**Prediction of Number of New single Family Houses Sold**

**Bala Surya Teja Nandamuri (wv7708)**

**Jani Pasha Mohammed (xw1139)**

**Leela Kali Manasa Mandadi (th7515)**

**Sadvilas Buddiga (jx8549)**

**Swathi Bommina (yr1610)**

**Executive Summary**

This project aims to predict yearly annual rate for new single-family houses sold for the upcoming year using data from the U.S. Census Bureau . The time series data from the year 1980-2022 is implemented. We have also found that that there is statistically significance all the lags from lag1 to lag 12 showing high auto correlation.

Several types of models were employed for the project, including regression-based models, advanced exponential smoothing models, and ARIMA (autoregressive integrated moving average model). Variations of regression and advanced exponential smoothing models were also developed to ensure accuracy. In cases where it was deemed appropriate, regression models were enhanced with a trailing moving average for residuals and an autoregressive model for residuals. The same enhancements were made for advanced exponential smoothing models. Model evaluation was based on two accuracy metrics: RMSE and MAPE.

Among the models tested, the best-performing model was the regression model auto. Arima model followed by linear trend and seasonality, enhanced with a autoregressive model for residuals.

**Introduction**

Time Series Analytics is an important area of study that enables businesses and individuals to predict future trends and make informed decisions. Real Estate Market is one of the key field in which we can use Time Series Analytics for where it is used to track trends and forecast future market conditions. In this report we will focusing on the annual rate for new single-family houses sold in the United States and how time series analytics can help us understand the trends in this market.

The annual rate for new single-family houses sold is a key metric and gauge the health of the housing industry, as it provides insight into the overall health of the housing market. This metric reflects the number of newly constructed single-family homes that are sold over the course of a year. This metric is closely watched by economists, policymakers, and investors, as it serves as an indicator of the current state of the housing market, the overall economic growth, and consumer confidence.

The value of this analysis extends beyond the real estate industry, as the housing market has a significant impact on the broader economy. The significance of this report lies in the fact that it provides valuable insights for various stakeholders in the housing market, including real estate developers, investors, policymakers, and homebuyers. A strong housing market can contribute to job growth, stimulate consumer spending, and drive economic expansion. Conversely, a weak housing market can have the opposite effect, leading to job losses, reduced consumer confidence, and economic downturns. By analyzing the data using time series analytics, we can better understand the trends and patterns in the housing market, which can help stakeholders make informed decisions.

We aim to provide an in-depth analysis for the annual rate of new single-family houses sold in the United States using time series analytics. We will explore the trends and patterns in the data over time, identify any seasonality or cyclicality, and develop forecasting models to predict future trends.

In conclusion, the annual rate for new single-family houses sold in the United States is a vital metric that provides insights into the health of the US economy and the housing market. Time series analytics can be used to analyze the data and predict future trends, which is of great value to various stakeholders in the housing market. Through this report, we aim to provide insights into this important metric and its trends using various statistical and machine learning models.

**Steps of Forecasting**

**Step 1: Defining the Goal**

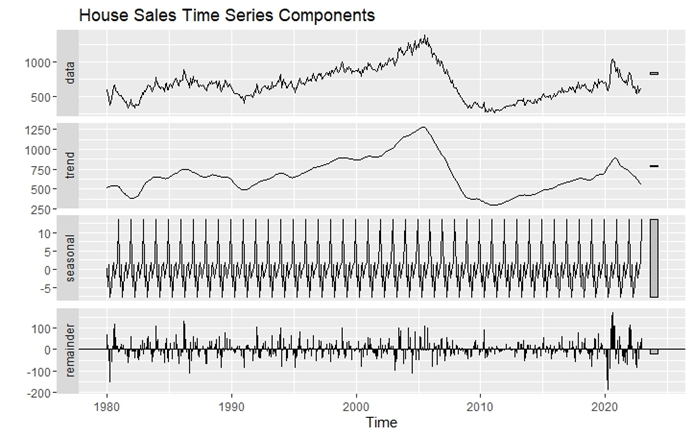
The objective of this project is to develop a predictive model to forecast the Annual Rate for New Single-family Houses sold in the US market for the upcoming year. The model should consider both the trend and seasonality components of historical data and provide accurate forecasts. Since the data for a year will be available for next year, the model should be reevaluated annually to incorporate new data and improve accuracy. The forecasts will be used to monitor the US housing market, specifically the number of new single-family houses sold per year. The project was completed using the R language.

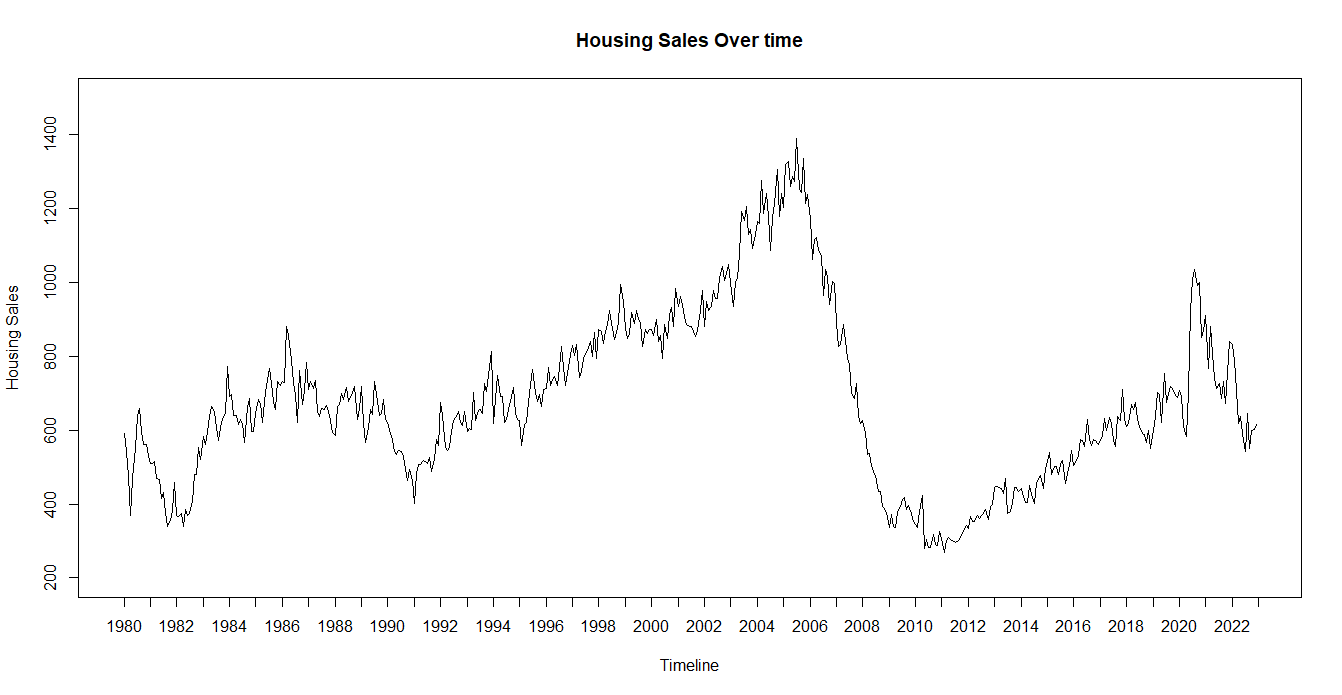
**Step 2: Gathering Data**

The data we gathered is from the source of U.S. Census Bureau which gathers data on new single-family houses sold in the United States through their New Home Sales program. The report is focused on the time series analysis on the dataset provided by U.S. Census Bureau recording the yearly house sales of new single-family houses. The time period for the dataset ranges from 1980 to 2022.

**Step-3 Exploring and Visualizing the Time Series.**

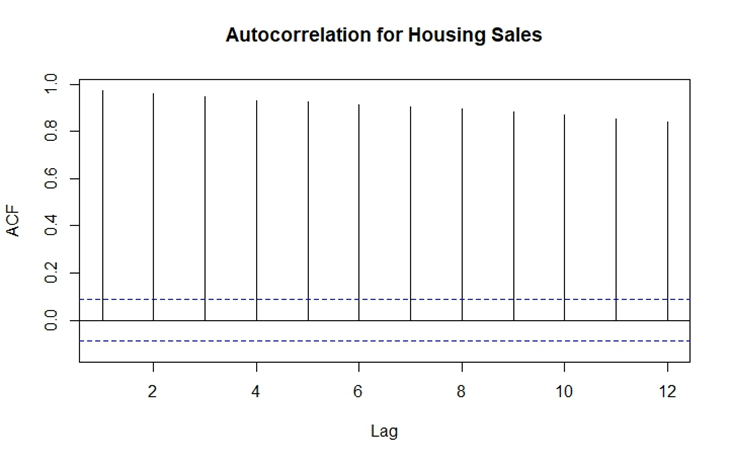
The plot below shows us the various components of the time series analytics, The x-axis represents time, and the y-axis represents different component. The seasonal component shows the repeating patterns in the data, the trend component shows the long-term changes in the data, and the residual component represents the random fluctuations or the noise in the data that cannot be defined or explained by the seasonal or trend components.





The above plot shows the time series of New Single Family Housing Sales for the year spanning from 1980 to 2022. The x-axis represents the timeline and the y-axis represents the housing sales over time. The plot also shows a clear seasonal pattern, where the housing sales peak during certain months of the year and decline during other months.Overall, the plot provides a visual representation of the fluctuations in housing sales over time and highlights the presence of seasonality and trend.

The below autocorrelation plot shows the correlation between the values of the time series at different lags. The x-axis represents the lags, and the y-axis represents the correlation values. We can observe from the graph that the there is correlation which is statistically significant with all the lags from lag 1 to lag 12. All the lags shows that there is positive correlation. We can say that the data is not just comprised of the level components.

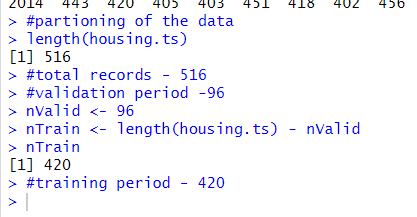


**Step 4: Data Preprocessing**

The original data is gathered from the U.S. Census Bureau contains the Annual Rate for New Single-family Houses Sold: United States. Every unnecessary rows that showed the irrelevant data is removed, except for housing sales . By doing so, only the relevant data will be considered for the analysis without any noise. Next, time series dataset is created, with data ranging from the span of 1980-2022. There are total of 516 observations in the time series that is created.

**Step 5: Partition of the Time Series.**

Usually in the time series the partition is done in random. The general assumption is that training partition consists of 70-85% of data and the validation consists of the 15-30% of the data. In our data set the validation partition consist of the 96 records and the training partition consists the total records of 420.



**Step 6 & 7: Apply Forecasting & Comparing Performance**

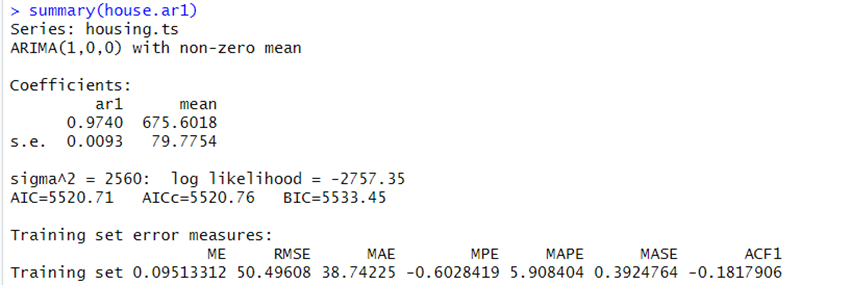
**Testing Of The Hypothesis**

To test the hypothesis we will be needing the values of ar1 and se which is standard error which can be obtained arima function and creating ar(1) model.

