FINAL PROJECT REPORT

ARIMA FITTING AND FORECASTING: US HOUSE SALES TIME SERIES

Team Members:

- Ha Do
- Priscille Koutouan
- Dimitrios Ligas

Class: ST534 – Time series analysis

To: Dr. Martin

Dated: 12.05.2022

I. Introduction

The purpose of the ST 534 project is to model a time series data and present the entire modeling process via a report. To complete our project, a relevant dataset was needed. We searched for a dataset with interesting features such as trend, seasonality, and non-stationarity due to a varying mean and/or variance. The dataset also needed to be big enough to allow for modeling without constraints. Keeping in mind these features, we focused on datasets that were of interest to the group such as ones from the housing and stocks markets. We ended up working with a dataset related to the housing market as some of the datasets from the stock market were not good enough for modeling based on preliminary analysis.

The final dataset that was used for this project contains historical data of the number of houses sold monthly from January 1963 to September 2022 in the U.S. This data is for new residential sales and specifically single-family houses. It does not contain multifamily buildings or existing homes. The data was collected using the Survey of Construction from the United States Census Bureau and was accessed by our team for this project on the United States Census Bureau website[1].

The chosen time series data is appropriate for the purposes of modeling for several reasons. First, house sales have been recognized to have seasonal characteristics. Miller et. al. (2011) [2] analytically suggested this feature of the data series using a least square regression model. A time series model with seasonality component to fit this dataset will provide a different view about this hypothesis. Secondly, it is a reliable and big enough dataset with 717 observations. Each observation is a month in a year, so we expect to see seasonality and preliminary analysis confirmed this expectation. As the dataset spans a long period, trend behavior is not expected when considering the data as whole. From the data, we learn that there is on average 54,644 single-family houses sold per month with standard deviation σ=18,678.

In this study, we filtered the dataset using an ARIMA model, then forecasted the number of house sales for the next twelve months. In the remaining sections of this report, we described: the steps undertaken to pre-process the data (*Part II*), the rationale

[1] https://www.census.gov/construction/nrs/historical_data/index.html

[2] Miller, Norm & Sah, Vivek & Sklarz, Michael & Pampulov, Stefan. (2011). Seasonality in Home Prices - Evidence from CBSAs. Journal of Housing Research. 22. 10.1080/10835547.2013.12092066.

in the model identification stage (*Part III*), the model fitting steps (*Part IV*) and forecasting using the best fitted model (*Part V*). We finally concluded the findings in *Part (VI)*.

