Time Series Forecasting with Walmart Case-Study

Author: Pranjal Pathak, Akhilesh Patil, Nikhil Madhu Belavinakodige, Prasenjeet Gadhe

Summary

This report provides an in-depth overview of time series analytics, a powerful tool for understanding data patterns and trends over time with a case study on Walmart Sales data. Time series analytics is widely used across various industries, such as finance, retail, healthcare, and manufacturing, to forecast future performance, optimize operations, and inform strategic decision-making. The report covers fundamental concepts, forecasting methods, applications, and case study to illustrate the practical implementation of time series analytics.

The report includes a case study on the Walmart Dataset, where major time series analytics models have been applied to gain insights into sales patterns, seasonal effects, and other factors that influence sales performance. By leveraging these models, Walmart can optimize inventory management, plan promotions, and make informed strategic decisions to drive growth and enhance overall performance. By exploring the basics of time series analytics, common forecasting methods, diverse applications, and real-world case studies, this report aims to equip readers with a comprehensive understanding of time series analytics and its practical implementation in various industries.

1. Introduction

1.1 Background

Time series analytics is the process of analyzing and modeling time-ordered data points to identify trends, patterns, and relationships. It plays a crucial role in various industries, such as finance, retail, healthcare, and manufacturing, for uncovering insights and predicting future outcomes. As data continues to grow exponentially and businesses become increasingly data-driven, the importance of time series analytics has become more pronounced. Organizations that can effectively leverage time series analytics gain a competitive advantage by making informed decisions, optimizing operations, and enhancing overall performance.

1.2 Objective

The primary objective of this project is to provide a comprehensive understanding of time series analytics to help businesses and practitioners harness the power of historical data for improved decision-making. By exploring the fundamental concepts, forecasting methods, applications, and case studies, this report aims to equip readers with a solid foundation in time series analytics, enabling them to:

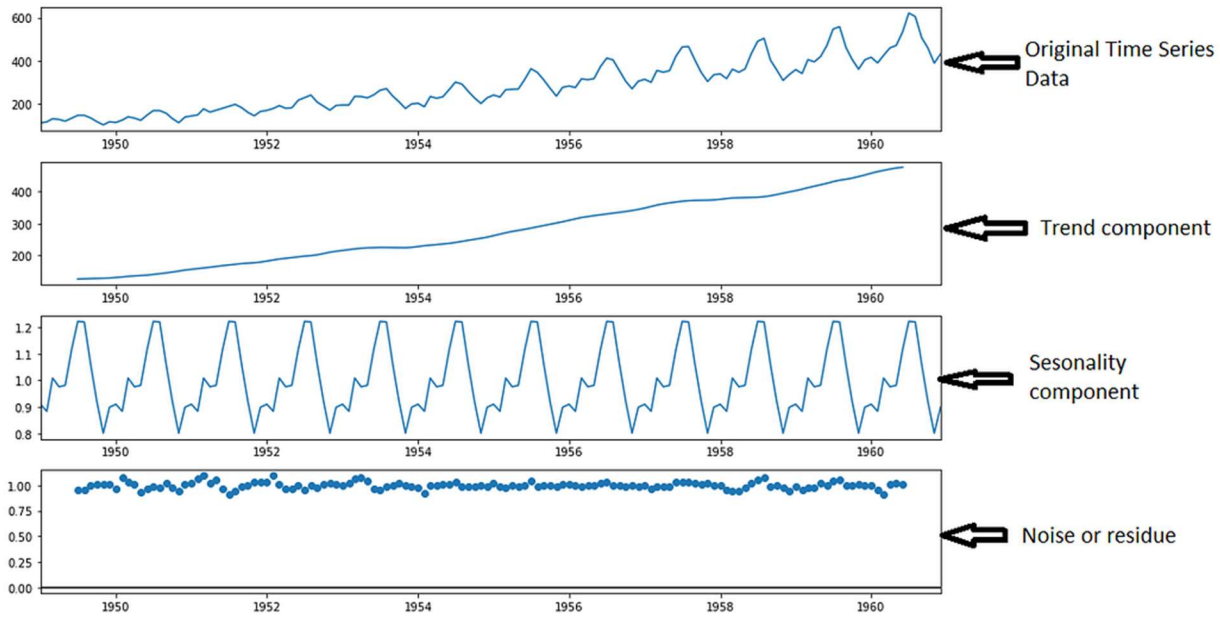
- Grasp the basic components of time series data and the importance of decomposition in analyzing the underlying structure
- Understand the strengths and limitations of common time series forecasting methods, including exponential smoothing, regression, and ARIMA
- Recognize the diverse applications of time series analytics across different industries, such as finance, retail, healthcare, and manufacturing

- Appreciate the practical implementation of time series analytics through real-world case studiy that is Walmart Sales prediction.

2. Time Series Analytics Basics

2.1 Components of Time Series

A time series consists of four main components, which, when combined, help to explain the underlying patterns and behavior of the data. Understanding these components is crucial for accurate time series analysis and forecasting. Below is the example for understanding the Temporal data components.



2.1.1 Trend

Trend represents the overall direction and movement of the data over a long period. It captures the general pattern of growth, decline, or stability in the data. For example, in the context of sales data, an upward trend would indicate increasing sales over time, while a downward trend signifies a decline in sales. Identifying trends helps businesses anticipate future developments and make strategic decisions accordingly.

2.1.2 Seasonality

Seasonality refers to regular fluctuations in the data that typically occur within fixed periods, such as daily, weekly, monthly, or yearly. Seasonal patterns are often related to factors like weather, holidays, and recurring events. For instance, in retail, sales may peak during the holiday season, while in the energy sector, electricity consumption might be higher during summer months due to increased air conditioning usage. Understanding seasonality is essential for businesses to plan and allocate resources effectively.

2.1.3 Cyclical

Cyclical components represent fluctuations in the data that do not follow a fixed pattern and often arise from broader economic cycles or industry-specific factors. These fluctuations can span multiple years and are typically longer than seasonal patterns. Cyclical components may include business cycles, market expansions, and contractions or fluctuations in demand due to external factors. Identifying and understanding cyclical patterns can help businesses prepare for potential downturns or capitalize on growth opportunities.

2.1.4 Irregular

Irregular components, also known as residual or random components, are unpredictable variations in the data that do not follow a discernible pattern. These fluctuations can result from one-off events, such as natural disasters, policy changes, or unexpected market shocks. Since irregular components are challenging to predict, they can introduce uncertainty and noise into time series forecasts. However, understanding the potential impact of irregular components helps businesses from various industry to develop contingency/ emergency planning and respond more effectively to unforeseen events.

2.2 Decomposition

Decomposition is a technique used to separate the different components of a time series to better understand and analyze the underlying structure. By breaking down the time series data into trend, seasonal, and residual components, analysts can more easily identify the primary drivers of the data and develop accurate forecasting models.

