

TIME SERIES ANALYSIS FOR RETAIL AND FOOD SERVICES AND SALES

**EXECUTIVE SUMMARY**

The analysis for Retail and Food Services Sales project, involves a detailed investigation into the historical data of retail and food services sales over an extended period. The primary objective is to identify recurring patterns and trends in sales behavior within this sector. By applying time series analysis techniques, we aim to reveal insights into the evolution of sales over time, including any seasonal variations or discernible trends.

This analysis serves the purpose of creating effective models for forecasting future sales trends in retail and food services. Through a thorough examination of historical data, we seek to uncover the underlying dynamics that influence sales fluctuations. The ultimate goal is to provide valuable insights that can aid businesses in making informed decisions and strategic plans based on a better understanding of how sales patterns may evolve in the future.

The analysis is done on data of 32 years. Models such as ARIMA, Auto ARIMA, and Regression based models with linear trend and seasonality followed by Two- level forecasting, Holt-Winter’s model has been implemented. From the visualization, we can uncover that the trends and seasonality observed has a subsequently upward and a combination of lower trends here and there. The data is statistically significant as all the 12 lags have positive trend observed.

Model Evaluation is highly based on the values of RMSE and MAPE.

The best and optimized results were obtained using the ***Auto ARIMA*** model. The evaluation is done on the basis of RMSE and MAPE metrics. By understanding historical sales patterns and forecasting future trends, businesses can make data-driven decisions to optimize operations and capitalize on emerging opportunities.

**INTRODUCTION**

The Retail and food services sectors are crucial contributors to the U.S. economy, and accurate forecasting of monthly sales within these industries is essential for business operations. Time series analysis can be used as a powerful tool for comprehending this data. By analyzing sales data over time, we gain insights into trends, patterns, and seasonal fluctuations, enabling us to make informed predictions and optimize future sales strategies. This analytical approach allows businesses to anticipate demand, adjust inventory levels, and design marketing efforts effectively. Ultimately, leveraging time series analysis empowers businesses in the retail and food services sectors to make data-driven decisions, enhancing their competitiveness and overall performance in the market.

The primary objective of this project is to conduct a comprehensive time series analysis of monthly sales data for retail and food services, spanning approximately 32 years of historical records. The aim is to analyze the dataset thoroughly and identify underlying patterns in monthly sales trends, to understand how these patterns evolve over time, uncover repeating seasonal fluctuations, long-term trends, and any irregularities or anomalies that may impact sales performance.

We can achieve our goal by following the eight steps of forecasting, in which we will run the dataset through various models such as Regression-based models, Holt-Winter’s Exponential smoothing model, autoregressive integrated moving average models (ARIMA) visualize and compare the forecasts using parameters such as RMSE, MAPE. Finally, we will conclude which model would be best for forecasting the sales and identifying the trends

In conclusion, this time series analysis project aims to empower retail and food services businesses with the tools needed to navigate the complexities of their market. By harnessing the power of temporal data, organizations can gain a competitive edge and adapt proactively to the ever-changing landscape of consumer behavior and market dynamics.

**STEPS USED IN FORECASTING**

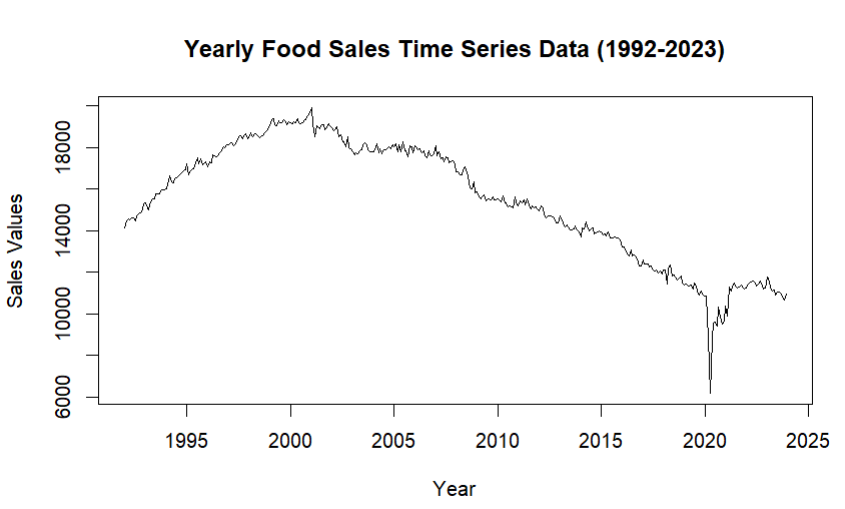
**STEP 1 – Goal Definition**

The primary goal of this time series analysis project is to gain insights into the trends and patterns of sales in the Retail and Food Services industry over the given time period. This analysis aims to provide valuable information for decision-making, forecasting, and strategic planning. The objective is used to identify and analyze any seasonal variations or patterns in sales data to help businesses anticipate and prepare for peak and off-peak periods. The anomalies are detected and any unusual spikes or drops in sales that may indicate exceptional events or factors affecting the Retail and Food Services industry. Understanding these anomalies can be crucial for proactive decision-making. The data is divided on monthly basis and is used for forecasting into the future. We shall summarize key findings and insights from the analysis, offering actionable recommendations for stakeholders to enhance decision-making and strategic planning.

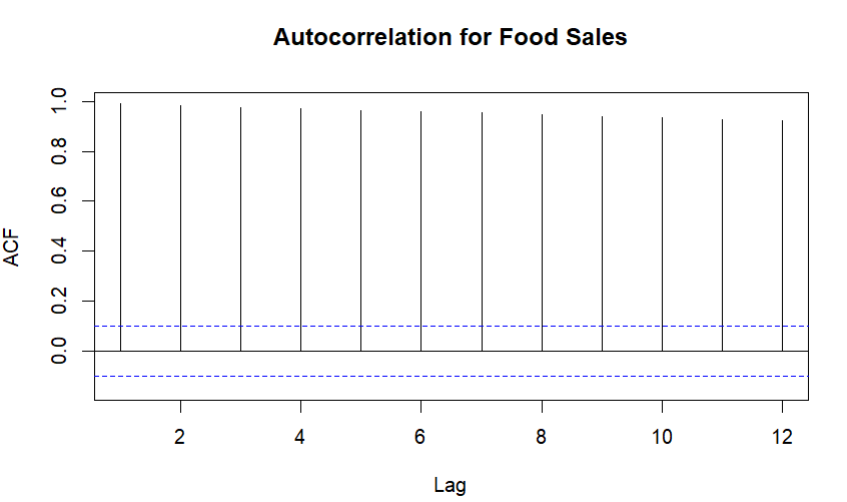
**STEP 2 – Information on Data and Dataset**

The provided data is organized on a monthly basis. Each row in the dataset corresponds to a specific month, and the "Formatted\_Date" column contains the month and year in a formatted text, such as "Jan-92" for January 1992. This monthly granularity allows for a detailed examination of trends, patterns, and fluctuations in the Retail and Food Services industry over the entire time period covered by the dataset, from January 1992 to February 2023.

**STEP 3 – Explore and Visualize Data**



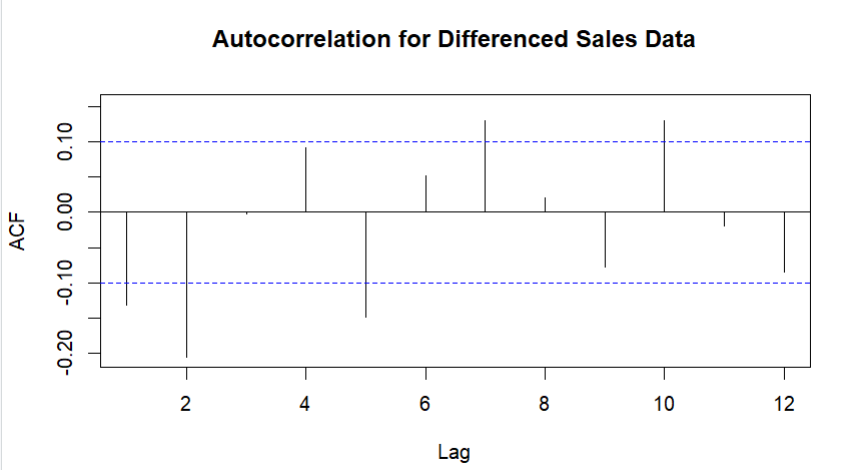
The data plots show the sales for the period of 32 years i.e.) from year 1992 to 2023. The time series trend is said to have a combination of upward and downward trends. There is a slight dip in sales from the year 2008, but the overall trend observed is said to be positive.



Autocorrelation function (ACF) tells you how much past values of a time series influence future values. The ACF graph in the above chart shows that there is a positive autocorrelation at lag 1. This means that there is a positive correlation between food sales at a given time and food sales at the previous time period. From the above chart, we can see that the data is highly correlated, all the autocorrelation coefficients for all the lags are higher than zero. The graph from lag 1 to lag 12 is said to have positive autocorrelation as they have crossed the blue dotted lines. For these, the autocorrelation values have a positive value. All of them are said to be statistically significant and at most of the points, there is a consistent trend and seasonality being shown in the graph.

First Differencing and null hypothesis

The primary objective of this analysis is to investigate the autocorrelation in first-differenced sales data, utilizing the Acf() function in R. Additionally, a z-test is conducted to assess the significance of the autocorrelation at lag 1. The calculated z-statistic was -1.333333, resulting in a p-value of 0.09121122. Since the p-value exceeds the chosen significance level (p-value > alpha), the decision is made to accept the null hypothesis.



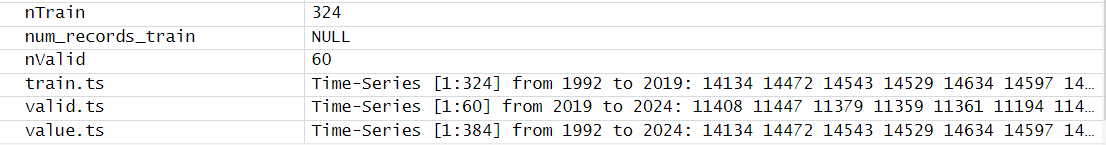
The fact that the autocorrelation function (ACF) for the differenced sales data is close to zero for all lags suggests that there is no significant autocorrelation in the data. This is a good thing for time series forecasting, because it means that the errors in the forecast are not correlated with each other.

**STEP 4 – Data Preprocessing**

To ensure the data integrity and suitability for the analysis, preprocessing steps are taken into consideration. The dataset comprises of ‘Value’ and ‘Formatted\_Date’ columns. Also, the data taken is for a span of January 1992 to December 2024. On close observation, we see that there are missing values denoted by ‘NA’ in the dataset for the span of February – December 2024, revealing the presence of missing data. Considering the significance of maintaining data completeness, the decision was made to remove rows containing missing values. Consequently, all data corresponding to the year 2024 was omitted from the dataset. Hence, the final dataset is said to be considered from 1992 to 2023.

**STEP 5 – Partition Series**

The dataset is divided into two parts – Training and Validation. The training is used to train the forecasting models and the set consists of 324 records from the period of January 1992 to December 2018. The validation set is used to validate the performance of the forecasting models and has 60 records from the period of January 2019 to December 2023.



**STEP & 7 - Apply Forecasting & Comparing Performance**

Advanced Exponential Smoothing

**The Holt- Winters Method**

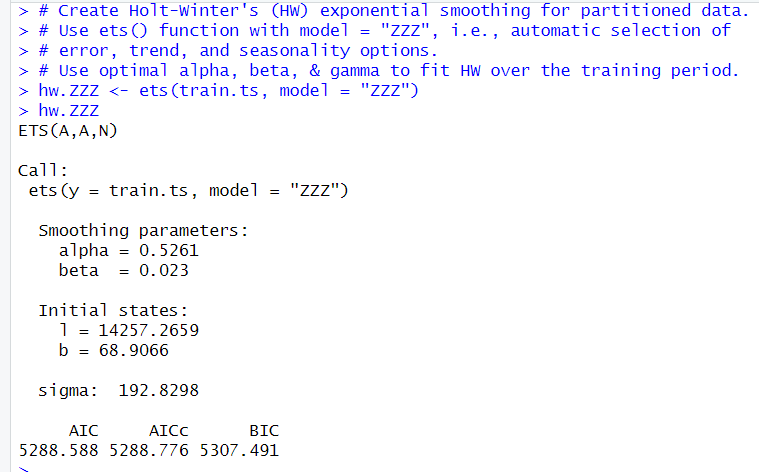
The Holt-Winters method, also known as the triple exponential smoothing method, is an advanced time series forecasting technique that incorporates seasonality and trend components. It is particularly useful for datasets with a strong seasonality pattern. This model was applied for a dataset of food sales with a span of 32 years. Much before this, it is being evaluated on the training and the validation partitions.

The Holt-Winters method is depicted by c (Z, Z, Z) parameter in the code which means it is a representation of error, trend and seasonality components respectively. As the alpha (error), beta (trend), or gamma (seasonality) values in the code are not being shown specifically, the model will generate optimized values during the fitting process. Here the model is taking additive error, additive trend and seasonality component.

|  |  |
| --- | --- |
| **Alpha** | **Beta** |
| 0.5261 | 0.023 |

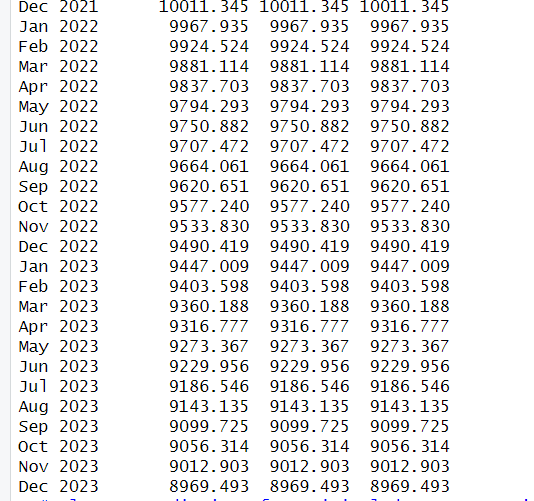
The alpha parameter is pretty much closer to zero, that represents there is a high level of smoothing value. It also depicts the weight of the most recent observation when updating the level component. The beta parameter represents the changes in the trend component is responsive to the changes.

