Forecasting Total Sales for Balkis: A Juice Manufacturing Company in Lebanon

MSBA317 Forecasting Analytics

Final Project

By:

Hadi Knaiber

Lara Baltaji

Nour Azakir

Shadi Youssef

Batoul Ramadan

American University of Beirut

April 202

Professor: Rabih Badran, Ph.D.

Contents

Abstract

This study aims to develop a statistical model for forecasting the monthly sales of fresh juice produced by Balkis Company. Different models were applied to analyze and forecast monthly sales data from January 2000 to October 2019 for a Juice Manufacturing Company in Lebanon. The data includes factors such as season of the month and branch location. The model's performance is evaluated based on its ability to accurately predict sales for up to 24 months ahead, from 2019 to 2021. The results demonstrate that the MAA ETS model produced the lowest RMSE, indicating excellent forecasting accuracy. The model's parsimonious use of parameters further enhances its suitability for forecasting fresh juice sales.

1 Introduction

Forecasting sales is a crucial function for businesses as it enables them to anticipate future demand and make informed decisions about production, inventory, and resource allocation. Companies can forecast future sales by examining historical data and identifying patterns and trends.

In this project, we will use forecasting methods to analyze actual sales data from Balkis, a fresh juice manufacturer in Lebanon. Our objective is to determine the best forecasting model to predict future sales that can help the company plan its inventory and make informed decisions regarding pricing, promotions, and marketing strategies.

We will begin by exploring and visualizing the data to identify any trends, seasonality, or other patterns. Then, we will use different statistical methods such as Time Series Decomposition, Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA), Exponential Smoothing, and other methods to accurately forecast future sales. Our analysis will provide Balkis with valuable insights that can help them optimize their business operations, enhance their forecasting accuracy, and plan for future growth.

2 Literature Review

In this literature review, we will explore some of the key studies that have been conducted on forecasting monthly sales of food and beverage products. One of the most commonly used methods of forecasting sales is time series analysis. This forecasting tool involves using statistical tools in order to interpret and analyze data that varies over time. It entails visualizing and modeling data which is collected at regular intervals of time. This method is strictly used to forecast future trends or patterns.

For example, in a study done by Suwanvijit, Lumley, Choonpradub, and McNeil (2011), the authors explored two popular time series forecasting methods, the Lee-Carter method and the Holt-Winters method, for long-term forecasting of sparkling beverages sales. Their data comprised monthly sales from January 2000 until December 2004 collected from a company which sells sparkling beverages in 14 provinces of Southern Thailand. They tried the classical nonlinear Lee-Carter forecasting approach. This method is widely used to forecast age-specific

mortality rates based on lengthy time series of historical data. It can also be applied for seasonal and nonlinear data. In addition, they applied an exponential smoothing Holt-Winters with additive seasonality method to the log-transformed data containing season of the month and branch location as factors. Exponential smoothing is a technique used to update a forecast based on new data. It involves assigning progressively lower weights to older observations and relatively higher weights to more recent ones. This means that the most recent observations have a greater impact on the forecast than older ones. Both models were able to accurately predict sales for up to 24 future months.

In a more recent study (Ensafi Y., Amin S., Zhang G., & Shah B., 2022), the authors investigate a public monthly dataset including the sales history of a retail store from 2014 to the end of 2017 in order to forecast future sales of furniture. The study employs several forecasting models, including classical time-series techniques like Seasonal Autoregressive Integrated Moving Average (SARIMA) and Triple Exponential Smoothing, as well as more advanced deep learning techniques like Facebook Prophet (FBProphet), Lon-Short Term Memory (LSTM) and Convolutional Neural Network (CNN). The performances of the different models are compared using different accuracy measurements, including Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The results indicate that the Stacked LSTM method is the most effective model, with Prophet and CNN models coming next in terms of performance.

In the study of predicting how much champagne would be sold each month (Saxena A., Nanda S., 2020), researchers used time series analysis to forecast sales for ten years. The study was carried out with the help of time-series models, ARIMA and SARIMAX, which are reliable and accurate yet assume that the data will be constant over time, which may not always be the case. Despite certain limits, the authors believe that this type of study could be useful in other areas as well, such as predicting sales of other products, forecasting weather, or predicting stock market fluctuations. It could assist firms in determining if they will generate a profit in the future.

In a four article, "Forecasting Monthly Sales of White Goods Using Hybrid Arimax and Ann Models" (Yücesan M., 2018), the author proposed hybrid ARIMAX-ANN and SARIMAX-ANN sales forecasting models for the white goods industry, which includes washing machines, dishwashers, refrigerators, and small home appliances. The research study employs 46-month

sales data from a white goods wholesaler as well as explanatory variables such as currency rate, holidays, consumer confidence index (CCI), producer price index (PPI), and residential sales in the region. The author aimed to increase the accuracy of the models by applying hybrid models to forecast sales data in the white goods industry. As this study was intended to be the first guide to further studies in sales forecasting for the white goods sector, it demonstrates that hybrid methods outperform single methods. It does, however, have certain drawbacks, such as a lack of regional data, which resulted in insufficient explanatory variables, and more observations are required to create a more accurate prediction.

3 Problem Description

Balkis Company, a fresh juice manufacturer in Lebanon, is currently facing trouble growing their sales, especially in light of the economic crisis and recession afflicting Lebanon. To address this issue, this study aims to develop an accurate sales forecasting model. By accurately predicting sales, the firm can identify trends and seasonal fluctuations, allowing them to optimize production and inventory management accordingly. They can also enhance marketing strategies, and make informed business decisions to promote development and profitability. This research hopes to assist Balkis in navigating the difficult economic climate and achieving sustained success.

4 Exploratory Data Analysis

This study examines a dataset containing monthly sales of fresh juice from January 2000 to October 2019. The data was obtained from Balkis with whom we established contact, in order to be able to access authentic data for this project. The data originally incorporated detailed demand information for every food product used in production, in addition to the total sales earned over the year.

In this research, we decided to focus on sales data as it is deemed more suitable for forecasting purposes than demand data. Using R, the data will be visualized to identify any seasonal, cyclic, or trend-related patterns, as well as to uncover potential insights. Then, all the models will be fitted and evaluated using different metrics such as the Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Finally, an interactive web application will be built using the Shiny package in R to automate the best-performing model.

We used a variety of methodologies to undertake exploratory data analysis on our time series dataset. As shown in Fig 1, the auto plot revealed a clear seasonality with an overall upward trend throughout the years. However, there was a noticeable spike in the year 2004, which could potentially be a data entry error.

