**Time Series Analysis and Forecasting of Housing Sales**

Kyle McGonigle, Adriel Naranjo, Joel Saetern, Jeremy Manalo

California State University, East Bay

BAN 673 Time Series Analytics

Dr. Zinovy Radovilsky

May 1, 2024

# Summary

In this time series project, we explored the feasibility of utilizing historical house sale data in order to develop a forecast of future housing sale volume. We investigated several time series forecasting models to identify the most accurate approach for this specific dataset. The analysis required filtering on a larger data set for records of monthly house sale volume between January 1963 to February 2024. We evaluated the performance of five total models:

* Two-Level Model (Regression + Trailing MA)
* Holt-Winter's Model (Automatic)
* Two-Level Model (Regression + AR(1))
* Auto ARIMA Model
* Seasonal Naive Model

The models were trained on training data (516 months) and evaluated on the validation data (218 months). Each models’ performance was dependent on the metrics Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) to properly assess their forecasting accuracy. The lowest error metric would be considered as the best/most accurate model for this data.

We found that the Two Level LinReg + AR(1) for residuals model was most accurate, with a MAPE of 7.098 and RMSE of 4.722. The second most accurate model was Two-Level: LinReg with Trailing MA for residuals with an MAPE of 16.943 and RMSE of 11.538.

# Introduction

**Data Set and Collection Source:**

This analysis looks into a time series dataset obtained from the United States Census Bureau (census.gov). The dataset encompasses housing market information from January 1963 to December 2024, spanning over six decades. It is a refined resource for exploring historical trends and potential future developments in the US housing market.

**The data offers a comprehensive overview of various aspects of the housing market:**

* Time Periods:
  + January 1963 to December 2024
* Categories:
  + Houses Sold
  + Annual Rate of Houses Sold
  + Houses listed for sale
* Data Types:
  + All Houses
  + Houses that are Not Started
  + Houses that are Under Construction
  + Houses that are Completed, Median Sales Price
  + Average Sales Price, Months' Supply at Current Sales Rate
  + Median Number of Months For Sale Since Completion
* Error Types:
  + Relative Standard Error for All Houses
  + Relative Standard Error for Houses that are Not Started
  + Relative Standard Error for Houses that are Under Construction
  + Relative Standard Error for Houses that are Completed
  + Relative Standard Error for Median Sales Price
  + Relative Standard Error for Average Sales Price
  + Relative Standard Error for Months' Supply at Current Sales Rate,
  + Relative Standard Error for Median Number of Months For Sale Since Completion
* Geo Levels:
  + United States
  + Northeast
  + Midwest
  + South
  + West

This data set contained various metrics related to housing. By analyzing specific fields, we can apply statistical techniques to draw valuable insights, explore trends, and develop appropriate forecast models.

# Main chapter

#### Table 1. Steps for the forecasting process

| **Step 1** | Define goal |
| --- | --- |
| **Step 2** | Get data |
| **Step 3** | Explore & visualize series |
| **Step 4** | Pre-Process Data |
| **Step 5** | Partition Series |
| **Step 6** | Apply Forecasting Methods |
| **Step 7** | Evaluate & Compare Performance |
| **Step 8** | Implement Forecasts / System |

## Step 1: Define goal

This project has a predictive time series forecasting goal. We aim to leverage historical census data on house sales to develop a model that can accurately forecast the total number of houses sold in the future. The forecast will predict the numerical value representing the total number of houses sold. We expect this to allow for long term forecasting/planning, with a rolling forecast as data continues to be available on a monthly basis. We can also apply yearly / bi annually model examination, and determine which approach best captures the seasonality in the data.