Holt-Winter’s exponential for Training and Validation

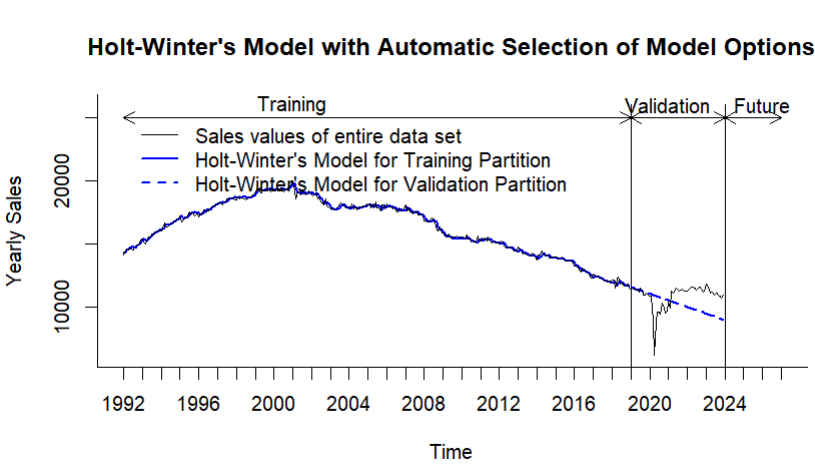


From the above values, we conclude that the statistics depict the efficient working of the model and it reasonably captures the underlying patterns of the data. The AIC, AICc, BIC values being 5288.588, 5288.776, and 5307.491 where the first two have lower values also show that the model fits in well. The sigma value =192.8298 indicates the randomness of the data.

Forecasting on the Training and Validation



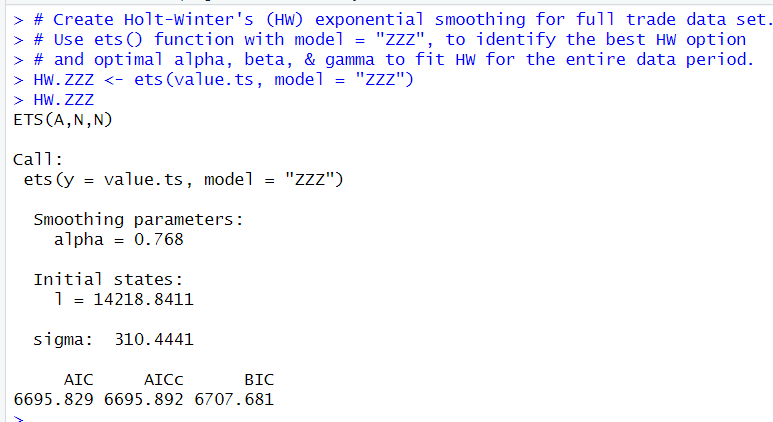
Plotting



The graph shows that the sales values of the model with automatic selection of model options, which is forecasted using Holt-Winter's model, have been increasing over time. The y-axis of the graph shows the yearly sales, and the x-axis shows the time. The validation partition of the data assess the models fit and by this we can conclude that the model is a robust and a good fit.

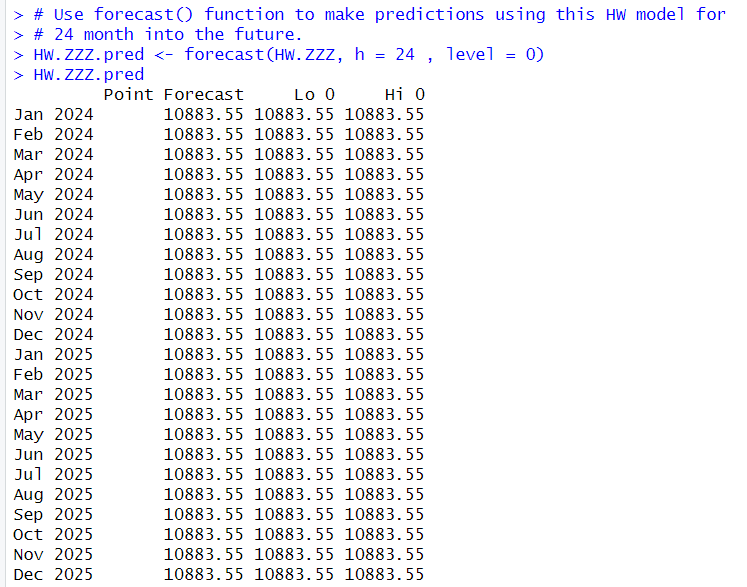
Holt-Winter’s exponential smoothing for entire data

The model here is being considered for the entire dataset.

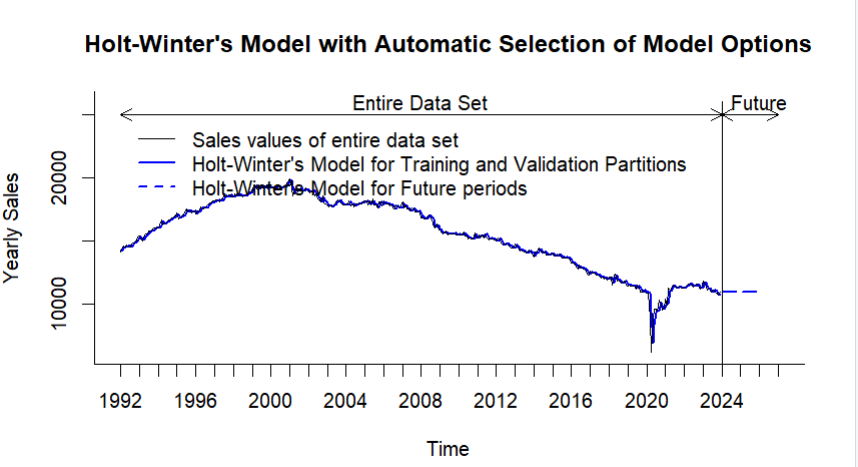


The ETS (A, N, N) model is representing the additive error, trend and seasonality components. The alpha value of 0.768 indicates that the model assigns relatively high weight to recent data points when making predictions. The sigma value is 310.4441 that shows the variability of the model used. The AIC, AICc, BIC values of 6695.829, 6695.892, and 6707.681 are all around the same range indicating lower values. In conclusion, the model indicates trend and seasonality showing a reasonably better fit. The optimized values in turn lead to more accurate predictions.

Forecasting on the entire data

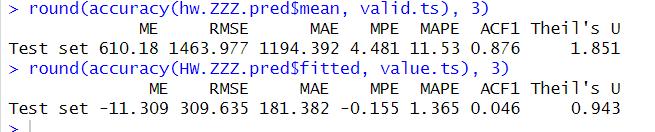


Plotting



From the above graph, the forecast sales values for future periods are also displayed on it. This is the core function of the Holt-Winters model, and it is likely generated by extrapolating the trends identified in the historical data. Overall, the graph suggests that the Holt-Winters model has been used to forecast sales of a product. The forecast shows the predicted sales values for future periods capturing the trends and components.

Accuracies



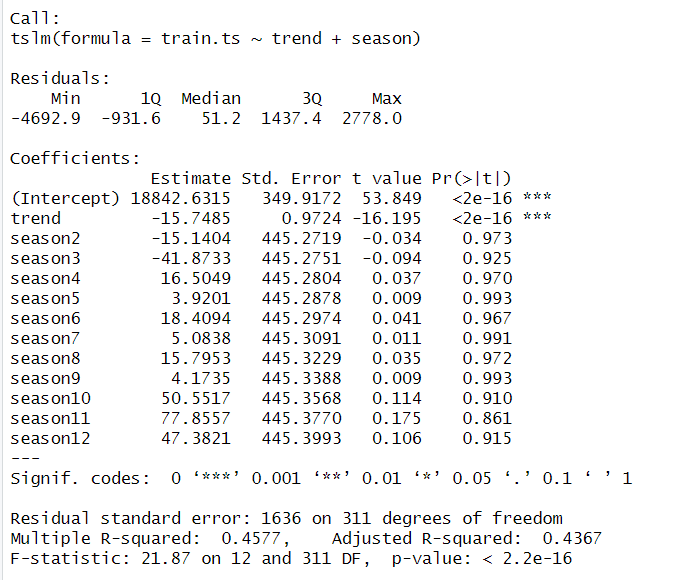
|  |  |  |  |
| --- | --- | --- | --- |
| Holt-Winter | RMSE | MAPE | ACF1 |
| Training & Validation | 1463.977 | 11.53 | 0.876 |
| Entire Data Set | 309.635 | 1.365 | 0.046 |

From the accuracies, the RMSE and MAPE values for the training and validation is 1463.977, 11.53 that gives the magnitude of the model. The ACF1 value is 0.876 that shows the degree of autocorrelation in the residuals. The RMSE and MAPE values of the entire data set being 309.635 and 1.365 has given better values and show the comprehensive understanding of the forecasting performance.

**Regression Models**

The next one used in the analysis is regression based models. This type of model is straightforward and usually gives good results because it considers both the overall trend and seasonal patterns. We can make it even better by adding autoregressive elements and looking at past errors. Before applying the model to the entire dataset, a preliminary assessment was conducted using distinct training and validation partitions.

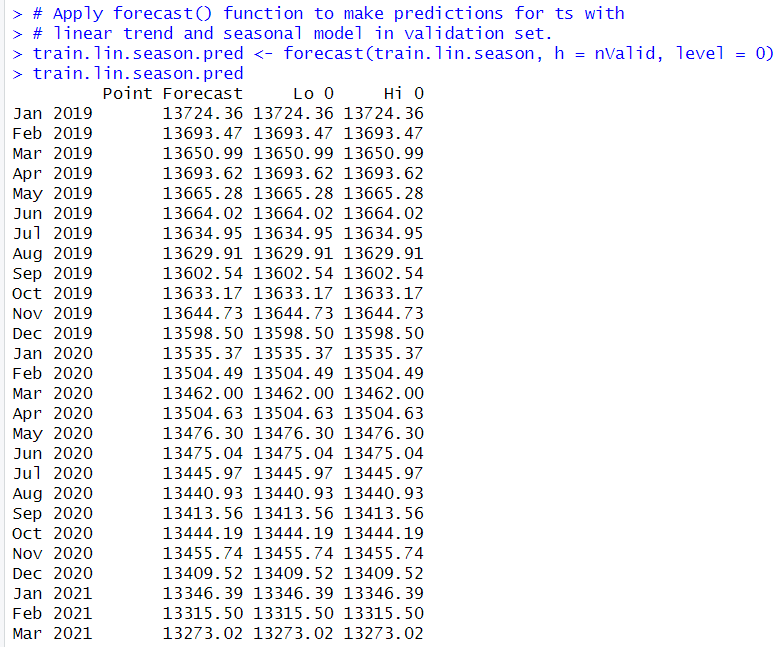
Regression model with Linear Trend and Seasonality



For the above time period, the model represents a regression model with linear trend and seasonality. The model is said to be statistically significant as the F-statistic p-value (2.2e-16) is a very low value, lower than that of the alpha of 5%. These also indicate a strong linear trend in the time series. This suggests that there is a consistent upward or downward movement in the data over time. The R-squared value is 45.7%. The model has one predictor and other dummy seasonal predictors. The intercept and trend coefficients are statistically significant, suggesting a strong linear trend in the time series and a good fit. Hence, it is also applied for forecasting.

*The Model Equation = Yt = 18842.6315 -15.7485 t - 41.8733D2 – 16.5049 D3+ . . . .+ 47.3821D12*

Forecasting

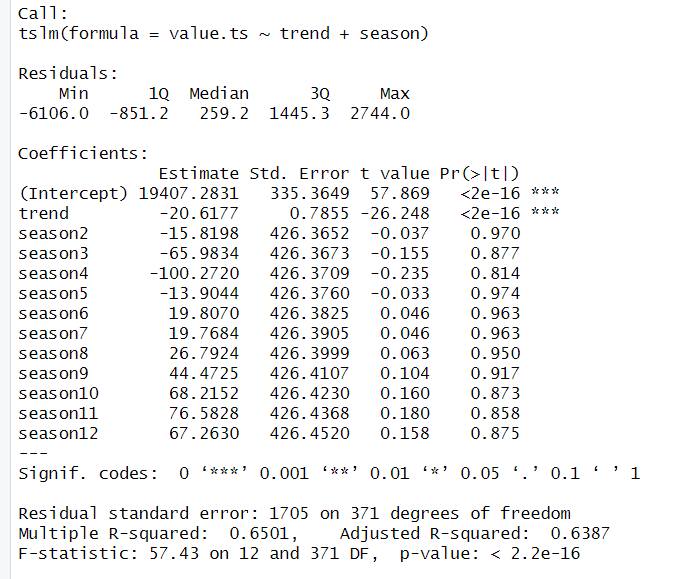
 

Plotting



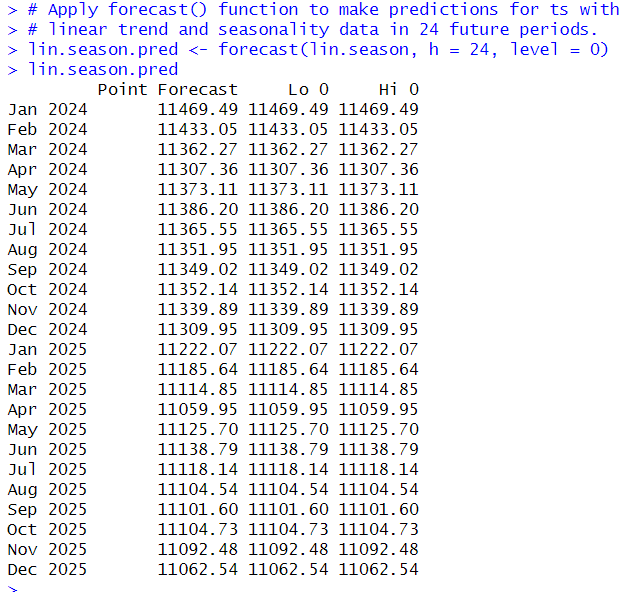
The graph above is a line graph showing a regression model for yearly food sales with both a linear trend and seasonality. The overall trend in food sales is upward and then declined after a point, which means that food sales have been increasing and then decreasing over time. This is consistent with the linear trend model that has been fit to the data. The validation data is used to assess how well the model generalizes to unseen data.