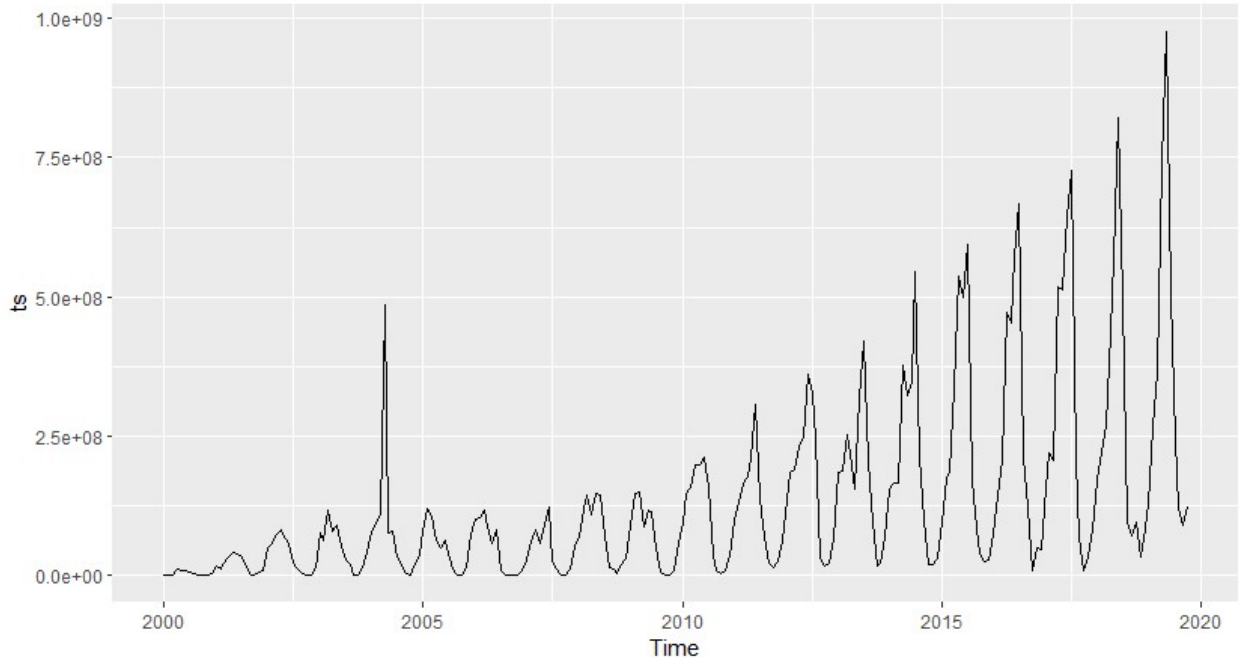

Fig1. Total Balkis sales over the years

To investigate the seasonality in more detail, we created a seasonal plot (Fig 2) which demonstrates that the values in the last three months of the year were similar, while the first three months of the year showed increasing seasonality over time. The summer months exhibits an upward trend, but the seasonality was less clear. This trend was confirmed by the subseries plot (Fig 3), which shows an increasing trend in all months, particularly in the summer months.

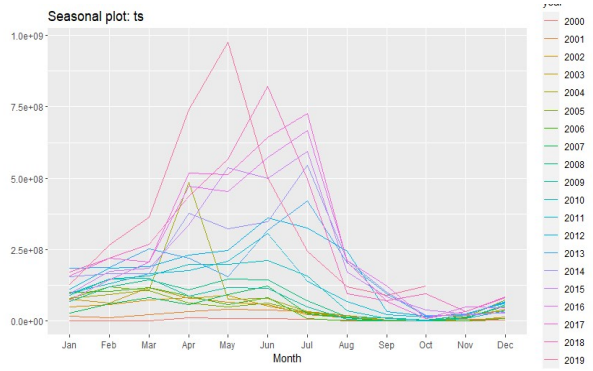

Fig2. Seasonal plot.

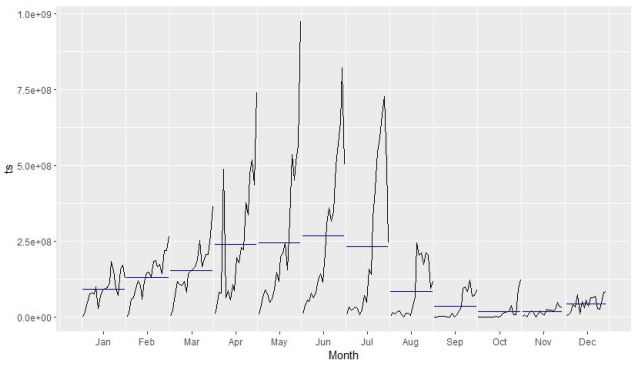

Fig3. Subseries plot

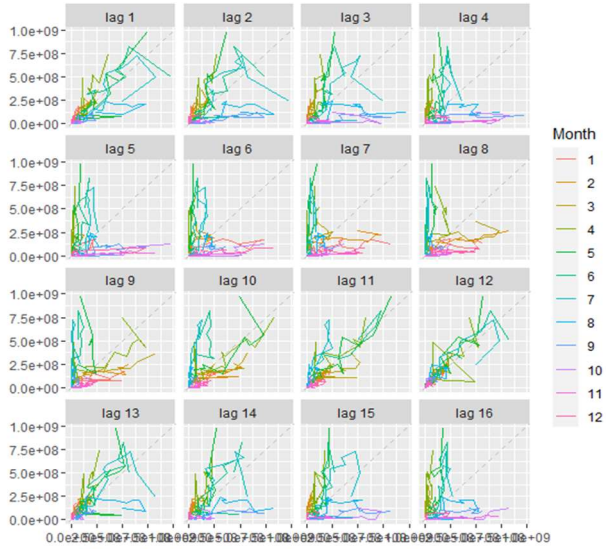

Fig4. Lag Plot

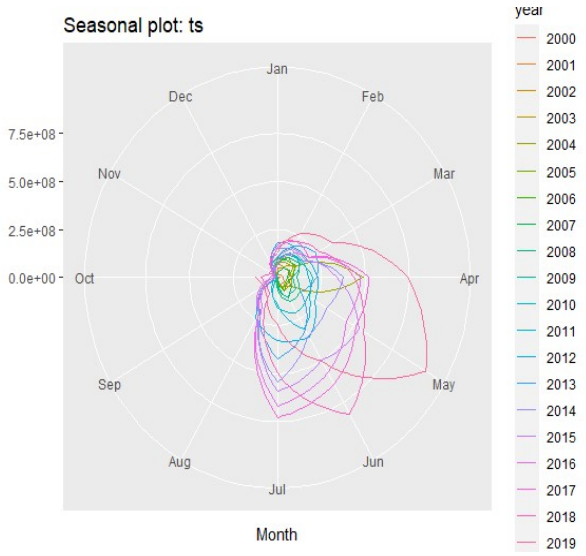

Fig5. Polar seasonal plot

The lag plot shows a strong positive correlation in lags 12, reflecting the strong seasonality in the data. That is, because in these two lags, the 12 quarters almost lie on the diagonal dotted line. This suggests that the data is influenced by the past values in a yearly manner (every 12 months). Finally, the polar seasonal plot shows that there was no significant change for September, October, November, and December but a clear increasing tendency for the other months, helped to better comprehend how the data varied during the different months.

5 Methodology and Results

In order to prepare our dataset for forecasting, we imported the dataset into a time series with frequency (12) to capture monthly patterns. However, we ran into a problem with the dataset's zero values, which can lead to mistakes when using specific forecasting methods. To solve this problem, we changed each value in the dataset by a negligible constant value of 0.001. This prevents errors brought on by zero values and guarantees that the input data is appropriate for the forecasting algorithms.

We divided the dataset into training and testing sets in order to model and test the data. The testing set is made up of monthly data from 2018 through October 2019, whereas the training set is made up of monthly data from September 2004 through December 2017. We have dropped the values before September 2004 as they were considered old and insignificant and

contained an outlier. This has allowed us to make sure that we are using more recent and pertinent data, which can improve the accuracy of our forecasts.

We started by using the simple naive methods as a benchmark which are naive, mean naive, seasonal naive, and drift naive forecasting techniques. By doing cross-validation on the dataset and computing the Root Mean Squared Error (RMSE), we then assessed their performance.

