Walmart Sales Forecasting Model Using SAS

Group 9

Rohit Akole, Ritika Ghosh, Sai Sri Movya Sonti, Harikrupa Vedere, Manikanta Chinta

University of Connecticut

OPIM 5671: Data Mining and Business Intelligence

Prof. Sudip Bhattacharjee

# Abstract

In this forecasting project we are using SAS Studio 3.8. This project demonstrates the seasonality and trend in the model and based on the time series exploration, we have created an ARIMAX model on Walmart Sales Dataset. (Ahmedov, n.d.)

*Keywords:* *SAS, Modeling, SAS Studio, ARIMAX, Time Series Exploration, Walmart Sales*

# Executive Summary

We had 3 files namely features, train, store. Using SQL left join, we merged the files using foreign keys.

1. For the Store table, we used Store as the foreign key.
2. For Features, we used Store and Date as the foreign key.
3. We imported the data from SQL as a .csv file and imported the .csv file in the JMP.
4. Recorded all the null values to 0 from MarkDown1 through MarkDown5.
5. We did the sum of all MarkDown1 through MarkDown5 and named it Sum\_MarkDown.
6. We created the new column month and year and aggregated the weekly data in the monthly data with Summary feature of JMP. We grouped the data using the Month\_Year column and did mean of all variables to get the average values and sum of weekly sales.
7. We have created 5 bins based on the incremental size of 50000 and created 2 models for each bin. One model includes sum\_MarkDown along with all other variables being constant and for the other model we used MarkDown1 through MarkDown 5 instead of Sum\_MarkDown.
8. Upon looking at the time series exploration, we found out that the sales are opposite to the seasonal trend of 2010 and 2011 from March through August.
9. There are spikes in February - March, Huge dip from August - September, again Rise in the Holiday season of September - December and again huge dip after holiday season in January.
10. Trend component shows declining trend, cross-correlation of Sum\_Sales and IsHoliday shows spikes for holidays. But since the IsHoliday is an intervention variable we didn’t prewhiten IsHoliday.
11. Temperature has seasonal spikes, so we had to prewhiten the temperature variable.
12. All other variables show no irregular or seasonal spikes in cross-correlation. So we didn’t apply prewhitening to other variables.
13. For all the stores in different bins, only temperature has seasonal spikes, so we did prewhitening on the temperature variable. For the prewhitening of Temperature, we had to do differencing to deal with seasonal spikes.

# 

# Business Case

In the highly competitive retail industry, understanding customer purchasing patterns is crucial for a company like Walmart. Seasonal fluctuations in sales, if not accurately predicted, could result in significant financial losses due to overstocking or understocking.

Walmart needs a robust strategy to predict future sales accurately. This strategy is pivotal for efficient inventory management, accurate revenue projections, and effective investment strategies. Moreover, achieving sales targets from the beginning of the season can positively influence Walmart’s revenue.

We propose to develop a sales forecasting model that takes into account various factors such as Temperature, Fuel price, Markdowns, Holidays, Unemployment, and the size of the store. This model will analyze the trend and seasonality of the sales based on these factors and provide accurate sales predictions.

## Benefits

1. Inventory Management: Accurate sales forecasting will help Walmart maintain optimal inventory levels, reducing the costs associated with overstocking or understocking.
2. Revenue Projections: With accurate sales forecasts, Walmart can make more precise revenue projections, aiding in financial planning and investment strategies.
3. Investment Strategies: Understanding sales patterns can guide Walmart’s investment strategies, ensuring resources are allocated where they will yield the most return..

By implementing this sales forecasting model, Walmart can stay ahead in the competitive retail industry, ensuring customer satisfaction while maximizing profitability.