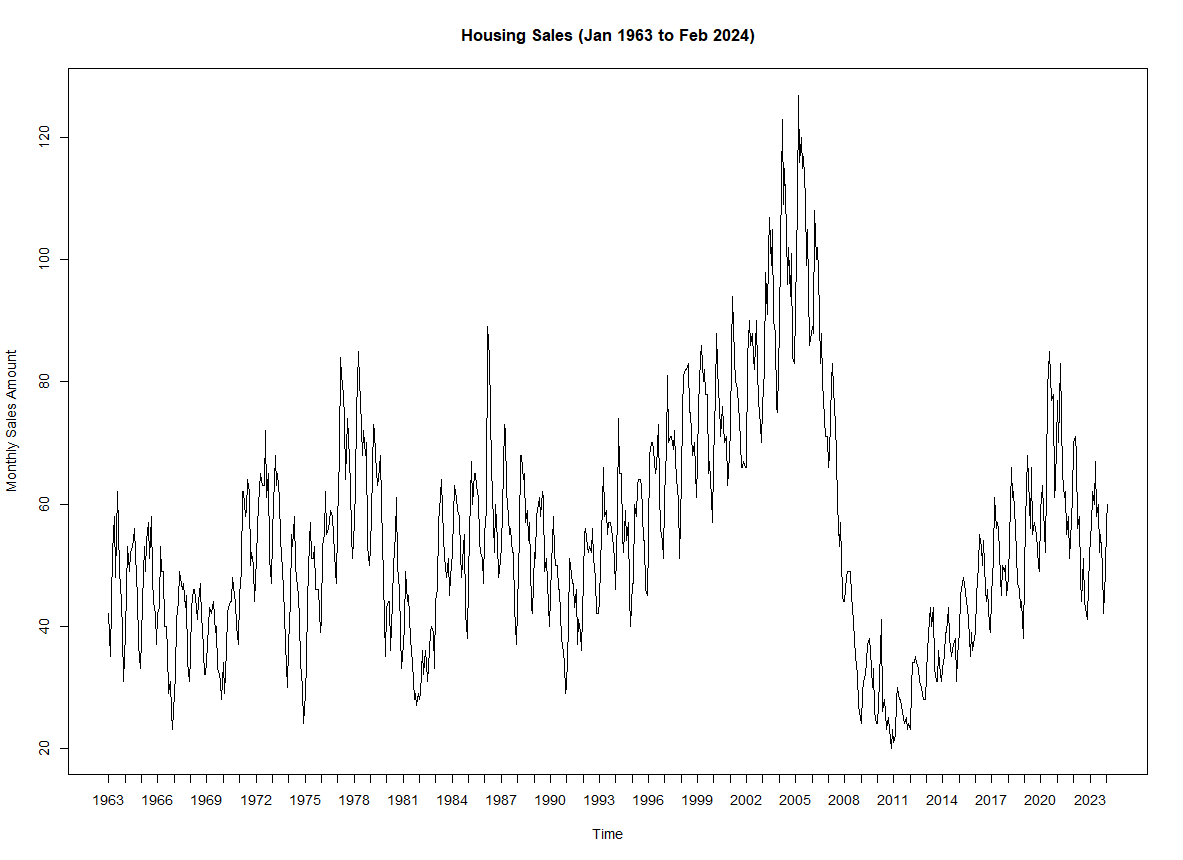
There are various potential stakeholders that may be interested in this forecast. For example, Realtors/Brokers can plan staff required and develop marketing strategies, construction material suppliers can anticipate demand and plan inventory levels, and insurance companies can adjust their pricing strategies for homeowner insurance.

## Step 2: Date collection

Data quality/characteristics:

* **Sample Size:** The sample data set covers a timeframe of 61 years and 2 months, spanning data from January 1963 to December 2023 with additional data for the first two months of 2024.
* **Source:** The data is sourced from a single, reliable source: the US Census Bureau (census.gov).
* Additional data will be collected in the future from census
* **Temporal frequency:** The data is recorded at a monthly frequency, providing granular insights into housing sales trends.
* **Series granularity**: The dataset contains detailed information on new single-family home sales across the US from 1963 to 2023 and the first two months of 2024.
* **Data Cleaning and Preparation:** To prepare our dataset for effective time series analysis, we engaged in multiple joins using the ID of each index value to understand the data better. Recognizing that our dataset was overly detailed and had multiple entries per time series index we concised the number of fields. We decided to go big picture by filtering for cat\_code = 1 to focus solely on sold data, geo\_idx = 1 to focus on the entire US and dt\_idx = 1 for total to aggregate information for all houses. This filtering yielded a single total value per time series index representing sold houses for each monthly time period.

## Step 3: Explore & visualize series



#### Figure 1. Total Monthly Housing Sales from January 1963 to February 2024

#### 

#### Figure 2. Seasonal Decomposition of Time Series

There are some significant patterns that can be observed in these plots. We can see that there was an overall upward non linear trend with additive seasonality for houses sold.There is a significant peak of monthly sales in 2005. However, it was followed by the lowest monthly sales in 2010 and 2011.

**Predictability evaluation using differenced acf function:**

This data set does not appear to be a random walk.

**Approach 1:**

#### 

#### Figure 3. Predictability check with AR(1)

The summary resulted in a z-statistic of -4.854839, and an extremely small p-value of 6.02e-7 (0.0000006024238). Since this p-value is much lower than the commonly used significance level of 0.05, we reject the null hypothesis. The observed data provides strong evidence against the null hypothesis, suggesting a statistically significant effect.

#### Approach 2:

#### 

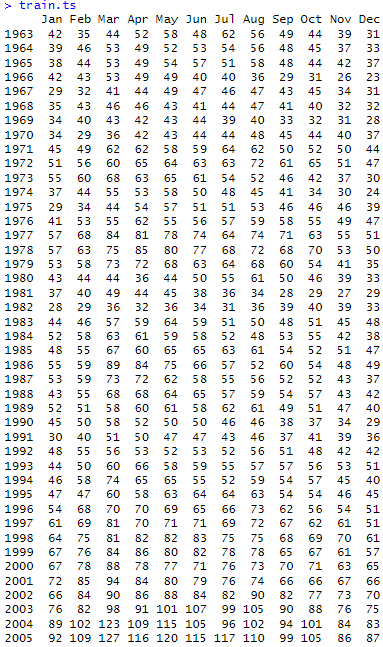
#### Figure 4. Predictability check with Autocorrelation of first differenced

We have significant lags in 4, 6, 8, 11 and 12 which also leads us to believe that this is not a random walk.