**AR(1) Model:**

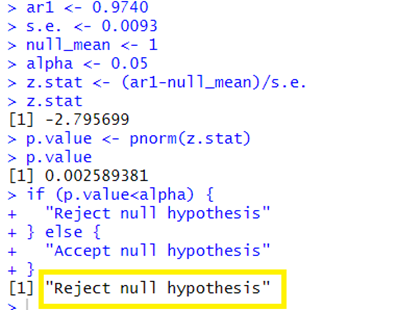
The ARIMA (1,0,0) model was fitted to the housing time series data. The model has one autoregressive term (AR1) with a coefficient of 0.9740, and a non-zero mean of 675.6018. The standard error of the AR1 coefficient is 0.0093 and the standard error of the mean is 79.7754.

The model's log likelihood is -2757.35 and the AIC (Akaike Information Criteria) is 5520.71, the AICc (Corrected AIC) is 5520.76, and the BIC (Bayesian Information Criteria) is 5533.45. These values can be used to compare different models and choose the one with the lowest AIC or BIC.



The performance of the model was evaluated using the mean error (ME), root mean squared error (RMSE), mean absolute error (MAE), mean percentage error (MPE), mean absolute percentage error (MAPE), mean absolute scaled error (MASE), and the first autocorrelation (ACF1) on the training set. The training set had ME of 0.09513312, RMSE of 50.49608, MAE of 38.74225, MPE of -0.6028419, MAPE of 5.908404, MASE of 0.3924764, and ACF1 of -0.1817906.

**Hypothesis Test**

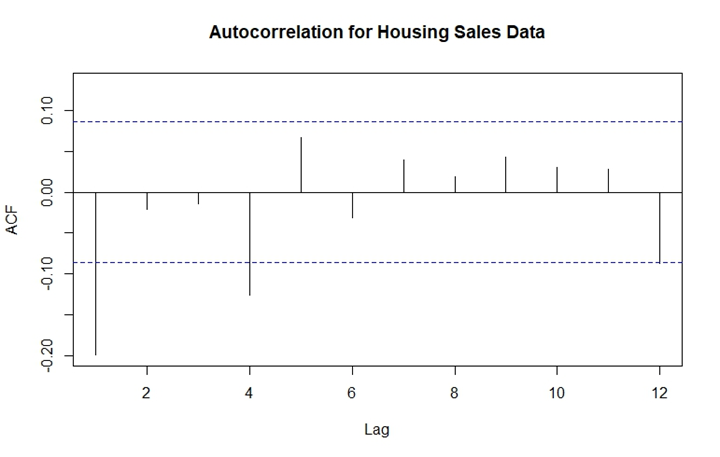


A null hypothesis test is performed to determine if the coefficient of the AR(1) is equal to zero (null\_mean).

In this case, the calculated p-value is 0.0026, which is less than the significance level (alpha) of 0.05. This means that there is less than a 5% chance of observing the coefficient as extreme or more extreme than the one calculated, if the null hypothesis is true.

Based on this, the conclusion of the hypothesis test is to reject the null hypothesis, and accept the alternative hypothesis that the coefficient of the AR(1) term is not equal to zero.

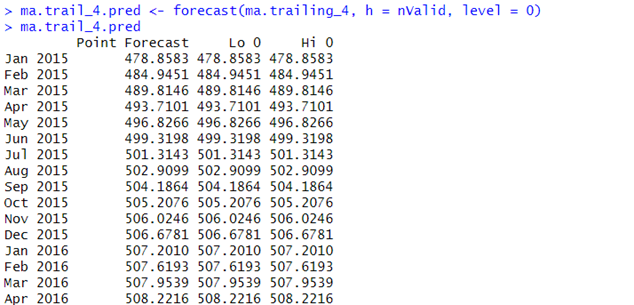
The first-differencing of the housing time series data with a lag of 1 is developed. The first-differenced time series can be interpreted as the change in the housing prices from one time period to the next. By performing first differencing, the time series data is transformed into a stationary series, which can be used to model the time series more effectively.



The ACF plot is created for the differenced Housing Sales Data, with a maximum lag of 12. The plot shows the correlation between the series and its lags, with the values ranging from -1 to 1. As per the plot, lag 1 and 4 are indicating a strong negative correlation between the residuals. The value for lag 12 indicates a weak negative autocorrelation. the rest of lags are not statistically significant and are close to being random.

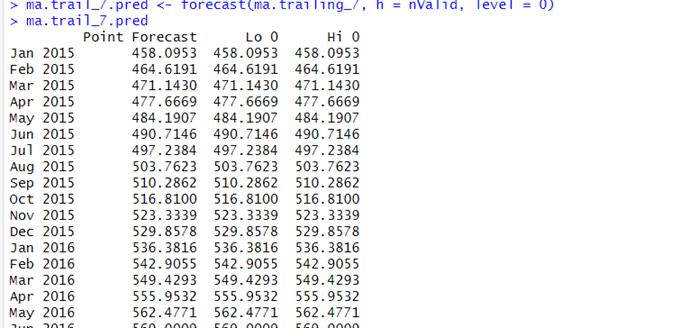
**Forecast Of MA Trailing Window Of 4**

The forecast consists of the point forecast, lower bound (Lo 0) and upper bound (Hi 0) of the forecast for each time step in the validation period. The forecast shows the time-series forecast from January 2015 to December 2022. The point forecast for each time point is the same and is equal to 509.2924. This indicates that the model is predicting a constant value for the time-series for all future time points.



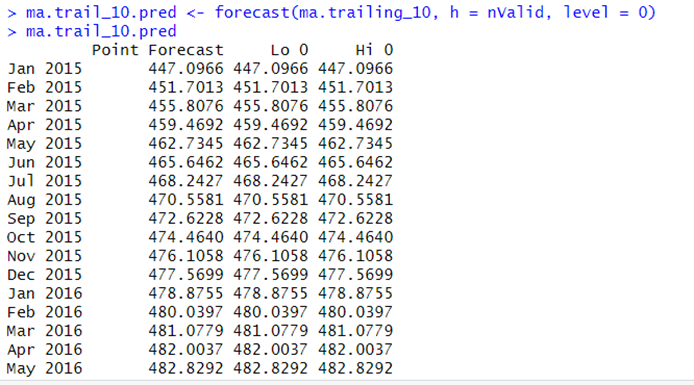
**Forecast Of MA Trailing Window Of 7**

This table shows the point forecast of a time series model based on moving average. The moving average is calculated using 7 time steps in the trailing period (ma.trailing\_7). The point forecast shows the predicted value of the time series for each time step in the validation period and the value changes for each time step. The predicted values seem to be increasing gradually and linearly.



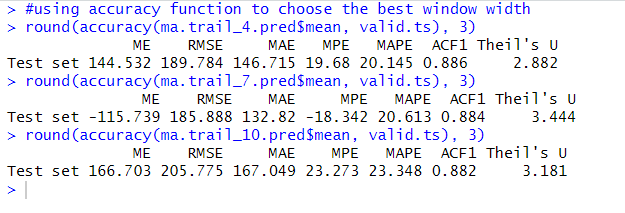
**Forecast Of MA Trailing Window Of 10**

This table shows the point forecast of a time series model based on a 10-step moving average (MA) method. The "Point Forecast" column in the table shows the predicted value of the time series for each time step in the validation period. Each value in the column is a constant value of around 489.62, which is the average of the previous 10 observations.



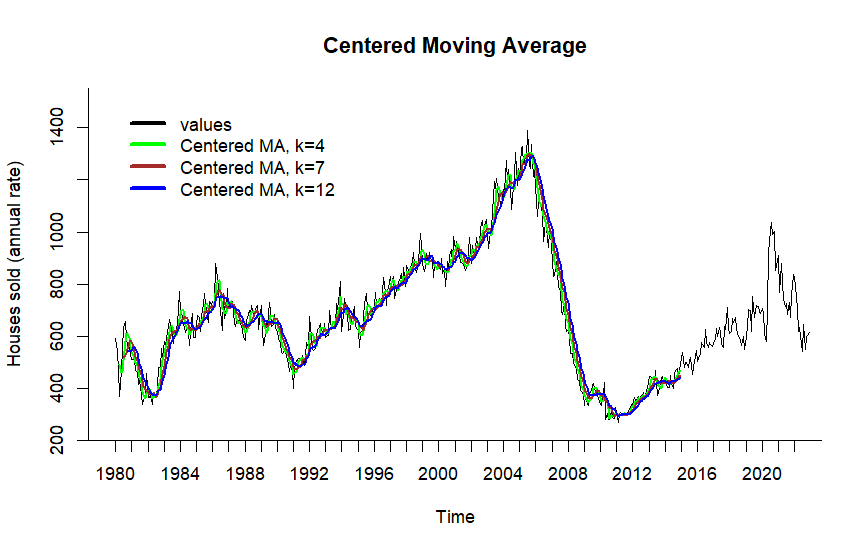
**Comparing the accuracy of MA 4, 7, 10:**

The values of accuracy measures for the three models indicate that the best model is the one with a window width of 4. The measures of MAPE (Mean Absolute Percentage Error) and Theil’S U Statitic are all lower for the model with a window width of 4 compared to the other two models with window widths of 7 and 10. Lower values of these accuracy measures indicate that the model is more accurate in predicting the time series. Additionally, the value of Theil's U, which measures the accuracy of the predictions relative to the actual values, is also lower for the model with a window width of 4 compared to the other two models. Thus, based on these values, the best model among the three is the one with a window width of 4.



This plot shows the original time series data for housing sales (annual rate) over time, from 1980 to 2022. The data is represented by a black line. Additionally, the plot shows three centered moving averages (CMA) for the time series, with window widths of 4, 7, and 10. These moving averages are shown as green, brown, and blue lines, respectively.

The fluctuations in the graph can be seen as having seasonality and a general upward trend over time. The seasonal fluctuations are apparent as the regular repeating patterns that occur over the course of a year. The upward trend can be seen as the overall increase in the number of houses sold over time. Additionally, there appears to be some irregular fluctuations in the data that don't seem to have a clear pattern, which could be due to various other factors.



**Regression Models**

Regression-based models are used for time series analysis, which vary based on the time series plot. This model is simple to use and provides reliable results by considering both trend and seasonality. They can be further enhanced by adding the Autoregressive components and a trailing moving average for the residuals. But before using the model on the entire dataset, it needs to evaluated by using training and validation partitions.