There are two main approaches to decomposition: additive and multiplicative. In additive decomposition, the observed data is the sum of the trend, seasonal, and residual components. In contrast, the multiplicative decomposition assumes that the observed data is the product of the three components. The choice between additive and multiplicative decomposition depends on the nature of the data and the presence of any interactions between the components.

Other information sources that drive the market are news articles, tweets and company press releases. It is often observed that stock performs exceptionally well for a brief period when the company declares quarterly profits. In a few other scenarios, tweets put out by prominent people in the industry also greatly influence the stock’s performance momentarily. By extracting historical data on such information, we analyze the correlation between the amount of swing observed in the market and the sentiment of the information.

Using neural networks and deep learning for this task adds to the solution’s complexity, thereby helping us to identify the complex patterns within the time series data from stock markets. However, the more complex the model is, the more difficult it becomes to interpret the model. Thus, enhances the need to explain such models to reduce computer-generated or data inherent biases. To address this issue, we look to introduce explanations for each selected feature and their contributions to making time series forecasts.

3. Time Series Forecasting Methods

There are several methods used for time series forecasting, each with its strengths and limitations. In this section, we will explore five common techniques: Naive method, Moving Average, Exponential Smoothing, Regression, and ARIMA.

3.1 Naïve Method

The Naive method is the simplest time series forecasting technique, assuming that the most recent observation is the best predictor of future values. This method can be useful as a baseline model, but it does not account for trends, seasonality, or other patterns in the data. Ex. Sales of April will be equal to sales of the March.

$y(n+1) = y(n)$

3.2 Moving Average

The Moving Average method calculates the average of a fixed number of past observations to generate a forecast. This approach smooths out short-term fluctuations and noise, making it easier to identify underlying trends. However, the Moving Average method is not well-suited for capturing seasonality or responding quickly to changes in the data. Ex. Sales of April month will be average of the sales of last three months i.e. March, February and January.

$$y_{n+1} = \frac{y_n + y_{n-1} + y_{n-2}}{4}$$

Month	Actual Shed Sales	3-Month Moving Average
January	10	
February	12	
March	13	
April	16	(10 + 12 + 13)/3 = 11 $^2/_3$
May	19	(12 + 13 + 16)/3 = 13 $^2/_3$
June	23	(13 + 16 + 19)/3 = 16
July	26	(16 + 19 + 23)/3 = 19 $^1/_3$

3.3 Exponential Smoothing

Exponential smoothing is a family of forecasting techniques that applies weighted averages to past observations to make forecasts. The weight assigned to each observation decreases exponentially as it moves further back in time. This method emphasizes the most recent data points, allowing it to capture trends and seasonality in time series data effectively.

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots,$$

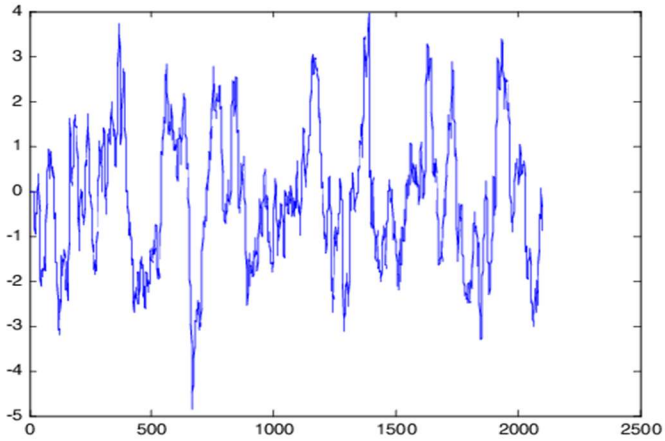
Smoothing Constant	Weight Assigned to Most Recent Period (α)	2nd Most Recent Period $\alpha(1-\alpha)$	3rd Most Recent Period $\alpha(1-\alpha)^2$	4th Most Recent Period $\alpha(1-\alpha)^3$	5th Most Recent Period $\alpha(1-\alpha)^4$
$\alpha = .1$	.1	.09	.081	.073	.066
$\alpha = .5$	.5	.25	.125	.063	.031

There are three main types of exponential smoothing:

3.3.1. Simple Exponential Smoothing (SES)

Simple Exponential Smoothing is suitable for time series data with no clear trend or seasonality basically random data. It calculates the weighted average of past observations, giving more weight to recent data points. SES produces a forecast by applying a smoothing factor (alpha) between 0 and 1.

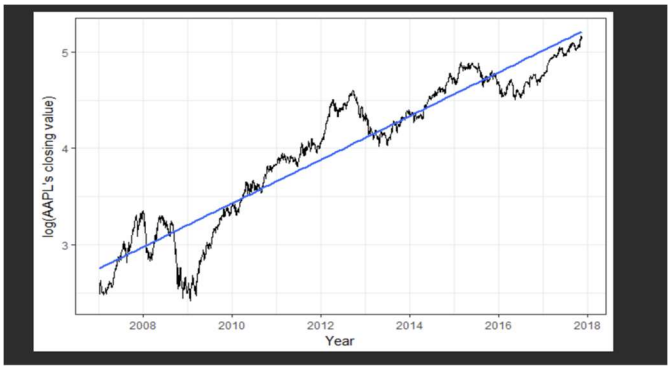
Random Data with no trend and no Seasonality



3.3.2. Holt's Linear Exponential Smoothing

Holt's Linear Exponential Smoothing extends SES to capture trends in the data. It introduces an additional smoothing factor (beta) for the trend component. This method calculates two equations: one for the level (local average) and another for the trend. The forecasts are generated by combining these two components.

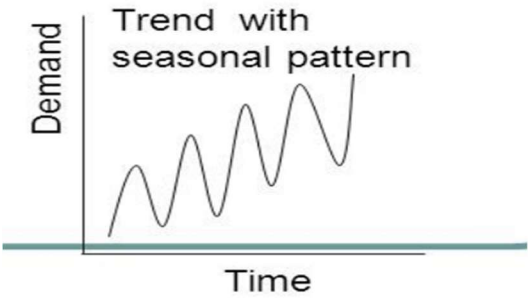
Linear Trend with no Seasonality



3.3.3. Holt-Winters Seasonal Exponential Smoothing

Holt-Winters Seasonal Exponential Smoothing is designed to capture both trend and seasonality in the time series data. It introduces a third smoothing factor (gamma) for the seasonal component. This method calculates three equations: one for the level, one for the trend, and one for the seasonal component. The forecasts are generated by combining all three components.

Linear Trend with Seasonality



3.4. Regression

Regression is a statistical method used to model the relationship between a dependent variable (e.g., sales) and one or more independent variables (e.g., time, season, or external factors). In the context of time series forecasting, regression can be applied to model trends and other patterns in the data.