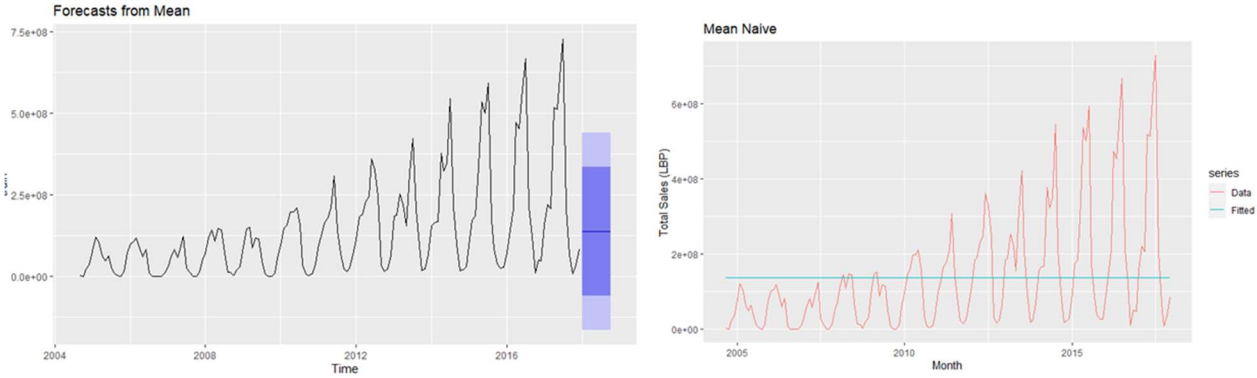

Fig 6. Mean Naive method

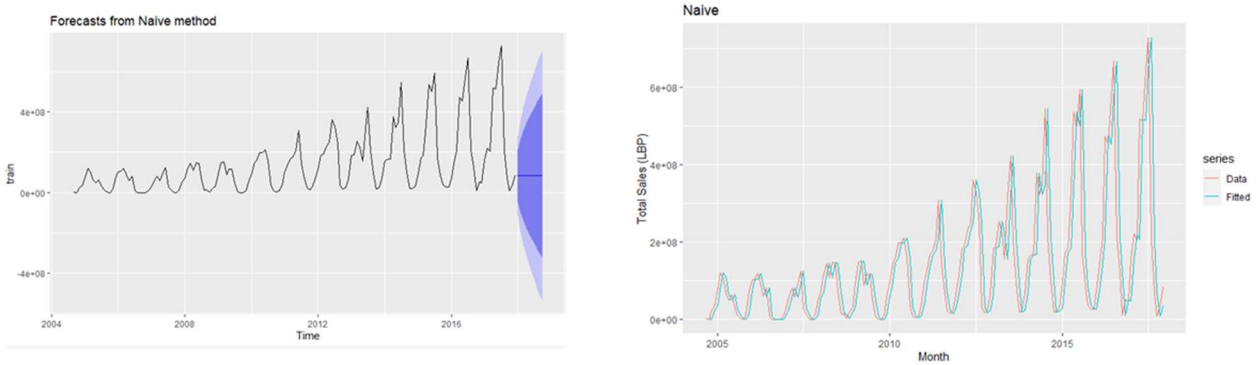

Fig 7. Naive method

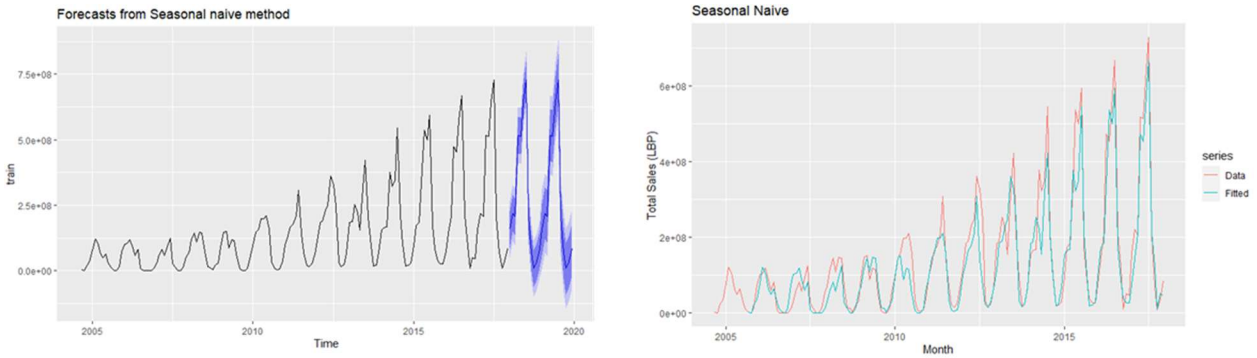

Fig 8. Seasonal naive method

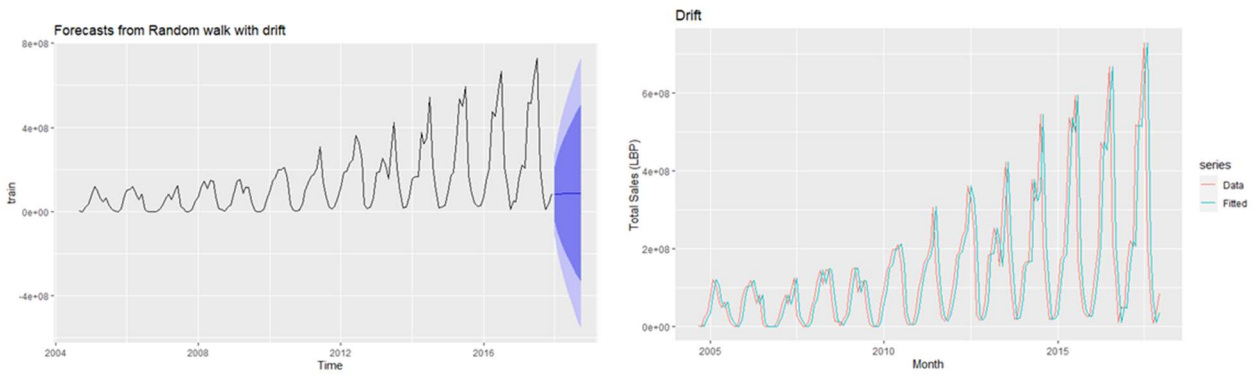

Fig 9. Drift method

Naive Forecasting Methods	RMSE	MAPE
Mean naive fit	300911410	5.861795e+01
Seasonal naive fit	178645502	3.916908e+01
Naive fit	337403176	5.631713e+01
Drift fit	335583421	5.566400e+01

According to the findings, the seasonal naïve technique achieved the lowest RMSE (178645502) and MAPE (3.916908e+01) which suggests that it provides the most accurate predictions among the tested forecasting methods. This means that the seasonal naïve technique is a suitable method to use as a baseline forecasting method, indicating that the seasonality in our data is strong.

ARIMA

Next, we applied time series decomposition to extract the trend components, seasonal component and remainder component. By decomposing a time series into its components, we can better understand the underlying patterns and behaviors in the data. This can help us improve our forecasting accuracy by allowing us to model each component separately and then combine them to produce a more accurate forecast. There are several methods for time series decomposition, including the classical decomposition method and the seasonal decomposition of time series by loess (STL) method.