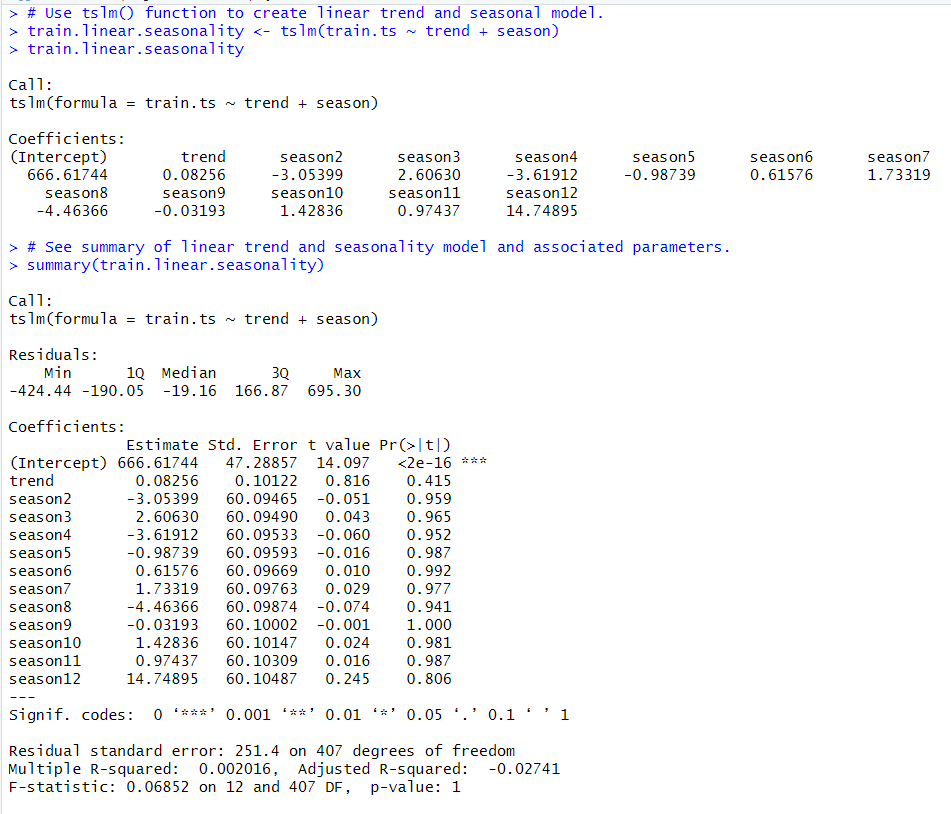
**Regression Model With Linear Trend And Seasonality**

The model was fit to the training partition of the housing sales time series data and includes a linear trend and seasonality as predictors.

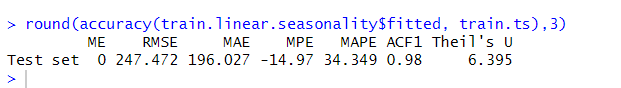
Based on the analysis, it appears that the model is not a good fit for the data. The model has a low adjusted R-squared value of -0.02741, which indicates that the model only explains a very small portion of the variability in the data. The F-statistic has a p-value of 1, indicating that the overall fit of the model is not significantly different from a random walk model.

Additionally, the coefficients for the trend and seasonal components of the model have low t-values and high p-values, indicating that these terms are not significant predictors of the time series.

Overall, based on the low adjusted R-squared value and insignificant coefficients, it can be concluded that the linear trend and seasonal model is not a good fit for the time series data.



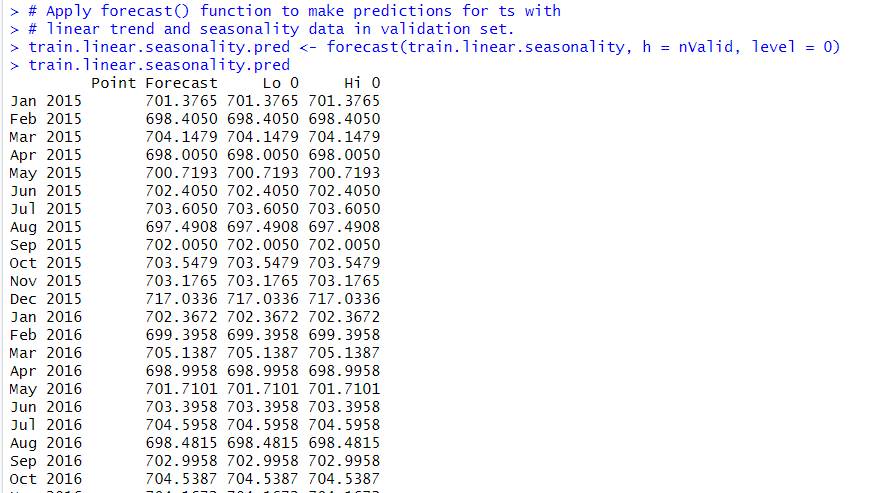
The accuracy measures for the model show that the mean error (ME) is 0, which means that on average the model is not over- or under-predicting the actual values. However, the RMSE value of 247.472 and MAE value of 196.027 indicate that the model has high prediction errors. The MPE and MAPE values of -14.97 and 34.349, respectively, suggest that the model has a significant bias towards under-predicting the actual values. The ACF1 value of 0.98 and Theil's U value of 6.395 further indicate that the model is not a good fit for the time series data.



The forecast function is used to make predictions for the time series data in the validation set with a linear trend and seasonality model. The forecast was made for a horizon of nValid time steps. The point forecast values are shown in the output table, with the corresponding lower and upper bounds of the 0 level forecast interval.

The point forecast values seem to be uniform and consistent, with a value of around 700-703 for most of the months. However, it is important to note that the lower and upper bounds of the forecast interval are not provided, which limits the assessment of the predictive uncertainty of the forecast.

Overall, it is difficult to determine the quality of the forecast based solely on the point forecast values. Further analysis and comparison with actual values would be necessary to evaluate the accuracy and usefulness of the forecast.

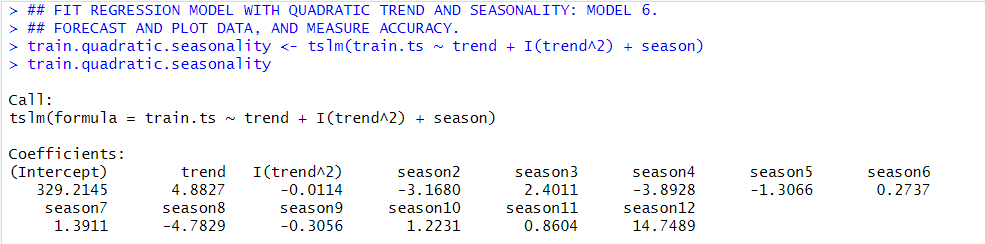


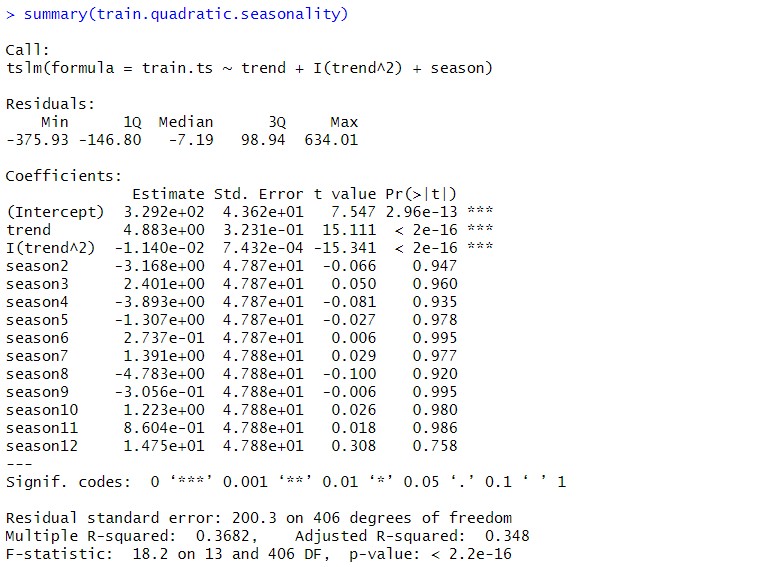
**Regression Model with Quadratic Trend and Seasonality**

As per the summary of the regression model with quadratic trend and seasonality developed, coefficients of the regression are presented in the table, with estimates, standard errors, t-values, and p-values. All of the variables representing the quadratic trend and the seasonal factors have significant coefficients, with a p-value less than 0.05.

The residuals of the model are summarized, with minimum, first quartile, median, third quartile, and maximum values. The residual standard error of the model is 200.3 with 406 degrees of freedom.

The goodness-of-fit of the model is measured by the R-squared value, which is 0.3682, and the adjusted R-squared value, which is 0.348. The F-statistic of the model is 18.2, with a p-value less than 2.2e-16, which indicates that the model as a whole is significant.

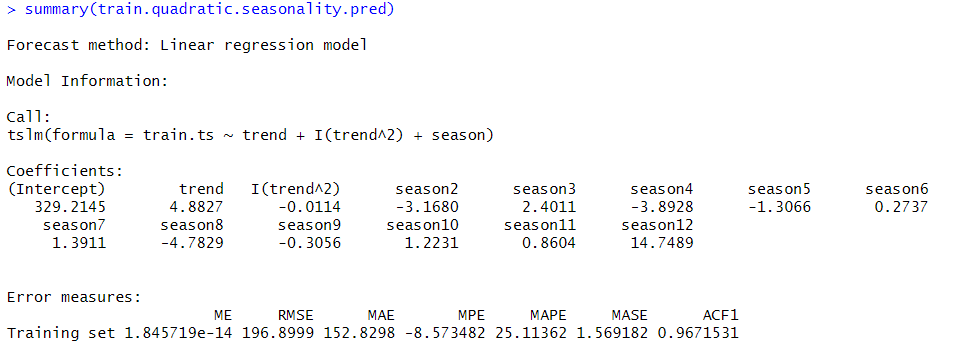




The accuracy of the model shows that the model has a relatively low error, with the RMSE and MAPE around 196 and 153, respectively. The negative MPE value indicates that the forecasts are biased, overestimating the actual values. The MAPE of 25.114% indicates that the average error is relatively high compared to the actual values. The ACF1 value close to 1 implies that there is still some autocorrelation present in the residuals. The Theil's U statistic of 4.466 is relatively low, indicating that the model is not too far off from the actual values.



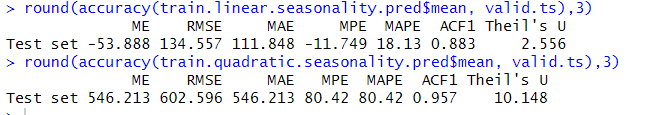
The results of the evaluation (forecast of the quadratic trend and seasonality) indicate that the model has a ME of 1.845719e-14, a RMSE of 196.8999, a MAE of 152.8298, a MPE of -8.573482, a MAPE of 25.11362, a MASE of 1.569182, and an ACF1 of 0.9671531. These results suggest that the model is able to fit the training data well, but has some limitations in terms of its ability to accurately forecast the test data.



As per the accuracy measures, it can be summarized that the accuracy of the linear seasonal forecast on the validation set has a mean error of -53.888, root mean squared error of 134.557, mean absolute error of 111.848, mean percentage error of -11.749, mean absolute percentage error of 18.13, first order autocorrelation of 0.883 and Theil's U of 2.556.

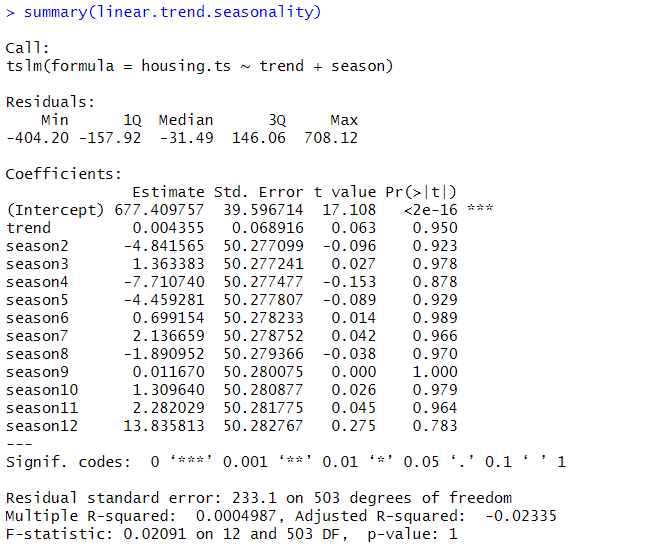
On the other hand, the accuracy of the quadratic seasonal forecast on the validation set has a mean error of 546.213, root mean squared error of 602.596, mean absolute error of 546.213, mean percentage error of 80.42, mean absolute percentage error of 80.42, first order autocorrelation of 0.957 and Theil's U of 10.148.