3.4.1. Linear Regression

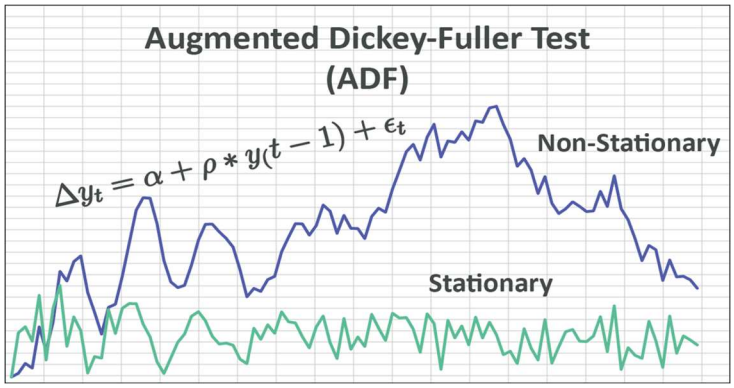
Linear regression is the most basic form of regression, modeling the relationship between the dependent and independent variables using a straight line. It is useful for capturing linear trends in time series data. However, linear regression has limitations in capturing complex patterns, seasonality, and cyclical components. In cases where these factors are present, more advanced regression techniques or other forecasting methods may be more appropriate.

3.4.2. Multiple Regression

Multiple regression is an extension of linear regression, allowing for the inclusion of additional independent variables in the model. This approach can help account for factors such as seasonality, cyclical components, or other external variables that may influence the dependent variable. Multiple regression can improve forecasting accuracy by incorporating more information into the model, but it can also become more complex and computationally intensive as the number of independent variables increases.

3.5 ARIMA

ARIMA (Autoregressive Integrated Moving Average) is a widely used technique for time series forecasting. It models the data as a combination of autoregressive (AR), moving average (MA), and differencing (I) components. One important assumption for ARIMA model is that data should be stationary that is it has constant mean and constant variance. Methods like Augmented Dickey-Fuller Test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test helps to find the stationariness of the temporal data.

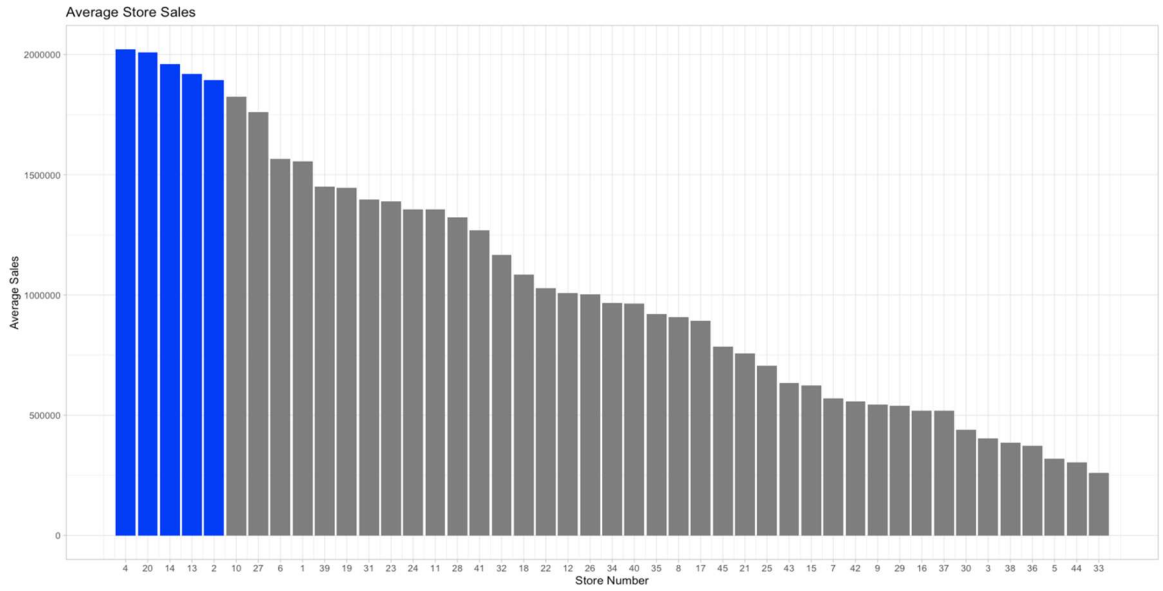


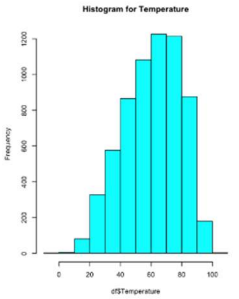
By understanding the strengths and limitations of these forecasting methods, analysts can choose the most appropriate approach for their specific time series data and use case. Combining or adjusting these methods based on the unique characteristics of the data can lead to improved forecasting accuracy and better-informed decision-making.

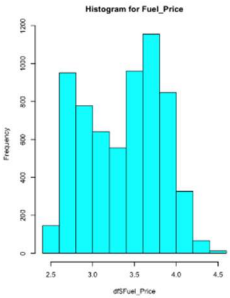
4. Case-Study Walmart Data

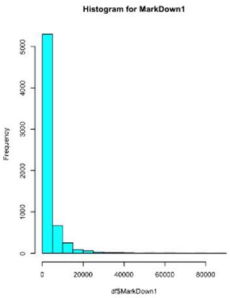
Three years data of the 45 stores of the Walmart was selected for analysis and prediction.

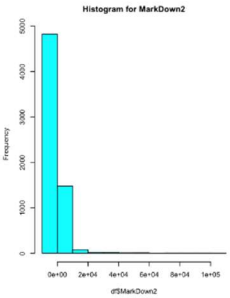
EDA: Top 5 stores with highest sales were selected for the analysis. Store number 4, 20, 14, 13 & 2 were selected for the analysis and prediction. Outliers were taken care by replacing them with the mean of the respective column. All the distributions were not normal hence were explored with transforming them to logarithmic, square, square root, etc. Most of the columns were not correlated, The highest correlation was 0.7 between MarkDown1 and MarkDown4.