# 

# Dataset Schema

|  |  |  |
| --- | --- | --- |
| Column Name | Datatype | Description |
| Month\_Year | Date | Monthly date |
| IsHoliday | Numerical | Holiday Yes or No |
| Sum\_Weekly\_Sales\_ | Numerical | Monthly Sale |
| Mean\_Fuel\_Price\_ | Numerical | Monthly Aggregated Average of Fuel Price |
| Mean\_Unemployment\_ | Numerical | Monthly Aggregated Average of Unemployment |
| Mean\_Tempurature\_ | Numerical | Monthly Aggregated Average of Temperature |
| Mean\_CPI\_ | Numerical | Monthly Aggregated Average of CPI |
| Mean\_MarkDown1\_ | Numerical | Monthly Aggregated Average of MarkDown1 |
| Mean\_MarkDown2\_ | Numerical | Monthly Aggregated Average of MarkDown2 |
| Mean\_MarkDown3\_ | Numerical | Monthly Aggregated Average of MarkDown3 |
| Mean\_MarkDown4\_ | Numerical | Monthly Aggregated Average of MarkDown4 |
| Mean\_MarkDown5\_ | Numerical | Monthly Aggregated Average of MarkDown5 |
| Mean\_Sum\_MarkDown\_ | Numerical | Monthly Aggregated Average sum of All MarkDowns |

### Dataset Link: https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast/?select=features.csv

# 

# Time Series Exploration

Time series exploration suggests there is a seasonal trend in the data. As per the seasonal cycle suggests, the sale cycle is reversed for 2012 as compared to 2010 and 2011 from March till August. Subsequently declining pre-holiday season and post holiday season. Correlations for Sum\_Weekly\_Sales\_ suggests there is possibility of the modeling since the white noise is statistically significant. Trend component suggests there is a slight declining trend but it’s almost flat. Dickey-Fuller Unit Root Test has p-values more than 0.05 for zero mean and single mean.

## Bin 1:

### Model with All MarkDowns:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Augmented Dickey-Fuller Unit Root Tests** | | | | | | | |
| **Type** | **Lags** | **Rho** | **Pr < Rho** | **Tau** | **Pr < Tau** | **F** | **Pr > F** |
| **Zero Mean** | 0 | -6.9169 | 0.0634 | -1.88 | 0.0573 |  |  |
|  | 1 | -1.9627 | 0.3309 | -0.92 | 0.3130 |  |  |
|  | 2 | -0.7100 | 0.5212 | -0.59 | 0.4557 |  |  |
| **Single Mean** | 0 | -45.5214 | 0.0003 | -6.86 | 0.0001 | 23.54 | 0.0010 |
|  | 1 | -44.2920 | 0.0003 | -4.63 | 0.0006 | 10.72 | 0.0010 |
|  | 2 | -19.1565 | 0.0068 | -2.62 | 0.0980 | 3.42 | 0.2265 |
| **Trend** | 0 | -45.5882 | <.0001 | -6.78 | <.0001 | 22.99 | 0.0010 |
|  | 1 | -44.3343 | <.0001 | -4.60 | 0.0035 | 10.74 | 0.0010 |
|  | 2 | -19.2368 | 0.0444 | -2.62 | 0.2759 | 3.55 | 0.4884 |

After looking at the Seasonal Cycles plot for Sum\_Weekly\_Sales\_, we can see that there is an opposite seasonal trend which starts from April 2012 till August 2012 as compared to the same seasons of 2010 and 2011. After the holiday season from September to December, there is a steep decline in December till February for the sales for bin 1 store and the sale again picks up in March.

As per the correlations for the Sum\_Weekly\_Sales plot, the white noise suggests there are variables in our model which need modeling and forecasting. The PACF suggests statistically significant values at lag 5 and 6. Also, for ACF, lag 12 suggests statistical significance.

As the Seasonal Component suggests, there is seasonality in the data with rises around holiday season and steep decline right after that. The Trend Component is relatively flat with no strong upward or downward trend. However, the actual data around the trend line is fluctuating a lot, which suggests there is variability in the Monthly Sales (Sum\_Weekly\_Sales\_ column). Despite short term highs and lows, the trend remains around the same level.

The Cross-Correlations plots suggests all variables fall under CCF with no statistically significant spikes except for Mean\_Temperature\_ and IsHoliday. There are seasonal high and low spikes in the temperature. It suggests there is a need for prewhitening for this variable.

Since the IsHoliday is an intervention variable, we don’t perform prewhitening operations on it. Though the cross-correlations suggests there are positive statistical significant seasonal spikes at lag 4, 6, 8 and negative statistical significant seasonal spikes at lag 0, and 12.