Based on the accuracy metrics, the linear seasonality model seems to perform better than the quadratic seasonality model. The linear seasonality model has a lower RMSE, MAE, and MPE, and a lower Theil's U, which are all metrics used to evaluate the accuracy of a forecast. The linear seasonality model also has a lower MAPE, which measures the average percentage error in the forecast.

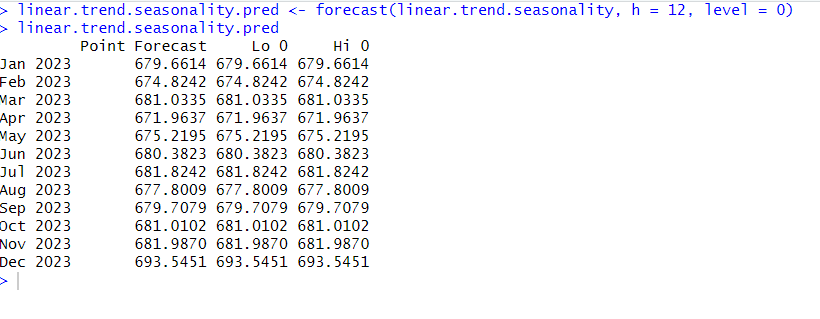


**Linear trend and seasonality on entire dataset:**

In this analysis, a time series linear regression model with a linear trend and seasonality was fit to the entire housing time series data. The model's summary showed that the coefficients for the linear trend and the seasonality components were not significant, as indicated by their high p-values. The multiple R-squared value was low (0.0004987), indicating a poor fit of the model to the data. The residual standard error was high (233.1), and the F-statistic had a high p-value (1), which suggests that the model does not provide a good explanation for the variation in the housing time series data.

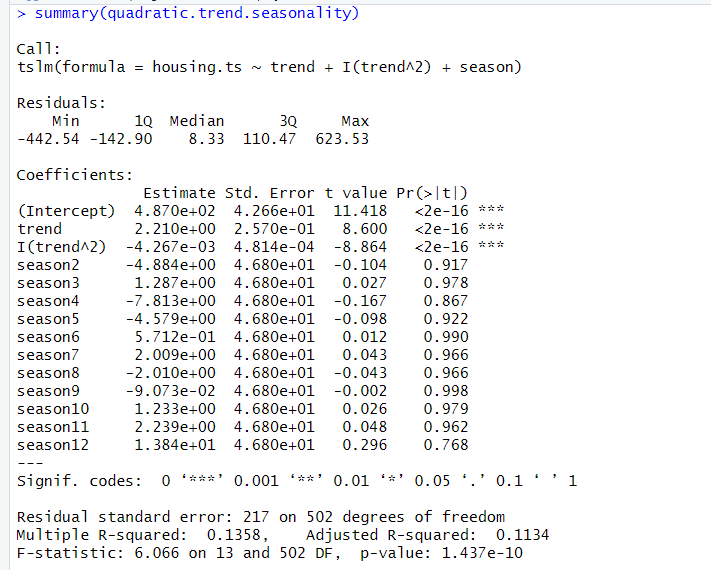


The forecast of the linear trend-seasonality model predicts the values of the housing prices for the next 12 months. According to the forecast, the housing prices are expected to have a steady increase from January 2023 to December 2023, with the highest forecasted value in December 2023 at 693.5451. The "Lo 0" and "Hi 0" columns represent the lower and upper bounds of the 95% prediction interval, which are both equal to the point forecast. This suggests that there is a high level of confidence in the forecast.

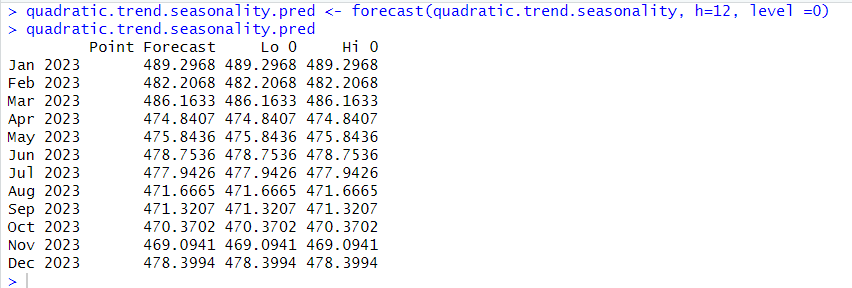


**Quadratic trend and seasonality on entire dataset:**

In this analysis, a time series regression model with a quadratic trend and seasonality was fit to the entire housing time series data. As per the summary, based on R-squared value of 0.1358 and the adjusted R-squared value of 0.1134, we can see that this quadratic trend and seasonality model provides a better fit to the data compared to the linear trend and seasonality model. Additionally, the F-statistic of 6.066 with a p-value of 1.437e-10 indicates that the model is significant and provides a good fit to the data.

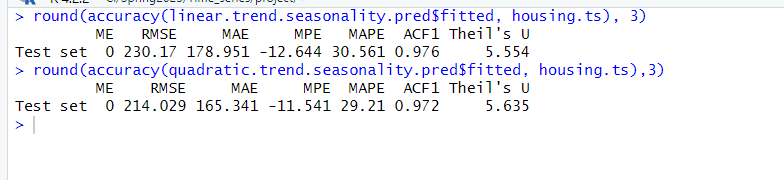


The forecast from the quadratic trend and seasonality model predicts that the median housing prices will remain relatively stable in the next 12 months. There will be a small increase in the median housing prices between January and December 2023, but the increase will be relatively small compared to the overall level of median housing prices.



**Accuracy comparison of Linear Trend and seasonality and Quadratic trend and seasonality on entire dataset:**

Both the linear trend-seasonality model and the quadratic trend-seasonality model were used to predict the housing prices. The linear trend-seasonality model had an RMSE of 230.17 and the quadratic trend-seasonality model had an RMSE of 214.03. The RMSE is a measure of the difference between the actual and predicted values and lower values indicate better performance. Based on the RMSE values, the quadratic trend-seasonality model performed better than the linear trend-seasonality model. Additionally, the quadratic trend-seasonality model also had a lower MAE and Theil's U value compared to the linear trend-seasonality model. These results indicate that the quadratic trend-seasonality model is the best model for predicting housing prices.

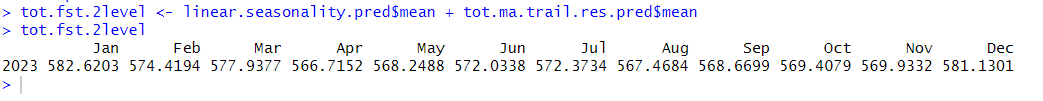


In above data we can see that the modes having high Theil’U statistic value which are above 1 which shows that models are non reliable for the forecast. We can enhance it by adding the Moving average and autoregressive residuals.

**Two level Forecast**

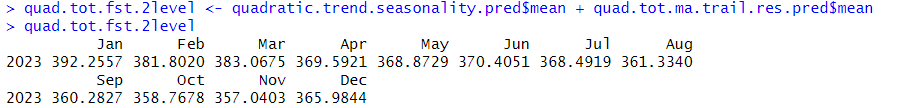
Trailing Moving Average for Regression Residuals

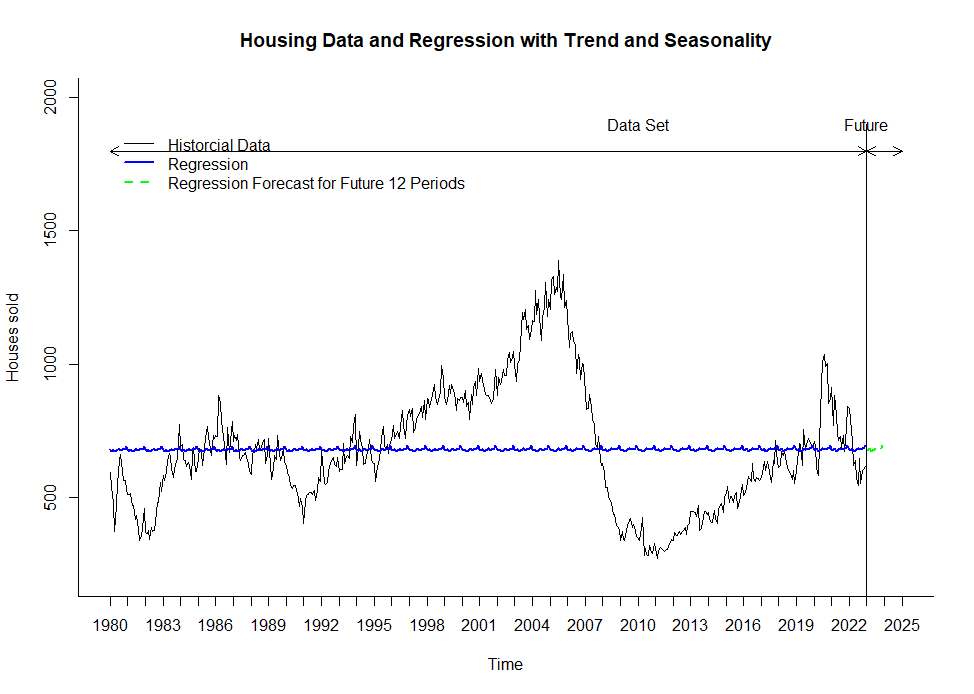
To improve the regression model with quadratic trend and seasonality, a trailing moving average was used to forecast residuals from the model. These components were then combined to create a two-level model. The window width is chosen as 4.



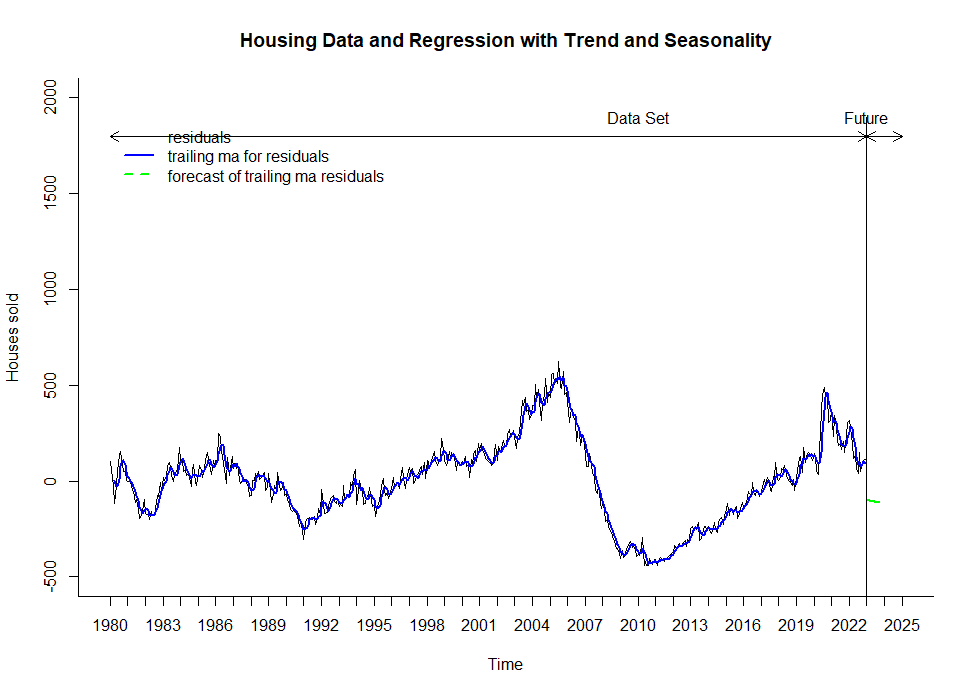
The above values are forecast values of entire data set for linear trend and seasonality with trailing ma with window width of 4

The below is forecast values of quadratic trend and seasonality with trailing ma with window width of 4 for entire data set.