## Step 4: Pre-process data

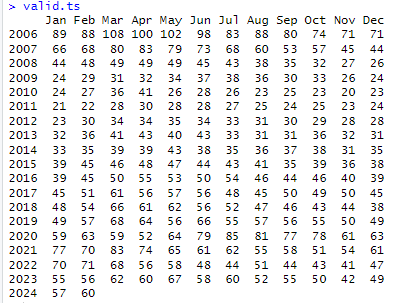
We applied various filters to get the subset of data that we intended to analyze and forecast on. We filtered the data set to keep the US region and the total houses sold value. This allowed us to observe a single total value per time series index, representing sold houses for each monthly time period. Lastly, we transformed the data into a time-series using the ts function.

## Step 5: Partition time series



#### Table 2. Periods of training data

Our training data is 516 months (43 years), starting from Jan 1963 to December 2005.



#### Figure 5. Period of validation data

Our validation data consists of the remaining 218 months (18 years and 2 months).

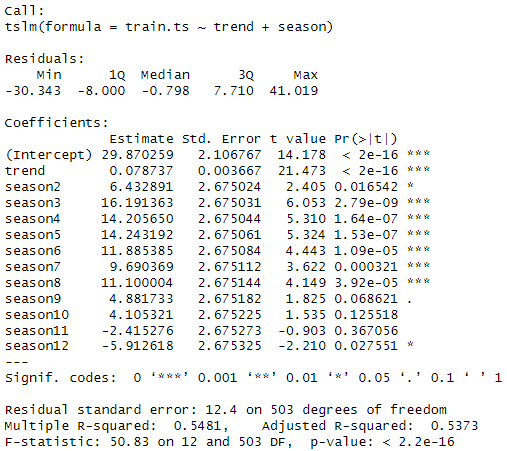
## Step 6: Apply Forecasting Methods

After developing a data partition, we continue by creating a trailing moving average with window widths of k = 4, 8, and 12 for the training partition. Rollmean and forecast data outputs for all window widths are available for view in **Appendix A1.**

#### Table 3. Window Width Accuracy Measures

| **Window Width** | **MAPE** | **RMSE** |
| --- | --- | --- |
| MA 4 | 131.07 | 52.902 |
| MA8 | 140.809 | 56.657 |
| MA 12 | 156.247 | 62.771 |

MA 4 was the best performing trailing moving average. We moved forward with this window width and developed the Two-Level Model (Regression with Linear Trend and Seasonality & Trailing MA).

Model Equation

| 29.870259 +  0.078737×trend +  6.432891×season2 +  16.191363×season3 +  14.205650×season4 +  14.243192×season5 +  11.885385×season6 +  9.690369×season7 +  11.100004×season8 +  4.881733×season9 +  4.105321×season10 −  2.415276×season11  −5.912618×season12+ϵ |
| --- |

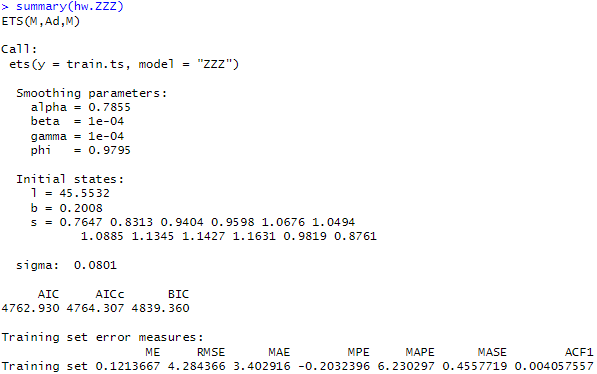
### Two-Level Model: Regression with Linear Trend and Seasonality + Trailing MA:

#### 

#### 

#### Table 4. Outputs for validation data, regression forecast, trailing MA forecast for residuals, and two-level (combined) forecast in the validation period.