As per the Augmented Dickey-Fuller Unit Root Tests Analysis (SAS Institute Inc., 2019), the table suggests there is some non-stationary data at Single Mean Lag 2 and Trend Lag 2 as per the p-value of Tau and p-value of F. This result suggests we will have to perform some differencing on the data to make it stationary.

### Model with Sum of MarkDowns variable:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Augmented Dickey-Fuller Unit Root Tests** | | | | | | | |
| **Type** | **Lags** | **Rho** | **Pr < Rho** | **Tau** | **Pr < Tau** | **F** | **Pr > F** |
| **Zero Mean** | 0 | -6.9169 | 0.0634 | -1.88 | 0.0573 |  |  |
|  | 1 | -1.9627 | 0.3309 | -0.92 | 0.3130 |  |  |
|  | 2 | -0.7100 | 0.5212 | -0.59 | 0.4557 |  |  |
| **Single Mean** | 0 | -45.5214 | 0.0003 | -6.86 | 0.0001 | 23.54 | 0.0010 |
|  | 1 | -44.2920 | 0.0003 | -4.63 | 0.0006 | 10.72 | 0.0010 |
|  | 2 | -19.1565 | 0.0068 | -2.62 | 0.0980 | 3.42 | 0.2265 |
| **Trend** | 0 | -45.5882 | <.0001 | -6.78 | <.0001 | 22.99 | 0.0010 |
|  | 1 | -44.3343 | <.0001 | -4.60 | 0.0035 | 10.74 | 0.0010 |
|  | 2 | -19.2368 | 0.0444 | -2.62 | 0.2759 | 3.55 | 0.4884 |

The 2nd model suggests the same as bin 1 model 1. All the plots and charts are pretty similar. The only difference in this model is, we have taken the sum of all 5 MarkDowns and created a new column. To see whether it suggests some changes in the trend not as separate MarkDown, but as a whole.

Augmented Dickey-Fuller Unit Root Tests again suggests the same thing as the previous model. The model needs to be stationary. We will need to perform Differencing (d) and Seasonal differencing (D) to make the data stationary (Yaffee & McGee, 2000).

## 

## Bin 2:

### Model with All MarkDowns:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Augmented Dickey-Fuller Unit Root Tests** | | | | | | | |
| **Type** | **Lags** | **Rho** | **Pr < Rho** | **Tau** | **Pr < Tau** | **F** | **Pr > F** |
| **Zero Mean** | 0 | -6.6856 | 0.0684 | -1.82 | 0.0655 |  |  |
|  | 1 | -1.5970 | 0.3773 | -0.79 | 0.3664 |  |  |
|  | 2 | -0.4676 | 0.5722 | -0.40 | 0.5306 |  |  |
| **Single Mean** | 0 | -46.3773 | 0.0003 | -7.01 | 0.0001 | 24.58 | 0.0010 |
|  | 1 | -40.2421 | 0.0003 | -4.52 | 0.0007 | 10.26 | 0.0010 |
|  | 2 | -18.8362 | 0.0075 | -2.73 | 0.0780 | 3.74 | 0.1497 |
| **Trend** | 0 | -47.0600 | <.0001 | -7.04 | <.0001 | 24.78 | 0.0010 |
|  | 1 | -41.3826 | <.0001 | -4.44 | 0.0053 | 10.05 | 0.0010 |
|  | 2 | -18.7695 | 0.0504 | -2.61 | 0.2804 | 3.62 | 0.4759 |

### Model with Sum of MarkDowns variable:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Augmented Dickey-Fuller Unit Root Tests** | | | | | | | |
| **Type** | **Lags** | **Rho** | **Pr < Rho** | **Tau** | **Pr < Tau** | **F** | **Pr > F** |
| **Zero Mean** | 0 | -6.6856 | 0.0684 | -1.82 | 0.0655 |  |  |
|  | 1 | -1.5970 | 0.3773 | -0.79 | 0.3664 |  |  |
|  | 2 | -0.4676 | 0.5722 | -0.40 | 0.5306 |  |  |
| **Single Mean** | 0 | -46.3773 | 0.0003 | -7.01 | 0.0001 | 24.58 | 0.0010 |
|  | 1 | -40.2421 | 0.0003 | -4.52 | 0.0007 | 10.26 | 0.0010 |
|  | 2 | -18.8362 | 0.0075 | -2.73 | 0.0780 | 3.74 | 0.1497 |
| **Trend** | 0 | -47.0600 | <.0001 | -7.04 | <.0001 | 24.78 | 0.0010 |
|  | 1 | -41.3826 | <.0001 | -4.44 | 0.0053 | 10.05 | 0.0010 |
|  | 2 | -18.7695 | 0.0504 | -2.61 | 0.2804 | 3.62 | 0.4759 |