Regression model with Linear Trend and Seasonality on Entire Dataset

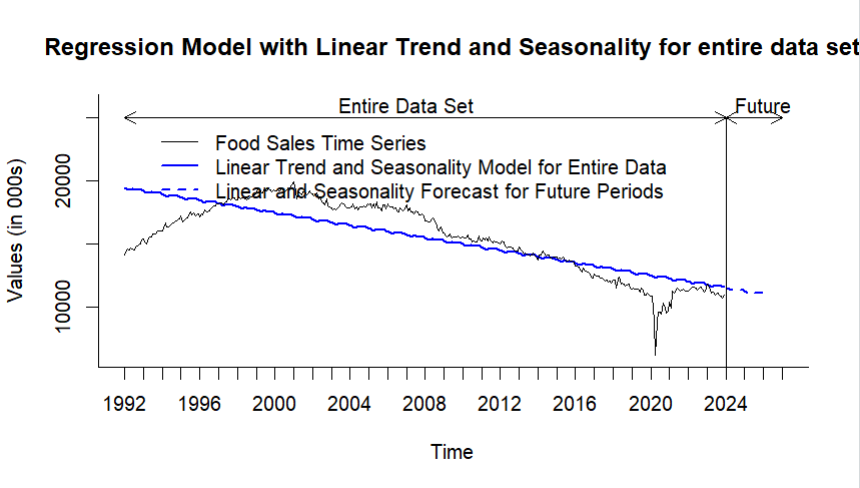


The model provides valuable insights into the trend and overall seasonality of the time series data. While the trend is statistically significant, caution should be exercised in interpreting the seasonal coefficients due to their lack of significance. The multiple R-squared value is 0.6501, indicating that the model explains approximately 65.01% of the variability in the response variable. The adjusted R-squared, accounting for the number of predictors, is 0.6387. This suggests a reasonably good fit. The F-statistic is 57.43 with a p-value less than 2.2e-16, indicating that the overall model is statistically significant. Hence, it a good fit for the forecasting.

Forecasting on Entire data

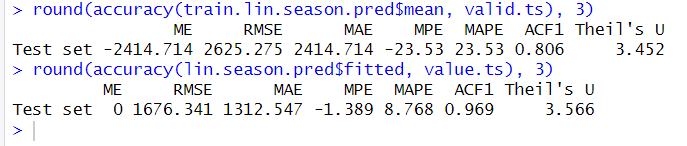


Plotting



There is a general upward trend in food sales over time. This means that food sales have been increasing on average over the period that the data covers. There is also a seasonal component to food sales. This means that there are predictable ups and downs in sales throughout the year. The forecasted trend shows that the model predicts that food sales may increase in the future.

Accuracies



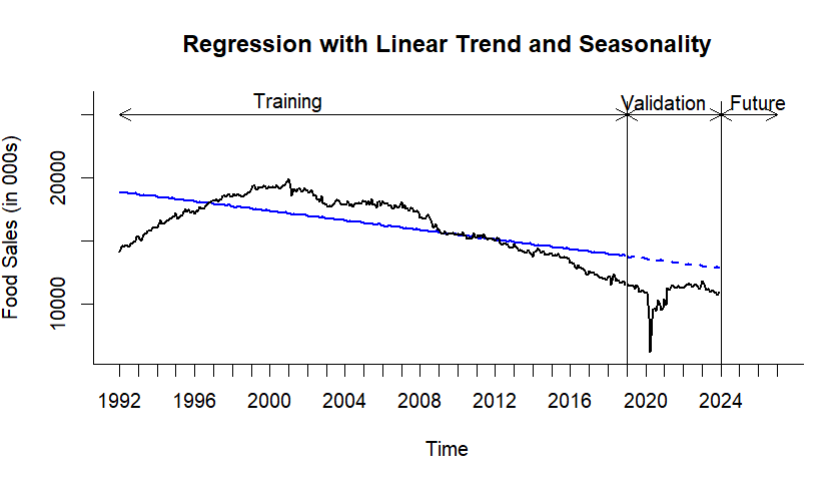
|  |  |  |  |
| --- | --- | --- | --- |
| Regression model with Linear Trend and Seasonality | RMSE | MAPE | ACF1 |
| Training & Validation | 2625.275 | 23.53 | 0.806 |
| Entire Data Set | 1676.341 | 8.768 | 0.969 |

From the accuracies, the RMSE and MAPE values for the training and validation is 2625.275, 23.53 that gives the magnitude of the model. The ACF1 value is 0.806 that shows the degree of autocorrelation in the residuals. The RMSE and MAPE values of the entire data set being 1676.341and 8.768 has given better values and show the comprehensive understanding of the forecasting performance.

**Two - Level Forecasting**

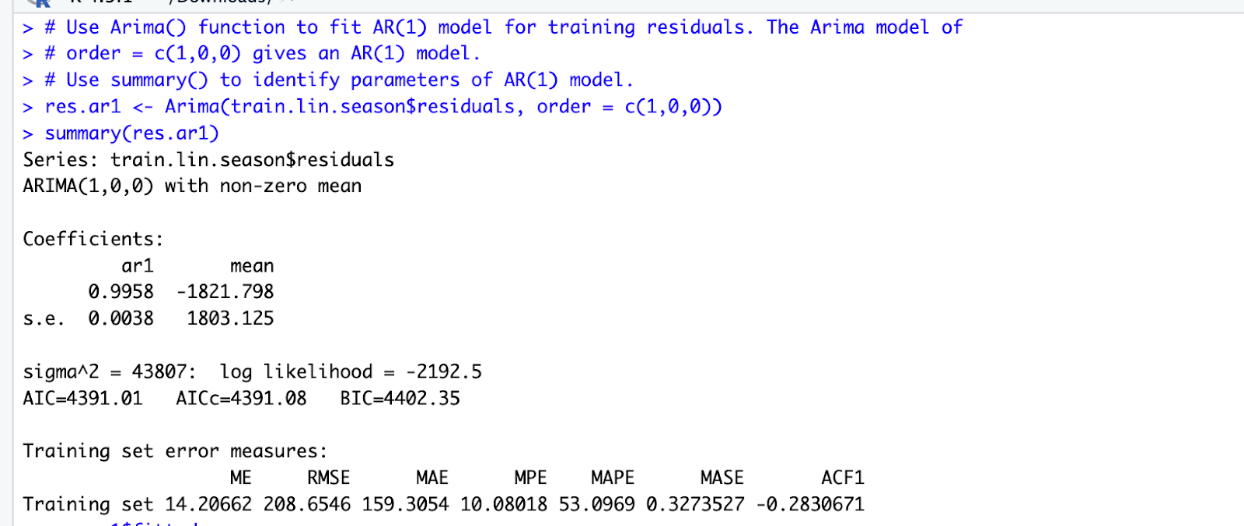
Develop a regression model with linear trend and seasonality – Autocorrelation for Residuals

The residuals exhibit a range of values, indicating potential variability in the data that is not explained by the model, here on the training part. The model explains approximately 45.77% of the variance in the training data, as indicated by the multiple R-squared value. The adjusted R-squared accounts for the number of predictors and suggests a slightly lower explanatory power of 43.67%. The F-statistic is highly significant, supporting the overall significance of the model. The p-values for certain components lack statistical significance.



Autoregressive Models for Regression Residuals

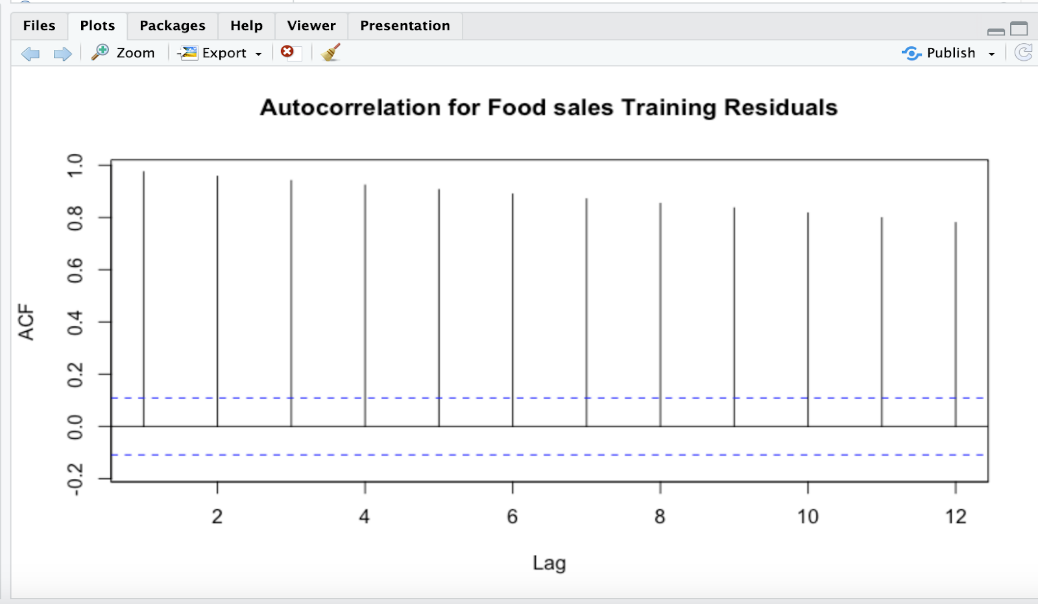
Prior to using an autoregressive model, a correlogram must be created which shows the autocorrelation between the residuals of the respective regression models.



*The Model Equation is yt = a + b1yt-1 + yt*

*We get the following model equation yt = -1821.798+ 0.9958 yt-1*

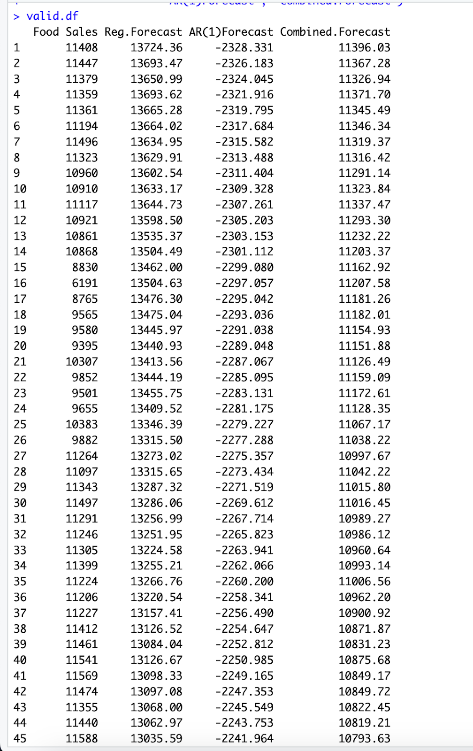
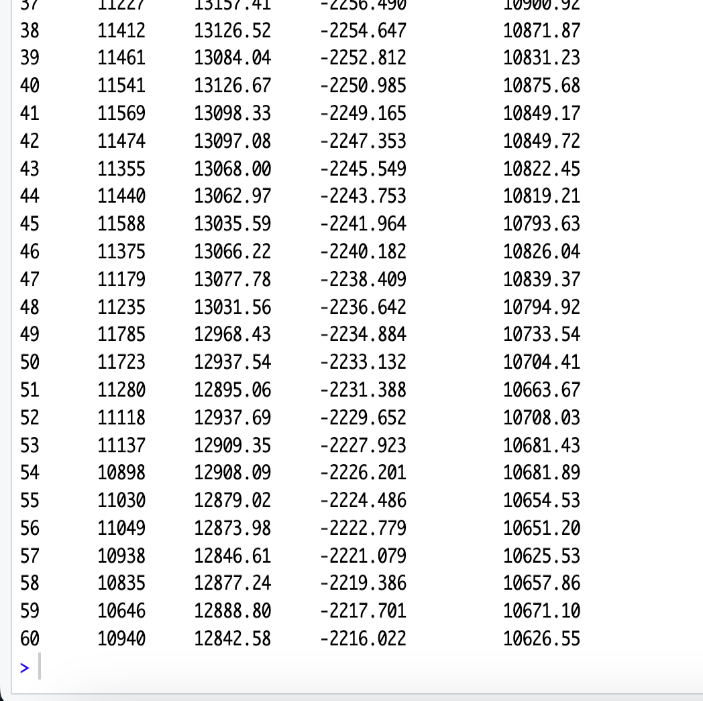
The coefficient of the ar1 (Yt-1) variable, β1 = 0.9958, and standard error of estimate, s.e. =0.0038. We will use these two parameters for hypothesis testing about the value of the AR(1) regression coefficient.