II. Variance Stabilization and Differencing Transformations

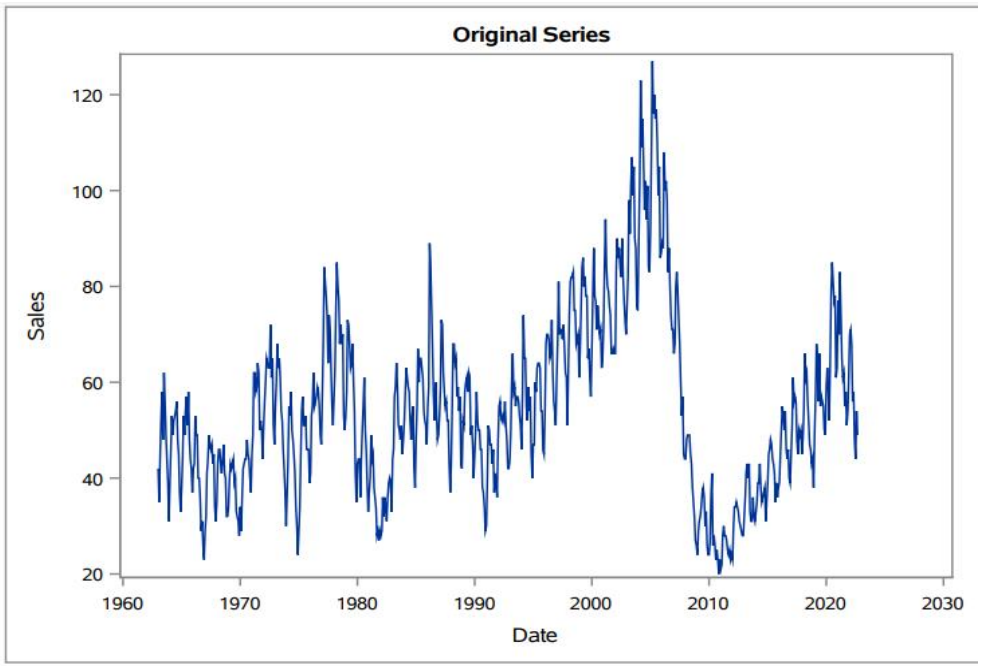


Figure 1 Monthly household sales in the U.S. from January 1963 - September 2022 (unit: thousand houses)

Figure 1 is the plot of the original series. Before we proceeded to time series modelling procedures, we decided to do a variance stabilizing transformation as it may seem from the plot that the variance might be slightly changing with time. We did a Box-Cox analysis:

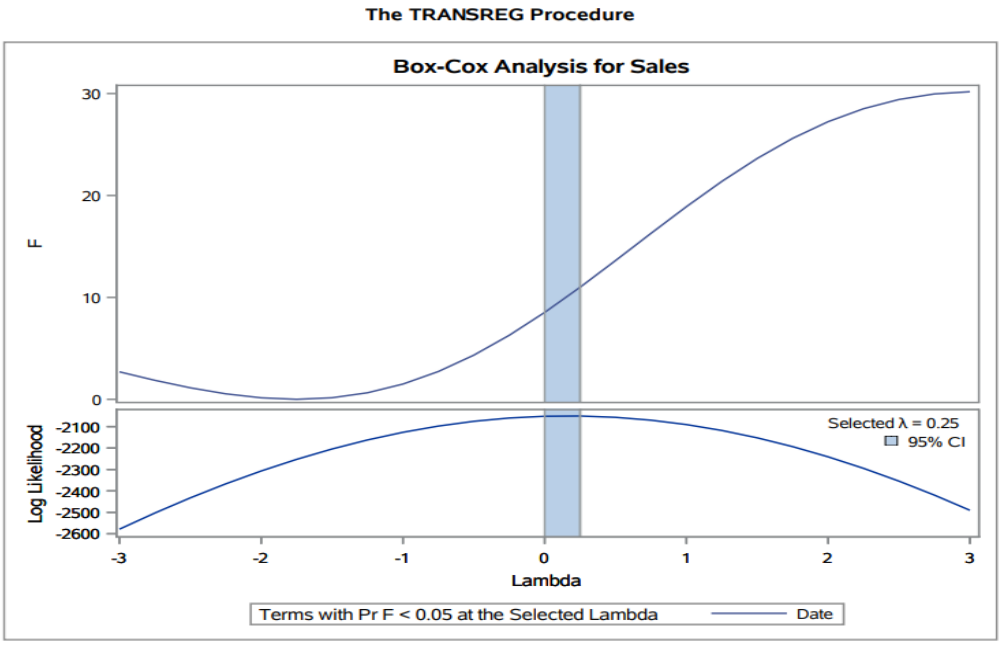


Figure 2 Result from the Box-Cox analysis

Results from the Box-Cox analysis (Figure 2) suggest that we should proceed to a transformation of the original series with λ=0.25 as follows:

$$Z_t = \frac{Y_t^{0.25}-1}{0.25},$$

where Y_t is the number of houses sold each month (the original series).