## Bin 3:

### Model with All MarkDowns:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

### 

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Augmented Dickey-Fuller Unit Root Tests | | | | | | | |
| Type | Lags | Rho | Pr < Rho | Tau | Pr < Tau | F | Pr > F |
| Zero Mean | 0 | -7.4955 | 0.0526 | -1.98 | 0.0462 |  |  |
|  | 1 | -2.0909 | 0.3160 | -0.96 | 0.2937 |  |  |
|  | 2 | -0.8186 | 0.5003 | -0.63 | 0.4360 |  |  |
| Single Mean | 0 | -46.5387 | 0.0003 | -7.05 | 0.0001 | 24.86 | 0.0010 |
|  | 1 | -45.1714 | 0.0003 | -4.66 | 0.0005 | 10.88 | 0.0010 |
|  | 2 | -23.5678 | 0.0015 | -2.82 | 0.0648 | 3.97 | 0.0986 |
| Trend | 0 | -46.7216 | <.0001 | -6.99 | <.0001 | 24.45 | 0.0010 |
|  | 1 | -45.6237 | <.0001 | -4.69 | 0.0027 | 11.21 | 0.0010 |
|  | 2 | -24.0401 | 0.0110 | -2.87 | 0.1825 | 4.37 | 0.3331 |

### Model with Sum of MarkDowns:

### 

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## 

## 

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Augmented Dickey-Fuller Unit Root Tests | | | | | | | |
| Type | Lags | Rho | Pr < Rho | Tau | Pr < Tau | F | Pr > F |
| Zero Mean | 0 | -7.4955 | 0.0526 | -1.98 | 0.0462 |  |  |
|  | 1 | -2.0909 | 0.3160 | -0.96 | 0.2937 |  |  |
|  | 2 | -0.8186 | 0.5003 | -0.63 | 0.4360 |  |  |
| Single Mean | 0 | -46.5387 | 0.0003 | -7.05 | 0.0001 | 24.86 | 0.0010 |
|  | 1 | -45.1714 | 0.0003 | -4.66 | 0.0005 | 10.88 | 0.0010 |
|  | 2 | -23.5678 | 0.0015 | -2.82 | 0.0648 | 3.97 | 0.0986 |
| Trend | 0 | -46.7216 | <.0001 | -6.99 | <.0001 | 24.45 | 0.0010 |
|  | 1 | -45.6237 | <.0001 | -4.69 | 0.0027 | 11.21 | 0.0010 |
|  | 2 | -24.0401 | 0.0110 | -2.87 | 0.1825 | 4.37 | 0.3331 |

### 

### 

### 

## Bin 4:

### Model with All MarkDowns:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Augmented Dickey-Fuller Unit Root Tests** | | | | | | | |
| **Type** | **Lags** | **Rho** | **Pr < Rho** | **Tau** | **Pr < Tau** | **F** | **Pr > F** |
| **Zero Mean** | 0 | -7.0749 | 0.0603 | -1.88 | 0.0584 |  |  |
|  | 1 | -1.9218 | 0.3358 | -0.88 | 0.3275 |  |  |
|  | 2 | -0.5620 | 0.5517 | -0.46 | 0.5084 |  |  |
| **Single Mean** | 0 | -45.9454 | 0.0003 | -6.94 | 0.0001 | 24.10 | 0.0010 |
|  | 1 | -45.0292 | 0.0003 | -4.77 | 0.0004 | 11.43 | 0.0010 |
|  | 2 | -20.7252 | 0.0040 | -2.82 | 0.0646 | 3.98 | 0.0979 |
| **Trend** | 0 | -47.0770 | <.0001 | -7.05 | <.0001 | 24.82 | 0.0010 |
|  | 1 | -48.2145 | <.0001 | -4.75 | 0.0023 | 11.48 | 0.0010 |
|  | 2 | -21.7120 | 0.0221 | -2.70 | 0.2423 | 3.89 | 0.4248 |