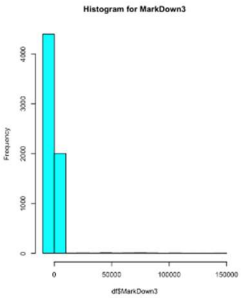


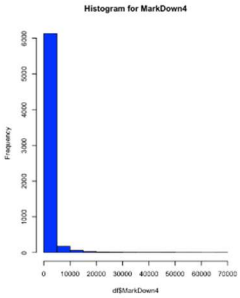

Histogram for Temperature
Frequency
df$Temperature

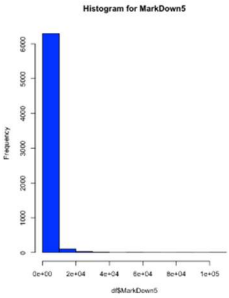

Histogram for Fuel_Price
Frequency
df$Fuel_Price

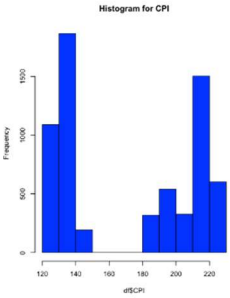

Histogram for MarkDown1
Frequency
df$MarkDown1

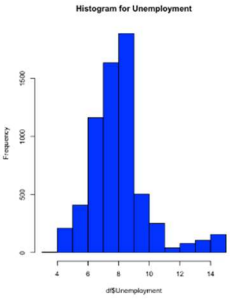

Histogram for MarkDown2
Frequency
df$MarkDown2

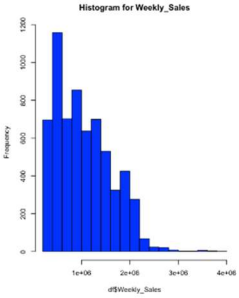

Histogram for MarkDown3
Frequency
df$MarkDown3

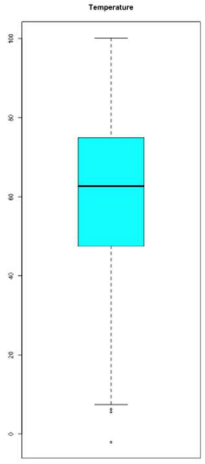

Histogram for MarkDown4
Frequency
df$MarkDown4

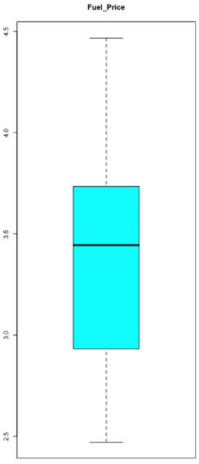

Histogram for MarkDown5
Frequency
df$MarkDown5

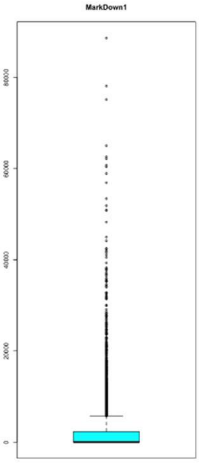