### Exponential Smoothing: Holt-Winter’s Model



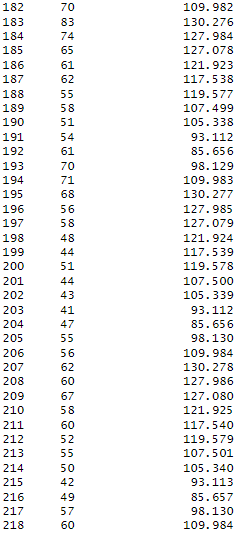
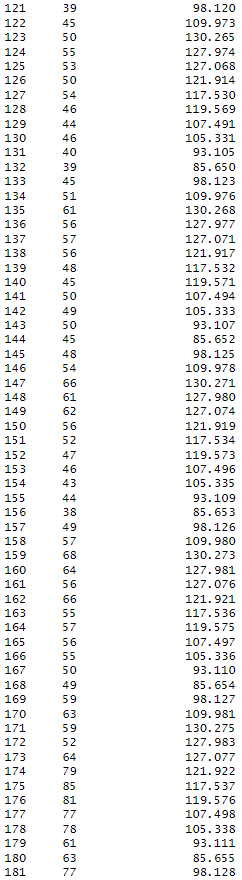
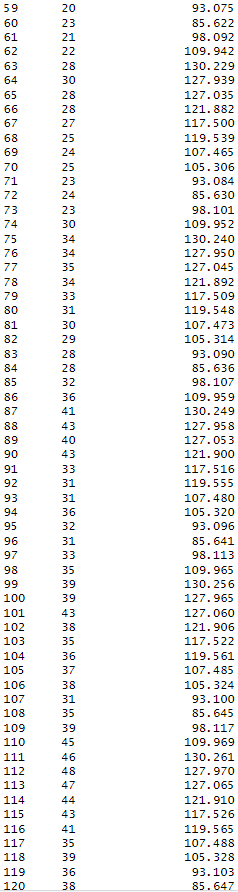
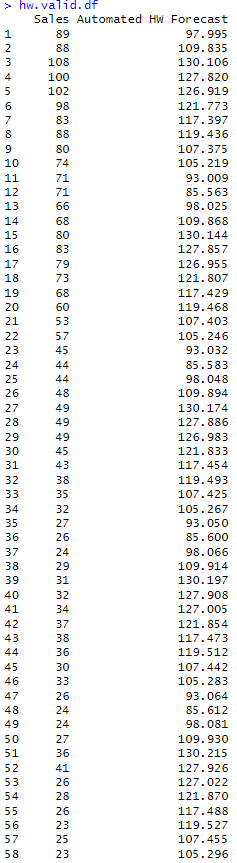
#### Figure 6. Summary of Holt-Winter’s Model with Automatic Model Options and Parameters

In their time series project with historical house sale data, the user has implemented the Holt-Winters (HW) model with specific options: multiplicative error (M), additive damped trend (Ad), and multiplicative seasonality (M). The model was fine-tuned by setting the exponential smoothing constant to 0.7855 for overall smoothing, while the smoothing constant for trend is set very low at 0.0001, indicating a slight trend damping effect, and similarly, the smoothing constant for seasonality is also set to 0.0001 for capturing seasonal variations in the data.

#### 

#### 

#### **Table 5.** Forecast outputs for HW model



#### Table 6. Validation data for HW model

### Autoregressive Model: Two-Level Model Regression with Linear Trend and Seasonality + AR(1) for Regression Residuals

#### 

#### Figure 7. Summary for Two-Level Model: Regression with Linear Trend and Seasonality + AR(1) for Regression Residuals

The two level model summary indicates a statistically significant coefficient for trend component (0.0787). This suggests an upward trend.

Model Equation

| 29.870259 + 0.078737×trend + 6.432891×season2 + 16.191363×season3 + 14.205650×season4 + 14.243192×season5 + 11.885385×season6 + 9.690369×season7 + 11.100004×season8 + 4.881733×season9 + 4.105321×season10 − 2.415276×season11  −5.912618×season12+ϵ |
| --- |

#### 

#### Figure 8. Autocorrelation with a maximum of 24 lags for Housing Sales Data training period

The autocorrelation plot shows positive ACF values (substantially higher than the horizontal threshold) for almost all 24 lags, with Lag 1 being the greatest, and Lag 24 being the lowest. There is a temporal relationship in the data thus patterns we can make predictions on. The gradual decay towards zero suggests a trend component might be present.

#### 

#### Figure 9. Summary for AR(1) model for the regression residuals

This model suggests that the current residual () is influenced by the previous residual () multiplied by a coefficient of 0.9328, with an intercept term of 0.8789 added.

Model Equation for residuals

|  |
| --- |

#### 

#### Figure 10. Autocorrelation of AR(1) for "residuals of residuals" model

The autocorrelation plot for the AR(1) for "residuals of residuals" model shows that most ACF values fall within the threshold, indicating weak autocorrelation. The only lags that fall outside the threshold are Lags 1, 12, 20, and 21.

### Create Dataframe for Validation period

#### 

#### 

#### 

#### Table 7. Two-level forecasting model (regression model with linear trend and seasonality & AR(1) model for residuals) for the validation period.

Next we developed a two level forecasting model for the validation period. The table above displays validation partition data (Sales), regression forecast (Regression.Fst), MA forecast for regression residuals (MA.Residuals.Fst), and combined (2-level) forecast (Combined.Fst) that combines the two previous forecast.

### Auto ARIMA Model:

#### 

#### Figure 11. Summary for Auto ARIMA model on training data set

Model Equation

|  |
| --- |

The auto ARIMA model utilizes a seasonal ARIMA model, with the following parameters:

* p = 2, order 2 autoregressive model AR(1)
* d = 0, No first differencing
* q = 1, order 1 moving average MA(1) for error lags
* P = 0, no autoregressive model for the seasonal part
* D = 1, first differencing for the seasonal part
* Q = 1 order 1 moving average model for the seasonal part’s error lags
* m = 12, for yearly seasonality.