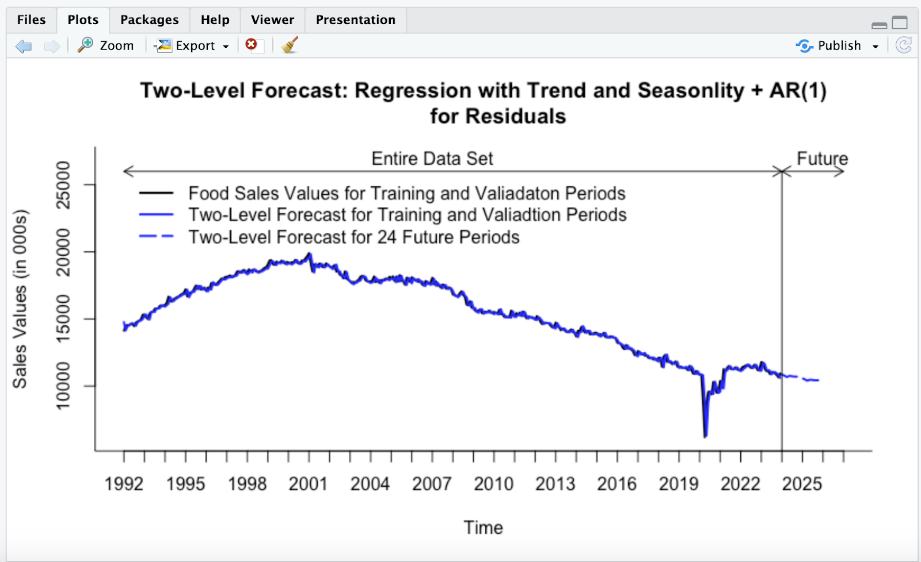
Most of the lags for Autocorrelation for Regression Model’s residuals are statistically significant with strong positive correlation in our dataset. So, it is a good idea to add it to our forecast an AR model for residuals.

Two-level model for linear trend and seasonality model and AR(1) model for residuals

The table describes the wholesale food sales data and forecasts in the validation partition (Valid. regression model’s forecast in the validation period (Reg.Forecast), AR(1) model’s forecast of the regression residuals in the validation period (AR(1)Forecast), and combined forecast (Combined. Forecast) as a sum of the regression and AR(1) models’ forecasts.

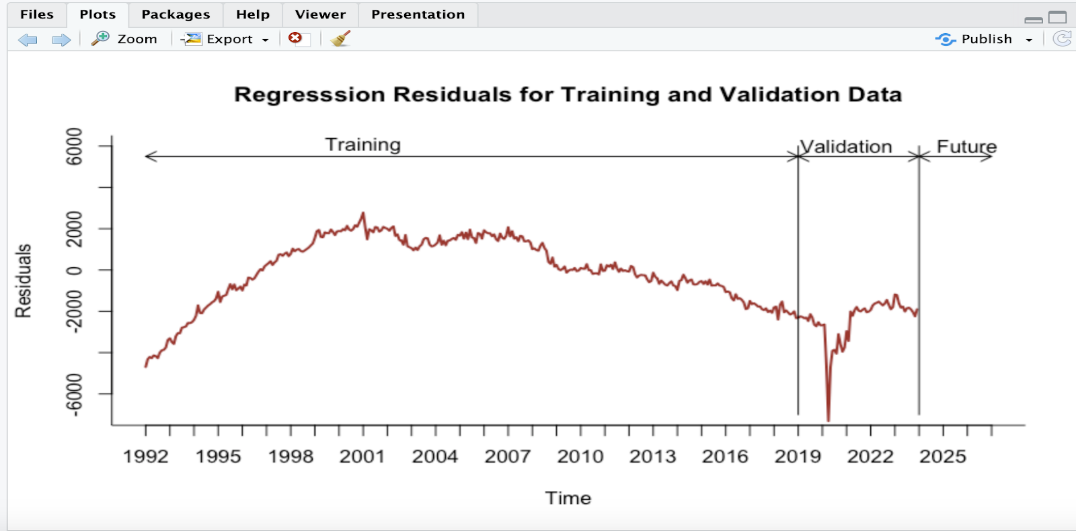
 

The plot of the 2-level model for linear trend and seasonality and AR(1) model for residuals for residuals for training and validation data set shows that the model is fitting well into data, in which we are capturing trend and seasonality components.

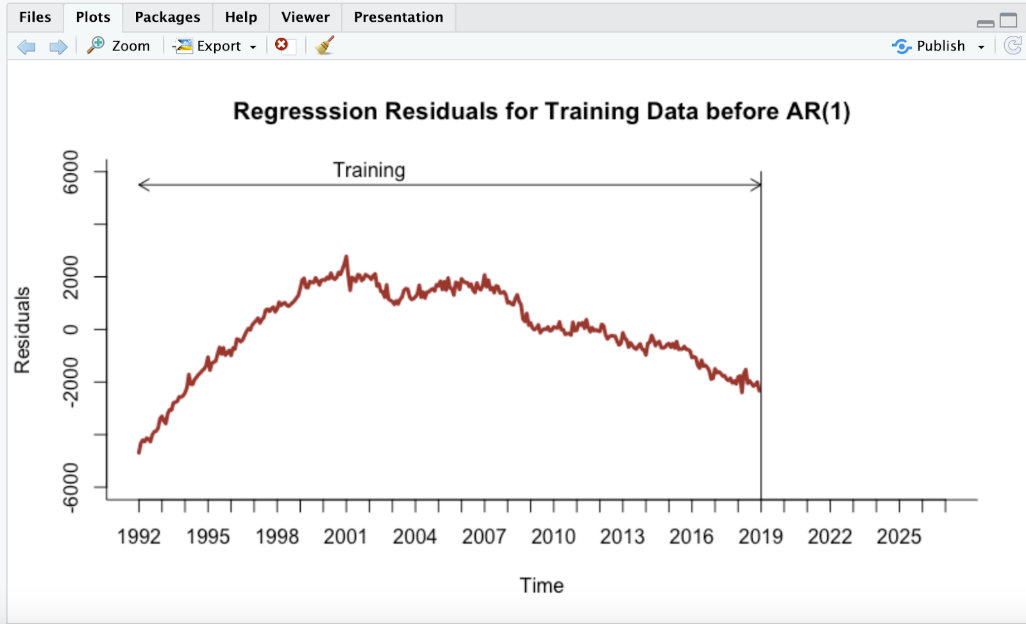


Two-level forecasting Model’s residuals plots

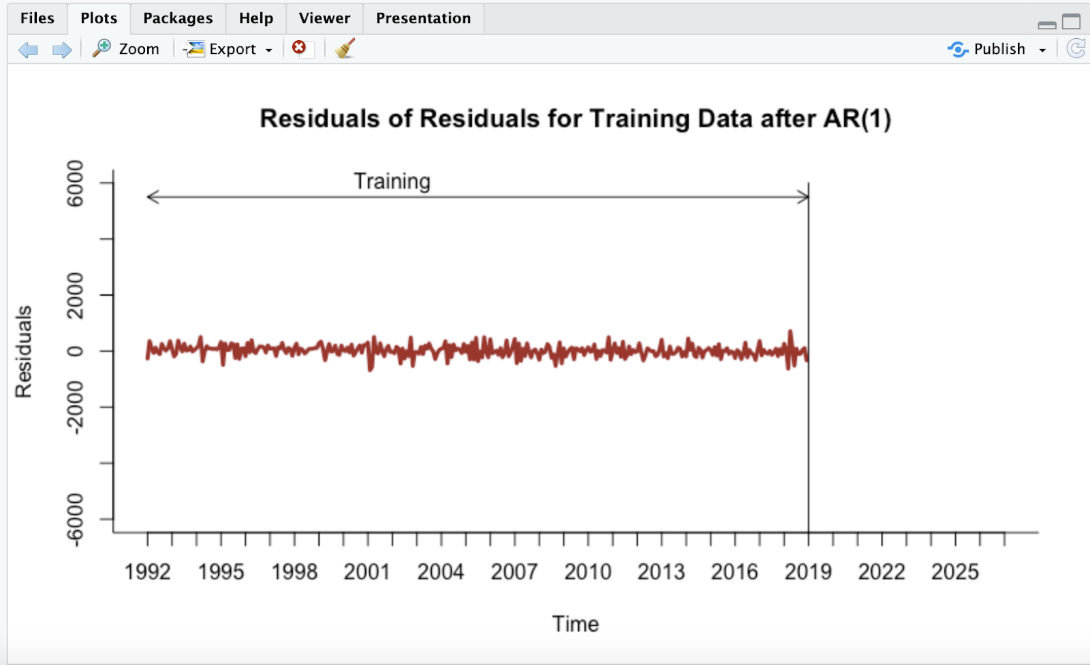
Regression residuals for the entire dataset (training and validation data).



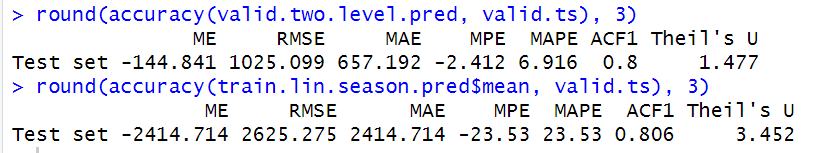
Regression residuals for training data before AR(1)

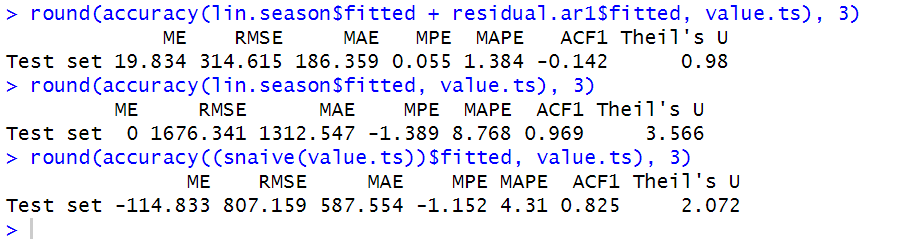


Regression residuals for training data after entire dataset



Accuracies





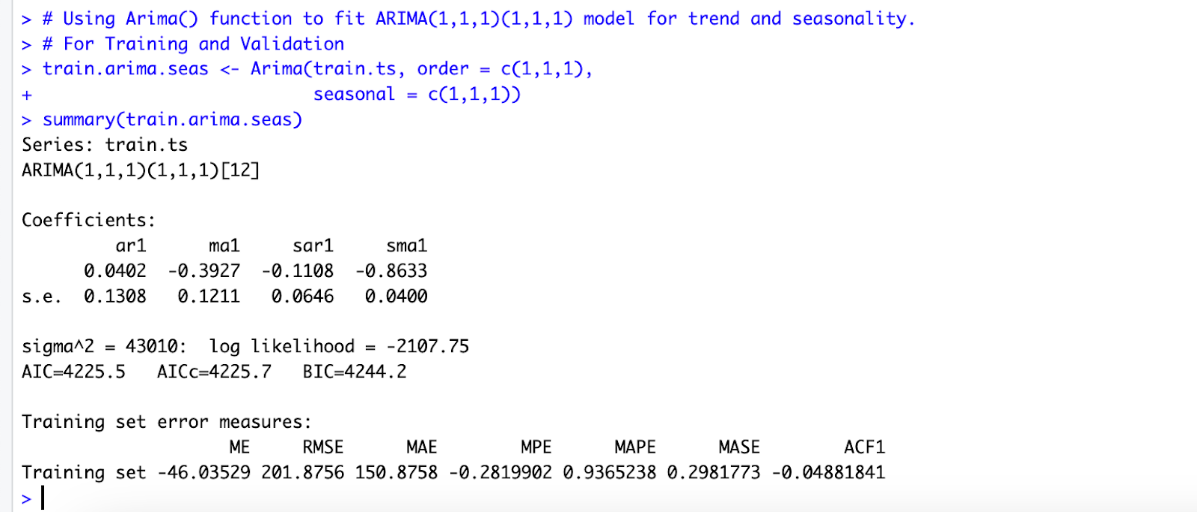
|  |  |  |  |
| --- | --- | --- | --- |
| Two-Level Forecast | MAPE | RMSE | ACF1 |
| Training and Validation | 6.916 | 1025.099 | 0.8 |
| Entire Data Set | 1.384 | 314.615 | -0.142 |

From the accuracies, the RMSE and MAPE values for Two-Level Forecast model for the entire dataset is 314.615, 1.384 that gives the magnitude of the model. The ACF1 value is -0.142 that shows the degree of autocorrelation in the residuals.

**Auto Regressive Integrated Moving Average Model**

The ARIMA (Auto Regressive Integrated Moving Average) model is a popular time series forecasting method that combines auto regression, differencing, and moving averages. This can be used for forecasting on data with various components such as level, trend and seasonality. Seasonal ARIMA (Seasonal Auto Regressive Integrated Moving Average), is an extension of the basic ARIMA model designed to handle time series data exhibiting seasonal patterns or fluctuations. Seasonality refers to the periodic variations in data that occur at regular intervals, here it is monthly.

