year. These observations provide important insights into the underlying patterns and dynamics of the weekly_sales time series, which can inform the development of forecasting models and the interpretation of future predictions.

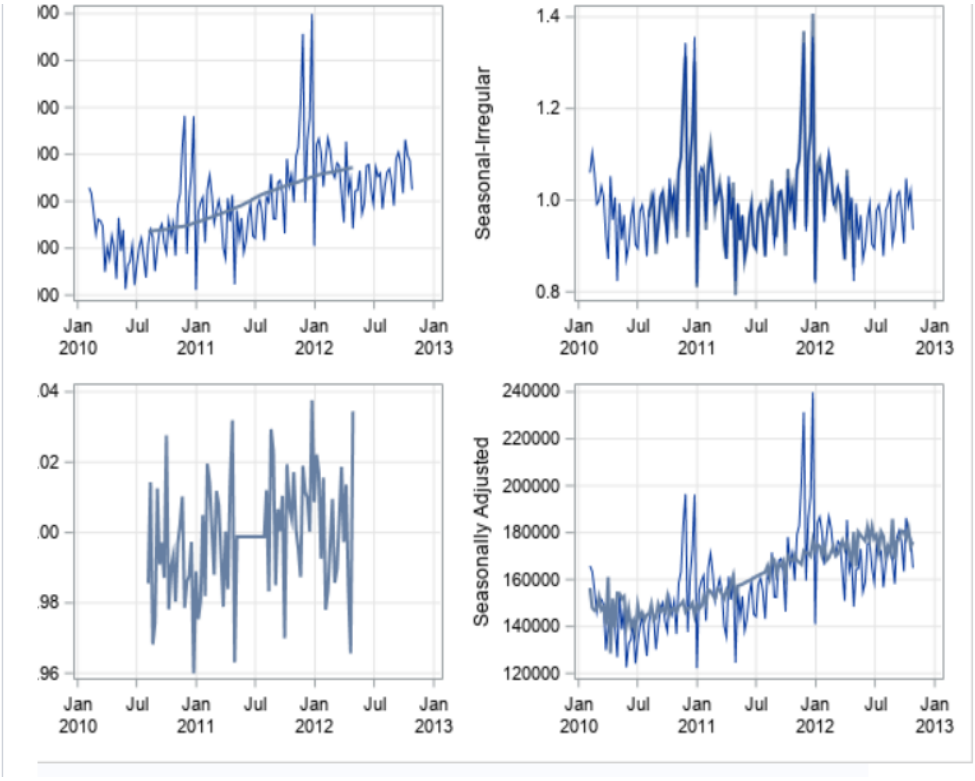

Fig 9: Decomposition analysis of weekly_Sales

Augmented Dickey-Fuller Test

By conducting an augmented Dickey-Fuller (ADF) test(**Fig 10**), we can determine whether a given time series is stationary or non-stationary. In the present case, the tau value derived from the ADF test is less than 0.05, indicating that the time series is stationary. These findings are significant because stationarity is a key assumption of many time series models, and deviations from stationarity can result in biased parameter estimates and unreliable forecasts. Thus, the ADF test results provide valuable information for selecting appropriate modeling strategies and interpreting the results of subsequent analyses.

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-1.0328	0.4660	-0.72	0.4042		
	1	-0.3276	0.6072	-0.38	0.5438		
	2	-0.1190	0.6546	-0.19	0.6170		
Single Mean	0	-71.2061	0.0012	-6.85	<.0001	23.43	0.0010
	1	-37.0903	0.0012	-4.25	0.0008	9.03	0.0010
	2	-23.0596	0.0042	-3.19	0.0230	5.09	0.0360
Trend	0	-107.102	0.0001	-9.24	<.0001	42.69	0.0010
	1	-77.2871	0.0005	-6.24	<.0001	19.52	0.0010
	2	-59.8574	0.0005	-5.00	0.0004	12.51	0.0010

Fig 10: ADF test

Cross Correlation of Independent Variables

We generated cross-correlation plots to examine the relationship between the dependent variable and various independent variables. Among the tested variables, only MarkDown3 exhibited a significant correlation at lag 4 with the dependent variable. Consequently, MarkDown3 was selected as the primary independent variable for forecasting purposes. These findings are significant because they highlight the importance of carefully selecting relevant independent variables for time series modeling, which can have a substantial impact on the accuracy and reliability of forecasts.