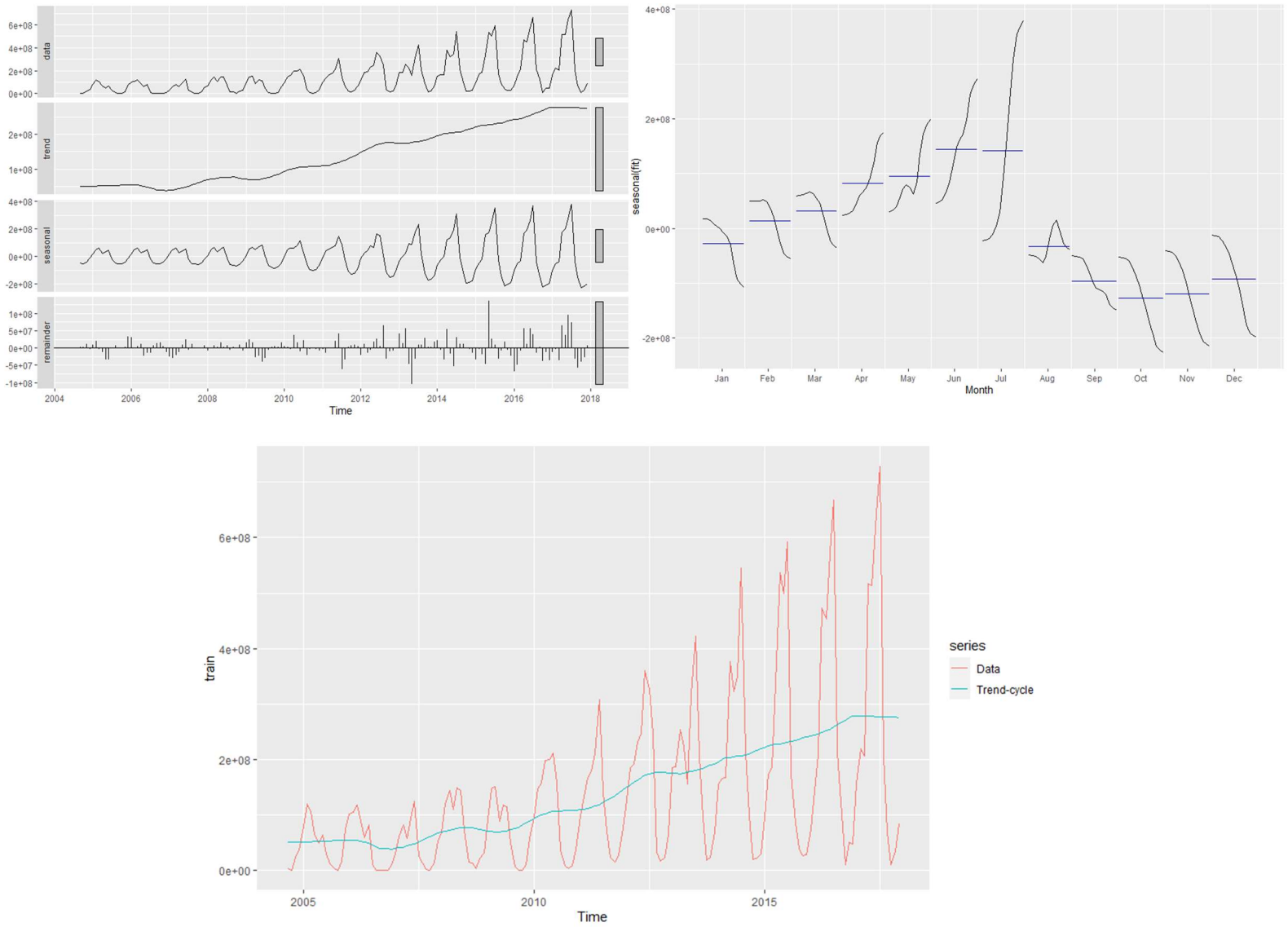

Fig 10. Time Series Decomposition

Moreover, to check whether the data is stationary or not, we used the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. Stationarity is an important property of time series data, as it enables the use of various statistical models and techniques for forecasting. A stationary time series is one whose statistical properties remain constant over time, such as its mean, variance, and autocorrelation. In contrast, a non-stationary time series exhibits trends, seasonal patterns, or other forms of non-random behavior that change over time. Based on the test results, we achieved a test statistic of 1.3979 which is higher than the critical value, so we reject the null hypothesis of stationarity, which suggests that the data is not stationary and requires differencing for ARIMA-related models.

Next, we applied non-seasonal ARIMA models. ARIMA modeling is a popular method for time series forecasting that combines Auto regression (AR), Moving Average (MA), and differencing techniques to capture the autocorrelation and seasonality in the data. In this section, we applied ARIMA modeling to our dataset to obtain accurate forecasts. The first step in ARIMA modeling is to identify the order of differencing (d), the order of the autoregressive term (p), and the order of the moving average term (q) through analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. We used ggtsdisplay() function to visualize the ACF and PACF plots and identified the values of the initial autocorrelation factor for our ARIMA model.

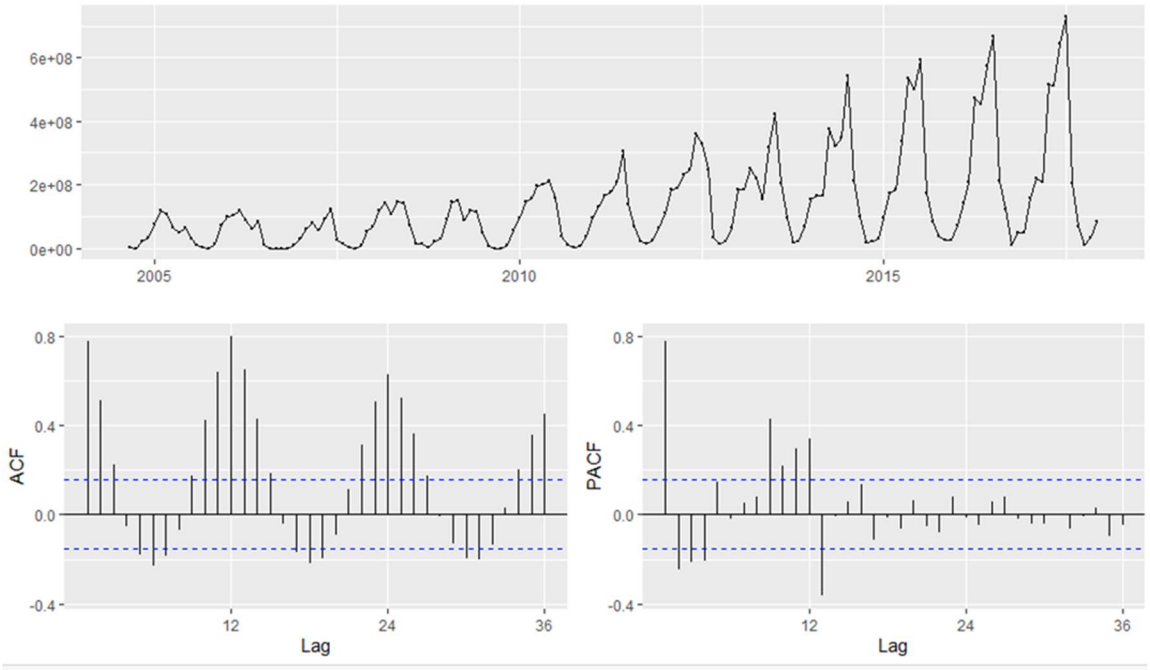

Figure 11

We noticed that the ACF is sinusoidal and exponentially decaying and the PACF has a significant spike at lag 1 but none beyond it which suggests that AR(1) with differencing of order 1(d=1) can be an acceptable starting point. Thus our first ARIMA model will be: ARIMA(1, 1, 0). Then, we iteratively changed the differencing factor in order to improve the performance. The results are as follows:

ARIMA model	AICc	RMSE	MAPE	Residuals are autocorrelated using Ljung-Box test?
ARIMA(1, 1, 0)	6314	345751508	57.30205 %	No
ARIMA(1, 2, 0)	6331	449711690	310.7801 %	No

Based on the AICc, RMSE, and MAPE values, the ARIMA(1,1,0) model appears to be the best choice among the three models tested. This model has the lowest RMSE and MAPE whereas it has the same AICc. Additionally, the Ljung-Box test indicates that the residuals are white noise for this model, which suggests that the model adequately captures the information in the data. Since we have the same AICc for two models therefore we focus on the lowest RMSE and MAPE indicating that the ARIMA(1,1,0) model should be selected for forecasting.

The following figures show the time series plot, ACF plot and distribution plot of the residuals from the chosen model ARIMA (1, 1, 0).

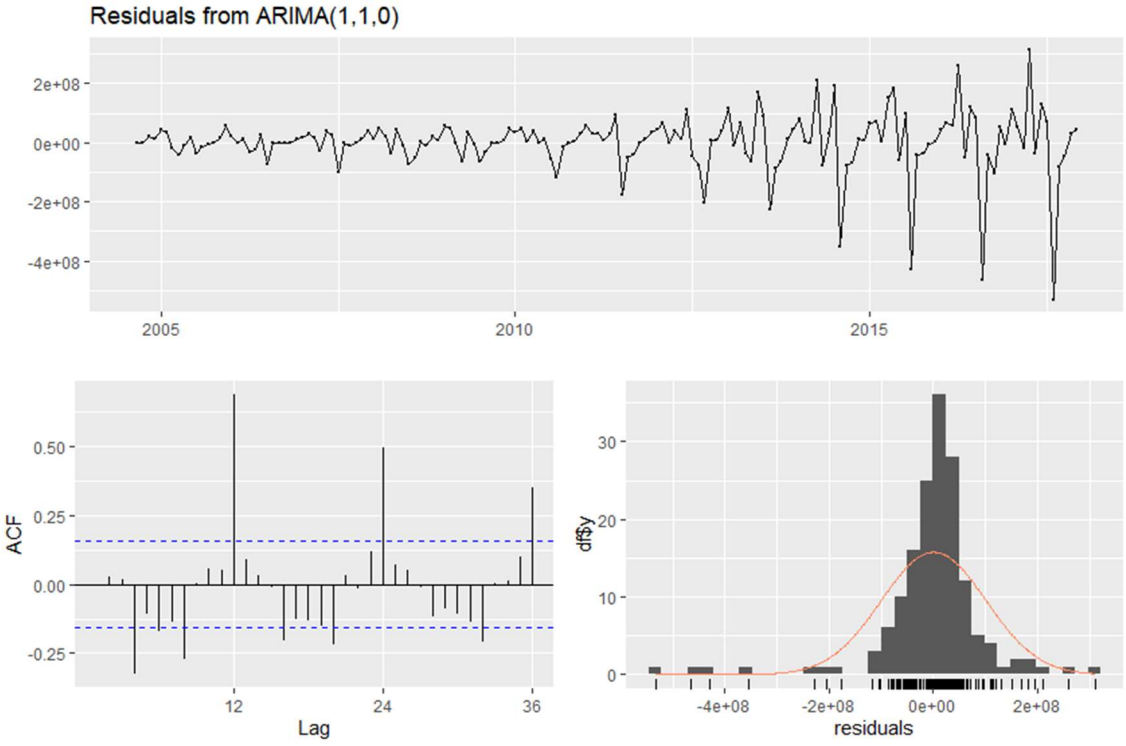

Figure 12

While this model did achieve a lower AICc, we can notice that the time plot of the residuals does not seem to be stationary. The ACF shows that there still exists autocorrelations between the residuals and the distribution is left skewed (not normally distributed). This indicates that the residuals are not white noise.