ARIMA for Training and Validation data:

****

This is a seasonal ARIMA model, ARIMA(p, d, q)(P, D, Q)[m], where:

• p = 1, order 1 autoregressive model AR(1)

• d = 1, first differencing

• q = 1, order 1 moving average MA(1) for error lags

• P = 1, order 1 autoregressive model AR(1) for the seasonal part

• D = 1, first differencing for the seasonal part

• Q = 1, order 1 moving average MA(1) for the seasonal error lags

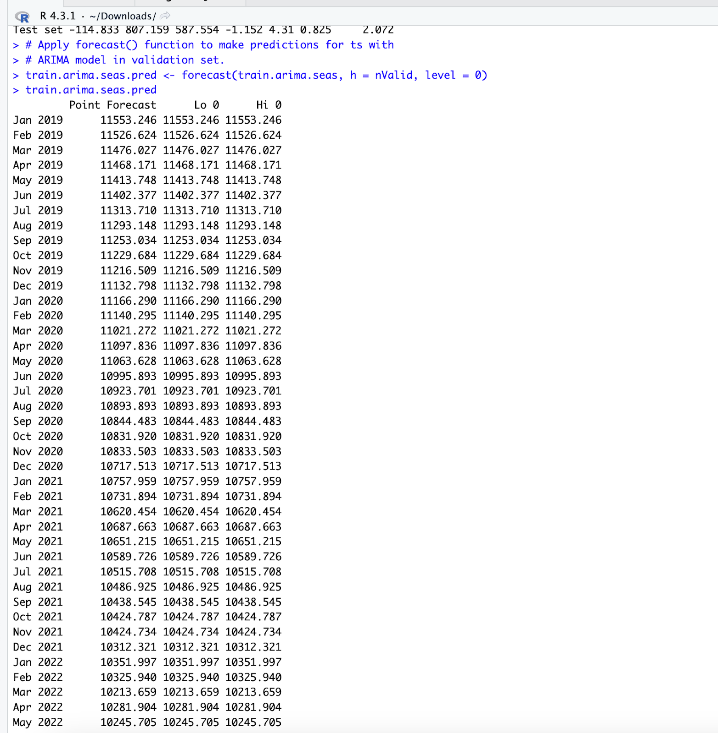
• m = 12, for yearly seasonality.

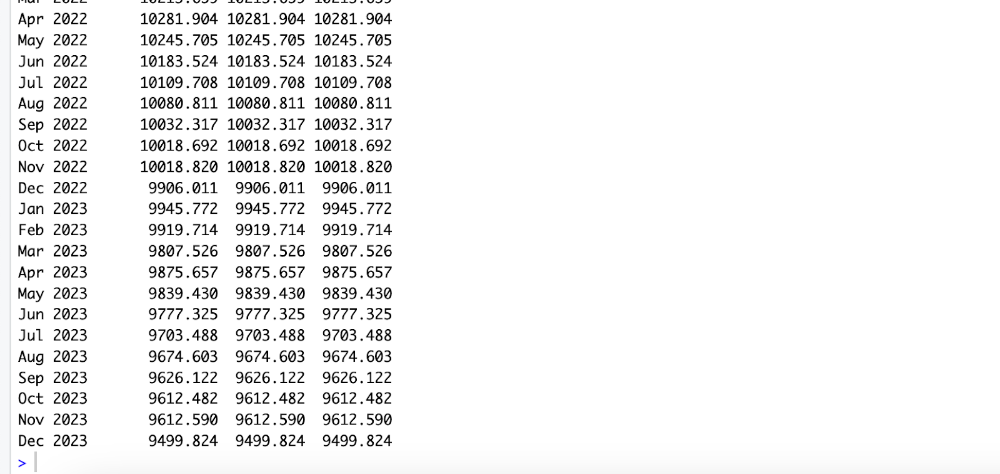
*The model’s equation is:*

*yt - yt-1 = 0.0402yt-1 - 0.3927εt-1 - 0.1108(yt-1 -yt-13) - 0.8633ρt-1.*

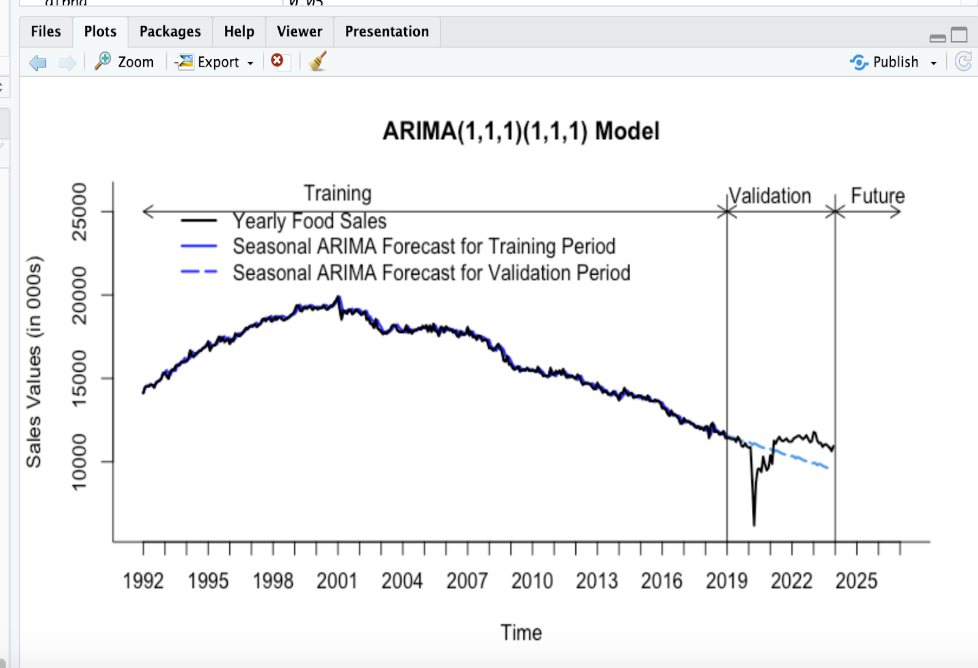
The variance of the error term (sigma^2) is 43010, and the log likelihood is -2107.75. The model is statistically significant. The model selection criteria AIC, AICc, and BIC are 4225.5, 4225.7, 4244.2, respectively.

ARIMA model’s forecast for the validation period

****

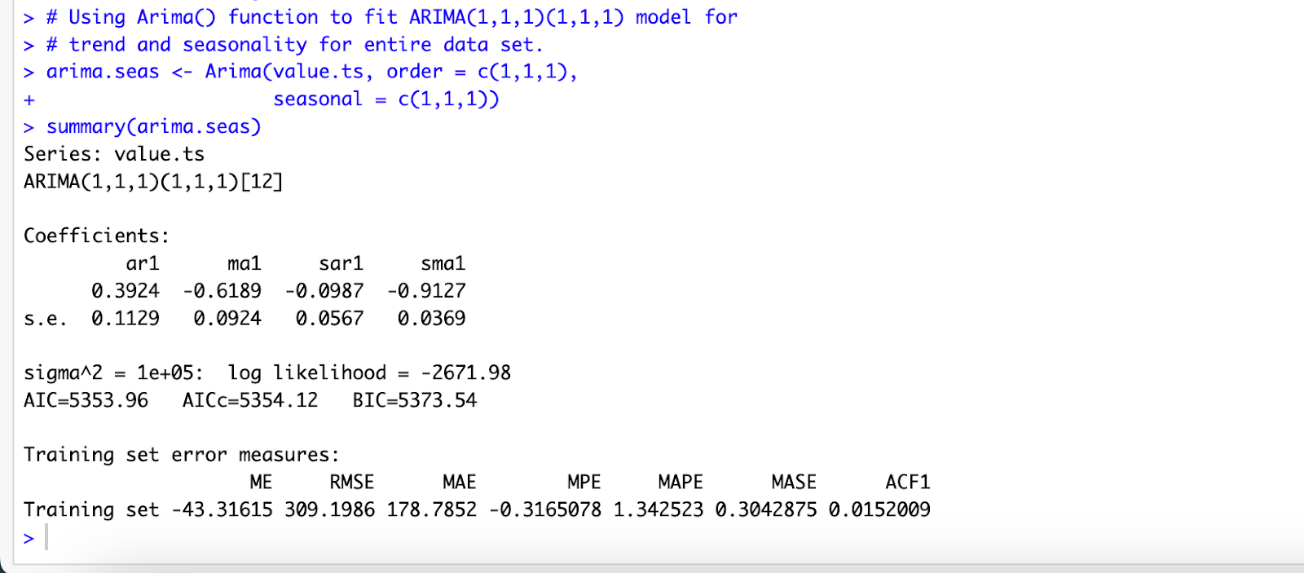


The plot validation data compared with forecast of validation data by ARIMA model

****

The plot of the ARIMA for training and validation data sets shows that the model is fitting well into data, capturing the trend components and is said to be statistically significant. So we can proceed for further forecasting and for the entire dataset.

ARIMA model for Entire Data Set

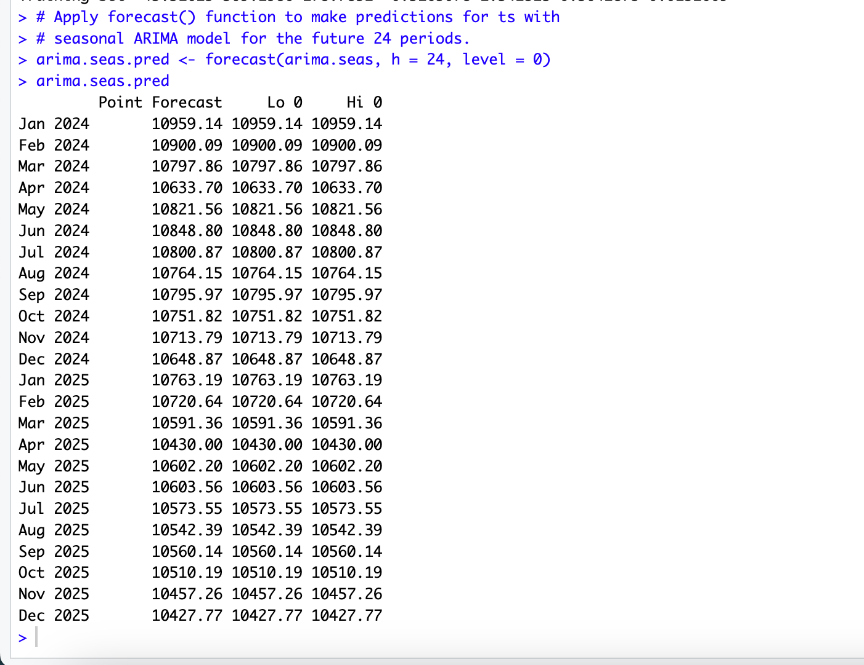


*The model’s equation is:*

*yt - yt-1 = 0.3924yt-1 - 0.6189εt-1 - 0.0987(yt-1 -yt-13) - 0.9127ρt-1.*

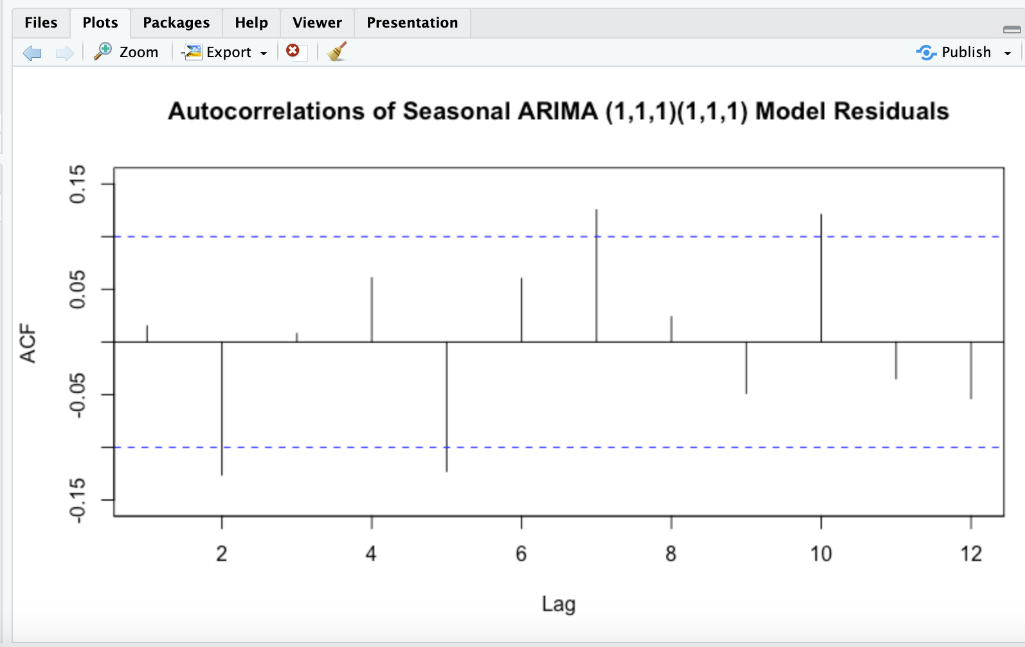
The variance of the error term (sigma^2) is 1e+05, and the log likelihood is -2671.98. The model selection criteria AIC, AICc, and BIC are 5353.96, 5354.12, 5373.54 respectively.