### Model with Sum of MarkDowns:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## 

### 

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Augmented Dickey-Fuller Unit Root Tests | | | | | | | |
| Type | Lags | Rho | Pr < Rho | Tau | Pr < Tau | F | Pr > F |
| Zero Mean | 0 | -7.0749 | 0.0603 | -1.88 | 0.0584 |  |  |
|  | 1 | -1.9218 | 0.3358 | -0.88 | 0.3275 |  |  |
|  | 2 | -0.5620 | 0.5517 | -0.46 | 0.5084 |  |  |
| Single Mean | 0 | -45.9454 | 0.0003 | -6.94 | 0.0001 | 24.10 | 0.0010 |
|  | 1 | -45.0292 | 0.0003 | -4.77 | 0.0004 | 11.43 | 0.0010 |
|  | 2 | -20.7252 | 0.0040 | -2.82 | 0.0646 | 3.98 | 0.0979 |
| Trend | 0 | -47.0770 | <.0001 | -7.05 | <.0001 | 24.82 | 0.0010 |
|  | 1 | -48.2145 | <.0001 | -4.75 | 0.0023 | 11.48 | 0.0010 |
|  | 2 | -21.7120 | 0.0221 | -2.70 | 0.2423 | 3.89 | 0.4248 |

### 

## 

## Bin 5:

### Model with All MarkDowns:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

### 

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Augmented Dickey-Fuller Unit Root Tests | | | | | | | |
| Type | Lags | Rho | Pr < Rho | Tau | Pr < Tau | F | Pr > F |
| Zero Mean | 0 | -7.0542 | 0.0607 | -1.93 | 0.0521 |  |  |
|  | 1 | -1.8166 | 0.3487 | -0.89 | 0.3239 |  |  |
|  | 2 | -0.6067 | 0.5423 | -0.57 | 0.4618 |  |  |
| Single Mean | 0 | -49.5550 | 0.0003 | -7.58 | 0.0001 | 28.71 | 0.0010 |
|  | 1 | -51.5421 | 0.0003 | -4.96 | 0.0003 | 12.31 | 0.0010 |
|  | 2 | -19.0124 | 0.0071 | -2.56 | 0.1101 | 3.27 | 0.2625 |
| Trend | 0 | -49.5823 | <.0001 | -7.48 | <.0001 | 28.01 | 0.0010 |
|  | 1 | -51.3651 | <.0001 | -4.91 | 0.0015 | 12.22 | 0.0010 |
|  | 2 | -18.7576 | 0.0506 | -2.53 | 0.3136 | 3.42 | 0.5140 |

### 

### Model with Sum of MarkDowns variable:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## 

### 

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Augmented Dickey-Fuller Unit Root Tests | | | | | | | |
| Type | Lags | Rho | Pr < Rho | Tau | Pr < Tau | F | Pr > F |
| Zero Mean | 0 | -7.0542 | 0.0607 | -1.93 | 0.0521 |  |  |
|  | 1 | -1.8166 | 0.3487 | -0.89 | 0.3239 |  |  |
|  | 2 | -0.6067 | 0.5423 | -0.57 | 0.4618 |  |  |
| Single Mean | 0 | -49.5550 | 0.0003 | -7.58 | 0.0001 | 28.71 | 0.0010 |
|  | 1 | -51.5421 | 0.0003 | -4.96 | 0.0003 | 12.31 | 0.0010 |
|  | 2 | -19.0124 | 0.0071 | -2.56 | 0.1101 | 3.27 | 0.2625 |
| Trend | 0 | -49.5823 | <.0001 | -7.48 | <.0001 | 28.01 | 0.0010 |
|  | 1 | -51.3651 | <.0001 | -4.91 | 0.0015 | 12.22 | 0.0010 |
|  | 2 | -18.7576 | 0.0506 | -2.53 | 0.3136 | 3.42 | 0.5140 |

### 

## 

# Modeling and Forecasting

## Bin 1:

### Prewhitening of Temperature:

|  |  |
| --- | --- |
| ARIMA Model with (0, 0, 0)(0, 0, 0) | |
|  |  |

|  |  |
| --- | --- |
| SARIMA Model with (2, 2, 0)(0, 1, 0) | |
|  |  |

|  |  |
| --- | --- |
| Autoregressive Factors | |
| Factor 1: | 1 + 1.05331 B\*\*(1) + 0.38836 B\*\*(2) |

## 

## 

As per the above ARIMA and SARIMA plots, we can see that the ARIMA model needed on the Temperature variable as it had lags at the beginning of the PACF and IACF as well as there were lags at irregular intervals and seasonal intervals as well. The above SARIMA model was the best model we got through different iterations for bin 1 data.