Histogram for CPI
Frequency
df$CPI

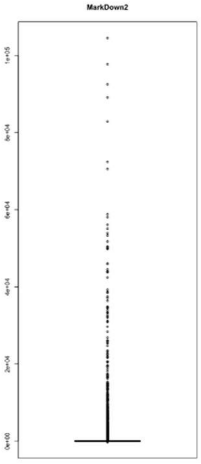

Histogram for Unemployment
Frequency
df$Unemployment

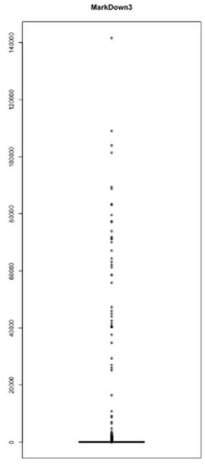

Histogram for Weekly_Sales
Frequency
df$Weekly_Sales

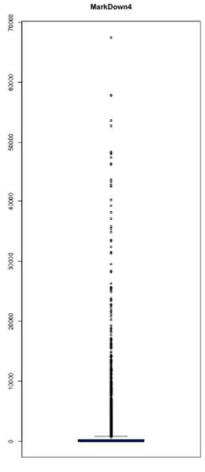

Temperature

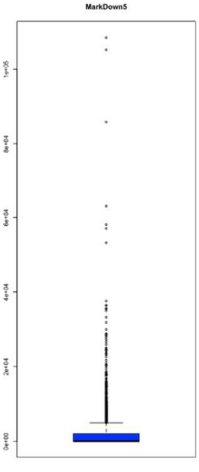

Fuel_Price

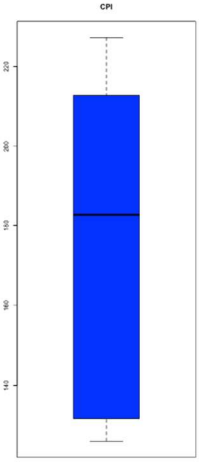

MarkDown1

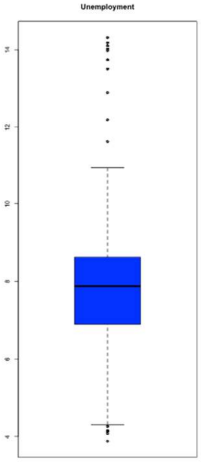

MarkDown2

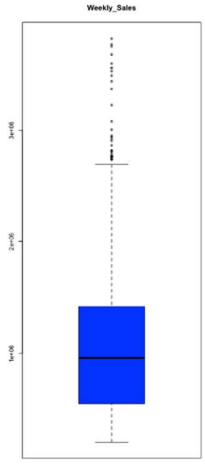

MarkDown3


MarkDown4

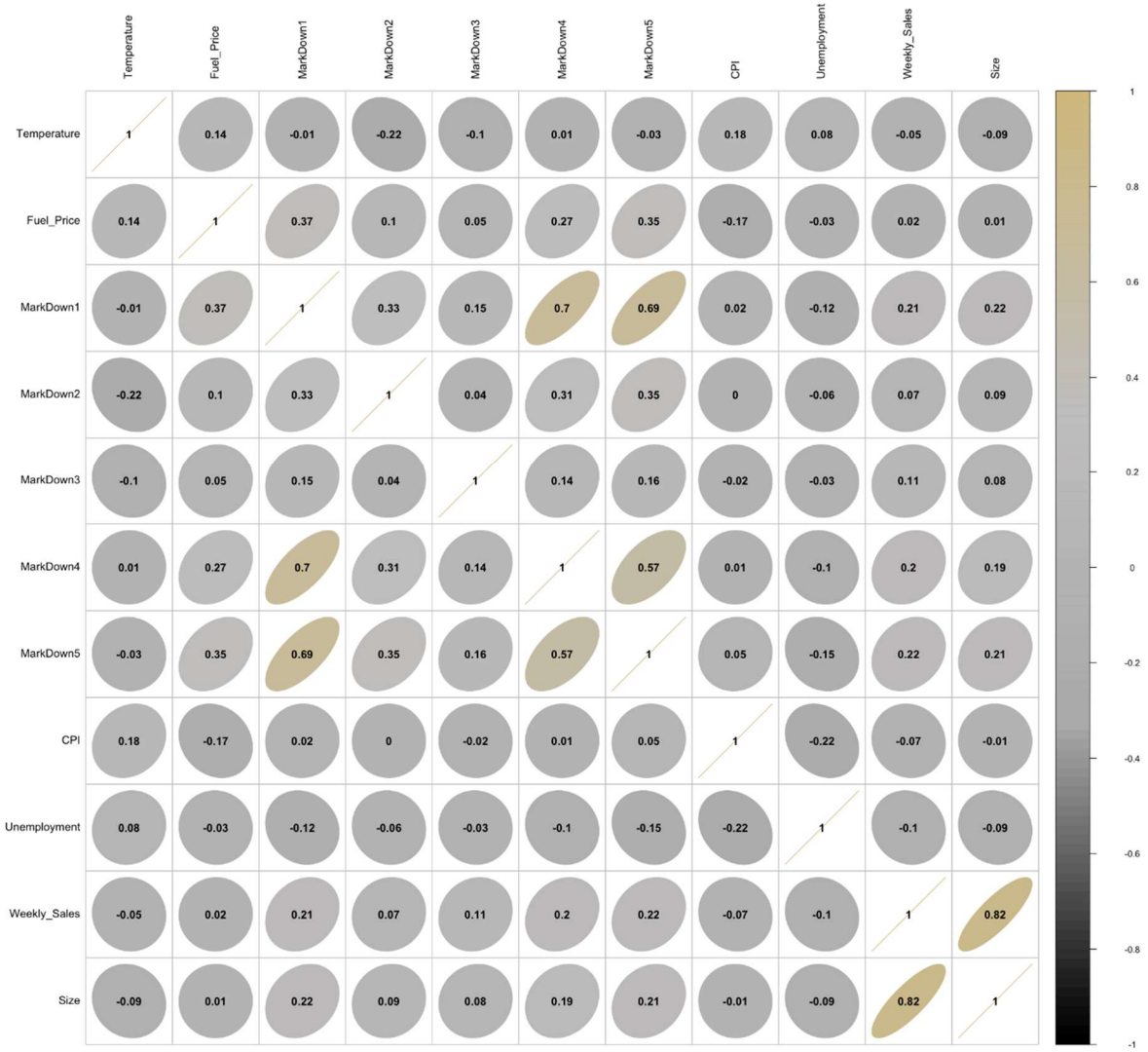

MarkDown5


CPI

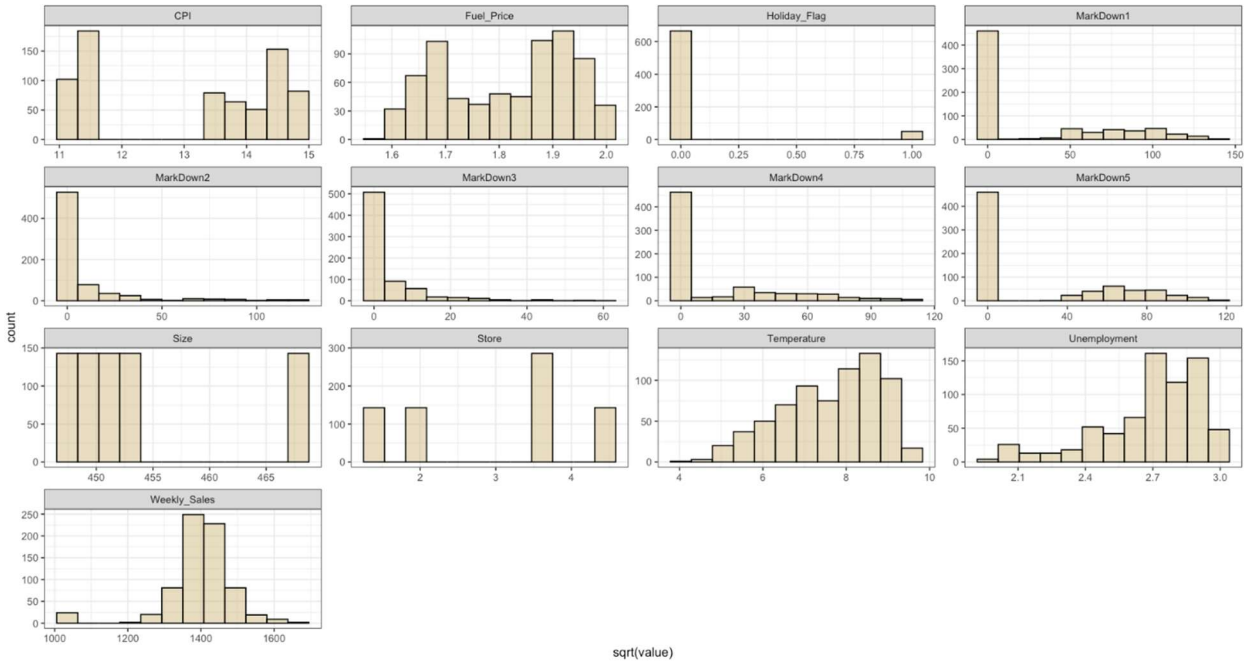

Unemployment


Weekly_Sales

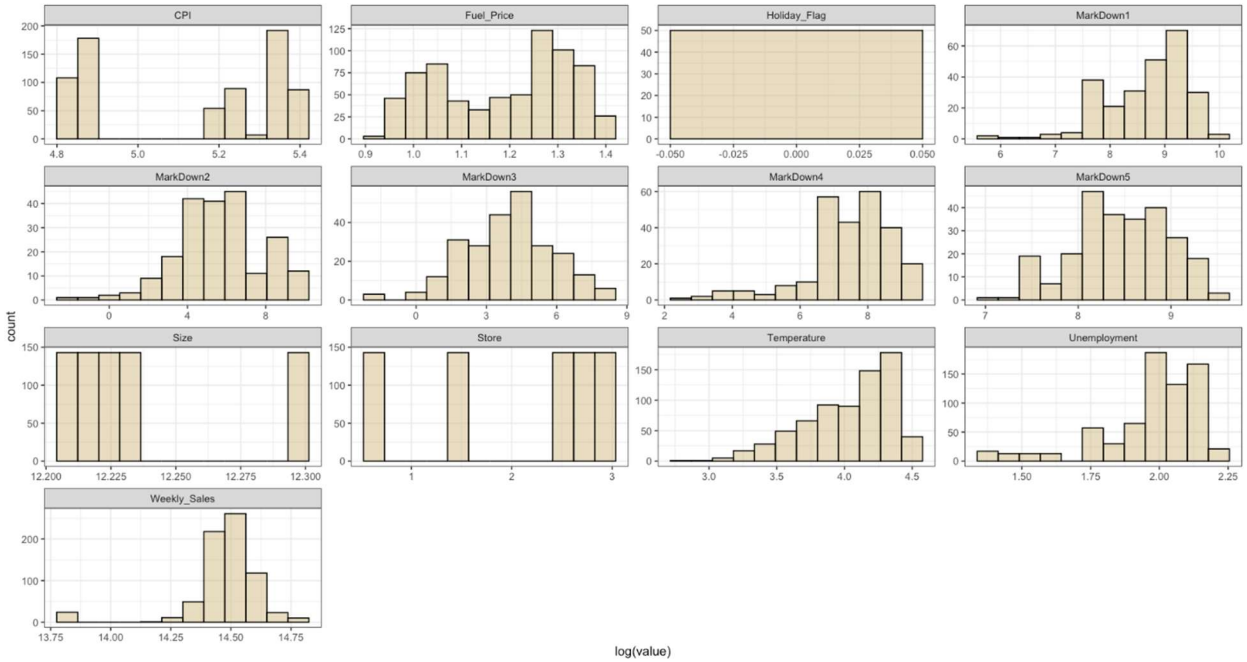
Correlation Matrix:



Log Transformed:

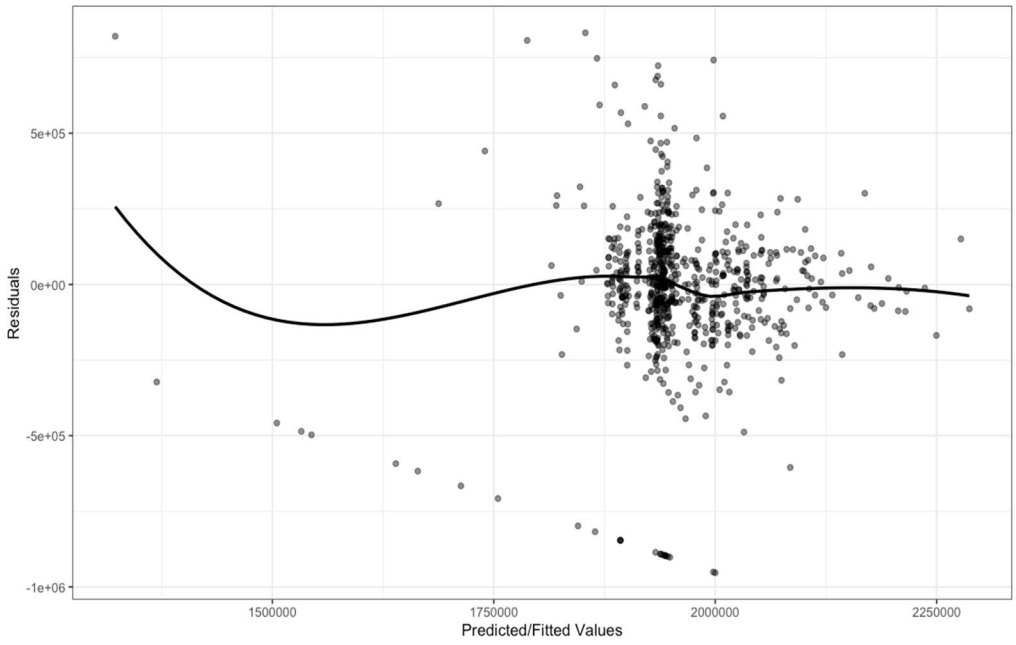


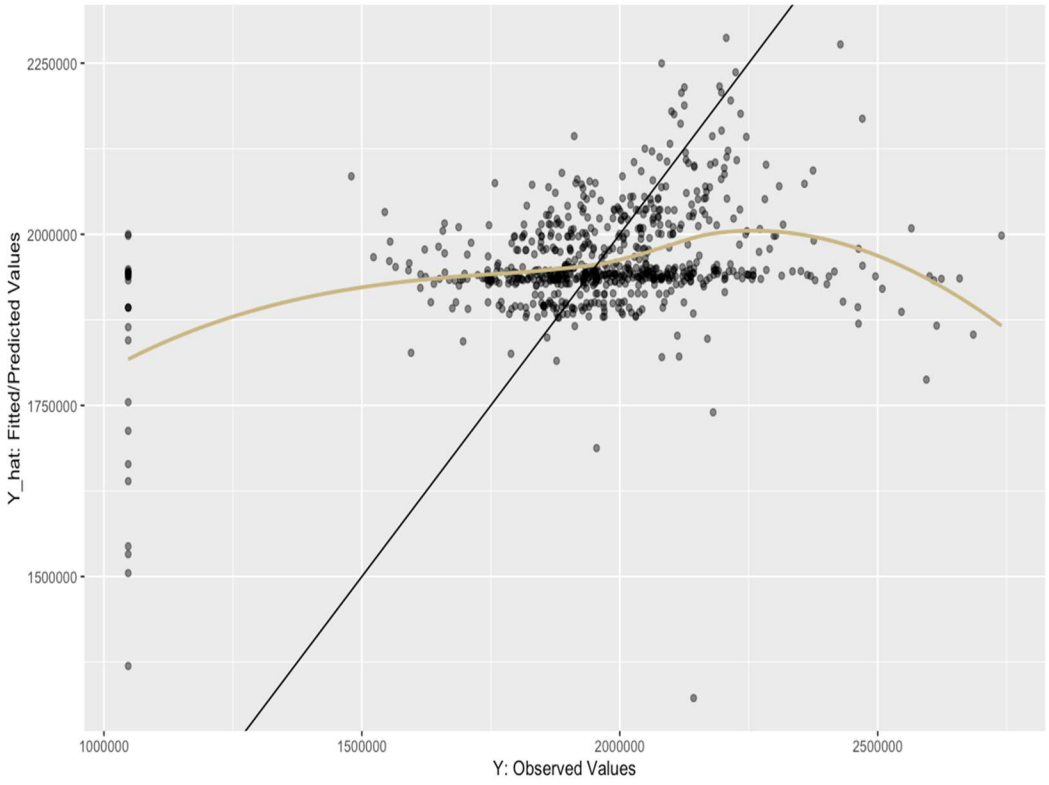
Square root transformation

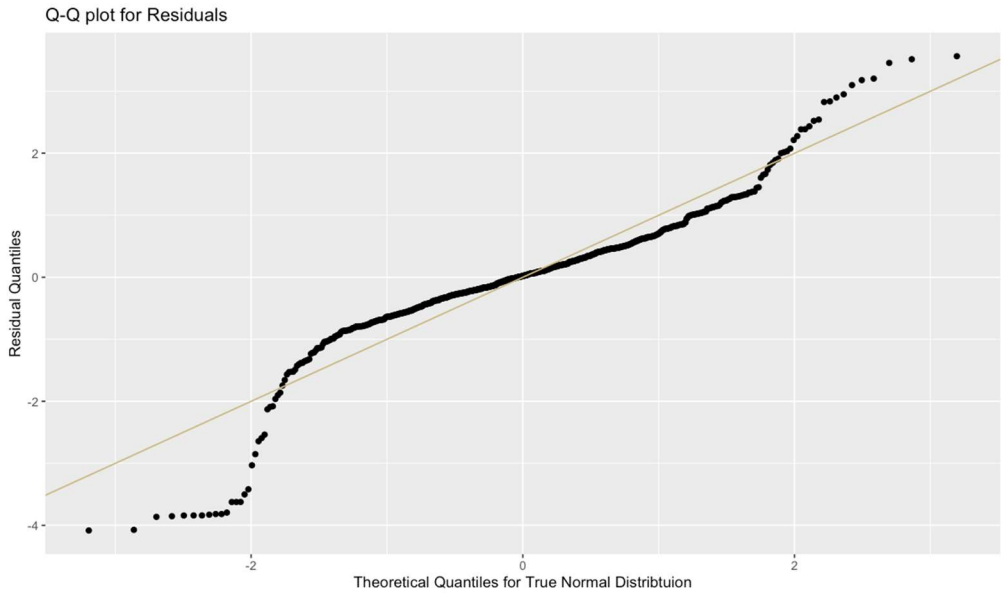


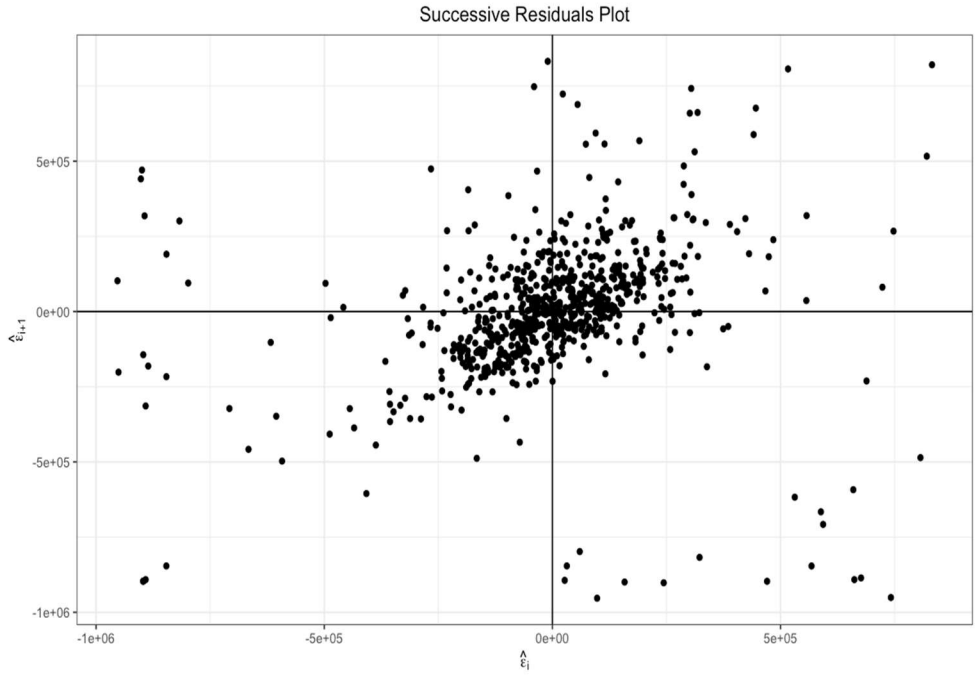