ARIMA model’s forecast for the future

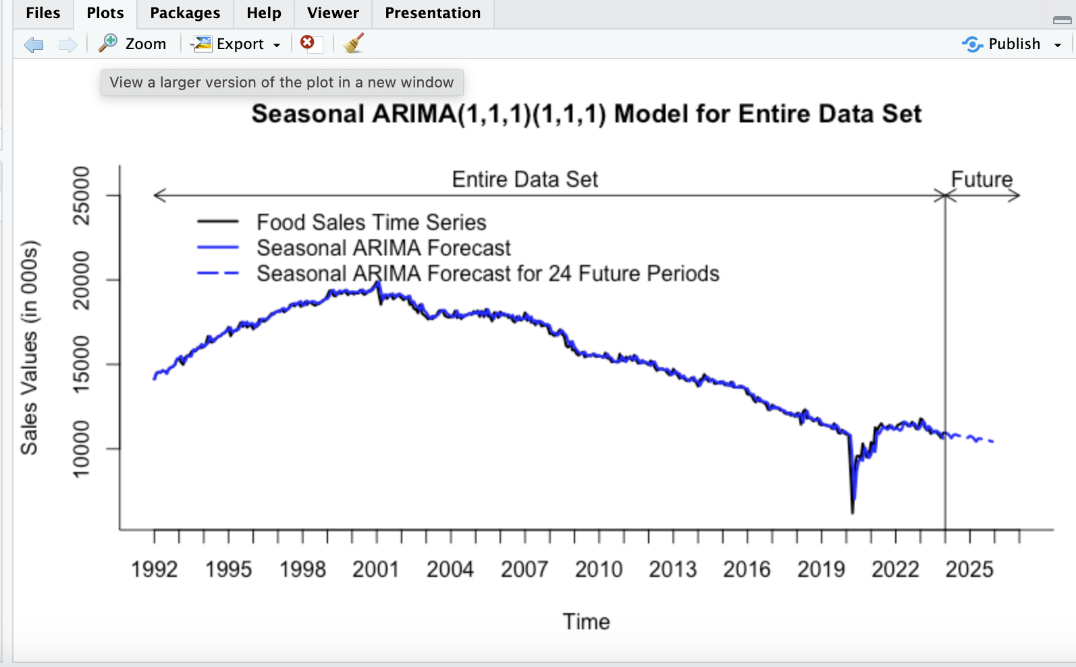


The autocorrelation plot of ARIMA model for the entire dataset

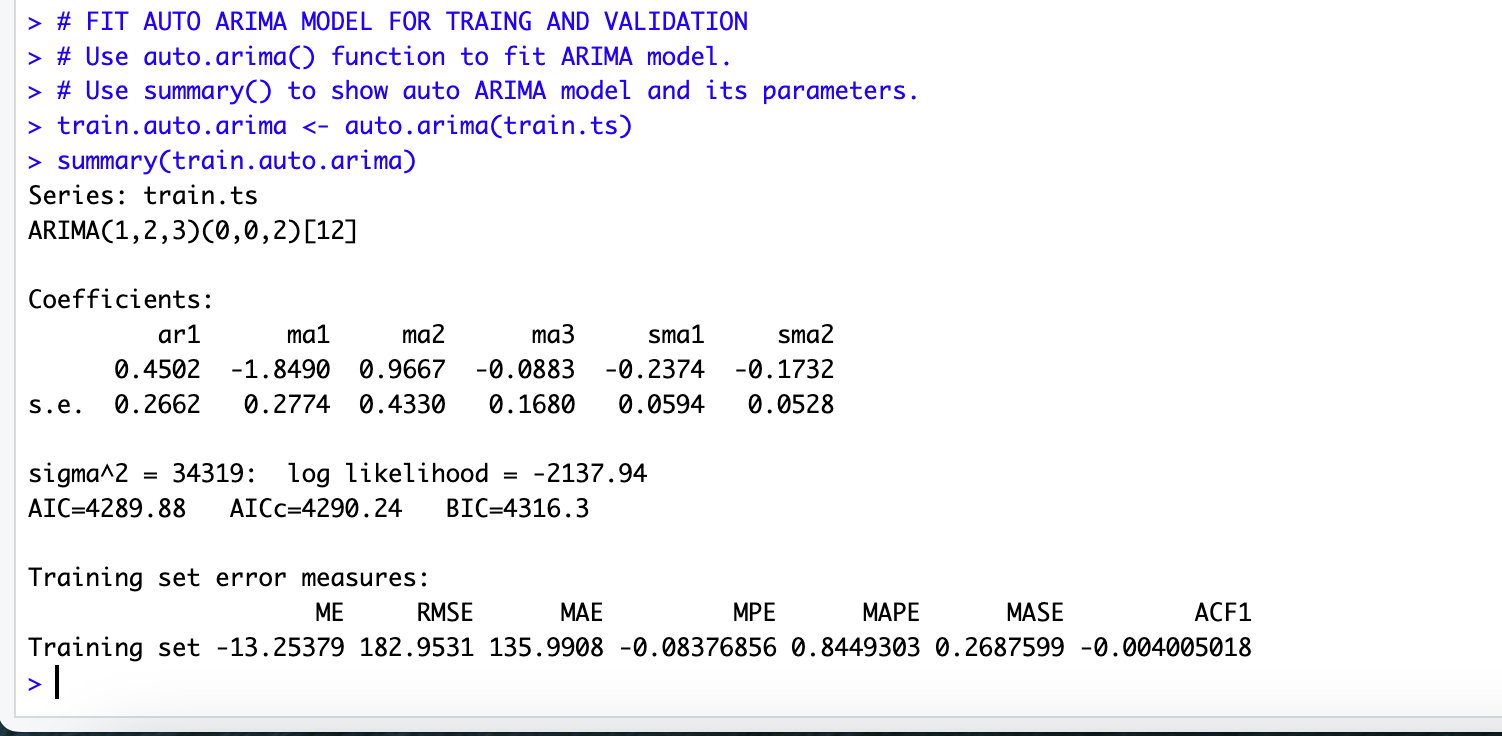
We can see that many lags autocorrelation is insignificant but for lag 2, 5,7,10 the autocorrelation is significant, so, we can say there is a room for improvement.



The future forecast of the year 2024-2025 by ARIMA model



AUTO ARIMA for Training and Validation data



This is a seasonal Auto ARIMA model, ARIMA(p, d, q)(P, D, Q)[m], where:

• p = 1, order 1 autoregressive model AR(1)

• d = 3, first differencing

• q = 2, order 1 moving average MA(1) for error lags

• P = 0, order 1 autoregressive model AR(1) for the seasonal part

• D = 2, first differencing for the seasonal part

• Q = 0, order 1 moving average MA(1) for the seasonal error lags

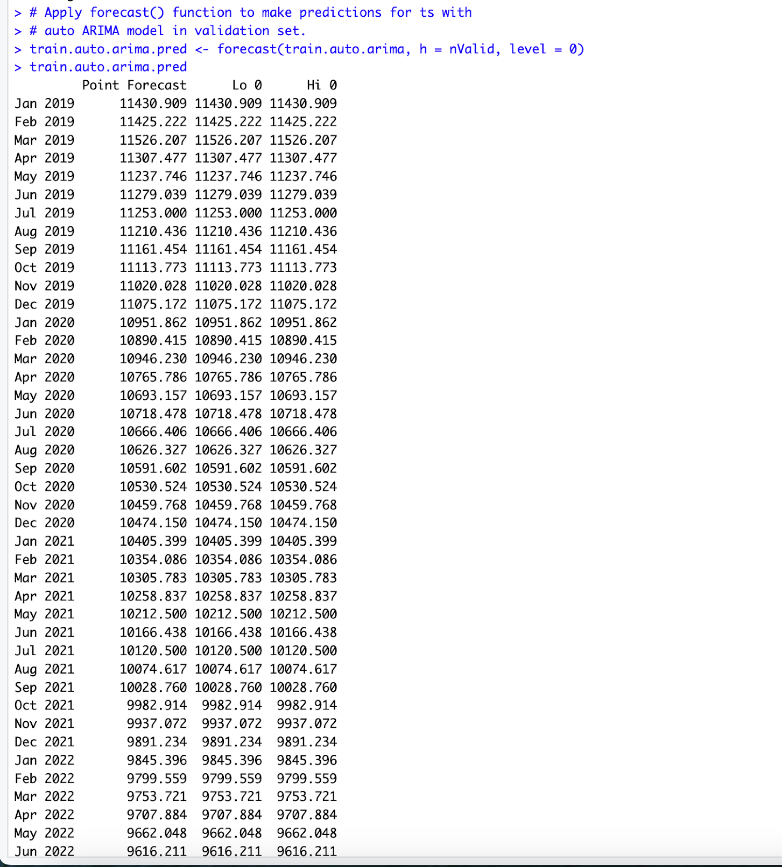
• m = 12, for yearly seasonality.

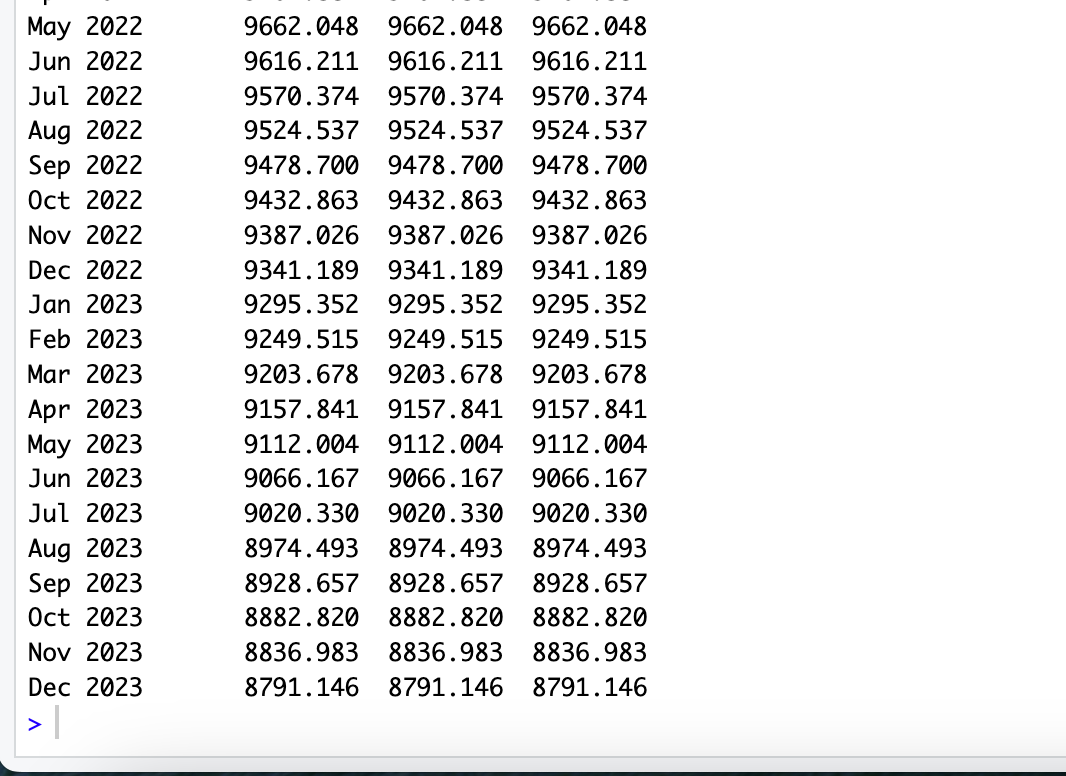
*The model’s equation is:*

*yt - yt-1 = 0.4502yt-1 - 1.849εt-1 + 0.9667εt-2 - 0.0883εt-3 - 0.2374ρt-1 - 0.1732ρt-2*

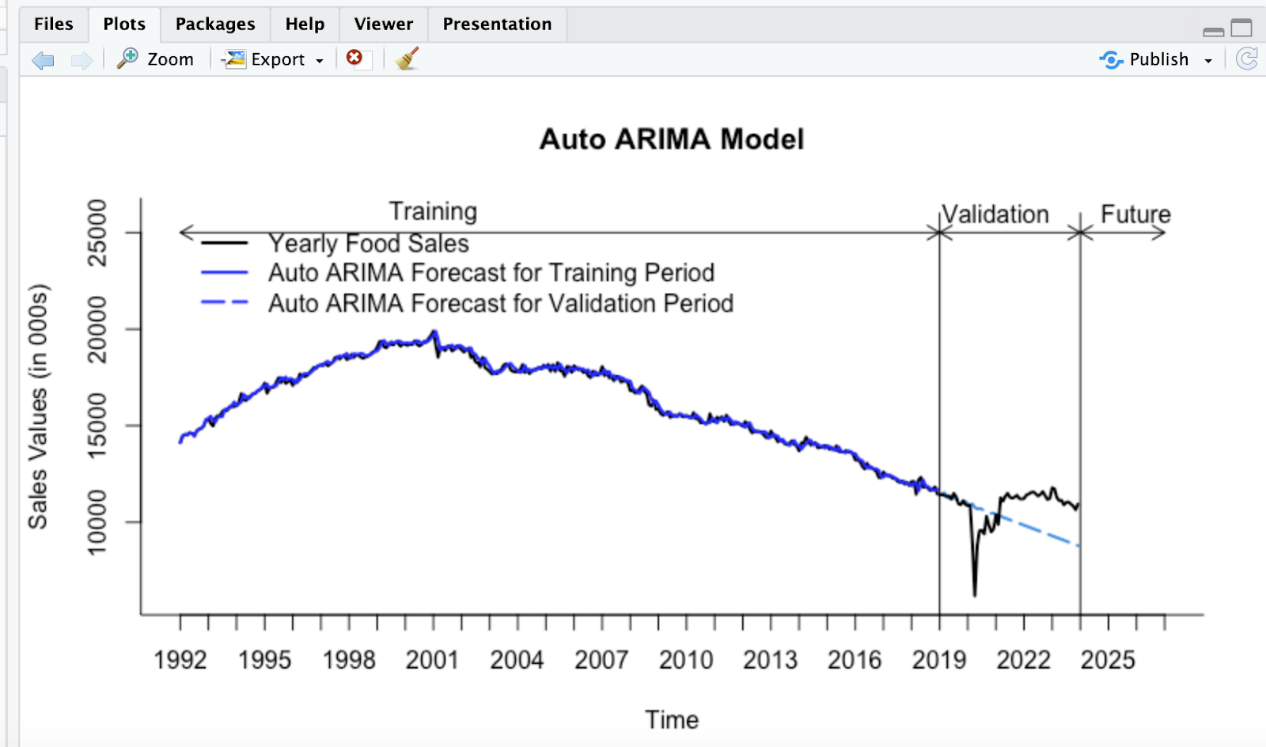
The variance of the error term (sigma^2) is 34319, and the log likelihood is -2137.94. The model selection criteria AIC, AICc, and BIC are 4289.88, 4290.24, 4316.3 respectively.