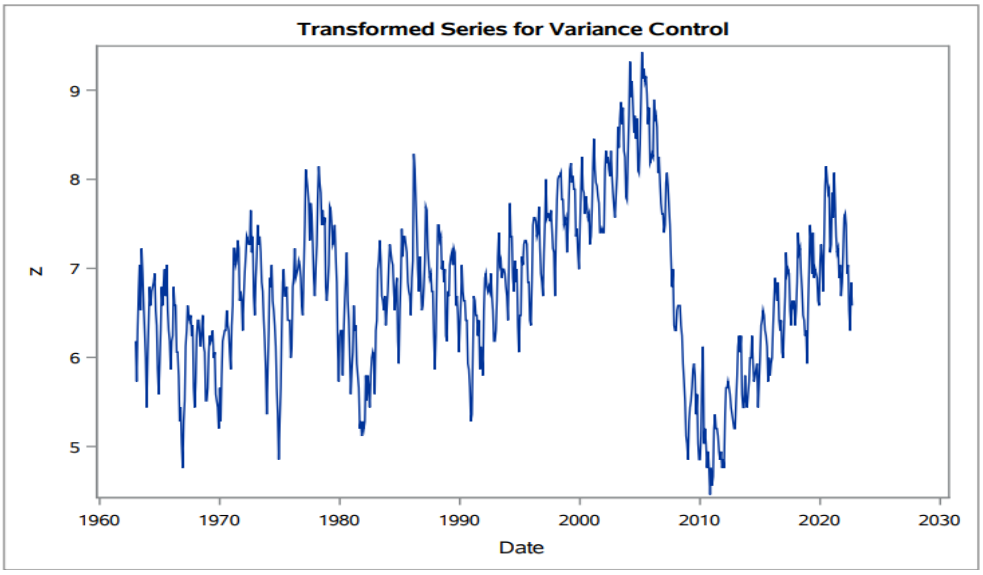


Figure 3 Series for monthly household sales in the U.S. from January 1963 - September 2022 after Box-Cox transformation

From Figure 3, we can observe the transformed series. The variance of the transformed series seems to be slightly improved compared to the original series, especially in the years 2007-2008.