#### 

#### 

#### Table 8. Forecast for Auto ARIMA model

#### 

#### Table 9. Forecast results for Auto ARIMA model validation period

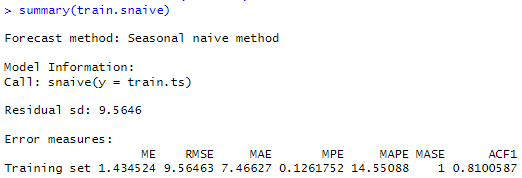
### 

### Naive

#### 

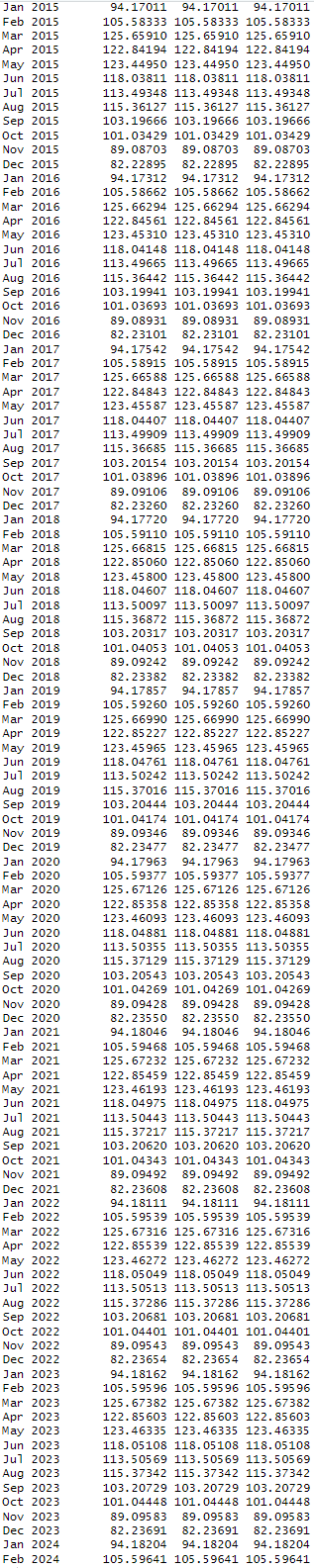
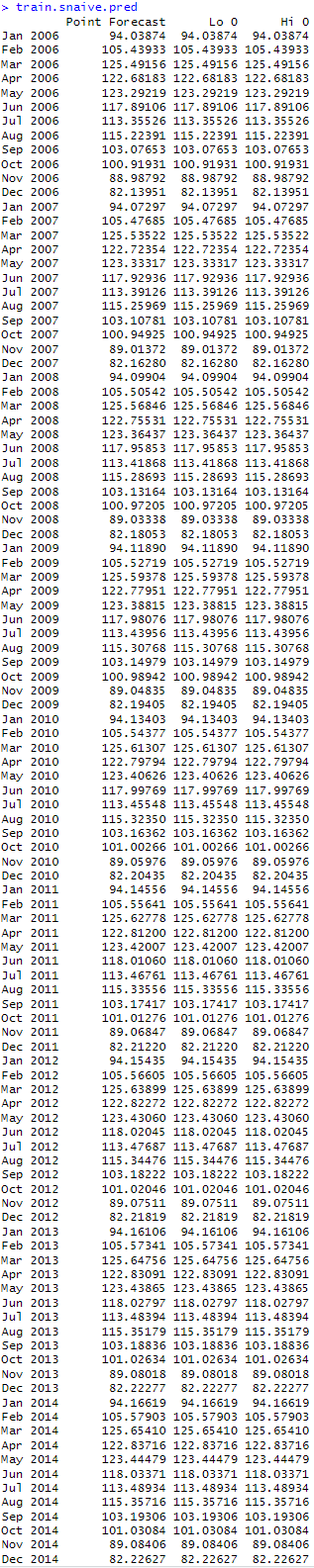
#### Table 10. Forecast results for Auto ARIMA model validation period

### Snaive



#### Figure 12. Summary for Seasonal Naive Model

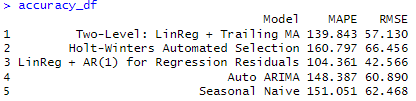
The summary statistics for the seasonal naive model returns a MAPE of 14.55 and RMSE of 9.565.



#### Table 11. Forecast results for Auto ARIMA model validation period

## Step 7: Evaluate & Compare Performance

With all of our models created, we then review the MAPE and RMSE accuracy measures in order to pick top 3 for running on the entire data set:



#### Figure 13. Accuracy measures for all selected models

**Two-Level: LinReg + Trailing MA** had the second best accuracy, based on MAPE (139.843) and RMSE (57.130). The trailing moving average component might not have been as effective in capturing the specific patterns compared to the AR(1) model.

**Holt-Winters Automated Selection** has the second-highest MAPE (160.797) and RMSE (66.456). While this model considers trend and seasonality, the specific parameters chosen by the automated selection may not have been optimal for the data set.