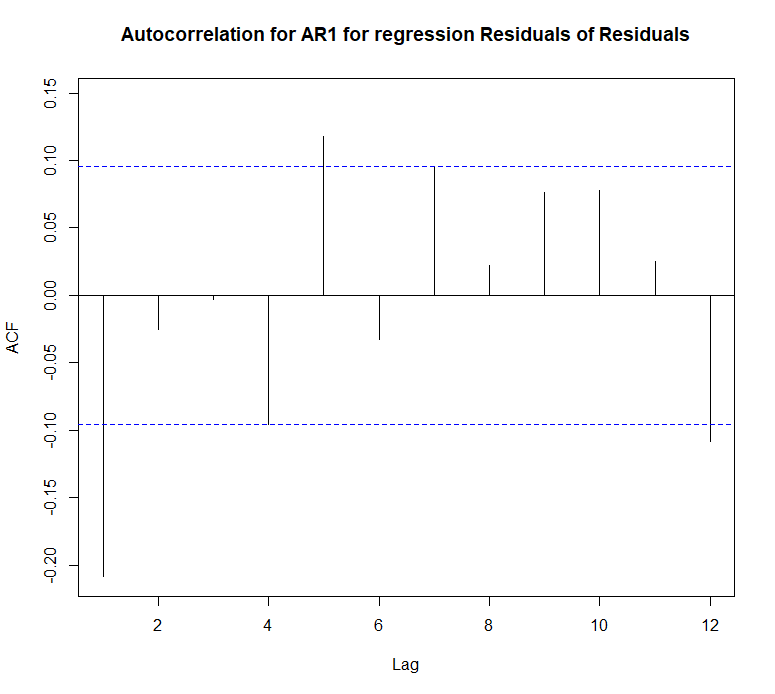
**Plot for the Linear Trend and Seasonality + Trailing Moving Average for Entire Data Set.**

**Plot for the Quadratic Trend and Seasonality + Trailing Moving Average for Entire Data Set.**

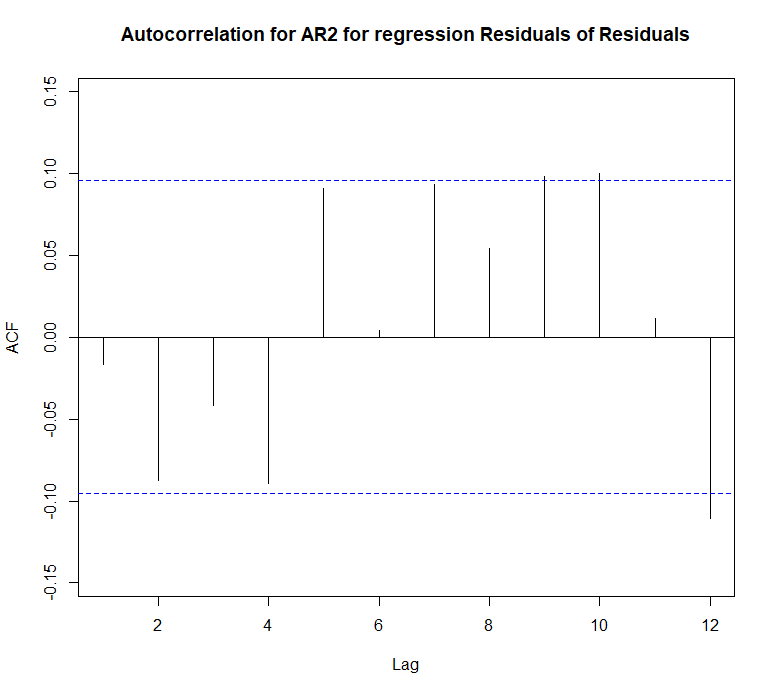


**Autoregressive Models for Regression Residuals**

The below is graph for AR 1 model for regression residuals of residuals.

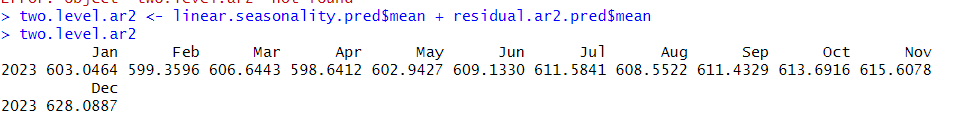


**The below is graph for AR2 model for regression residuals of residuals.**

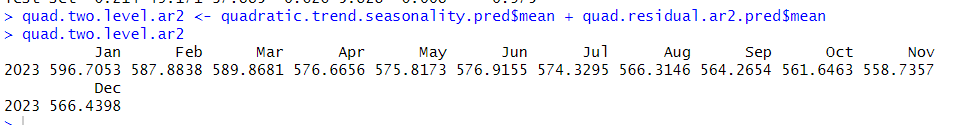


From the both the above graphs which shows the auto correlation of Residuals of residuals the AR2 model shows the better results in properly handling the dependencies between successive periods. Also the statistically insiginfincance of residuals of residuals i.e the leftovers is better shown in the AR2 which makes it a better choice.

**Two level Model for Linear Trend and Seasonality with AR(2) Model.**

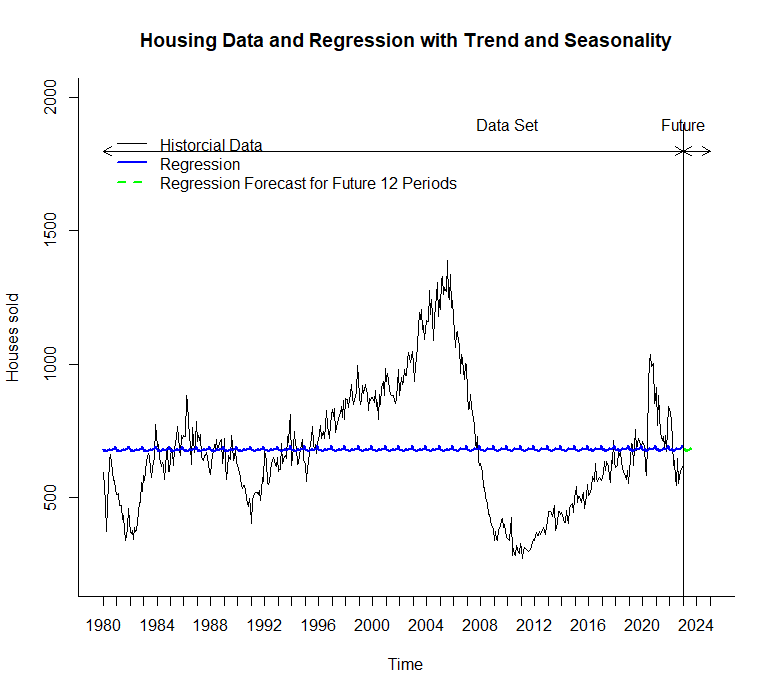


**Two level Model for Quadratic Trend and Seasonality with AR(2) Model.**

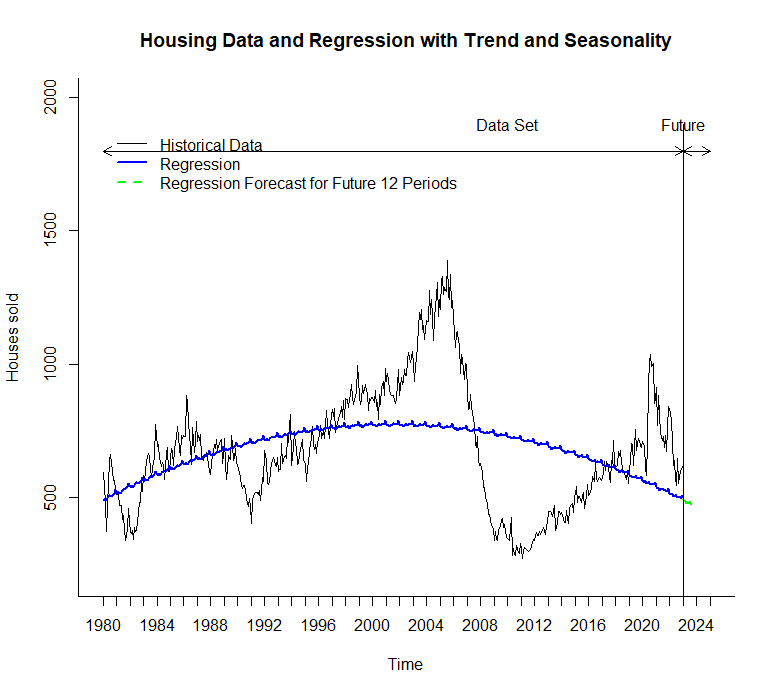


**Plots of Two Model with AR2 Model**

**Linear Trend and Seasnality for Entire Data Set**



**Quadratic Trend and Seasonalitu for Entire Data Set**



**Accuracy Performance of Regression Models**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | RMSE | MAPE | Theil’S U |
| Regression Model Linear Trend and Seasonality | 230.17 | 30.561 | 5.54 |
| Regression Model Quadratic Trend and Seasonality | 214.029 | 29.21 | 5.635 |
| **Regression with Linear Trend and Seasonality with AR(2) Model(TWO Level Model with AR)** | **49.197** | **5.813** | **0.973** |
| **Regression with Quadratic Trend and Seasonality with AR(2)Model(TWO Level Model with AR)** | **49.171** | **5.828** | **0.975** |
| Regression with Linear Trend and Seasonality with Trailing Moving Average(TWO Level Model with Trailing MA) | 55.411 | 6.451 | 1.086 |
| Regression with Quadratic Trend and Seasonality with Trailing Moving Average(TWO Level Model with Trailing MA | 100.839 | 12.54 | 2.128 |

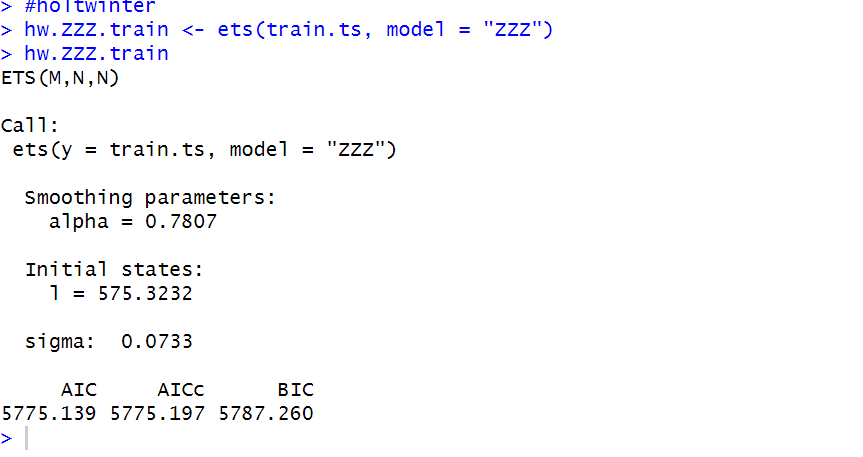
In above table we can see that Two-Level models with linear trend and seasonality with ar2and quadratic trend and seasonality with ar2 are the top two models. They both are quite similar in forecasting reliability. The RMSE value of Two Level Model with AR2 for Linear trend and Seasonality has the best MAPE with 5.813, Theil’U Statistic with 0.973, but Two-level mode with AR2 model for quadratic trend and seasonality has got better RMSE which is 49.171.