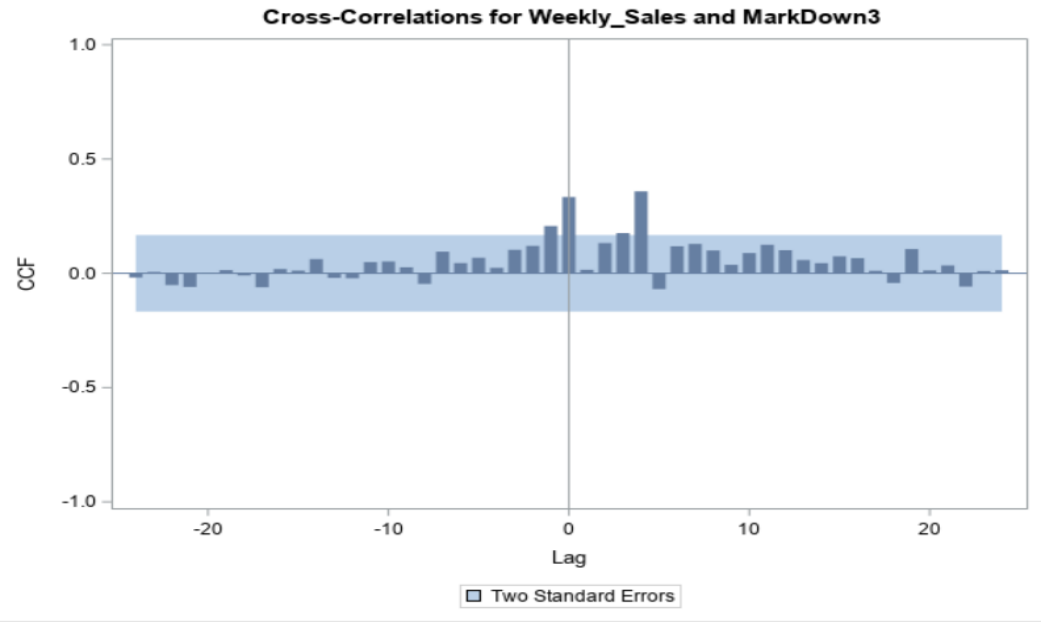

Fig 11: Cross-Correlation graph

6. Time Series Modeling

Since our dataset has both trend and seasonality component, also the seasonality component is periodic so we will use the following models to do the forecasting:

- Exponential smoothing (Winter additive and Winter multiplicative)
- ARIMA
- ARIMAX

Exponential smoothing (Winter additive Model):

We did an Exponential smoothing Winter additive model and found that there were still correlations between residuals and residuals were not spread as white noise. So it is necessary that we should check for some other model where the residuals are spread as white noise.