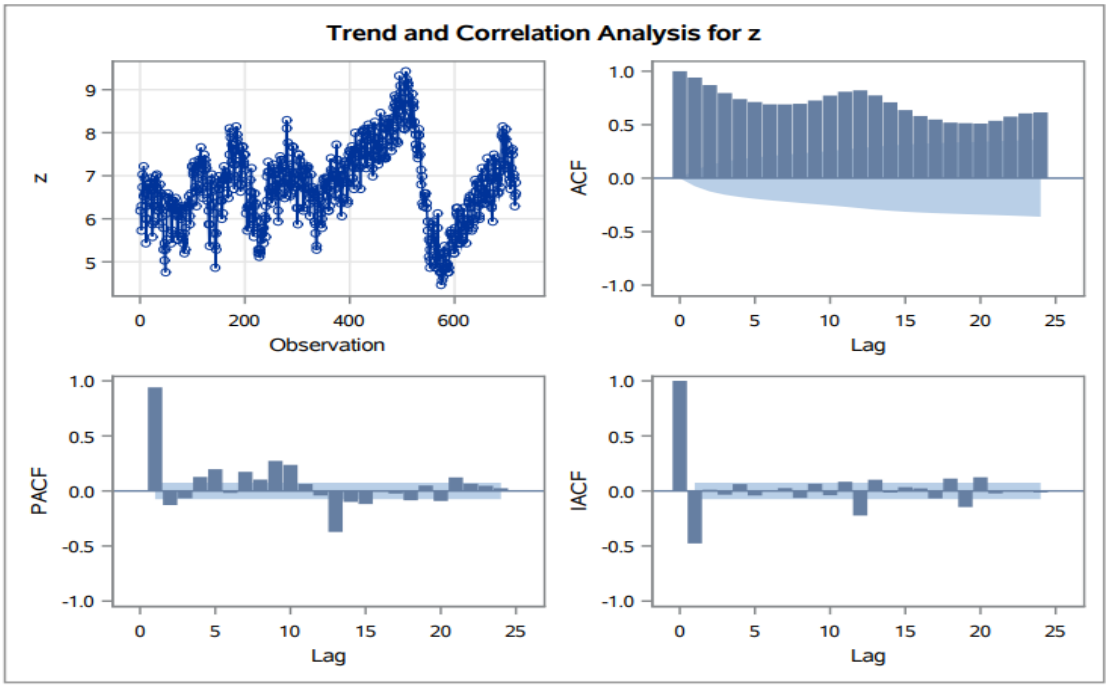


Figure 4 Examining the transformed data

Furthermore, most probably it contains a unit root since the mean seems to be non-constant across time and a slight linear trend is possible with drifts in higher levels (See Figure 4). In 2007, there is a major structural shift due to the housing market crisis in the U.S and the upcoming global financial crisis that is reflected in the series as an abrupt drop to very low levels. Following the economic recovery, it gradually starts to return to its previous levels, all the while maintaining a linear upward trend. Moreover, the ACF decreases very slowly, which adds to our initial intuition that the series is non-stationary (See Figure 4). For this reason, we took a regular difference on $\{Z_t\}$.

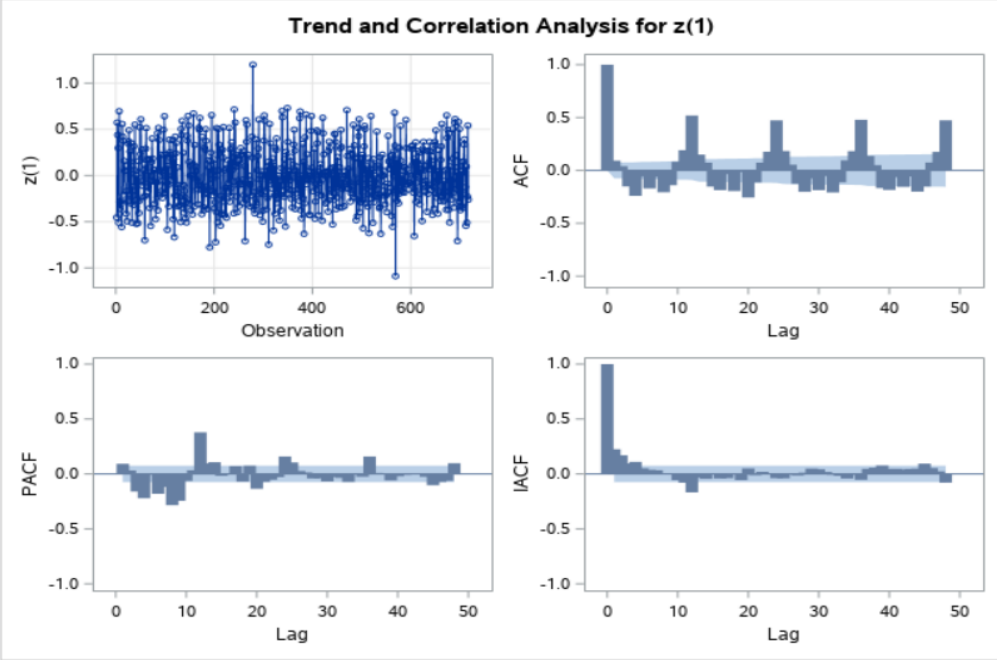


Figure 5 Examining the data after regular differencing

By examining the ACF plot of the differenced series in Figure 5, we see a pattern that previously was not obvious. There seems to be a persistent autocorrelation at the lags 12, 24, 36 ,48, etc. The ACF at those lags decreases rather slowly, suggesting that a seasonal difference with s = 12 is needed. We proceed to take a seasonal difference at lag s=12 and examine the resulting plots in Figure 6:

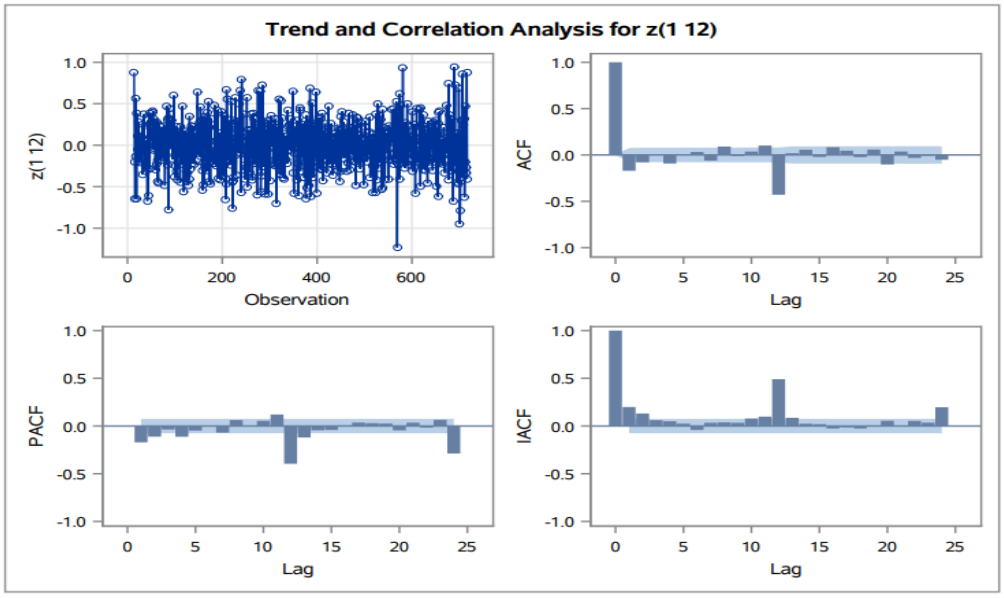


Figure 6 Examining the data after regular and seasonal differencing

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	31.81	6	<.0001	-0.171	-0.079	-0.000	-0.092	-0.003	0.031
12	181.17	12	<.0001	-0.062	0.091	-0.011	0.036	0.101	-0.429
18	190.95	18	<.0001	0.018	0.056	-0.022	0.083	0.046	-0.023
24	204.74	24	<.0001	0.058	-0.103	0.037	-0.032	-0.008	-0.052

Figure 7 Result of Ljung-Box test

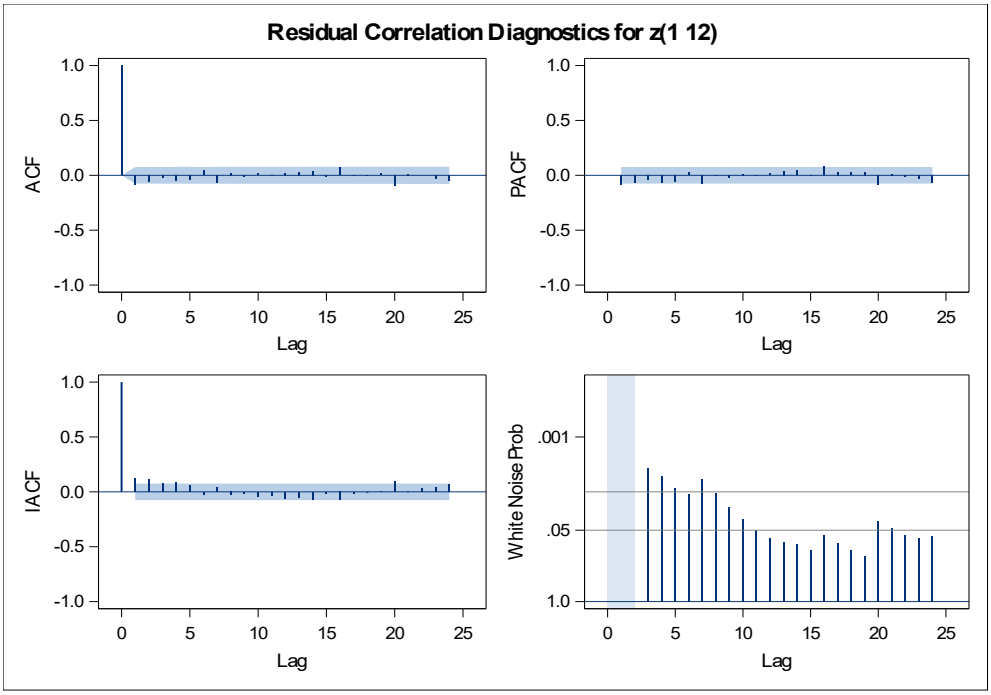
The resulting series is stationary, and the seasonal effect has been eliminated (see Figure 6). Therefore, no additional transformations are needed, and we can now proceed to the model identification and estimation stage. As the Ljung-Box test rejected the white noise hypothesis (Figure 7) , we toned to identify an appropriate model to filter the data.