### Old ARIMAX Model vs. Prewhitened ARIMAX Model with All MarkDowns (Model 1):

|  |  |
| --- | --- |
| **Old ARIMAX Model** | **Prewhitened ARIMAX Model** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |
|  |

As per the AIC and SBC, both models are similar. But the cross-correlation plot suggests that the Temperature variable needs prewhitening for better accuracy. There is seasonality in the Tempurature which we took care of using the SARIMA (2, 2, 0)(0, 1, 0) model. We included the prewhitening settings into the ARIMAX and it gave back better results for the cross-correlations of temperature. Now there are no statistically significant spikes which are crossing the CCF for all variables except IsHoliday.

### Old ARIMAX Model vs. Prewhitened SARIMAX Model with Sum MarkDowns (Model 2):

|  |  |
| --- | --- |
| **Old ARIMAX Model** | **Prewhitened ARIMAX Model** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |
|  |

## Bin 2:

### Prewhitening of Temperature:

|  |  |
| --- | --- |
| Temperature ARIMA Model with (0, 0, 0)(0, 0, 0) | |
|  |  |

|  |  |
| --- | --- |
| Temperature SARIMA Model with (1, 1, 0)(0, 1, 0) | |
|  |  |

### 

### Prewhitening of Sum\_MarkDown variable:

|  |  |
| --- | --- |
| Sum\_MarkDown ARIMA Model with (0, 0, 0)(0, 0, 0) | |
|  |  |

|  |  |
| --- | --- |
| Sum\_MarkDown ARIMA Model with (1, 1, 0)(0, 0, 0) | |
|  |  |

### 

### Old ARIMAX Model vs. Prewhitened ARIMAX Model with All MarkDowns (Model 1):

|  |  |
| --- | --- |
| **Old ARIMAX Model** | **Prewhitened ARIMAX Model** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |
|  |

### 

### 

### Old ARIMAX Model vs. Prewhitened SARIMAX Model with Sum MarkDowns (Model 2):

|  |  |
| --- | --- |
|  |  |
|  |
|  |

## 

## Bin 3:

### Prewhitening of Temperature:

|  |  |
| --- | --- |
| Temperature ARIMA Model with (0, 0, 0)(0, 0, 0) | |
|  |  |

|  |  |
| --- | --- |
| Temperature SARIMA Model with (1, 1, 0)(0, 1, 0) | |
|  |  |

### 

### 

### Old ARIMAX Model vs. Prewhitened ARIMAX Model with All MarkDowns (Model 1):

|  |  |
| --- | --- |
| **Old ARIMAX Model** | **Prewhitened ARIMAX Model** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |
|  |

### ARIMAX Model vs. Prewhitened SARIMAX Model with Sum MarkDowns(Model 2):

|  |  |
| --- | --- |
| **Old ARIMAX Model** | **Prewhitened ARIMAX Model** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |

|  |  |
| --- | --- |
| **Old ARIMAX Model** | **Prewhitened ARIMAX Model** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |
|  |

## 

## Bin 4:

### Prewhitening of Temperature:

|  |  |
| --- | --- |
| Temperature ARIMA Model with (0, 0, 0)(0, 0, 0) | |
|  |  |

|  |  |
| --- | --- |
| Temperature SARIMA Model with (0, 1, 0)(0, 2, 0) | |
|  |  |

### 

### 

### 

### Old ARIMAX Model vs. Prewhitened ARIMAX Model with All MarkDowns (Model 1):

|  |  |
| --- | --- |
| **Old ARIMAX Model** | **Prewhitened ARIMAX Model** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |
|  |