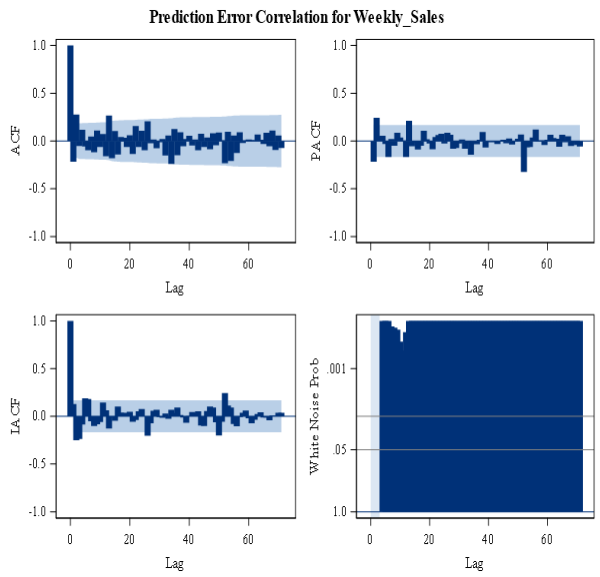

Fig 12.1: Error Correlation plot

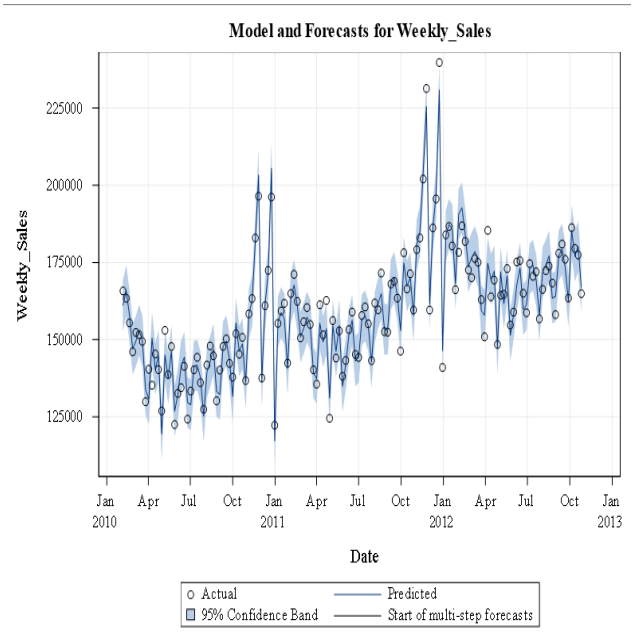

Fig 12.2: Forecast graph

Exponential smoothing (Winter multiplicative Model):

We found that when we ran the Winter multiplicative model again the residuals were not spread as white noise and there were correlations between noise as per ACF and PACF plots.

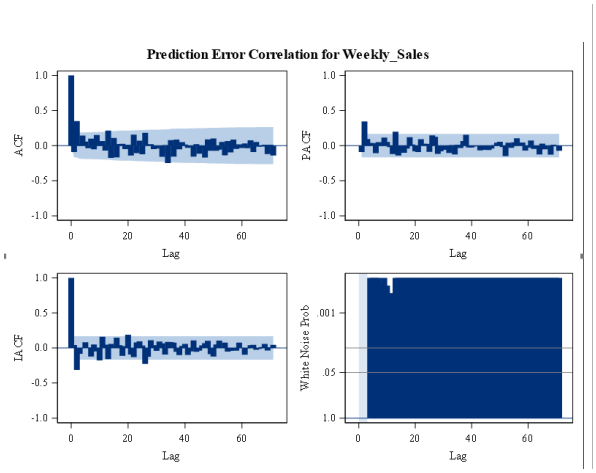

Fig 13.1: Error Correlation plot

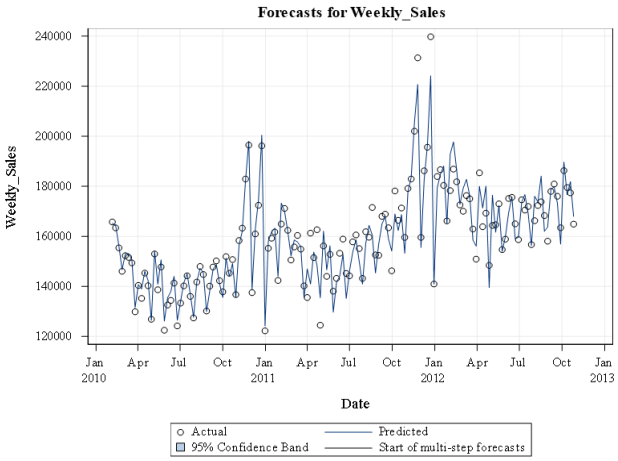

Fig 13.2: Forecast plot

ARIMA model:

While performing time series exploration and analyzing ACF and PACF plots, we got to know that we could fit in both AR and MA models. Also, our dataset has trends and seasonality. We will be running an ARIMA model by doing some trend and seasonal differencing . Based on the relations between different lags of ACF and PACF values, we have tried various combinations, and were able to get combinations as:

1) A(p,d,q)(P,D,Q)s = (4,0,4)(0,1,0)

2) A(p,d,q)(P,D,Q)s = (4,1,5)(0,1,0)

Also,we took the number of periods for forecast as 20 and number of holdbacks as 12.

ARIMA (A(p,d,q)(P,D,Q)s = (4,0,4)(0,1,0)): We found that when doing first order seasonal differencing in ARIMA the model performed pretty well, also the error spread as white noise.