The following figure shows the fitted and forecasted plot of our chosen non-seasonal ARIMA(1, 1, 0). This forecasting method seems to be naïve for our data.

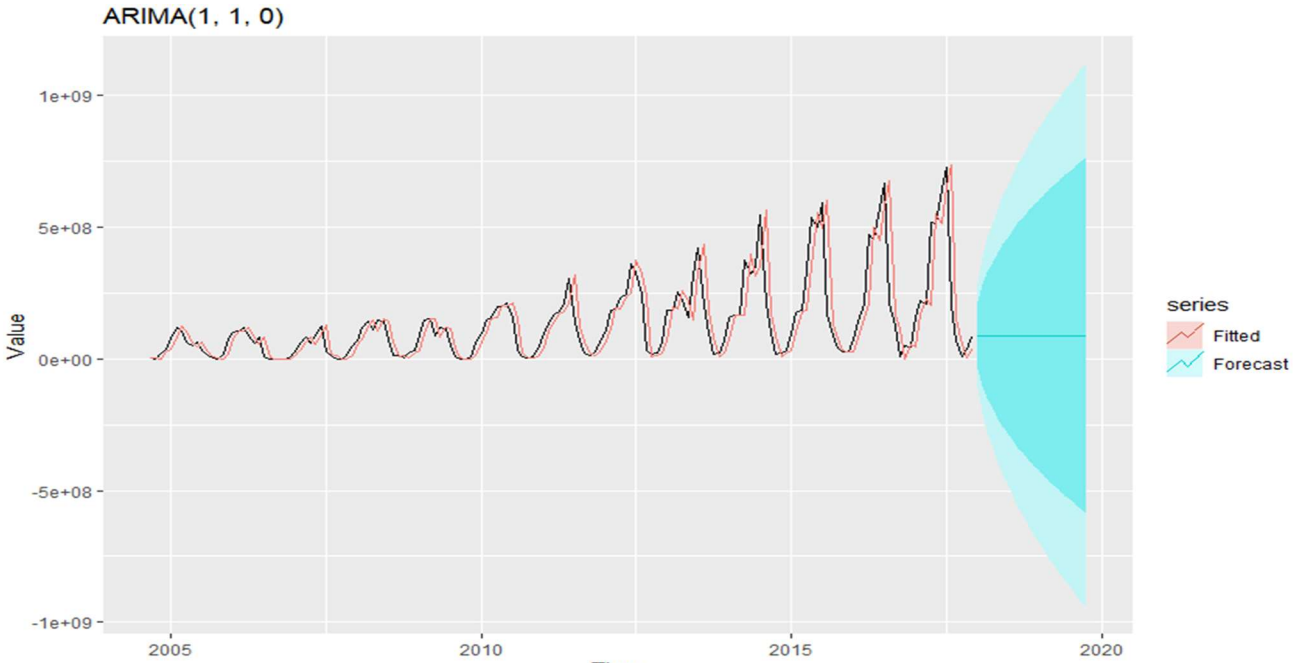

Figure 13

Seasonal ARIMA

Next, we applied Seasonal ARIMA, which takes into account both autoregressive (AR) and moving average (MA) components also, as well as seasonality. The SARIMA model is expected to work better on our data, given its strong seasonal component. To apply the SARIMA model, we first identified the appropriate values for the seasonal components, which are denoted as (P, D, Q) in addition to the non-seasonal components (p, d, q). The seasonal components represent the autoregressive, differencing, and moving average components at the seasonal frequency, while the non-seasonal components represent the same for the non-seasonal frequency.

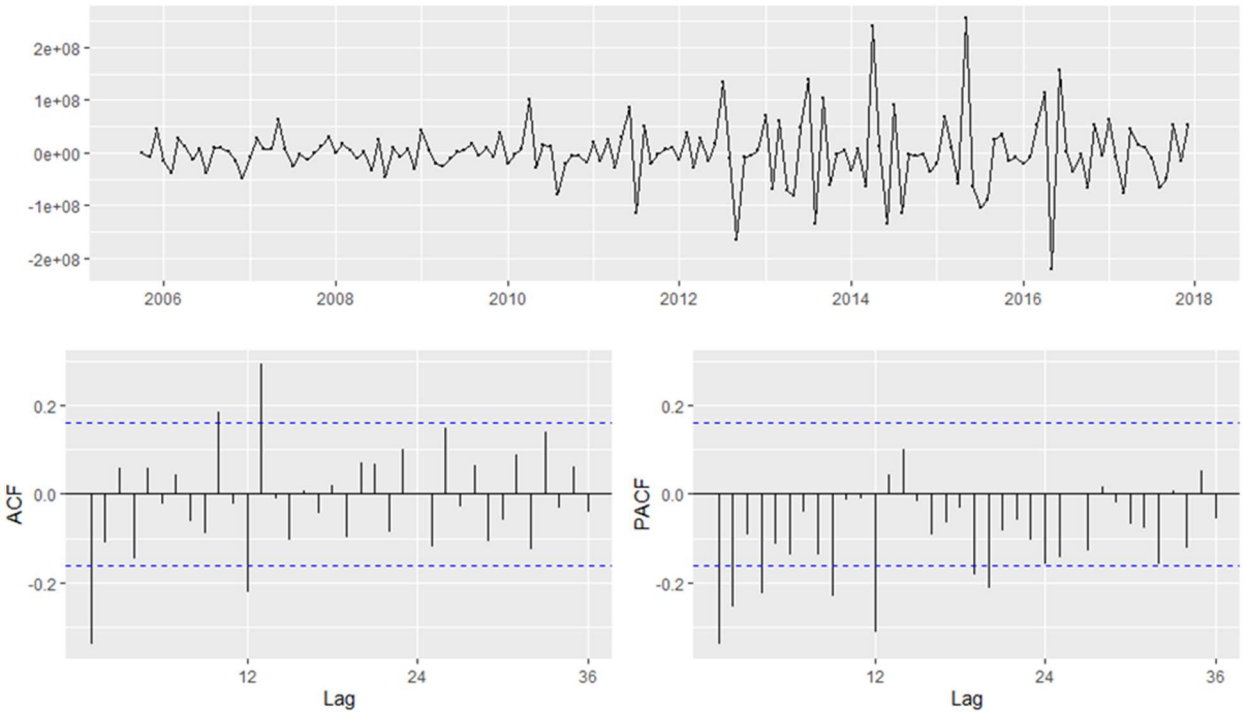

Figure 14

In the above plots, we initially identified a significant spike at lag 1 in the ACF and an exponentially decaying PACF for every lag, suggesting a non-seasonal MA(1). In addition, we identified significant spike at lag 12 in the ACF and an exponentially decaying PACF for every season (lag 12, lag 24, lag 36), suggesting a seasonal MA(1). We then proceeded to fit a SARIMA (0, 1, 1)(0, 1, 1)[12] model to the time series. The SARIMA model has a non-seasonal autoregressive order of 0, a non-seasonal differencing order of 1, and a non-seasonal moving average order of 1. Additionally, it has a seasonal autoregressive order of 0, a seasonal differencing order of 1, and a seasonal moving average order of 1, with a seasonal period of 12. We will use this model as the initial model and make iterative changes to improve the model performance.