Since the MAPE and Theil’s U statistic is better for the Two-Level Model with AR2 for Regression Model with Linear Trend and Seasonality is the better model for regression model

**Holts Winter Model**

**Two level Forecast using Holt’s Winter’s Model and AR Model**

**Summary of the Holt’s winter’s model is as shown below:**



**The model’s optimal parameters for training data are (M,N,N)**

**The first M stands for multiplicative error.**

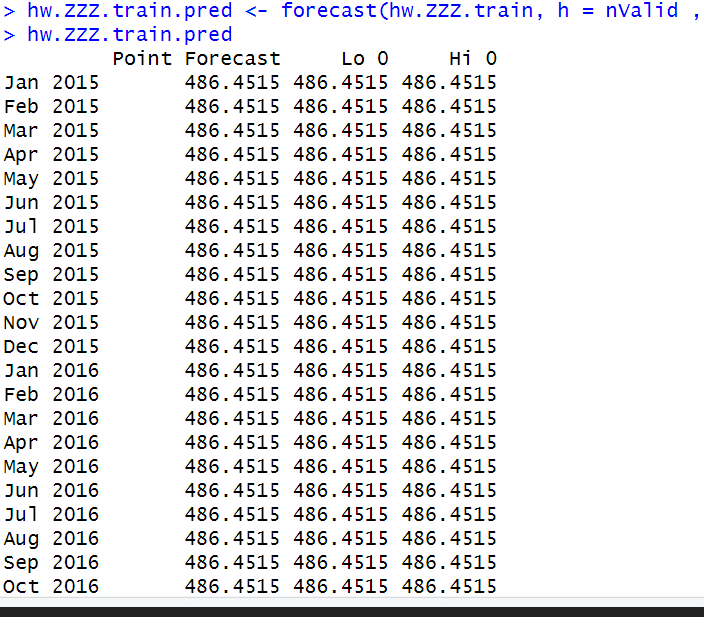
**N stands for no trend and**

**N stands for no seasonality.**

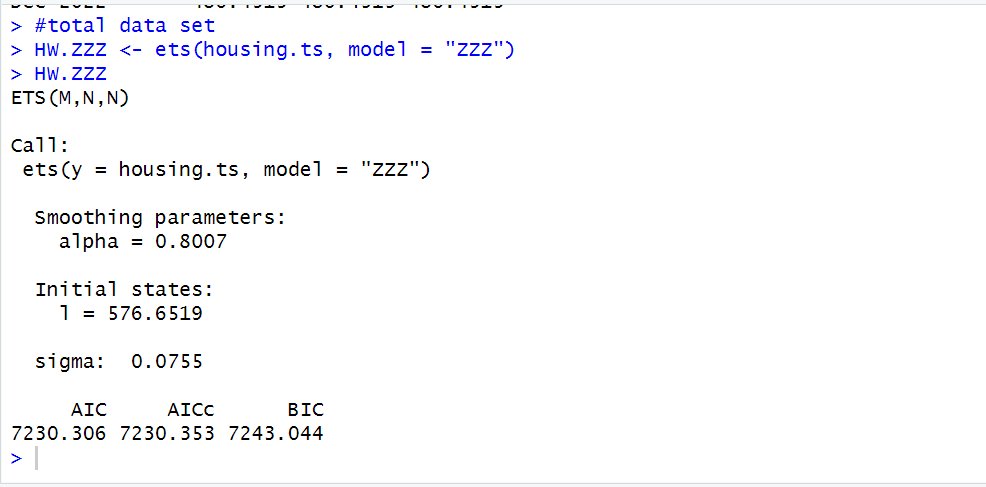
Smoothing parameter alpha is 575.3232.

And the model equation for the given model will be

**The predictions of the validation data are as follows:**



**The Model summary of the Holt’s winter’s model developed on the entire data set is as following:**



**The optimal parameters for the entire data set is also the same as that of the training data.**

**The model is ETS(M,N,N):**

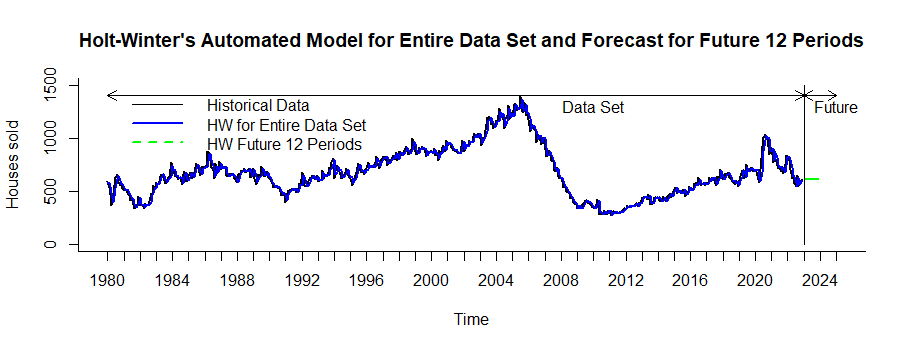
**Where M -multiplicative error,**

**N-no trend**

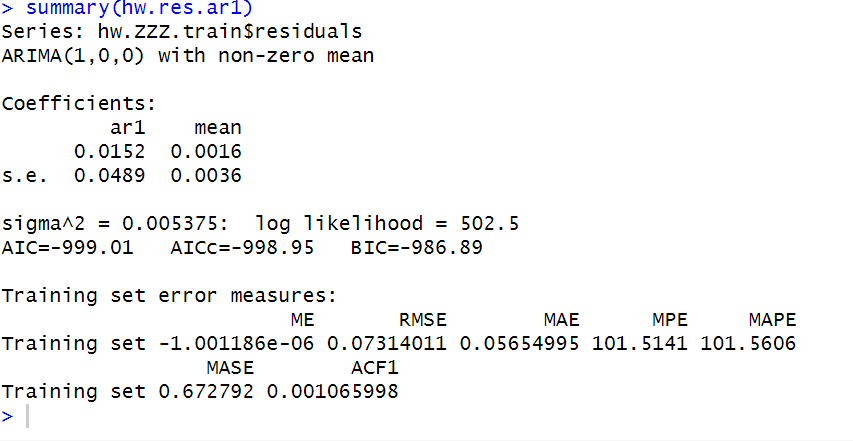
**N- no seasonality.**

**The Smoothing parameter of the model is alpha is 0.8007.**

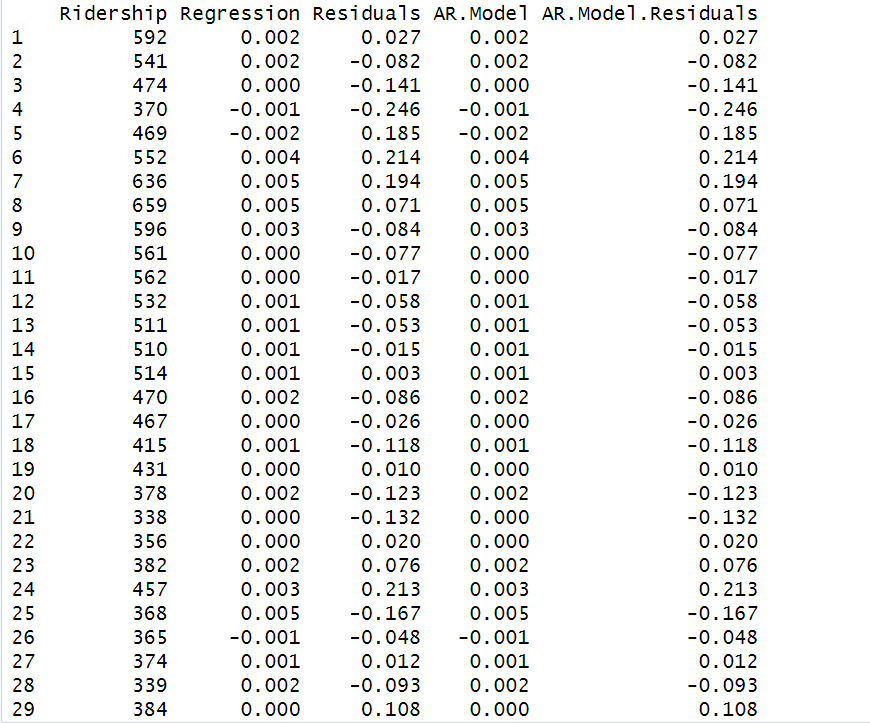
**The graph of the entire data with holt’s winter’s model is as following:**



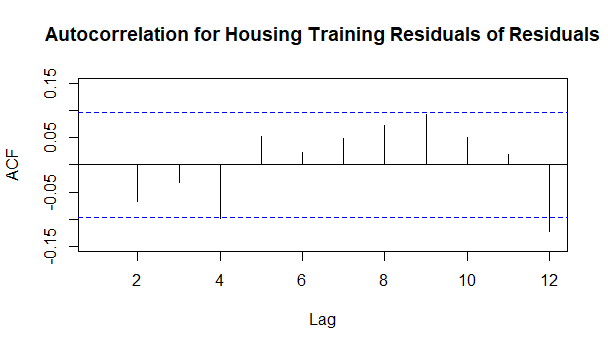
**The Arima model for the residuals of the Holt’s winter’s model are as following:**



The table with given data, Holt's winter’s model, residuals of HW’s model, AR model and AR model residuals.

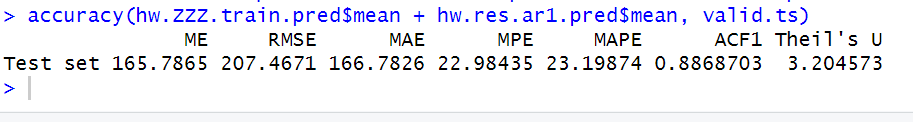


Following is the Autocorrelation plot of the training residuals of residuals

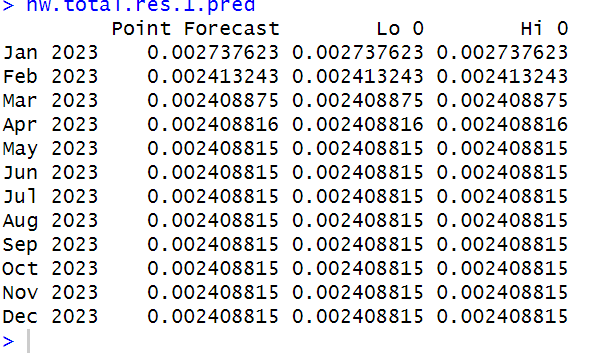


In the above plot the correlations at all the lags are in the blue lines except at lag 12, At lag12 also the correlation is very little more than the threshold which says that it is weak and we can consider the residuals as random walk and they are not predictable.