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	11668.3	3519.2	3.32	0.0009	0
MA1,1	0.26463	16.30300	0.02	0.9870	1
MA1,2	1.19632	11.92579	0.10	0.9201	2
MA1,3	-0.22814	7.49434	-0.03	0.9757	3
MA1,4	-0.23309	3.71381	-0.06	0.9500	4
AR1,1	0.28826	0.48031	0.60	0.5484	1
AR1,2	1.59521	0.32239	4.95	<.0001	2
AR1,3	-0.20676	0.43350	-0.48	0.6334	3
AR1,4	-0.67949	0.30611	-2.22	0.0264	4

Fig 14.1: Maximum Likelihood Estimation

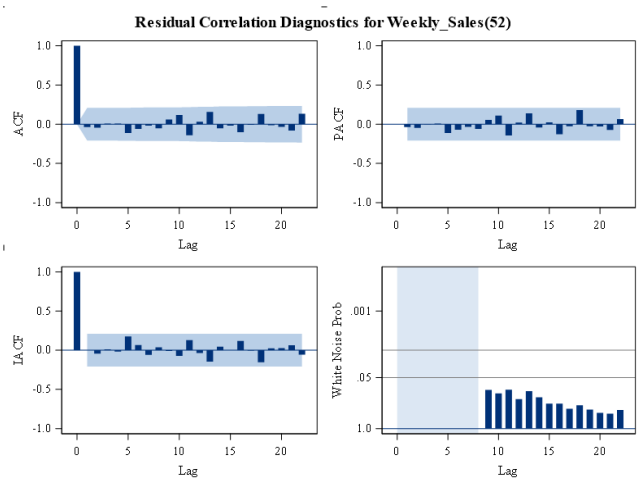

Fig 14.2: Residual Correlation

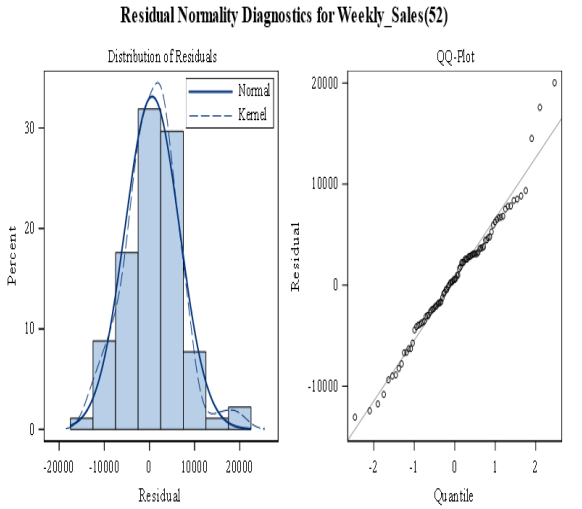

Fig 14.3: Residual Plots

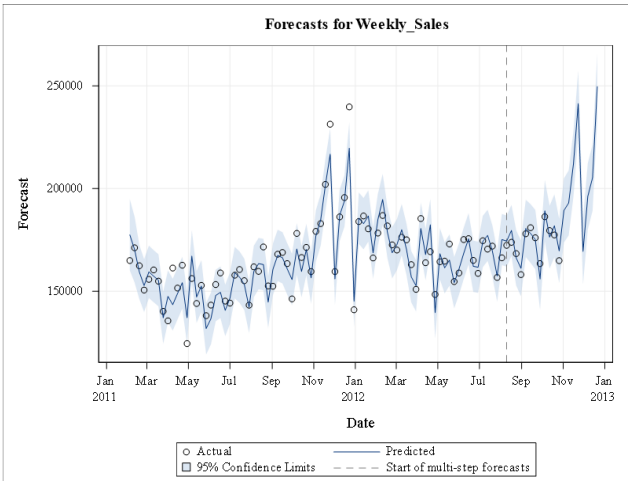

Fig 14.4: Forecasts for weekly_Sales

	Dept	_TYPE_	_STAT_	_VALUE_
1	92	ML	AIC	1862.4772924
2	92	ML	SBC	1885.0750279
3	92	ML	LOGLIK	-922.2386462
4	92	ML	SSE	3297549123.5

Fig 14.5: Output Statistics

2) A(p,d,q)(P,D,Q)s = (4,1,5)(0,1,0): We tried running ARIMA model again with some changes in q value(5) and did first order differencing to take care of trend in model. We found that residuals were spread as white noise and residual plots were also pretty well.

Also we found that this model performs better than the previous model when compared with AIC,SBS and SSE.