SARIMA model	AICc	RMSE	MAPE	Residuals are autocorrelated using Ljung-Box test?
SARIMA(0, 1, 1)(0, 1, 1)[12]	5646.44	176320711	48.98173 %	No
SARIMA(0, 1, 2)(0, 1, 1)[12]	5632.6	175998085	49.31869 %	No
SARIMA(0, 1, 2)(0, 1, 2)[12]	5633.89	177449473	49.50472 %	No
SARIMA(0, 2, 2)(0, 1, 1)[12]	5626.9	177982231	48.25881 %	No

The Seasonal ARIMA(0, 2, 2)(0, 1, 1)[12] model has the lowest AICc. Additionally, its residuals pass the Ljung-Box test for white noise, indicating that the model captures the remaining patterns in the data. Therefore, based on these metrics, SARIMA(0, 2, 2)(0, 1, 1)[12] is the best model among the ones tested.

The following figures show the time series plot, ACF plot and distribution plot of the residuals from the chosen seasonal SARIMA(0, 2, 2)(0, 1, 1)[12] model.

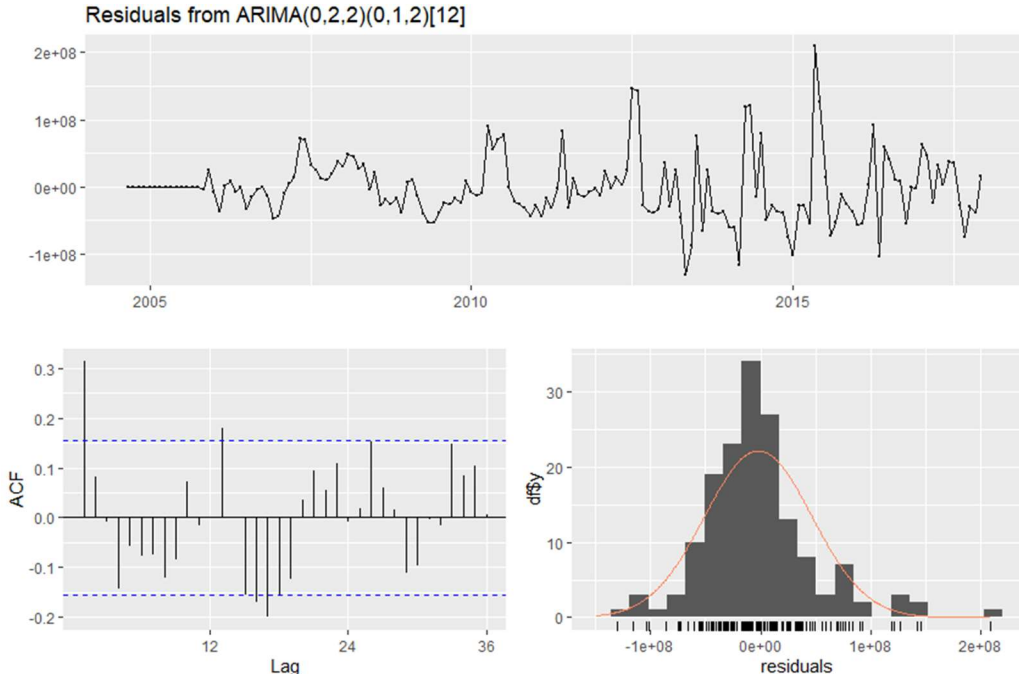

Figure 15

As seen in the figure, the time series plot of residuals does not exhibit clear patterns and the ACF does not show a clear autocorrlation between residuals at different lags. Additionally, the distribution is almost normally distributed.

The following figure shows the fitted and forecasted plot of our chosen seasonal SARIMA(0, 2, 2)(0, 1, 1)[12].

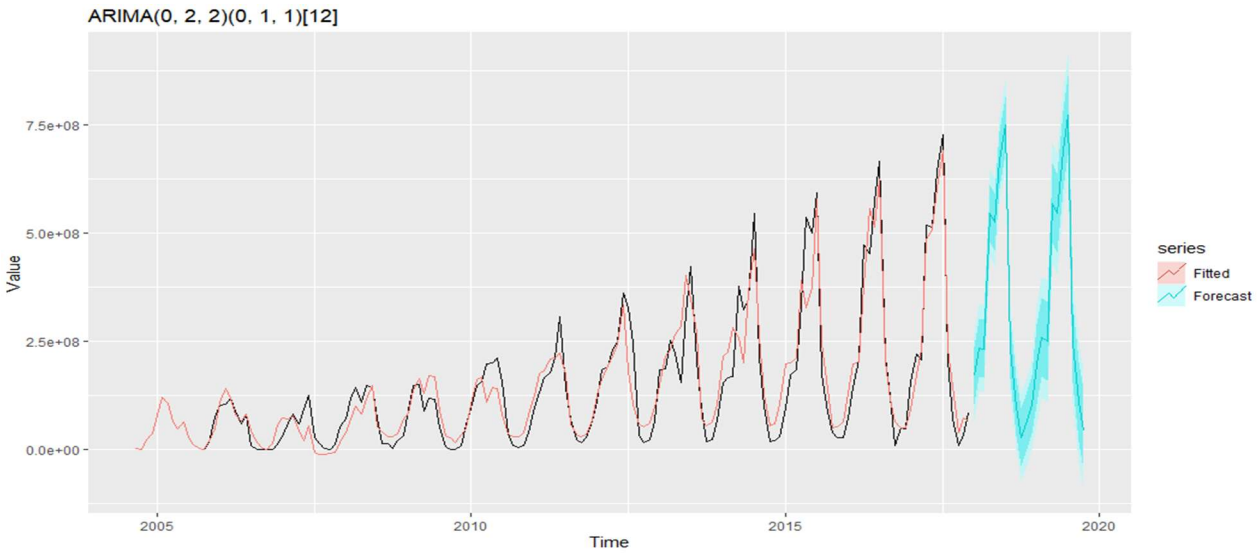

Figure 16

As seen in the above figure, this fitted model was able to capture the trend and the seasonality components of the trained data. Additionally, the forecast seems to be visually convincing.

Exponential Smoothing

Finally, we applied exponential smoothing models. Exponential smoothing is a popular and widely-used method for forecasting time series data. Unlike the ARIMA and SARIMA models, which rely on identifying patterns in autocorrelation and seasonality, exponential smoothing techniques apply a simple yet powerful principle: recent observations are more informative than older observations. In exponential smoothing, we compute a smoothed estimate of the series based on a weighted average of past observations. The weights decay exponentially as we move further back in time, giving more weight to recent observations and less weight to older ones. We will try to apply the exponential smoothing by using the most important exponential smoothing methods and by trying different variations in the combinations of the trend and seasonal components.

Using some of the most important exponential smoothing methods which are:

1 Simple Exponential Smoothing

2 Holt's Linear Exponential Smoothing

3 Holt-Winters Exponential Smoothing

We fitted each of these models on our time series data and compared their accuracy using various metrics such as AICc, RMSE, and MAPE.

Model	RMSE	MAPE	Residuals are autocorrelated using Ljung-Box test?
Simple Exponential Smoothing SES	184619733	41.17558 %	No
Holt's Exponential Smoothing	310918987	71.29565 %	No
Holt's Damped Exponential Smoothing	337435527	56.32823 %	No
Holt-Winters Additive Method	174421358	47.09397 %	No
Damped Holt-Winters Additive Method	174097546	43.72916 %	No
Holt-Winters Multiplicative Method	189919658	47.90618 %	No

It is important to note that the above models cannot generate AICc values due to their simplicity since they do not involve estimating a large number of parameters. Therefore, an alternative will be either the RMSE or the MAPE to compare between the models. However, the

MAPE can lead to division by the values that are zero or close to zero. Our dataset contains very small values (almost 0), therefore, the better choice for evaluation is the RMSE.

The Damped Holt-Winters Additive Method appears to be the best fit for the data among the exponential smoothing models based on the RMSE. It has the lowest RMSE value of 174097546 and an MAPE of 43.72916%. Additionally, the Ljung-Box test suggests that the residuals are white noise, indicating a good fit for the data.

The following figures show the time series plot, ACF plot and distribution plot of the residuals from the Damped Holt-Winters Additive Method model.