Linear Regression Diagnostics:

Most of the Linear Regression assumptions are violated.





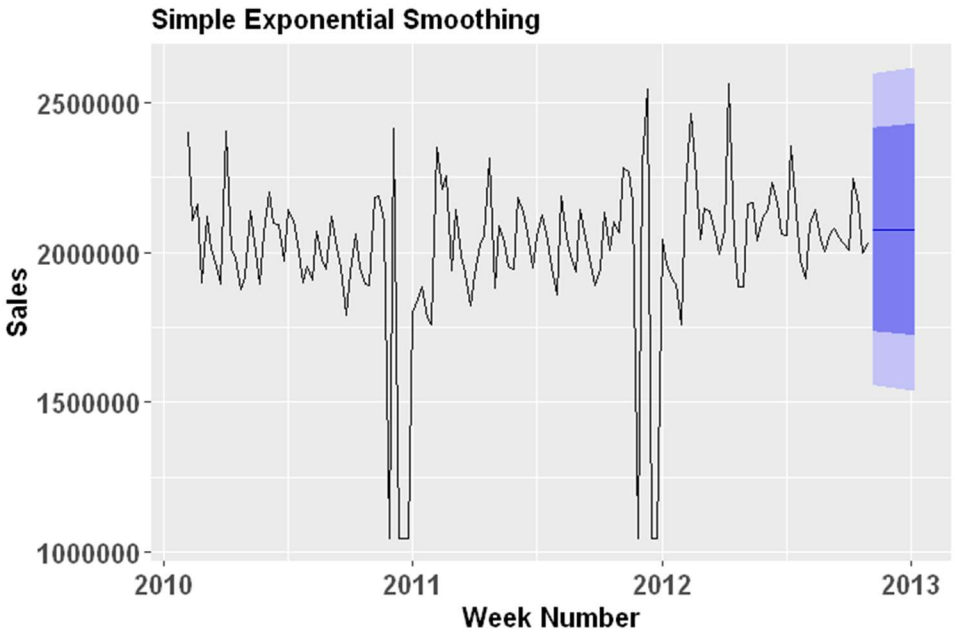




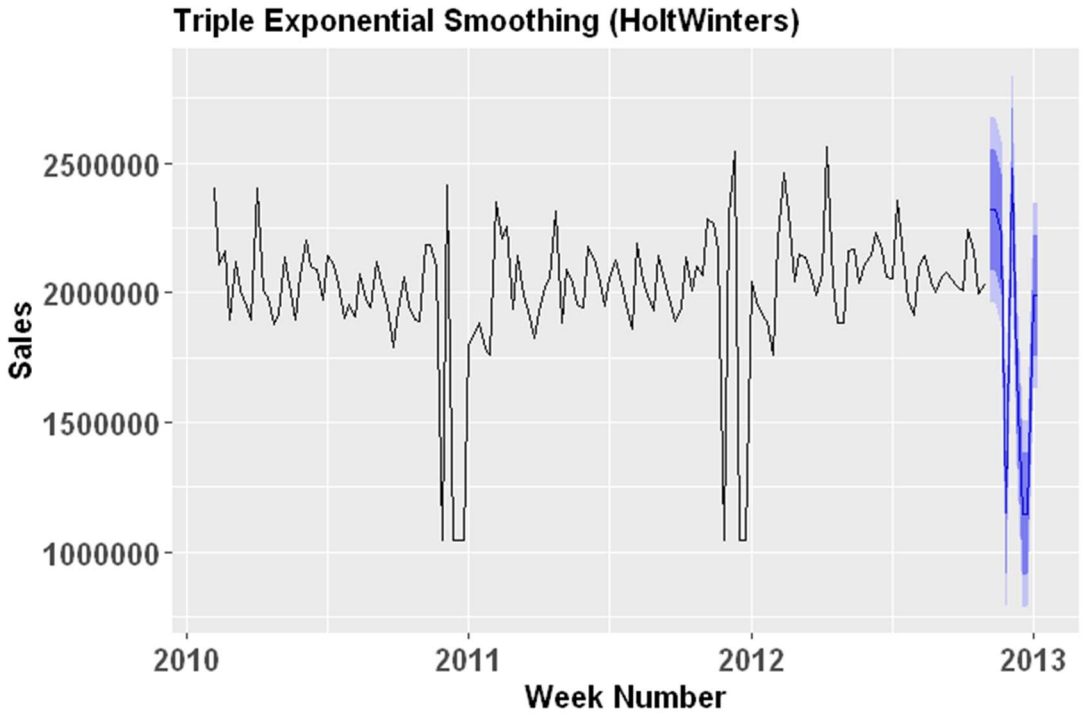
Model Selection:

Model	CV	AIC	BIC	AdjR2
Weekly_Sales ~ Temperature	151531	3823.38	3832.269	0.042855
Weekly_Sales ~ Fuel_Price	154223.3	3828.417	3837.306	0.008541
Weekly_Sales ~ Temperature + Fuel_Price	148918.5	3820.407	3832.258	0.068971
Weekly_Sales ~ CPI	151434.2	3823.198	3832.086	0.044077
Weekly_Sales ~ Temperature + CPI	146314.8	3815.362	3827.213	0.101242
Weekly_Sales ~ Fuel_Price + CPI	151046.8	3824.465	3836.316	0.042169
Weekly_Sales ~ Temperature + Fuel_Price + CPI	146292.7	3817.319	3832.133	0.095051
Weekly_Sales ~ Unemployment	154686.9	3829.276	3838.164	0.002571
Weekly_Sales ~ Temperature + Unemployment	149948.8	3822.379	3834.23	0.056043
Weekly_Sales ~ Fuel_Price + Unemployment	154100.8	3830.19	3842.041	0.003045
Weekly_Sales ~ Temperature + Fuel_Price + Unemployment	148633.4	3821.859	3836.673	0.06586
Weekly_Sales ~ CPI + Unemployment	149704.2	3821.912	3833.763	0.059121
Weekly_Sales ~ Temperature + CPI + Unemployment	145287.1	3815.346	3830.16	0.107449
Weekly_Sales ~ Fuel_Price + CPI + Unemployment	148683	3821.954	3836.768	0.065237
Weekly_Sales ~ Temperature + Fuel_Price + CPI + Unemployment	145054.6	3816.888	3834.665	0.103856

Simple Exponential Smoothing:



Triple Exponential Smoothing:

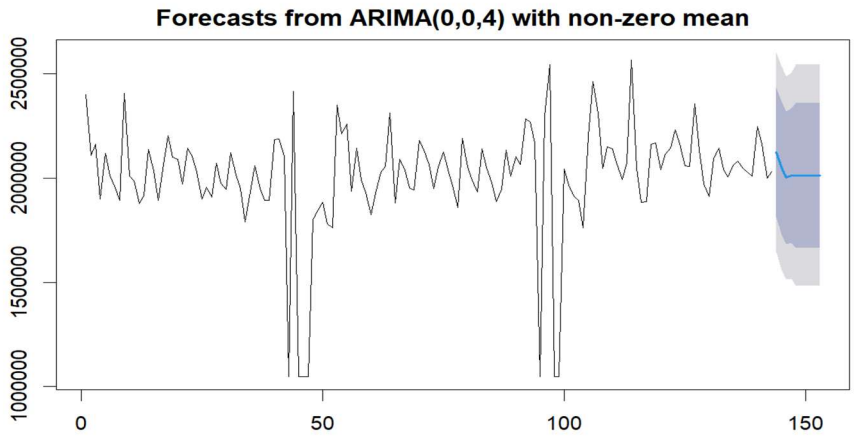


Predicted results are with 80% and 95% significance level.

```
Error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set 18904.43 182613.3 74022.95 0.548128 3.492867 0.6788869 -0.1072611

Forecasts:
         Point Forecast     Lo 80   Hi 80     Lo 95   Hi 95
2012.846        2323029 2088968.9 2557090 1965064.7 2680994
2012.865        2320785 2086683.4 2554886 1962757.6 2678812
2012.885        2231950 1997807.8 2466093 1873860.2 2590041
2012.904        1151621  917436.7 1385805  793467.2 1509774
2012.923        2480678 2246452.2 2714904 2122460.6 2838895
2012.942        1691426 1457158.5 1925694 1333144.8 2049707
2012.962        1146224  911913.9 1380533  787877.9 1504569
2012.981        1147713  913360.8 1382065  789302.4 1506123
2013.000        1989457 1755062.5 2223851 1630981.5 2347932
2013.019        1989199 1754761.3 2223636 1630657.8 2347739
```

ARIMA Model




```
                    ME      RMSE       MAE        MPE     MAPE      MASE       ACF1
Training set -1990.148 252416.24 157018.48 -2.598475 9.581823 0.8019994 -0.0282338
Test set     68386.692  92022.66  68585.91  3.212659 3.222623 0.3503146         NA
```

Accuraccy: RMSE is 92022.66, which means that on average, the forecasted values are off by around $92,022.66 from the actual values. The Mean Absolute Percentage Error(MAPE) is 3.22, which means that on average, the forecasted values are off by around 3.22% from the actual values. The Mean Absolute Scaled Error(MASE) is 0.35, which means that the forecast accuracy is 35% better than a naive forecast that simply uses the previous value as the forecast.

5. Conclusion

In conclusion, this project has provided a comprehensive overview of time series analytics, covering the fundamental concepts, various forecasting methods, applications across different industries, and a case study on the Walmart dataset. Through this analysis, we have demonstrated the practical implementation of time series analytics and its potential to drive growth and enhance overall performance.

Among the various forecasting models applied to the Walmart dataset, the ARIMA model emerged as the best performer. The model's accuracy was assessed using the following metrics:

- RMSE (Root Mean Square Error): 92,022.66, indicating that on average, the forecasted values are off by around $92,022.66 from the actual values.
- MAPE (Mean Absolute Percentage Error): 3.22, indicating that on average, the forecasted values are off by around 3.22% from the actual values.
- MASE (Mean Absolute Scaled Error): 0.35, indicating that the forecast accuracy is 35% better than a naive forecast that simply uses the previous value as the forecast.

These results show that the ARIMA model provides a reasonably accurate forecast, making it a valuable tool for predicting sales and informing strategic decision-making at Walmart.

In summary, time series analytics is a powerful tool for understanding data patterns and trends, empowering businesses to make data-driven decisions and improve operational efficiency. By leveraging historical data and selecting the most appropriate forecasting model, organizations can optimize resources, drive growth, and adapt to the rapidly changing market landscape.

6. References

https://otexts.com/fpp2/intro.html

https://medium.com/analytics-vidhya/time-series-forecasting-a-complete-guide-d963142da33f

https://www.google.com/books/edition/Time_Series_Forecasting/PFHMBQAAQBAJ?hl=en&gbpv=1&dq=time+series+forecasting&pg=PR7&printsec=frontcover

https://www.timescale.com/blog/what-is-time-series-forecasting/