The future forecast the validation period using Auto ARIMA model

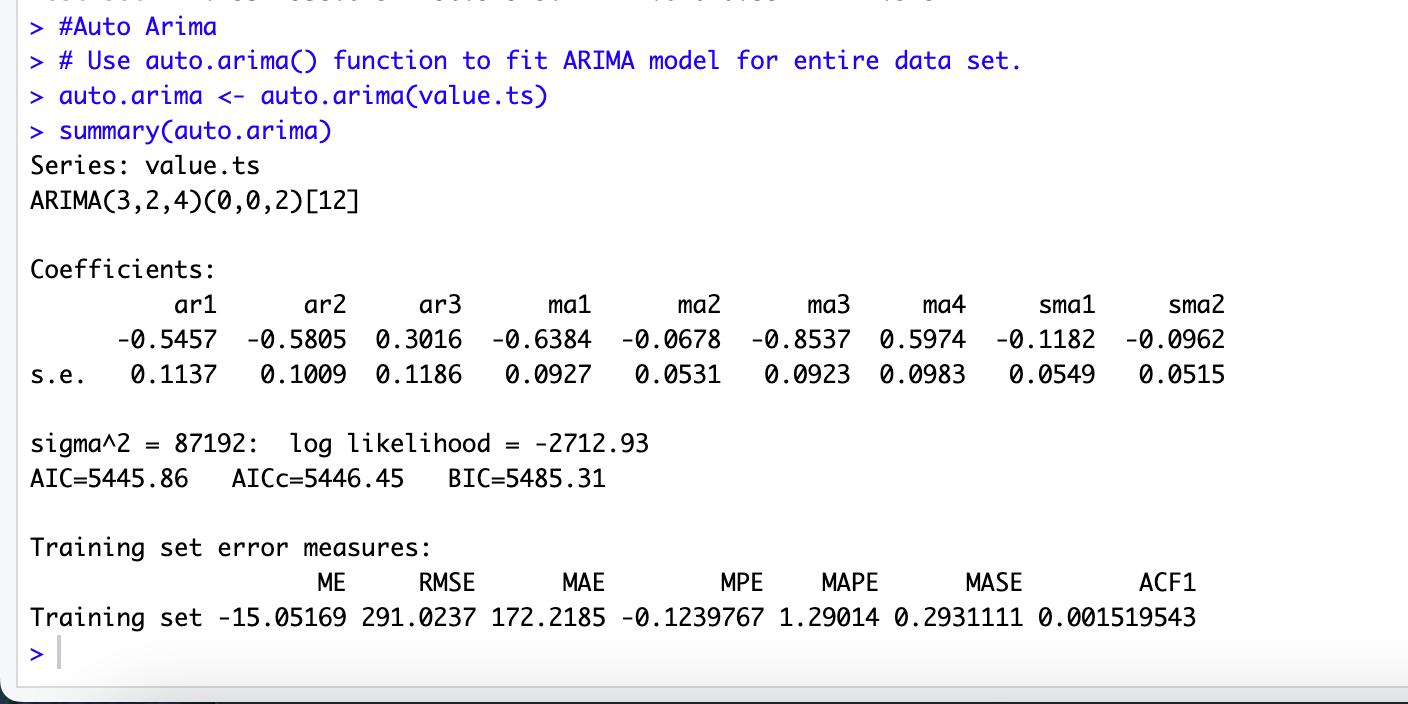




The plot validation data compared with forecast of validation data by Auto ARIMA model



Auto ARIMA model for Entire Data Set

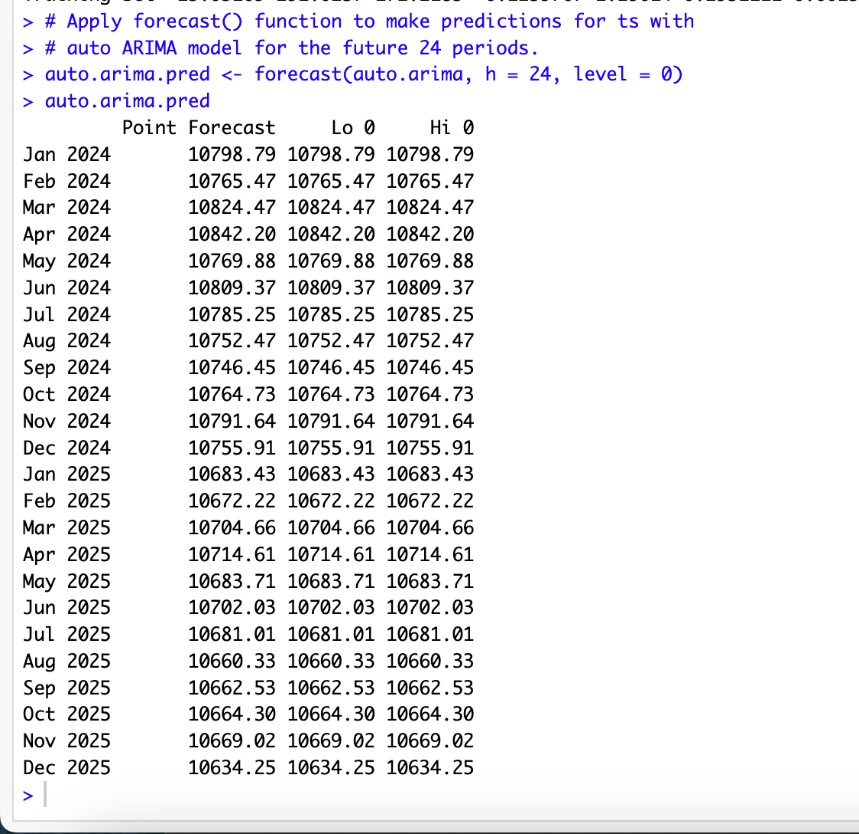


*The model’s equation is:*

*yt - yt-1 = - 0.5457yt-1- 0.5805yt-2 + 0.3016yt-3 - 0.6384εt-1 - 0.0678εt-2 - 0.8537εt-3 - 0.5974εt-4 -0.1182ρt-1 - 0.0962ρt-2*

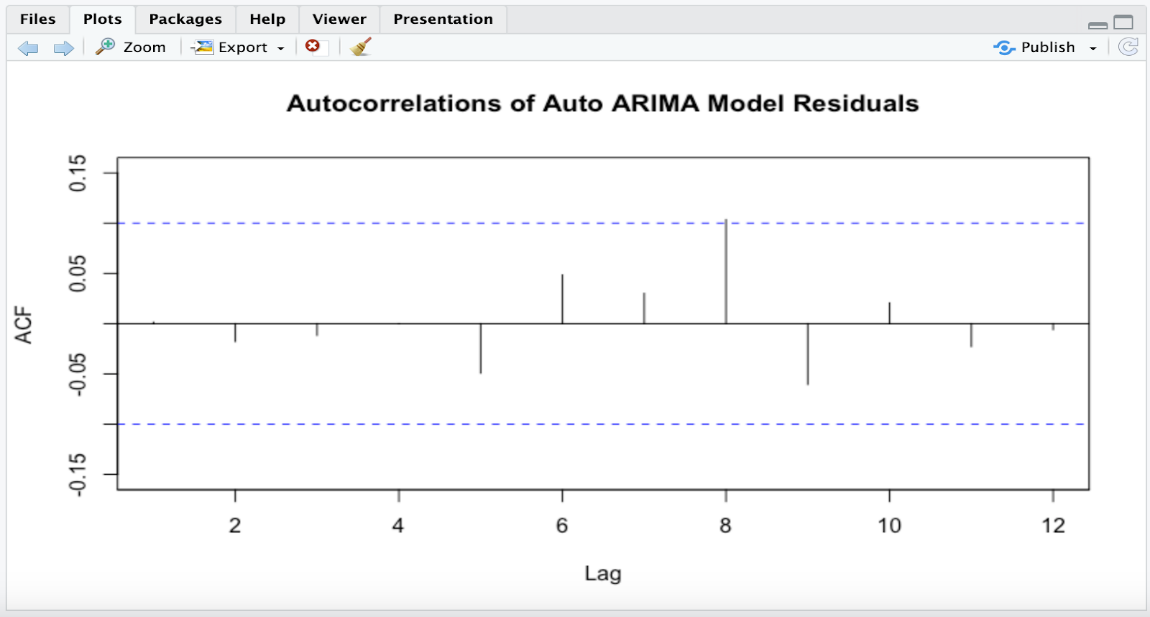
The variance of the error term (sigma^2) is 87192, and the log likelihood is -2712.93. The model selection criteria AIC, AICc, and BIC are 5445.86, 5446.45, 5485.31 respectively.

Future Forecast by the Auto ARIMA model



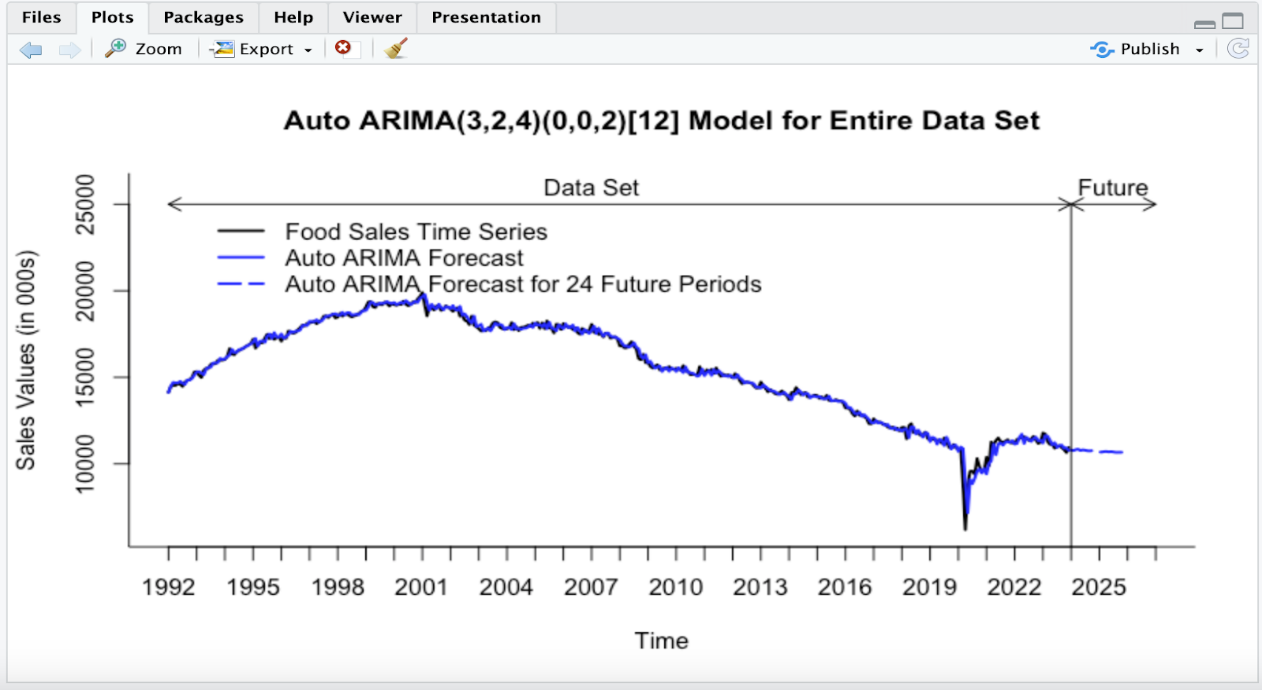
Autocorrelation plot for the Auto ARIMA model for entire dataset

We can see that the autocorrelation for the Auto ARIMA models residuals is insignificant. Hence we can say that the model is good.

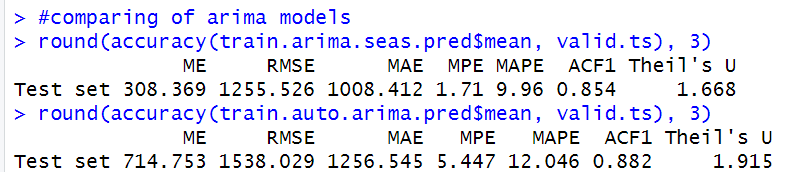


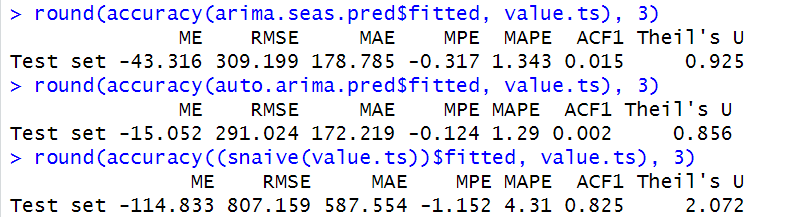
The future forecast of the year 2024-2025 by ARIMA model

The plot of the Auto ARIMA for entire data sets shows that the model is fitting well into data, capturing the trend components.