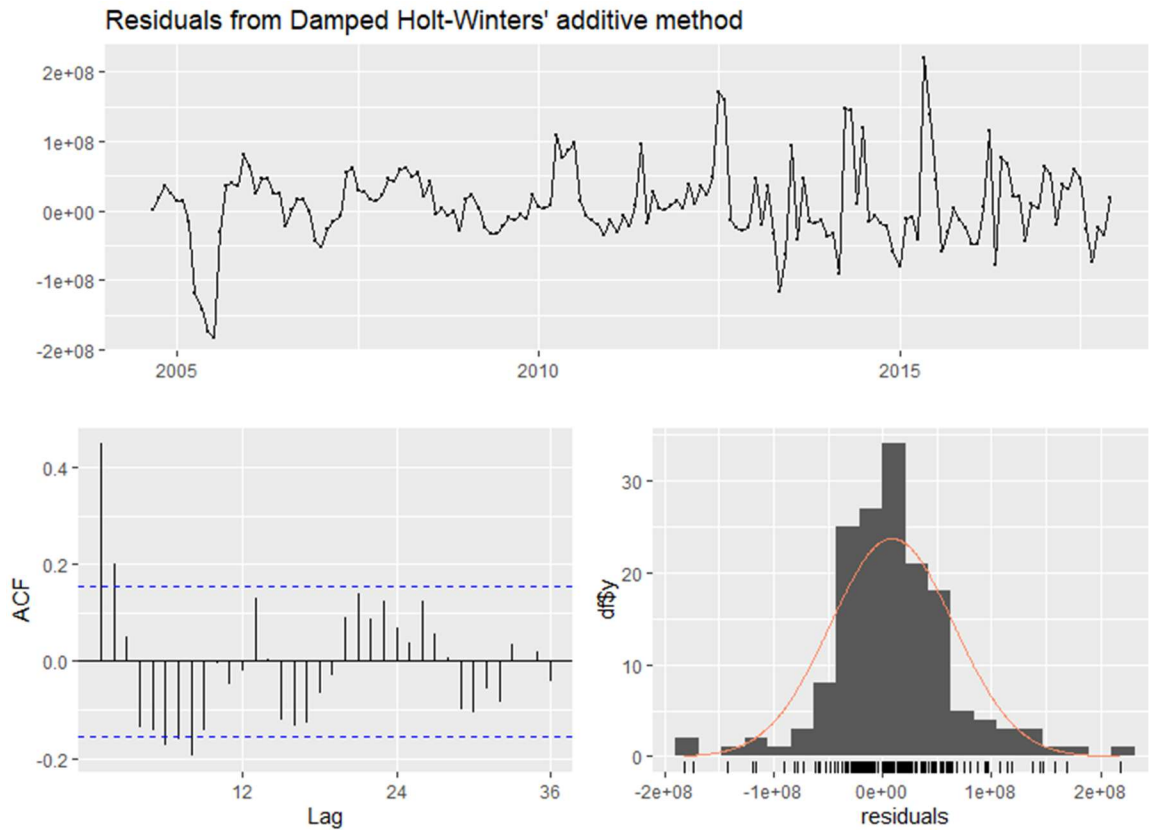

Figure 17

As seen in the figure, the plot's residuals are similar to that of a white noise indicating that they are stationary. This is also confirmed in the Ljung-Box test which resulted in a p-value<0.05. Additionally, the distribution of the residuals seem to follow a normal distribution.

The following figure shows the fitted and forecasted plot of the Damped Holt-Winters Additive Method model.

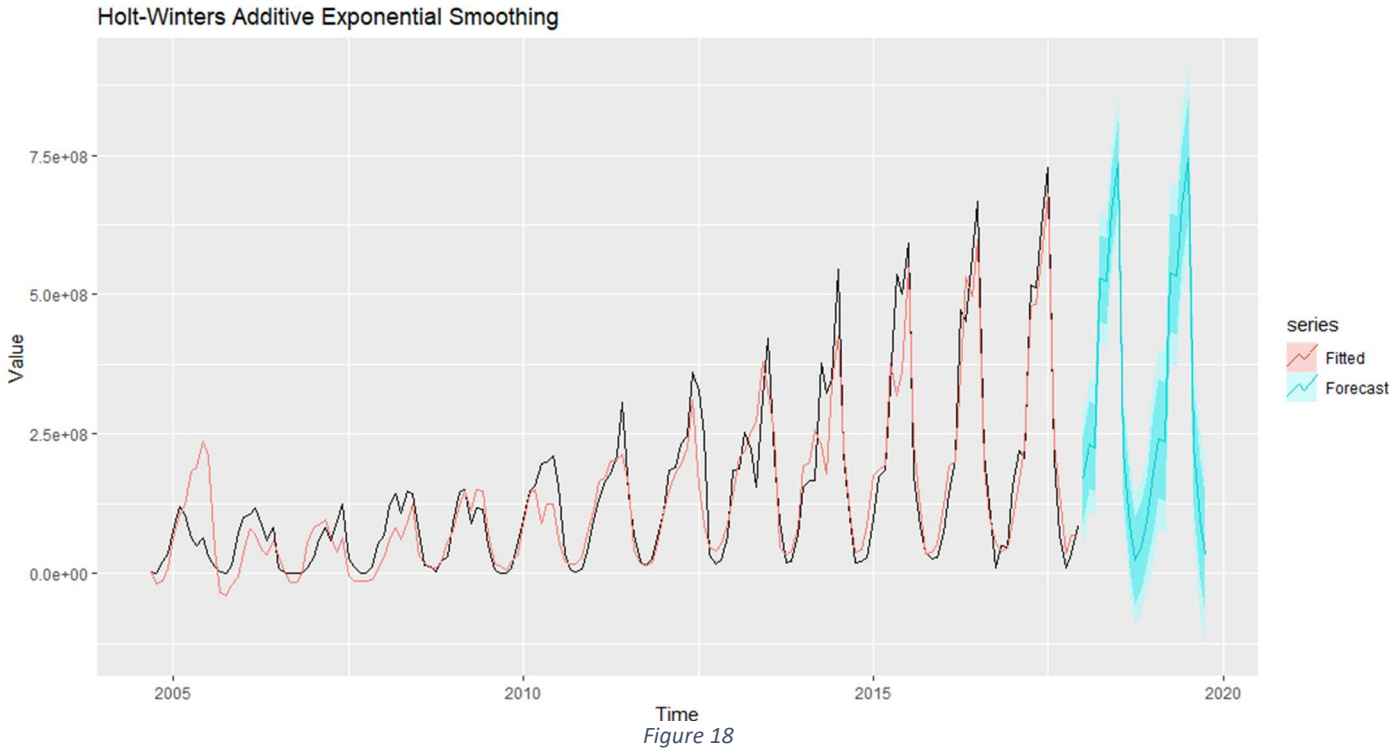

Figure 18

This model also seems to be convincing visually as it captures most of the fitted data and the forecast follows the same trend and seasonality.

1. **Using variations in the combinations of the trend and seasonal components (AAA, MAA, MAM)**

Exponential smoothing models can be further extended by including different combinations of trend and seasonal components. The AAA model includes all three components (level, trend, and seasonality) and is appropriate when all three are present and significant. The MAA model includes a level component and a multiplicative seasonality component and is appropriate when the seasonal fluctuations increase or decrease proportionally with the level. The MAM model includes a level component and an additive seasonality component and is appropriate when the seasonal fluctuations remain constant over time. These variations allow for a more flexible modeling approach that can capture different patterns in the data.

Model	AICc	RMSE	MAPE	Residuals are autocorrelated using Ljung-Box test?
AAA ETS Model	6567.975	174418178	47.08403 %	No
Damped AAA ETS	6568.595	174097159	43.74084 %	No
MAM ETS	6472.369	184619733	41.17558 %	No
Damped MAM ETS	6483.130	182679865	39.466 %	No
MAA ETS	6653.172	170371276	46.03703 %	No
Damped MAA ETS	6564.338	170725278	41.50926 %	No

Based on the metrics provided, the MAM ETS has a lower AICc than the MAA ETS model but it has a higher RMSE. The model with the lower AICc is the better the model at explaining the data whereas the RMSE is a measure of model predictions and since we are concerned with the predictions we will focus on the lowest RMSE. That's why we will take MAA ETS as the best model among the options provided.

The following figures show the time series plot, ACF plot and distribution plot of the residuals from the MAA ETS model.