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	42.30917	185.77394	0.23	0.8198	0
MA1,1	-1.03845	0.90648	-1.15	0.2520	1
MA1,2	-0.83955	1.43929	-0.58	0.5597	2
MA1,3	0.37742	0.52258	0.72	0.4702	3
MA1,4	0.27049	0.38207	0.71	0.4790	4
MA1,5	0.26430	0.55970	0.47	0.6368	5
AR1,1	-1.96751	0.76743	-2.56	0.0104	1
AR1,2	-2.26068	1.22505	-1.85	0.0650	2
AR1,3	-1.28122	1.09950	-1.17	0.2439	3
AR1,4	-0.38538	0.38664	-1.00	0.3189	4

Fig 15.1: Maximum Likelihood Estimation

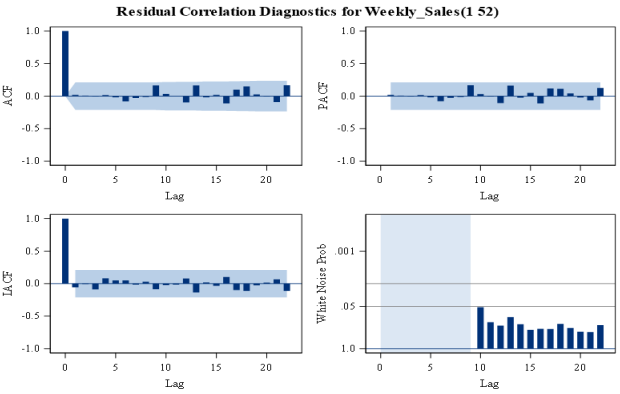

Fig 15.2: Residual Correlation

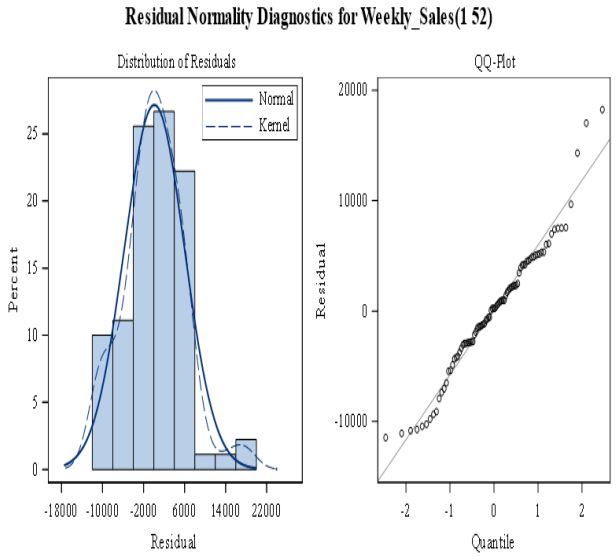

Fig 15.3: Residual Plots

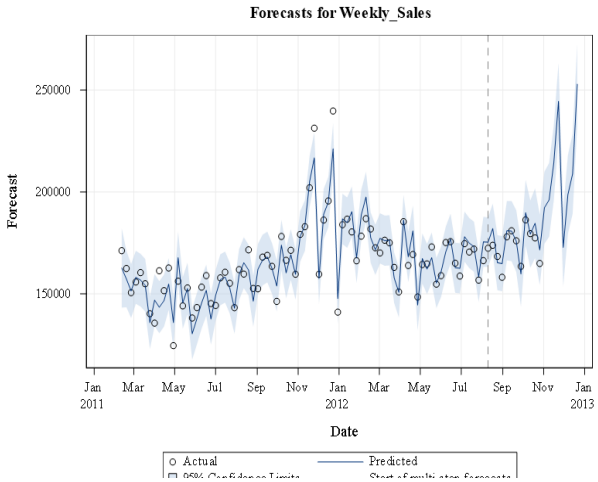

Fig 15.4: Forecasts for weekly_Sales

	Dept	_TYPE_	_STAT_	_VALUE_
1	92	ML	AIC	1841.7766891
2	92	ML	SBC	1866.7747858
3	92	ML	LOGLIK	-910.8883446
4	92	ML	SSE	3073522932.6

Fig 15.5: Output Statistics

ARIMAX: As we have seen in the time exploration plots, there is an independent variable markdown_3 that has lag 4 effect on weekly sales.So, we will be running an ARIMA model by doing some trend and seasonal differencing . Based on the relations between different lags of ACF and PACF values, we have tried various combinations, and were able to get combinations as:

1) A(p,d,q)(P,D,Q)s = (4,1,4)(0,1,0)

2) A(p,d,q)(P,D,Q)s = (4,1,5)(0,2,0)

Also, we took the number of periods for forecast as 20 and number of holdbacks as 12.