### 

### 

### ARIMAX Model vs. Prewhitened SARIMAX Model with Sum MarkDowns(Model 2):

|  |  |
| --- | --- |
| **Old ARIMAX Model** | **Prewhitened ARIMAX Model** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |
|  |

## 

## 

## Bin 5:

### Prewhitening of Temperature:

|  |  |
| --- | --- |
| Temperature ARIMA Model with (0, 0, 0)(0, 0, 0) | |
|  |  |

|  |  |
| --- | --- |
| Temperature SARIMA Model with (0, 2, 0)(1, 0, 0) | |
|  |  |

### 

### 

### Old ARIMAX Model vs. Prewhitened ARIMAX Model with All MarkDowns (Model 1):

|  |  |
| --- | --- |
| **Old ARIMAX Model** | **Prewhitened ARIMAX Model** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |

### 

### 

### ARIMAX Model vs. Prewhitened SARIMAX Model with Sum MarkDowns(Model 2):

|  |  |
| --- | --- |
| **Old ARIMAX Model** | **Prewhitened ARIMAX Model** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

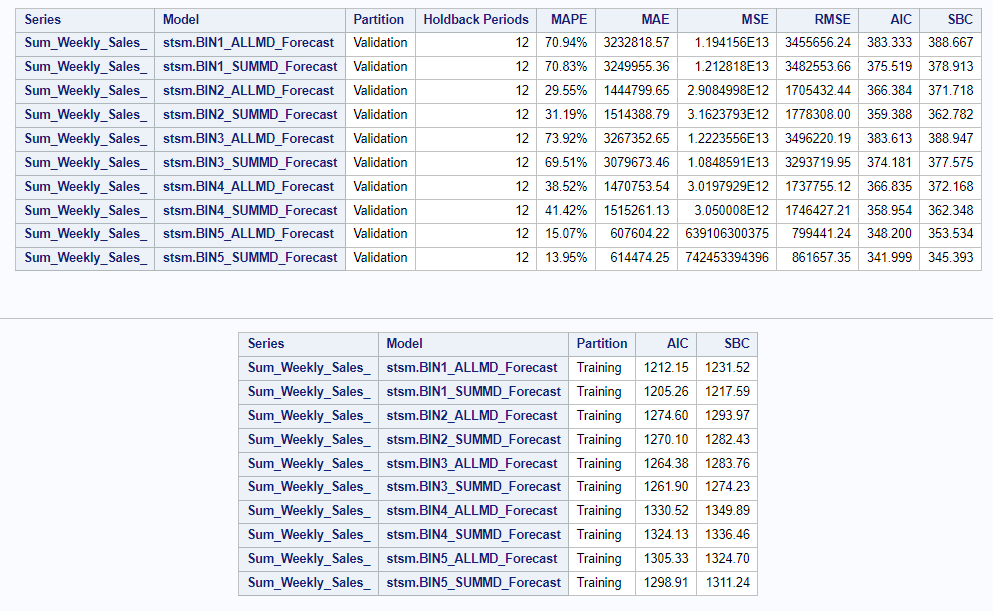
## 

# 

# 

# 

## Statistics



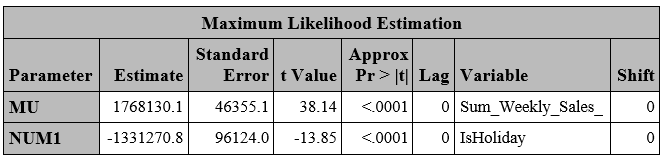
As per the statistics, SUM\_of\_MarkDown worked better as compared to the separated All MarkDowns. Similarly, as per the estimated likelihood table, the p-value of independent suggested they are not statistically significant except for the IsHoliday variable, which was our intervention variable. Hence we came up with the final models for 5 bins focused on the Sum\_of\_MarkDown.