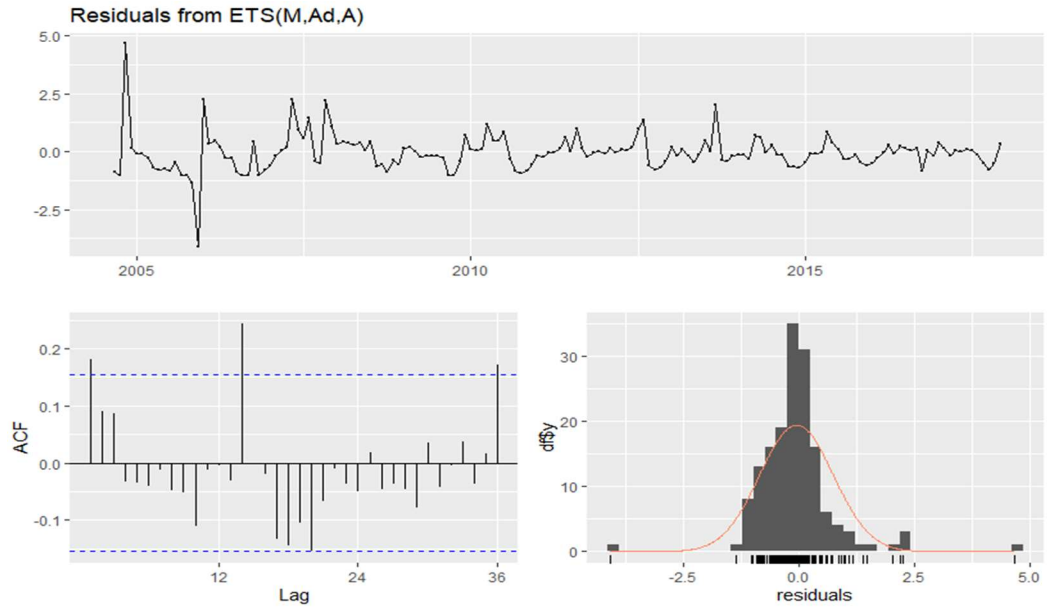

Figure 19

As seen in the figure, the residuals are similar to that of a white noise indicating that they are stationary. The time series plot shows no clear pattern (trend or seasonality) and the ACF shows no clear autocorrelation between residuals. Finally, the residuals follow a normal distribution. This is also confirmed in the Ljung-Box test which resulted in a p-value<0.05.

The following figure shows the fitted and forecasted plot of the MAA ETS model.

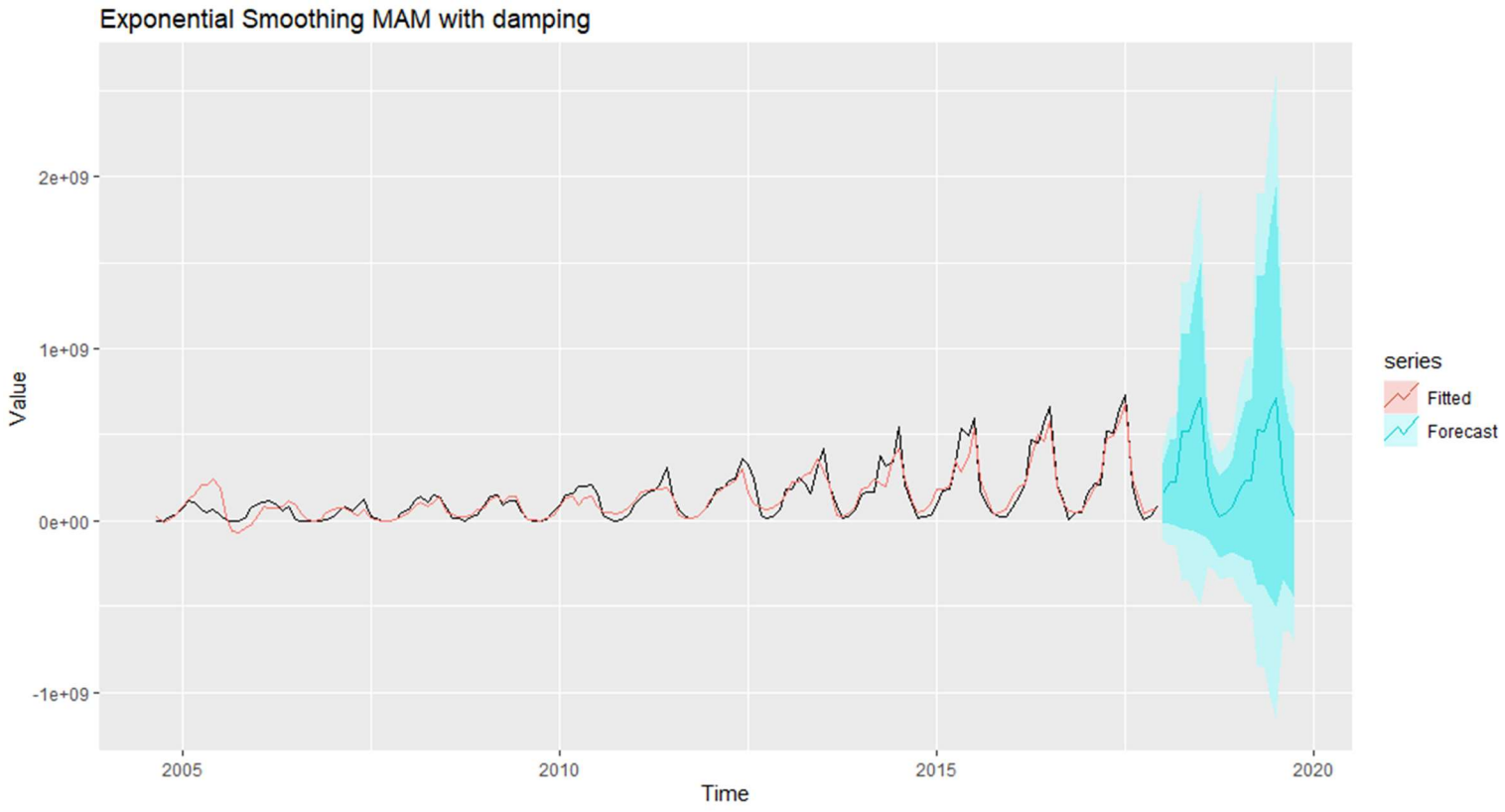

Figure 20

Model	**RMSE**
ARIMA(1, 1, 0)	345751508
SARIMA(0, 2, 2)(0, 1, 1)[12]	177982231
Damped Holt-Winters Additive Method	174097546
MAA ETS	170371276

Based on the above table, the **MAA ETS** model seems to be the best fit as it has the lowest RMSE, and its residuals are white noise as confirmed by the Ljung-Box test.

6 Conclusion and Recommendations

As a conclusion, after fitting and comparing different time series models on the data, we found that the MAA ETS model provided the best fit. This was based on the model's relatively low RMSE value and the confirmation of white noise residuals using the Ljung-Box test. This indicates that the model was able to capture the underlying patterns in the data and make accurate predictions.

However, it's important to note that the performance of the model may be impacted by external factors that are not captured by the data. Therefore, it is recommended to constantly monitor and update the model as new data becomes available. Additionally, it is preferred to incorporate expert judgment and domain knowledge to complement the statistical analysis and enhance the accuracy.

Note: Use the following Shiny website link for an interactive data visualization dashboard.
https://theforecasters776.shinyapps.io/forecastingproject/

7 References

Suwanvijit, W., Lumley, T., Choonpradub, C., & McNeil, N. (2011). Long-Term Sales Forecasting Using Long-Term Sales Forecasting Using. *The Journal of Applied Business Research*.

YÜCESAN , M. (2018). Forecasting Monthly Sales of White Goods Using Hybrid Arimax. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi Aralık* , 2603-2617.

Ensafi, Y., Amin, S., Zhang, & Shah. (2022). Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. *International Journal of Information Management Data Insights*.

Saxena, A., & Nanda, S. (2020). AN INNOVATIVE TIME SERIES BASED METHOD OF FORECASTING MONTHLY SALES OF CHAMPAGNE. *Samvakti Journal of Research in Information Technology*, 77-88.