ARIMAX (A(p,d,q)(P,D,Q)s = (4,1,4)(0,1,0)) : After running the ARIMAX model we found that residuals performed pretty well on white noise also we got to know how the markdown_3 affects the forecasted sales when compared to the ARIMA model.

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	18.75862	172.22559	0.11	0.9133	0	Weekly_Sales	0
MA1,1	-0.56663	3.58985	-0.16	0.8746	1	Weekly_Sales	0
MA1,2	0.06800	2.13449	0.03	0.9746	2	Weekly_Sales	0
MA1,3	0.71750	5.14451	0.14	0.8891	3	Weekly_Sales	0
MA1,4	-0.07600	1.47552	-0.05	0.9589	4	Weekly_Sales	0
AR1,1	-1.51179	1.71648	-0.88	0.3785	1	Weekly_Sales	0
AR1,2	-0.98803	2.90680	-0.34	0.7339	2	Weekly_Sales	0
AR1,3	0.03959	2.37407	0.02	0.9867	3	Weekly_Sales	0
AR1,4	0.13852	0.59403	0.23	0.8156	4	Weekly_Sales	0
NUM1	0.21869	0.06982	3.13	0.0017	0	MarkDown3	4

Fig 16.1: Maximum Likelihood Estimation

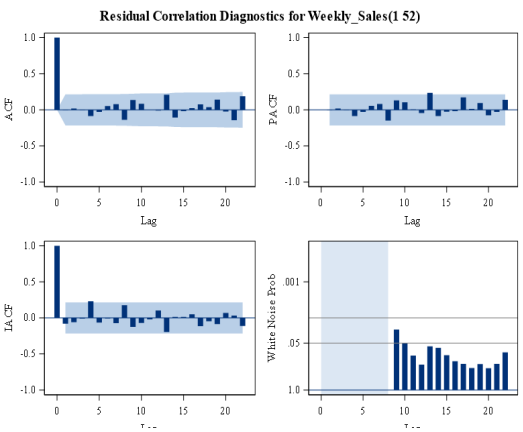

Fig 16.2: Residual Correlation

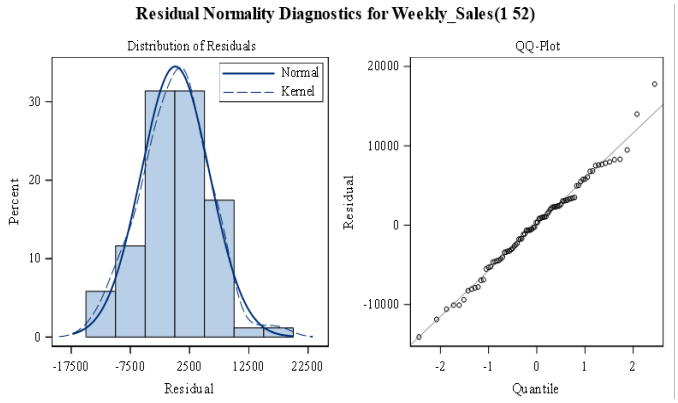

Fig 16.3: Residual Plots

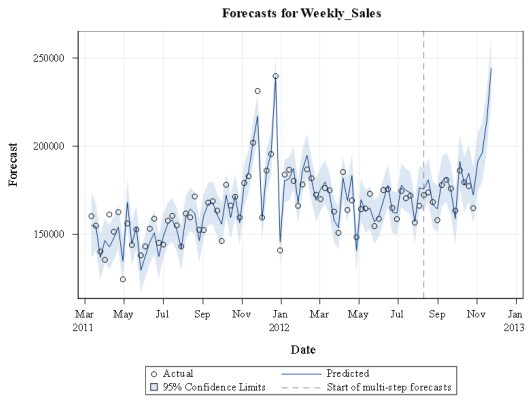

Fig 16.4: Forecasts for weekly_Sales

Also, we found that the AIC, SBC and SSE for this model has decreased in comparison with other previous models.