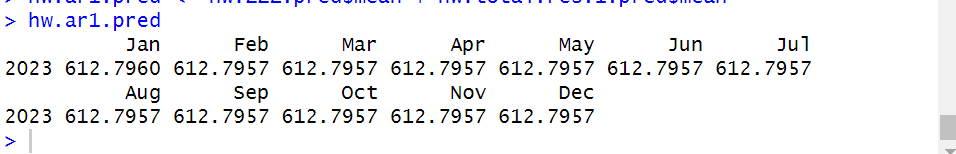
The Accuracy measures of the two level HW model are:



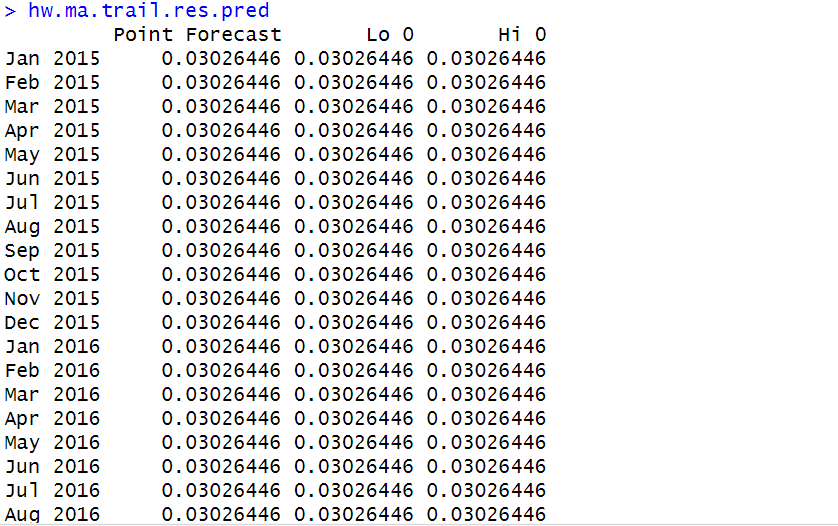
The Predictions of the residuals of the HW model using AR model for entire dataset is as following:



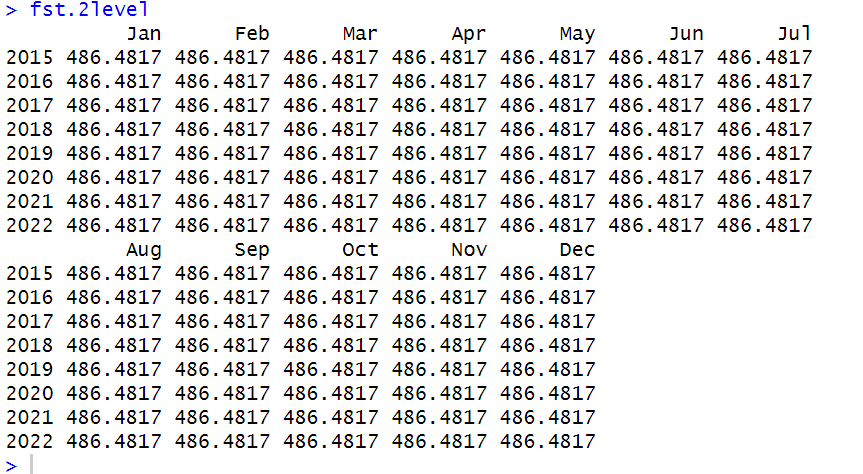
The prediction of the future with two level model(HW model and AR model is as shown below):

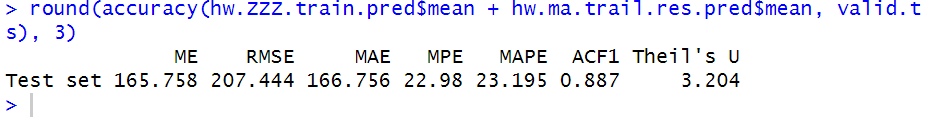


**The prediction of validation period using HW model and trailing MA with window width 4 is as following:**

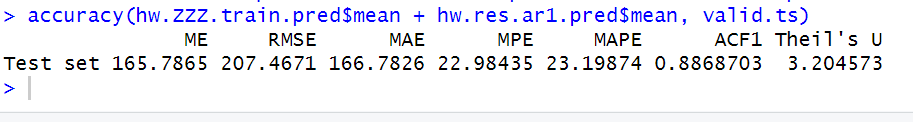


The two level forecast on the validation period using HW model and trailing MA is as follows:



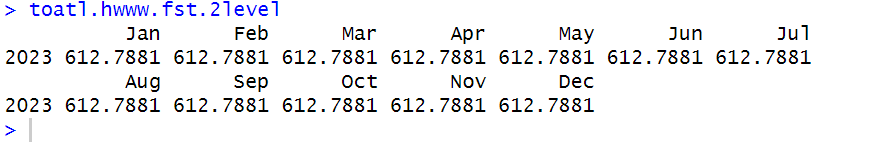
**The accuracy metrics of the two level forecast using Trailing ma and HW model is** 

The Accuracy measures of the two level HW model and AR model are:

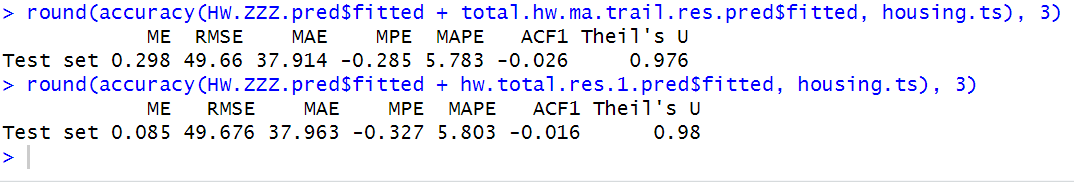


Both models are performing equally well as the RMSE, MAPE and Theil’s values are equal for both the models.

**The two level forecast for the entire data set using HW’s model and trailing ma is:**

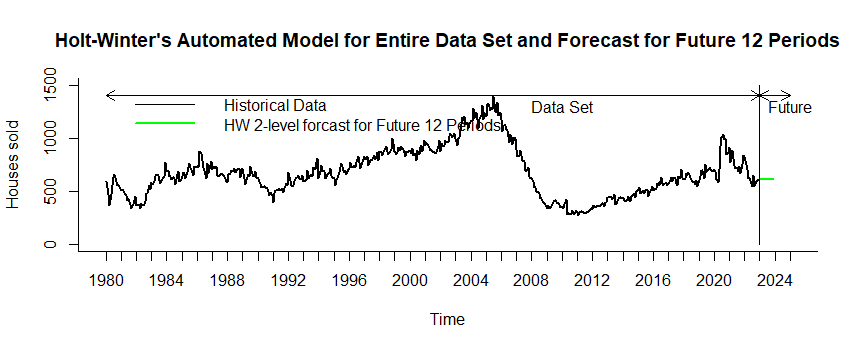


The accuracy measure of the above two model, HW’s model with trailing Ma and HW’s model with ARmodel are as following:



The RMSE and Theils’ u values are the same for both the models, whereas the MAPE value is little less for the two level model with trailing ma. So, we can conclude that both the model are performing equally well, but the two level model with HW’s and MA is performing little better than the other model.

The Plot for th Holt’s winter’s model with trailing MA for the entire dataset is as given below:



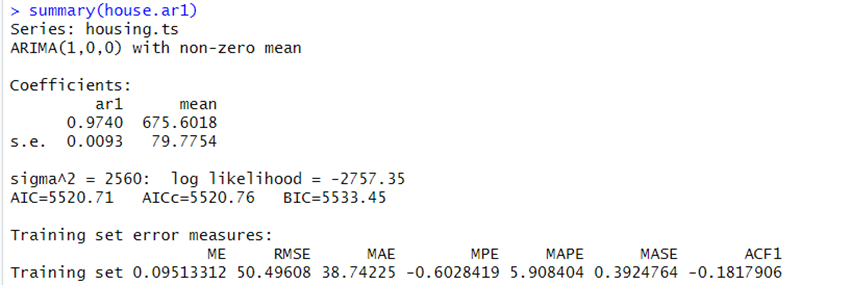
**Testing Of The Hypothesis**

To test the hypothesis we will be needing the values of ar1 and se which is standard error which can be obtained arima function and creating ar(1) model.

**AR(1) Model:**

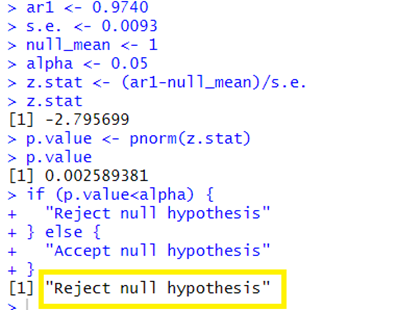
The ARIMA (1,0,0) model was fitted to the housing time series data. The model has one autoregressive term (AR1) with a coefficient of 0.9740, and a non-zero mean of 675.6018. The standard error of the AR1 coefficient is 0.0093 and the standard error of the mean is 79.7754.

The model's log likelihood is -2757.35 and the AIC (Akaike Information Criteria) is 5520.71, the AICc (Corrected AIC) is 5520.76, and the BIC (Bayesian Information Criteria) is 5533.45. These values can be used to compare different models and choose the one with the lowest AIC or BIC.



The performance of the model was evaluated using the mean error (ME), root mean squared error (RMSE), mean absolute error (MAE), mean percentage error (MPE), mean absolute percentage error (MAPE), mean absolute scaled error (MASE), and the first autocorrelation (ACF1) on the training set. The training set had ME of 0.09513312, RMSE of 50.49608, MAE of 38.74225, MPE of -0.6028419, MAPE of 5.908404, MASE of 0.3924764, and ACF1 of -0.1817906.

**Hypothesis Test for Arima**

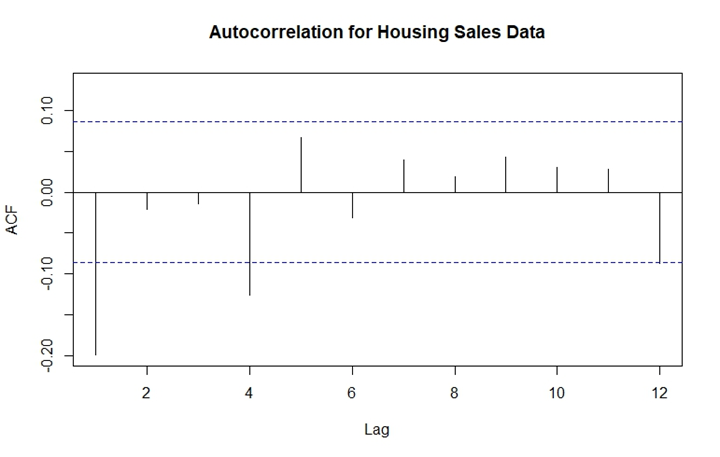


A null hypothesis test is performed to determine if the coefficient of the AR(1) is equal to zero (null\_mean).

In this case, the calculated p-value is 0.0026, which is less than the significance level (alpha) of 0.05. This means that there is less than a 5% chance of observing the coefficient as extreme or more extreme than the one calculated, if the null hypothesis is true.

Based on this, the conclusion of the hypothesis test is to reject the null hypothesis, and accept the alternative hypothesis that the coefficient of the AR(1) term is not equal to zero.