**LinReg + AR(1) for Regression Residuals** achieved the lowest MAPE (104.361) and RMSE (42.566). Using an AR(1) model to capture remaining autocorrelation in the residuals of a linear regression with trend and seasonality seemed to have improved the model's ability to learn the underlying patterns in the data.

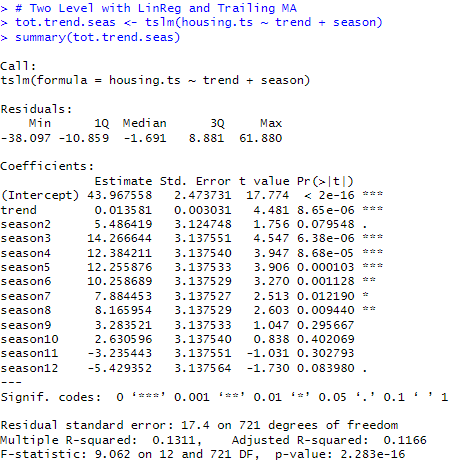
**Auto ARIMA** fell in the middle with a MAPE of 148.387 and RMSE of 60.890. Automatic model selection may be convenient, but it might not always outperform more refined approaches (such as LinReg + AR(1) model in this case).

We used **Seasonal Naive** as the baseline model, which simply predicts the average value for each month. It has the highest MAPE (151.051) and RMSE (62.468). As expected, it performs the worst since it doesn't account for any trends or underlying patterns in the data.

## **Step 8:** Implement Forecasts / System

After determining the best two performing models (Two Level with LinReg and Trailing MA and Two Level using LinReg and AR(1)), we run them on the entire dataset:

### Two Level with LinReg and Trailing MA



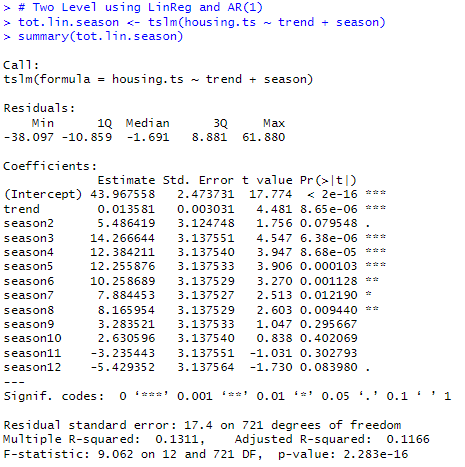
#### Figure 14. Summary for Two Level with LinReg and Trailing MA on entire dataset

Model Equation

43.967558 + 0.013581×trend + 5.486419×season2 + 14.266644×season3 + 12.384211×season4 + 12.255876×season5 + 10.258689×season6 + 7.884453×season7 + 8.165954×season8 + 3.283521×season9 + 2.630596×season10 − 3.235443×season11 − 5.429352×season12+ϵ

This model incorporates trend and seasonal variables alongside a trailing moving average (MA) with a window of 4. The trailing MA accounts for short-term fluctuations in the data, providing a smoothed representation of the series. Significant coefficients for each season indicate the impact of seasonal patterns on the outcome. The inclusion of both trend, seasonal variables, and the trailing MA reflects a comprehensive approach to capturing variations in the data for accurate forecasting.

### Two Level using LinReg and AR(1)



#### Figure 15. Summary for Two Level using LinReg and AR(1) for entire dataset

Model Equation

| 43.967558 + 0.013581×trend + 5.486419×season2 + 14.266644×season3 + 12.384211×season4 + 12.255876×season5 + 10.258689×season6 + 7.884453×season7 + 8.165954×season8 + 3.283521×season9 + 2.630596×season10 − 3.235443×season11 − 5.429352×season12+ϵ  This model integrates trend and seasonal variables to forecast the dependent variable. Additionally, it utilizes an autoregressive component (AR(1)) on the residuals, implying that the model accounts for the correlation between consecutive residual values. The significant coefficients for each season highlight the seasonal patterns' influence on the outcome, underscoring the model's ability to capture both trend and seasonal variations for accurate forecasting. |
| --- |

# Conclusion

In this time series project, we delved into the potential of using historical house sale data to predict future housing sale volumes. Our analysis spanned a significant period, filtering through monthly records from January 1963 to February 2024 to capture the essence of the housing market's trends and patterns.

We rigorously evaluated five distinct time series forecasting models, each with its own methodology and strengths. These models included the Two-Level Model (Regression + Trailing MA), Holt-Winter's Model (Automatic), Two-Level Model (Regression + AR(1)), Auto ARIMA Model, and the Seasonal Naive Model. Our evaluation criteria were rooted in real-world applicability, focusing on accuracy metrics such as Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE).

Among these models, the Two-Level LinReg + AR(1) for residuals emerged as the most accurate predictor for our dataset, boasting a MAPE of 7.098 and an RMSE of 4.722. This model's ability to combine regression analysis with autoregressive components for residual handling proved instrumental in capturing the nuances of the housing market's dynamics.