III. Brainstorming possible models

After taking a regular and seasonal difference at lag s=12, the process seems to be stationary with mean zero, μ=0. For this reason, we did not include a constant in our model estimations, a decision which was justified by the fact that the constant parameter turned out to be not statistically insignificant. From Figure 6, there is a big spike at lag 12 in the ACF plot, suggesting that a seasonal MA of order 1 is a strong possibility. Furthermore, we see that the ACF has a spike at lag 1 while the PACF decreases in an exponential manner before spiking again at 12. This suggests that a regular MA(1) component should be considered. There are also some small spikes at other lags, which were examined after some preliminary model fitting. Additionally, there was always the possibility of perceiving the ACF as decreasing exponentially and the PACF cutting-off after lag 2, although this might not be the case since, we have a similar spike also at lag 4. Nonetheless, some regular AR components were tested for earlier lags.

IV. Model fitting results

A. ARIMA(0,1,2)x(0,1,0)$_{12}$

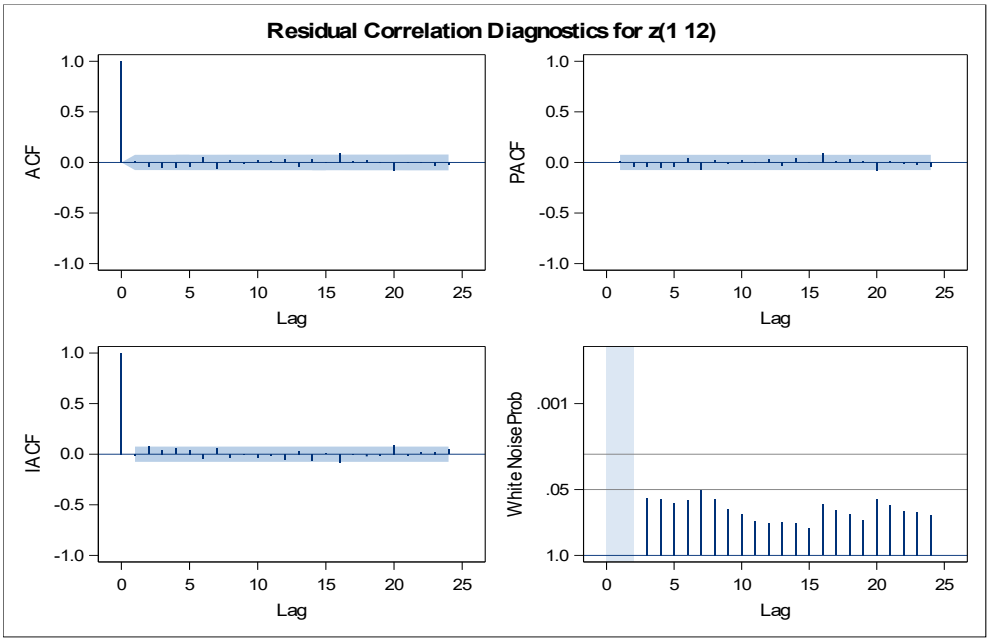


Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	13.04	4	0.0111	-0.089	-0.058	-0.026	-0.054	-0.040	0.044
12	17.23	10	0.0694	-0.068	0.022	-0.009	0.015	-0.002	0.023
18	22.95	16	0.1150	0.031	0.038	-0.017	0.072	0.001	0.003
24	32.76	22	0.0654	0.016	-0.100	0.012	0.000	-0.031	-0.046
30	43.88	28	0.0285	0.087	0.024	0.035	0.021	-0.071	-0.018