	TYPE	_STAT_	_VALUE_
1	ML	AIC	1756.4477392
2	ML	SBC	1780.9912122
3	ML	LOGLIK	-868.2238696
4	ML	SSE	2840226278.2

Fig 16.5: Output Statistics

ARIMAX(A(p,d,q)(P,D,Q)s = (4,1,5)(0,2,0)):

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	-753.02188	269.63356	-2.79	0.0052	0	Weekly_Sales	0
MA1,1	-1.17055	28.16137	-0.04	0.9668	1	Weekly_Sales	0
MA1,2	-0.91079	123.13280	-0.01	0.9941	2	Weekly_Sales	0
MA1,3	0.32991	51.26379	0.01	0.9949	3	Weekly_Sales	0
MA1,4	0.82235	65.51263	0.01	0.9900	4	Weekly_Sales	0
MA1,5	0.69631	72.56119	0.01	0.9923	5	Weekly_Sales	0
AR1,1	-2.20010	1.53979	-1.43	0.1531	1	Weekly_Sales	0
AR1,2	-2.63335	2.11415	-1.25	0.2129	2	Weekly_Sales	0
AR1,3	-1.76976	1.66202	-1.06	0.2870	3	Weekly_Sales	0
AR1,4	-0.65641	0.72237	-0.91	0.3635	4	Weekly_Sales	0
NUM1	-2.90956	11.65407	-0.25	0.8029	0	MarkDown3	4

Fig 17.1: Residual Plots

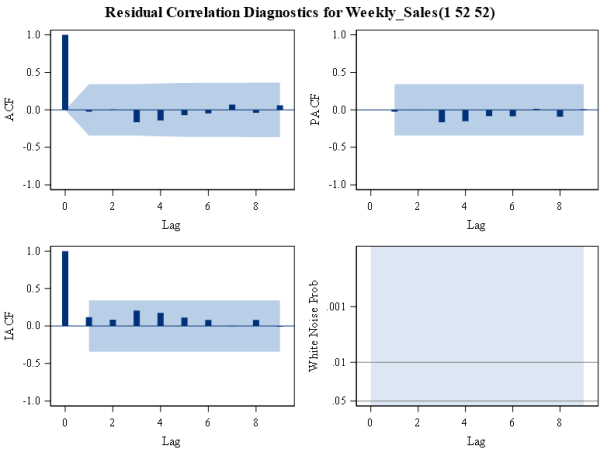

Fig 17.2: Forecasts for weekly_Sales

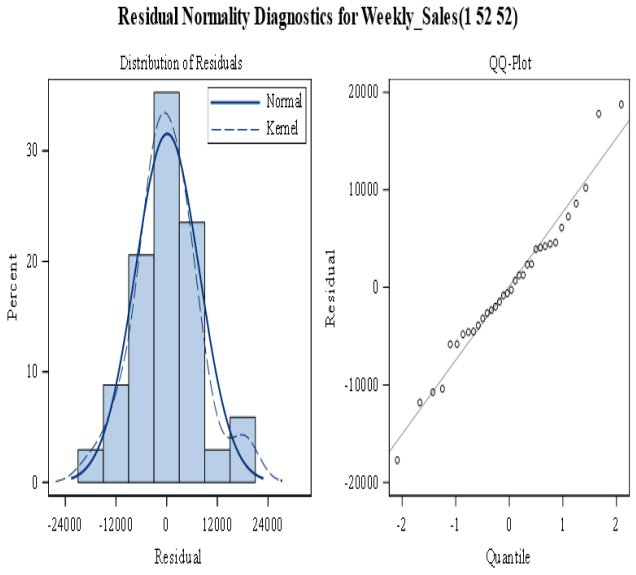

Fig 17.3: Residual Plots

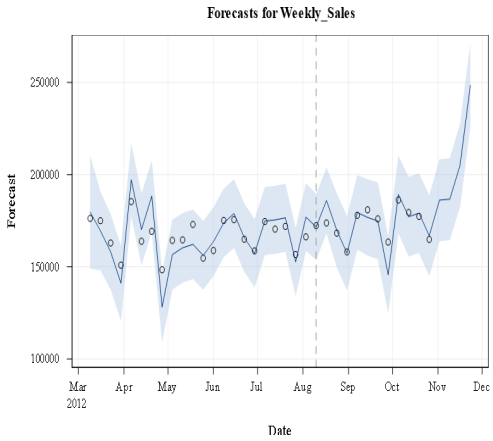

Fig 17.4: Forecasts for weekly_Sales

In this model as checked from Cross-correlation graph we took into consideration the lag 4 effect of markdown_3 on weekly_sales. After running the model we found that there was not significant correlation between residuals and we achieved complete white noise. Also the model performed much better than any other model on both training data and validation data.