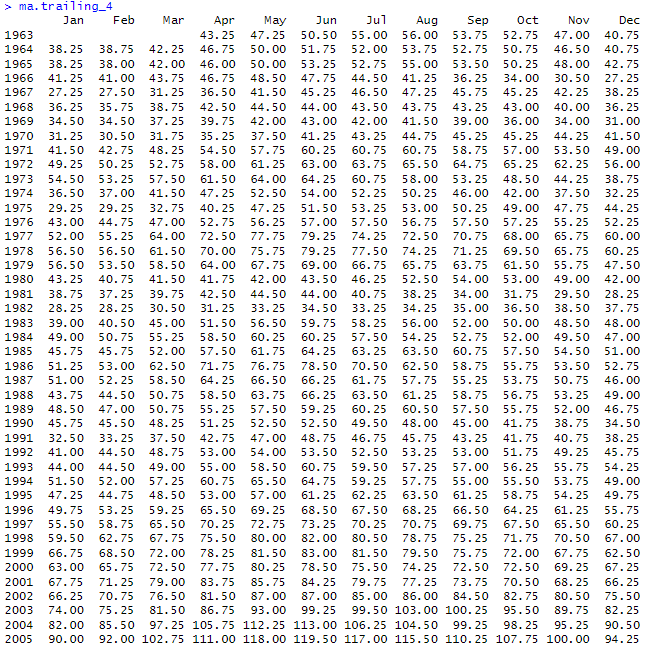
However, it's worth noting that the Two-Level: LinReg with Trailing MA for residuals also demonstrated respectable performance, albeit slightly less accurate than the top-performing model, with an MAPE of 16.943 and an RMSE of 11.538. This highlights the importance of considering various approaches and trade-offs in model selection, especially when dealing with complex and dynamic datasets like housing sales.

One key takeaway from this project is the significance of incorporating both regression analysis and autoregressive components, as seen in the top-performing model. This hybrid approach effectively leverages historical trends while accounting for residual variations, leading to more robust and accurate forecasts.

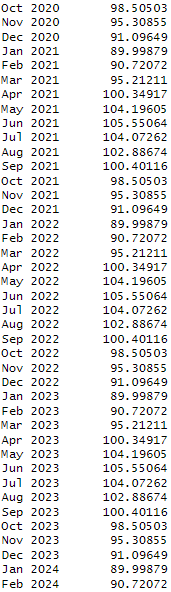
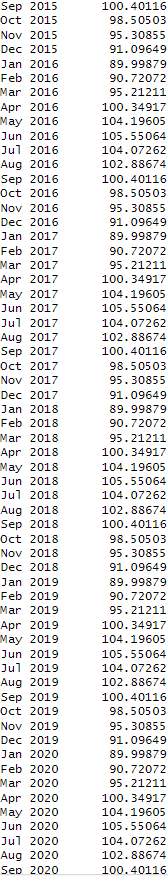
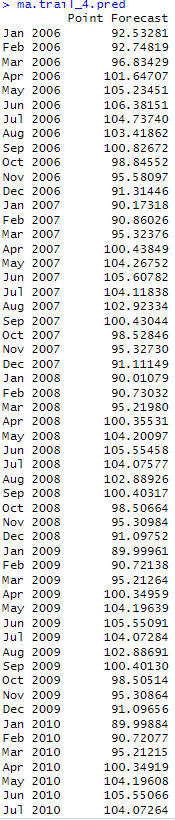
In terms of recommendations, we strongly advocate for continued monitoring and fine-tuning of the chosen model as the housing market evolves. Additionally, exploring ensemble techniques or hybrid models that integrate the strengths of multiple approaches could further enhance predictive accuracy. Overall, this project underscores the power of time series analysis in extracting actionable insights from historical data, offering valuable guidance for decision-makers in the housing industry and beyond.

# Appendices

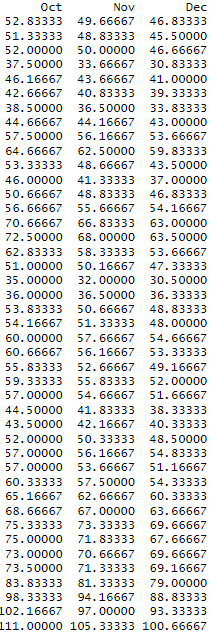
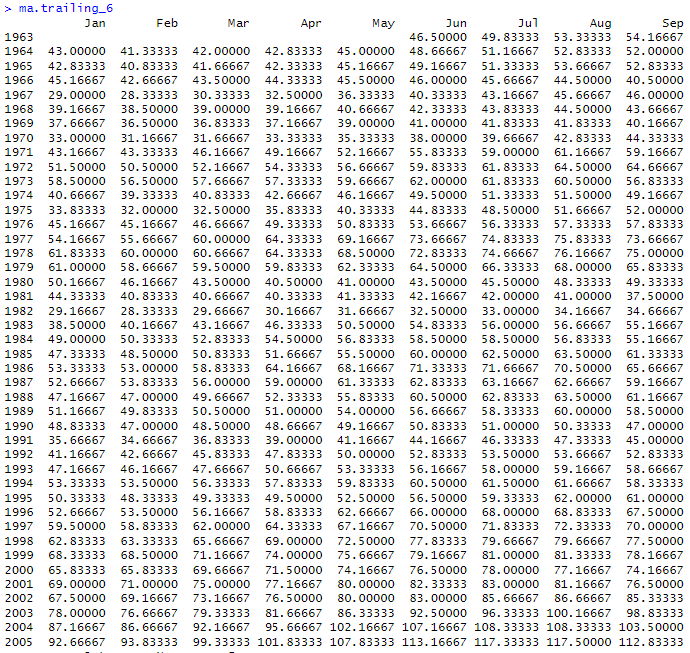
**Appendix A1:** Moving Average (Trailing MA) results