Accuracies





|  |  |  |  |
| --- | --- | --- | --- |
| Arima Model | MAPE | RMSE | ACF1 |
| Training and Validation | 9.96 | 1255.526 | 0.854 |
| Entire Data Set | 1.343 | 309.199 | 0.015 |

|  |  |  |  |
| --- | --- | --- | --- |
| Auto Arima Model | MAPE | RMSE | ACF1 |
| Training and Validation | 12.046 | 1538.029 | 0.882 |
| Entire Data Set | 1.29 | 291.024 | 0.002 |

From the accuracies, the RMSE and MAPE values for ARIMA model for the entire dataset is 309.199, 1.343that gives the magnitude of the model. The ACF1 value is 0.015 that shows the degree of autocorrelation in the residuals. The RMSE and MAPE values of the Auto ARIMA entire data set being 291.024 and 1.29 has given better values and show the comprehensive understanding of the forecasting performance. In both, auto ARIMA has the best values.

**STEP 8 – Implementing Forecast**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | MAPE | RMSE | ACF1 |
| Holt Winter | 1.365 | 309.635 | 0.046 |
| Two Level | 1.384 | 314.615 | -0.142 |
| Arima | 1.343 | 309.199 | 0.015 |
| Auto Arima | 1.29 | 291.024 | 0.002 |

The best model to be implemented on the analysis of food services is Auto Arima because of its accurate and apt values of RMSE and MAPE.

**Conclusion**

Upon analyzing both Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) values, the Auto ARIMA model emerges as the optimal choice among the evaluated models. The MAPE for the Auto ARIMA model stands at 1.29%, which is notably lower than the MAPE values for both the ARIMA model (1.343%) and the Seasonal Naive model (4.31%). A lower MAPE suggests a more accurate prediction on the test set.

Furthermore, considering the RMSE values, the Auto ARIMA model exhibits a RMSE of 291.024, which is lower than the RMSE for the ARIMA model (309.199) and substantially lower than the Seasonal Naive model's RMSE (807.159). A lower RMSE indicates smaller prediction errors and better overall model performance.

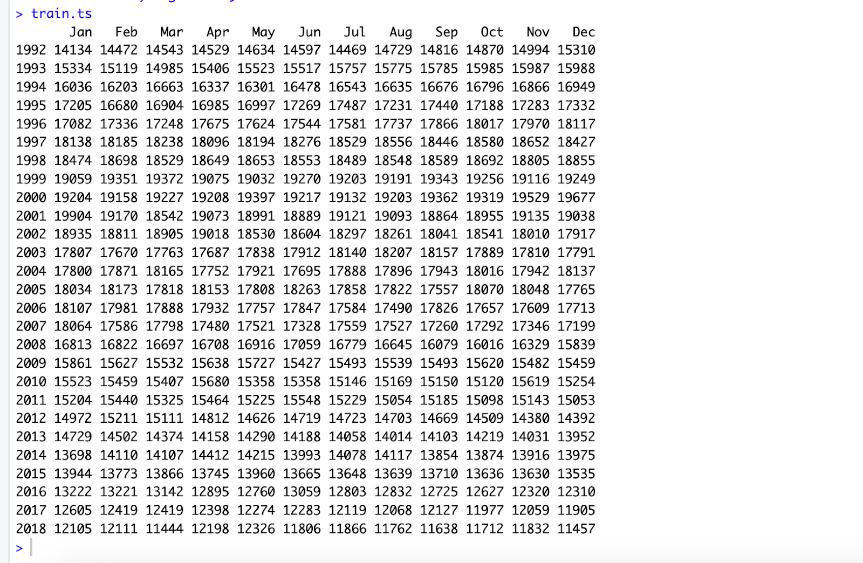
In summary, based on the combined assessment of MAPE and RMSE values, the ***Auto ARIMA*** model demonstrates superior predictive accuracy and is recommended as the most suitable model for this forecasting task compared to the other models in this analysis.

**APPENDICES**

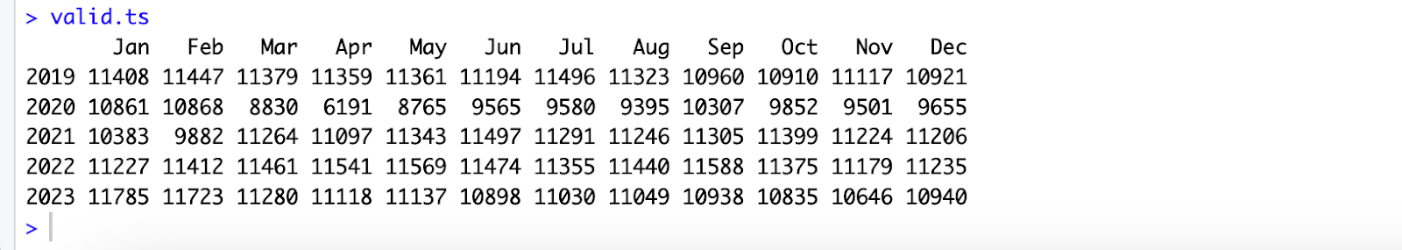
We took the dataset from following website:

<https://www.census.gov/econ/currentdata/datasets/>

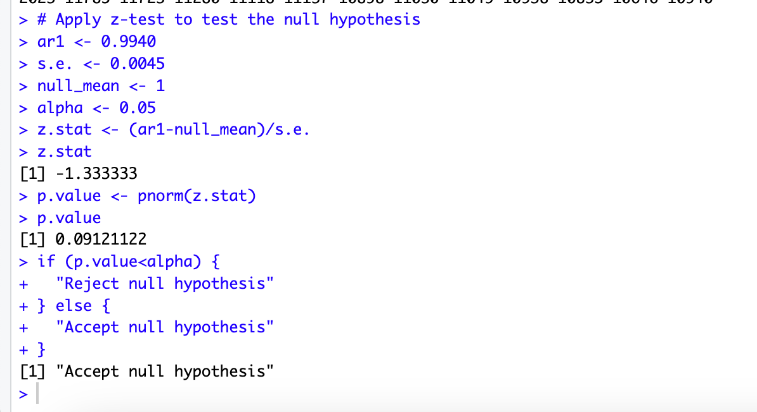
1. Training data



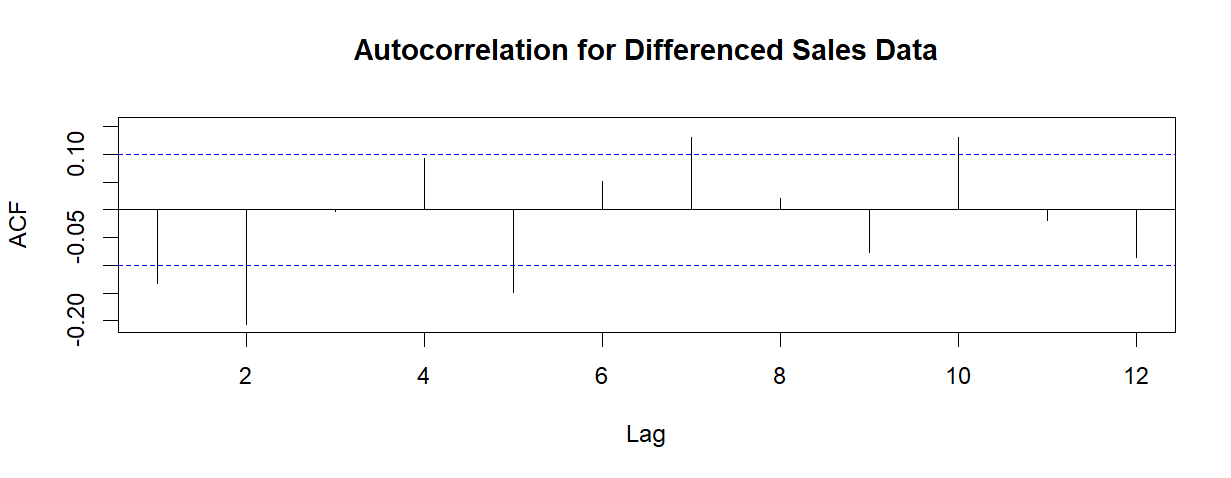
1. Validation data

****

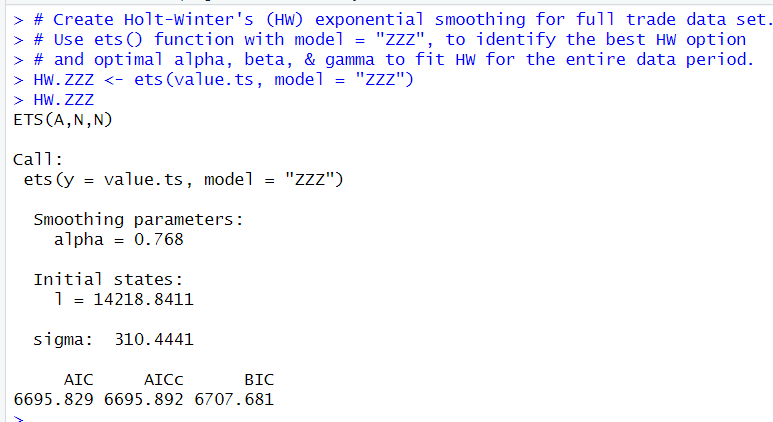
1. To check the dataset we perform 2 tests

****

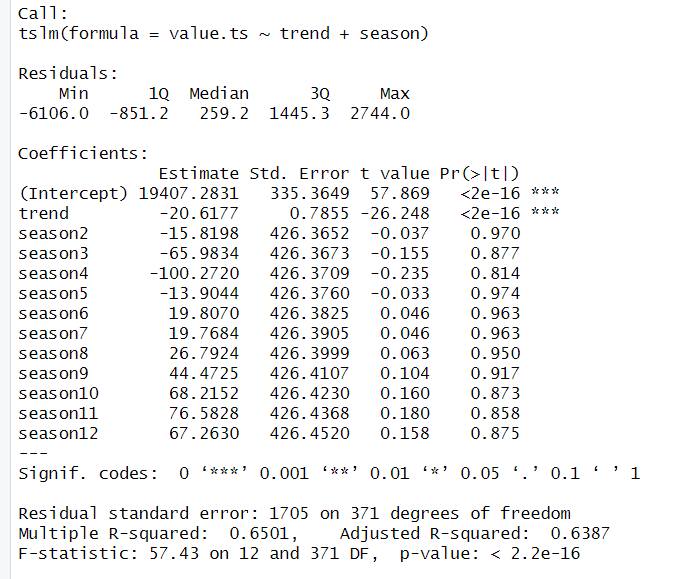
Another approach is using ACF plot for differenced series with different lags:In this we are get acf plot with significant autocorrelation.Which means we can predict the forecast using this dataset.

****

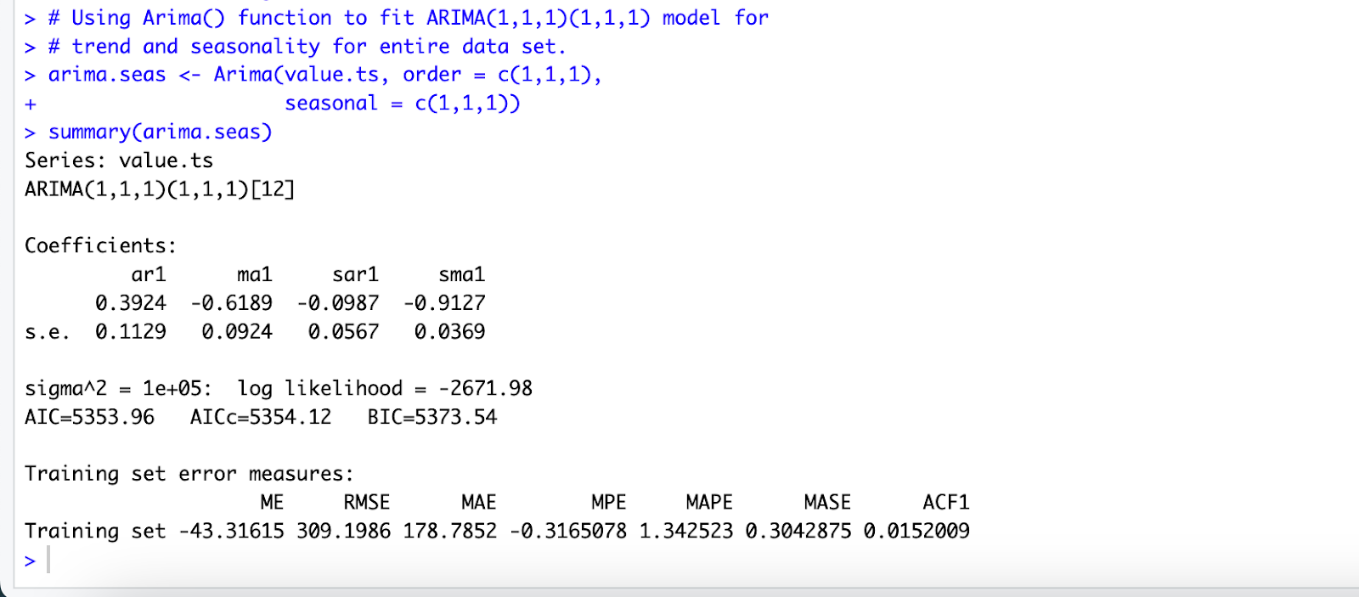
1. Holt – Winters Model



1. Regression model with Linear Trend and Seasonality on Entire Dataset



1. ARIMA Model for entire dataset



**Team Members**

|  |  |
| --- | --- |
| Veera Venkata Sai Kalyan Kaparaju | SS4785 |
| Sai Sarala Kavya Mudunuru | QP1501 |
| Arshad Shaik | MX9993 |
| Fayaz Aashiq Shaik | NU9413 |
| Mohammed Abdul Azeez Ansari | lr5894 |