	TYPE	_STAT_	_VALUE_
1	ML	AIC	729.45247182
2	ML	SBC	746.24243759
3	ML	LOGLIK	-353.7262359
4	ML	SSE	1899734230.8

Fig 17.5: Output Statistics

Model Comparison:

Model	p	d	q	P	D	Q	AIC	SBC	SSE	MAPE(%)
ARIMAX	4	1	5	0	2	0	729	746	1899734230	1.71
ARIMAX	4	1	4	0	1	0	1756	1780	2840226278	1.72
ARIMA	4	1	5	0	1	0	1841	1866	3073522932	1.75
ARIMA	4	0	4	0	1	0	1862	1885	3297549123	1.82

Fig 18 : Model Comparison

After comparing all the four models(**Fig 18**) we found the **ARIMAX(A(p,d,q)(P,D,Q)s = (4,1,5)(0,2,0)) with lag 4 effect** has performed well on both training(AIC,SBS) and validation data(SSE) also the model accuracy(98.3%) is highest amongst all.

7. Business Insights and Recommendations:

Markdown Variable :

- Markdown_3 has a positive impact on weekly sales of Store 4- Department 9.

Based on the ARIMAX model that was used to forecast Walmart sales for Store number 2 and department 92, the following business insights and recommendations can be made:

- Managing Inventory: The sales forecasting model can help Department 92 of Store 4 in managing their inventory effectively by providing insights into the expected demand for their products during different periods. By using the predicted values, the department can plan their inventory accordingly and ensure that they have sufficient stock to meet customer demand during the holiday season. This will reduce the risk of stockouts and overstocking, leading to lower costs and better customer satisfaction.
- Setting Sales Target: The sales forecasting model can also be used to set sales targets for Department 92 of Store 4. By analyzing the historical data and forecasting future sales, the department can set realistic and achievable sales targets. This will help them focus their efforts on achieving these targets and improving their performance. By doing so, they can increase their revenue and profitability.
- Improving Financial Planning: The sales forecasting model can be used to improve financial planning for Department 92 of Store 4. By predicting the expected sales for different periods, the department can plan their expenses and investments accordingly. They can also identify areas where they can reduce costs or increase efficiency to improve their profitability. This will help them make better financial decisions and achieve their financial goals.
- Responding to Market Changes: The sales forecasting model can help Department 92 of Store 4 in responding to market changes. By analyzing the predicted values and comparing them with actual sales, the department can identify any deviations and take corrective actions. This will help them adapt to changing market conditions and stay competitive. They can also use the insights from the model to identify new market opportunities and expand their product offerings.
- In summary, the ARIMAX model used for Walmart sales forecasting can provide valuable insights for Department 92 of Store 4. By leveraging these insights, the department can improve their inventory management, set realistic sales targets,

improve financial planning, and respond to market changes effectively. These actions can lead to lower costs, higher revenues, and better decision making for the department.

8. References:

- Walmart Sales Dataset of 45stores | Kaggle
- https://pandas.pydata.org
- https://chrisgrannan.medium.com/time-series-analysis-with-exponential-smoothing-d3ad82d47ab0
- https://medium.com/@lawrence.may/comparing-holt-winters-exponential-smoothing-and-arima-models-for-time-series-analysis-659d6f7738c1
- https://towardsdatascience.com/time-series-in-python-exponential-smoothing-and-arima-processes-2c67f2a52788
- EDA on whole data - https://colab.research.google.com/drive/19ACPjMbYiiWSe0oWnj0ti8ESwy2gChzH?usp=sharing
- EDA (Store 4 & Department 92) - https://colab.research.google.com/drive/15dtmkohsRJ5xSsEwFjXBs91BxEL7a3wg?usp=sharing