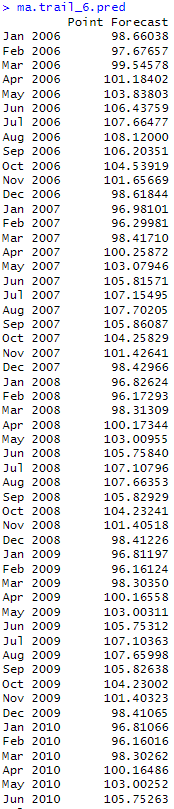
**Figure A1-1.** ma.trailing\_4



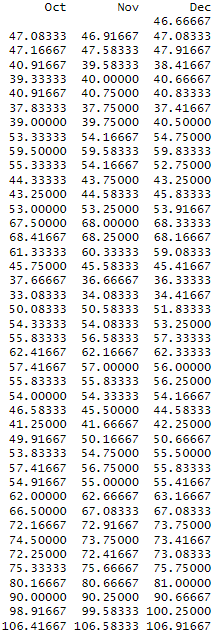
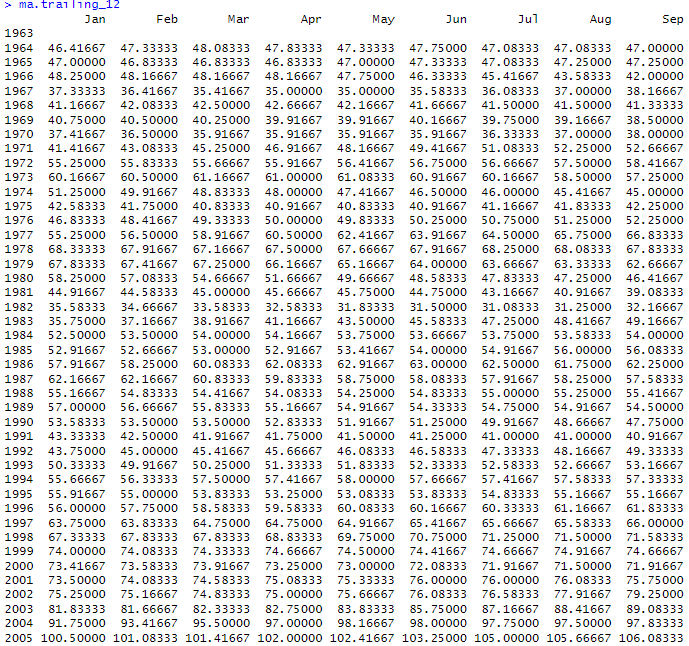
**Figure A1-2.** ma.trailing\_4.pred



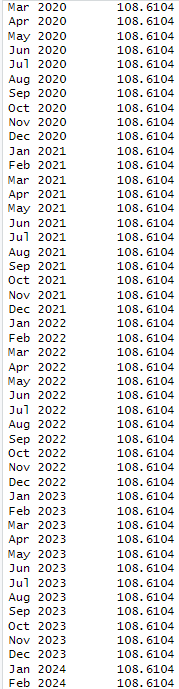
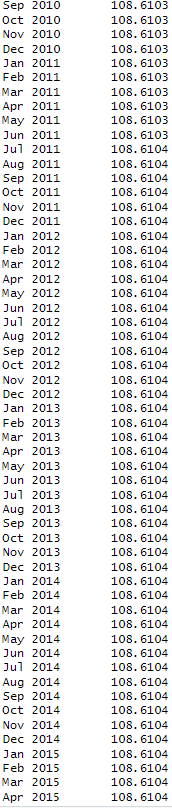
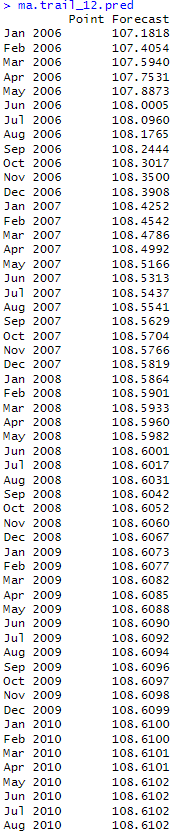
**Figure A1-3.** ma.trailing\_6



**Figure A1-4.** ma.trailing\_6.pred



**Figure A1-5.** ma.trailing\_12



**Figure A1-6.** ma.trailing\_12.pred