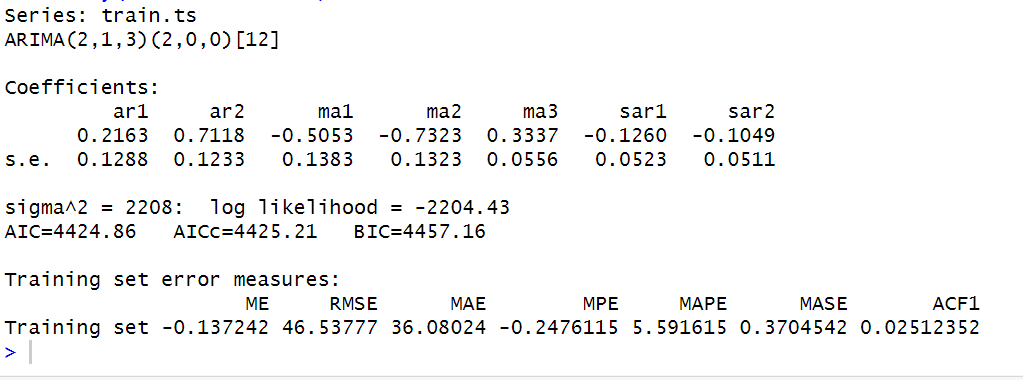
The first-differencing of the housing time series data with a lag of 1 is developed. The first-differenced time series can be interpreted as the change in the housing prices from one time period to the next. By performing first differencing, the time series data is transformed into a stationary series, which can be used to model the time series more effectively.



The ACF plot is created for the differenced Housing Sales Data, with a maximum lag of 12. The plot shows the correlation between the series and its lags, with the values ranging from -1 to 1. As per the plot, lag 1 and 4 are indicating a strong negative correlation between the residuals. The value for lag 12 indicates a weak negative autocorrelation. the rest of lags are not statistically significant and are close to being random.

**AUTO ARIMA Model**

The summary of the Arima model created by using automatic selection of the parameters and options is as follows:



**The Model with optimal parameters and options selected using auto.arima() function is**

**ARIMA(2,1,3)(2,0,0)[12].**

**2 - order 2 autoregressive model (ar2) for trend component.**

**1 - order 1 differencing for removing linear trend.**

**3- order 3 moving average for the error lags (lag3).**

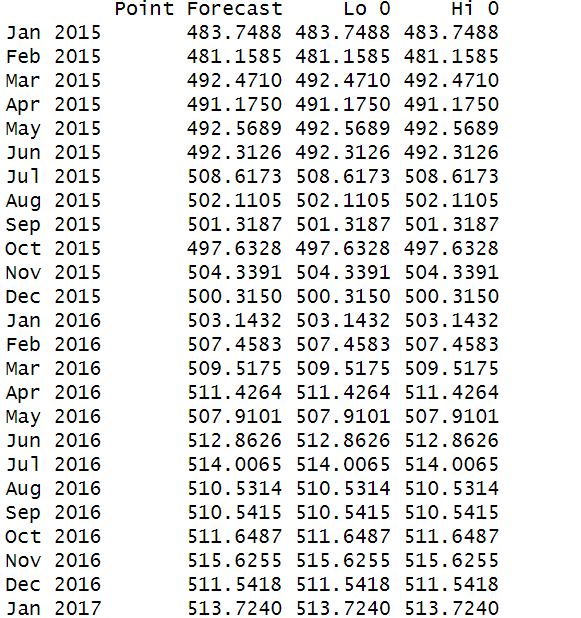
**2 - order 2 autoregressive model (ar2) for seasonality component.**

**0 - order 0 differencing for removing trend.**

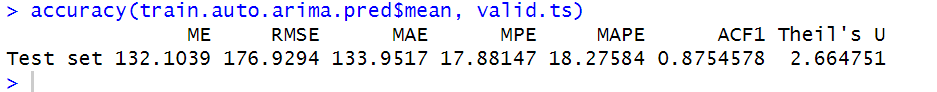
**0 - order 0 moving average for the error .**

**12 - represents the type of the data.**

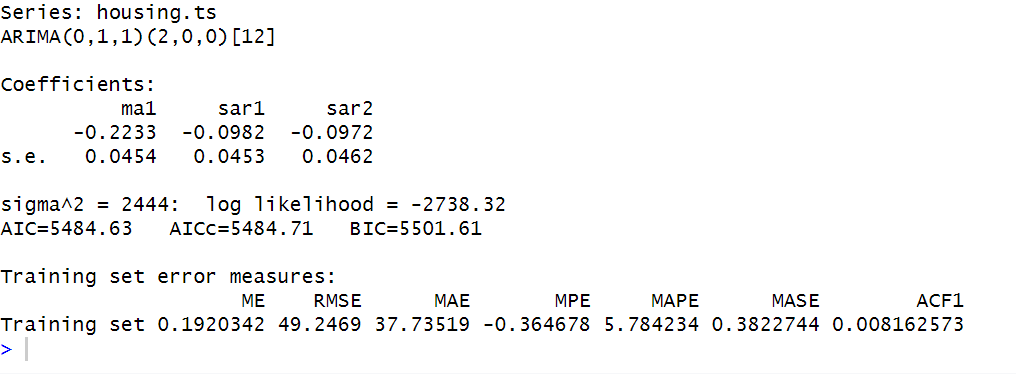
The forecast of the above Arima model on the validation period is as following.



**The accuracy measures of the Auto Arima model on the validation data is:**



**The summary of the auto arima model developed on the entire dataset is as following:**



**The Model with optimal parameters and options for selected using auto.arima() function is**

**ARIMA(0,1,1)(2,0,0)[12].**

**0 - order 0 autoregressive model for trend component.**

**1 - order 1 differencing for removing linear trend.**

**1- order 1 moving average for the error lags (lag1).**

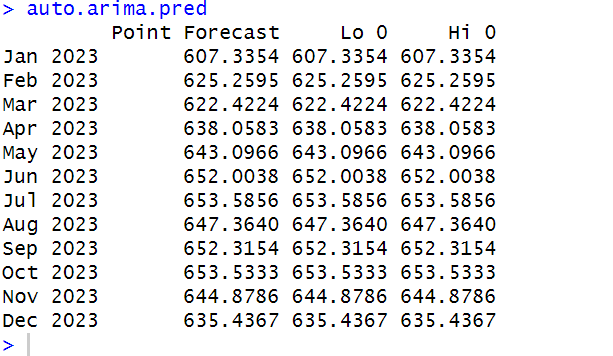
**2 - order 2 autoregressive model (ar2) for seasonality component.**

**0 - order 0 differencing for removing trend.**

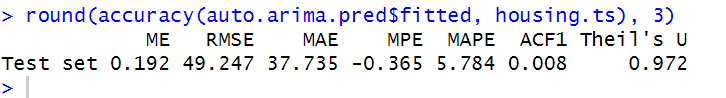
**0 - order 0 moving average for the error .**

**12 - represents the type of the data.**

**The forecast of the above Arima model on the future period is as following:**

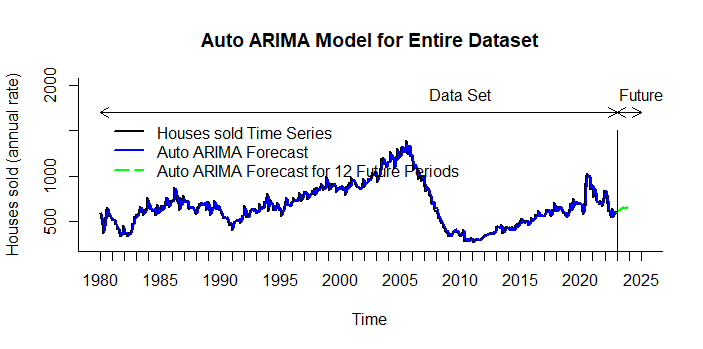


**The accuracy measures of the auto model on the given data is as follows:**



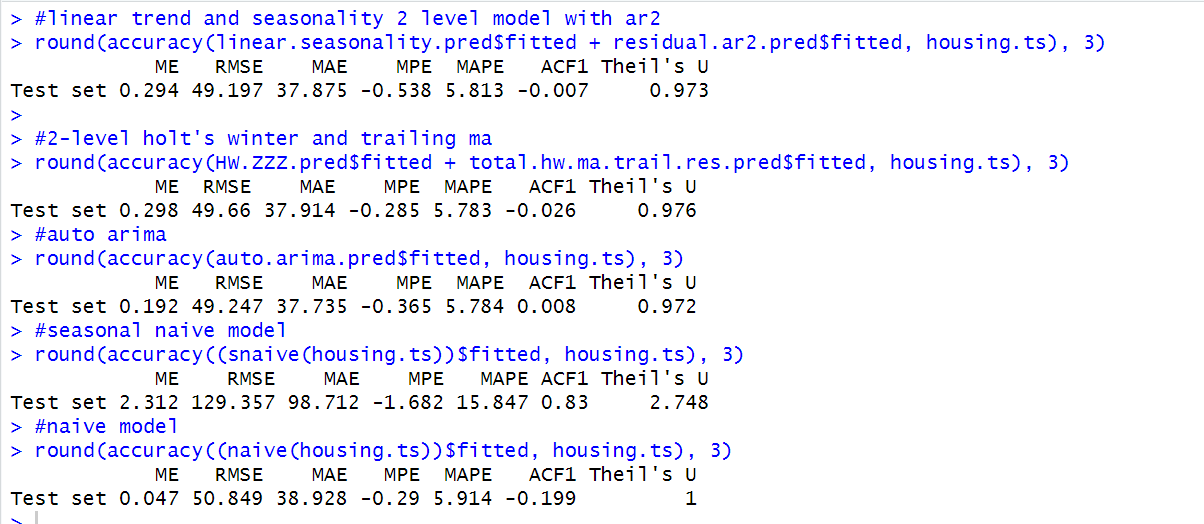
This model is performing comparative well on the entire data set rather than the part of the data. As we can see that the RMSE,MAPE, and theils’ U values are less for the entire data set rather than the validation period and training period.

**The plot for the given data and the auto arima model with the forecast to the future is as follows:**



**Step:8 Implementing Forecast**

**The accuracy measure of the all the models on the given data set is as follows:**



**Choosing Best of Best Models**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **RMSE** | **Mape** | **Theil’s U** |
| **Regression with Linear Trend and Seasonality with AR(2) Model(TWO Level Model with AR)** | **49.197** | **5.813** | **0.973** |
| **Holt’s winter model with trailing MA(Two Level Model with Holts Winter and AR)** | **49.66** | **5.783** | **0.976** |
| **Auto Arima Model** | **49.247** | **5.784** | **0.972** |
| **Seasonal Naive** | **129.357** | **15.847** | **2.748** |
| **Naive** | **50.849** | **5.914** | **1** |

**Among all the models,**

The RMSE,MAPE and Theil’s U are less for the **auto.arima model**, The RMSE value is 49.247, MAPE value is 5.784 and Theil’s U value is 0.972.

The second best model is two-level with **linear trend and seasonality** and AR model, RMSE, is 49.197, 5.813 and 0.973.

The Third Best Model is **Holt’s winter model with trailing MA** is RMSE 49.66, 5.783 is MAPE, Theil’s U is 0.976.

The fourth best model is the **Naive model** and the least accurate model among the **above models is seasonal naive model** .

**Conclusion**

The model of choice is the with Auto Arima Model gave the best predictions which is followed by Linear Trend and Seasonality with an Auto Regressive model gave the best predictions making it the second best model. As stated, prior, the model of choice should be reevaluated annually to ensure that accurate forecasts can be achieved for the subsequent periods.