# 

## Final ARIMAX Models for all separate bins

### Bin 1 Model:

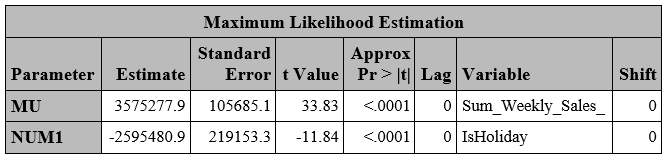
|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |



### Bin 2 Model:

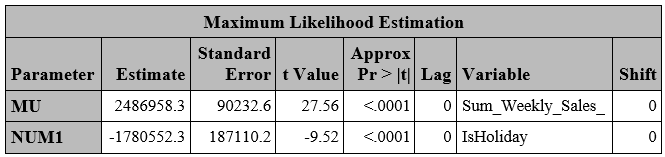
### 

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |



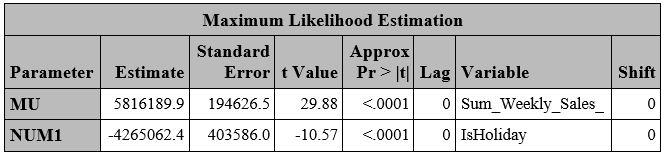
### Bin 3 Model:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |



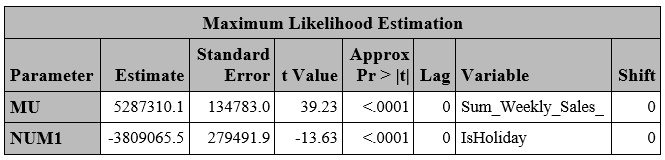
### Bin 4 Model:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |



### Bin 5 Model:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |



## Final Model Statistics

## 

## 

## 

## 

## 

As per the Accuracy(1-MAPE), the accuracy of the model is around 80-90% for all bins. These are the best store models for each Size Bin of the Stores.

# Findings

The data merged from three different files is processed to handle null values and aggregated to monthly figures. The time series exploration revealed seasonal trends and significant sales spikes during holiday seasons, along with a generally flat but fluctuating trend line. The findings emphasize the necessity of prewhitening certain variables like temperature and the importance of including holiday data due to its significant impact on sales. The final ARIMAX models for various store bins demonstrate an accuracy of 80-90%, suggesting robust predictability in sales forecasting. The detailed analysis and modeling efforts provide Walmart with actionable insights for inventory management, revenue projection, and investment strategies, aimed at maximizing profitability and maintaining competitive advantage.

* Seasonal Sales Impact: Department and store sales vary by season, peaking at holidays.
* Holiday Sales Surge: Sales spike on holidays, with Thanksgiving leading the total sales.
* Weekly Sales Patterns: 22nd week and pre-Christmas (51st week) mark sales peaks.
* Post-Holiday Sales Drop: January sees lowest sales following November and December highs.
* Variable Influence: CPI, temperature, unemployment, fuel prices don't predict weekly sales.

# Business Case Recommendations:

* **Seasonal Promotion Optimization:** Capitalize on seasonal variations by tailoring marketing and stock levels to anticipated high sales periods like Thanksgiving, Black Friday, and pre-Christmas (51st week), ensuring product availability aligns with demand spikes.
* **Holiday Sales Planning:** Given the significant sales increase during holiday periods, especially Thanksgiving, planning promotions, and stock levels should be prioritized around these dates. Re-evaluate the assignment of Christmas sales to accurately reflect shopping behavior, focusing on the 51st week rather than the last days of the year.
* **January Sales Strategy:** Develop strategies to counteract the traditional sales dip in January, such as post-holiday sales promotions or loyalty programs that encourage year-start shopping.
* **Data-Driven Inventory Management:** Leverage insights from sales patterns to manage inventory more effectively, reducing overstock and understock situations, especially during peak and trough sales periods.
* **Variable Analysis for Localized Strategies:** Despite the lack of a clear pattern with CPI, temperature, unemployment rates, and fuel prices on a weekly basis, consider deeper analysis or alternative data sources that may offer localized insights for tailored store strategies.

Implementing these recommendations can help in refining marketing efforts, optimizing inventory management, and enhancing overall sales performance throughout the year, adapting to the nuanced needs of different store types and seasonal trends.

**References**

Ahmedov, A. (n.d.). *Walmart Sales Forecast*. www.kaggle.com. https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast/?select=features.csv

SAS Institute. (2019). *Time Series Modeling Essentials, Course Notes*. SAS Institute Inc. LWSTSM51\_001

Yaffee, R. A., & McGee, M. (2000). *An Introduction to Time Series Analysis and Forecasting: With Applications of SAS® and SPSS®*. Elsevier Science.