Figure 8 Results of model ARIMA(0,1,2)x(0,1,0)$_{12}$

We first try a regular MA of order 2 with parameters at lags 1 and 12. Both parameters are statistically significant and estimated at θ_1=0.112 and θ_2=0.755. They also are not significantly correlated. However, the Ljung-Box test rejects the white noise hypothesis with strong evidence against Ho. In addition, though the PACF and ACF seem to look okay, the white noise probabilities for the model residuals are very low as shown in Figure 8. We conclude that the unfactored model does not fit well and cannot be considered further. **AIC=3.17** and **SBC=12.284.**

B. ARIMA(0,1,1)x(0,1,1)$_{12}$



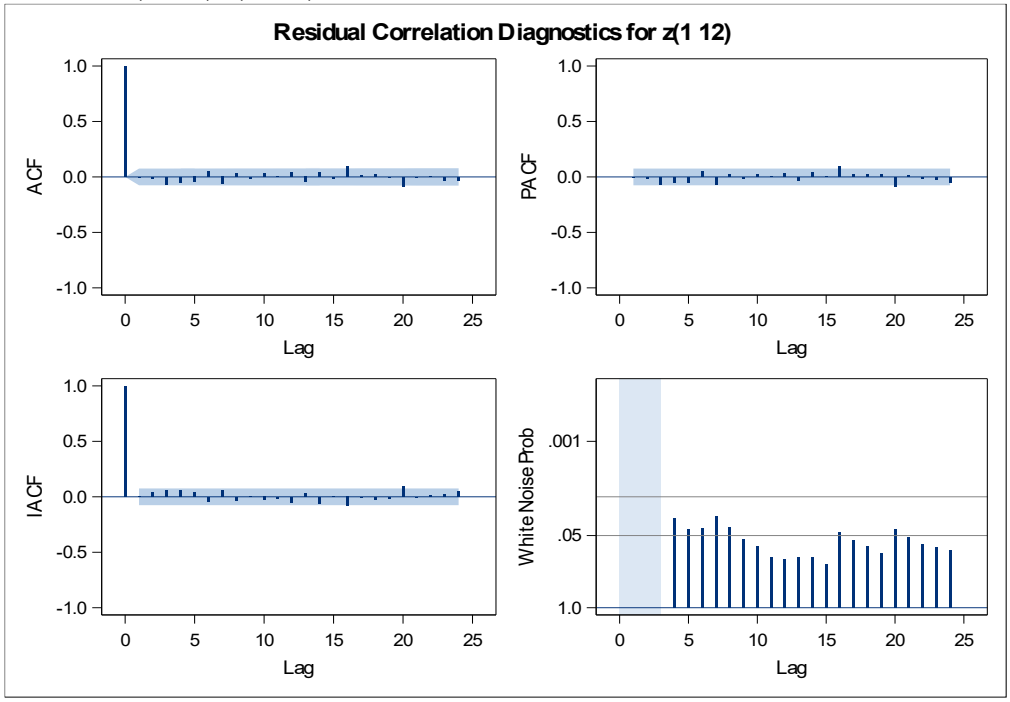
Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	8.27	4	0.0821	0.013	-0.043	-0.050	-0.052	-0.043	0.050
12	12.96	10	0.2260	-0.061	0.024	-0.012	0.026	0.009	0.036
18	21.69	16	0.1535	-0.042	0.038	-0.007	0.090	0.013	0.025
24	28.57	22	0.1577	-0.003	-0.087	-0.006	0.003	-0.033	-0.028
30	36.31	28	0.1348	0.027	0.035	0.034	0.028	-0.080	-0.016

Figure 9 Results of model ARIMA(0,1,1)x(0,1,1)$_{12}$

We next try the factored model, with a regular MA Component of order 1 at lag t=1 and a seasonal MA component of order 1 at lag s=12. The parameter estimates are θ_1=0.226 and Θ_1=0.773 and are both statistically significant and there is almost zero correlation between them. The Ljung-Box test does not reject the white noise Ho with robust probabilities, and the white noise probabilities are high as shown in Figure 9. Furthermore, the kernel for normality approximation of the residuals and the QQ-Plot look very good. We conclude that the model fits the data well. **AIC= -9.89** and **SBC= -0.777.**

C. ARIMA(2,1,0)x(0,1,1)$_{12}$

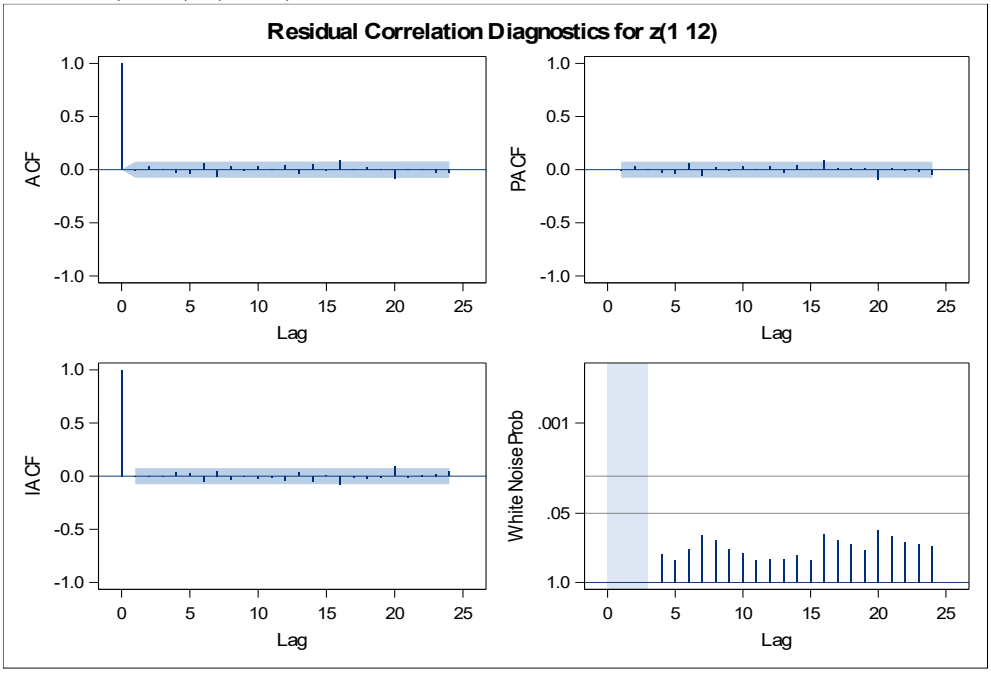


Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	8.50	3	0.0367	-0.004	-0.014	-0.067	-0.049	-0.045	0.053
12	13.74	9	0.1320	-0.063	0.029	-0.016	0.029	0.005	0.037
18	23.35	15	0.0769	-0.045	0.042	-0.011	0.093	0.012	0.025
24	29.98	21	0.0923	-0.002	-0.085	-0.006	-0.000	-0.031	-0.030
30	37.78	27	0.0814	0.027	0.036	0.031	0.029	-0.081	-0.014

Figure 10 Results of ARIMA(2,1,0)x(0,1,1)$_{12}$

As an alternative approach, we try to fit a seasonal ARIMA with a regular AR component of order 2 by considering the possibility of the PACF cutting-off after lag t=2 (See Figure 10). The parameter estimates are φ_1= -0.208, φ_2= -0.077 and Θ_1=0.770 and all of them are statistically significant. However, φ_2 came close to be considered insignificant at α=5% statistical significance level. The model fits reasonable, nevertheless the white noise Ho of the residuals gets rejected for the first 6 lags and the white noise probabilities are not that high. The **AIC** and **SBC** are **-7.816** and **5.854** respectively. The models is not an adequate fit as white noise is rejected.

D. ARIMA(1,1,1)x(0,1,1)$_{12}$



Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.30	3	0.2310	-0.009	0.029	-0.004	-0.025	-0.035	0.057
12	9.89	9	0.3598	-0.063	0.032	-0.013	0.028	0.007	0.042
18	19.53	15	0.1906	-0.038	0.050	-0.009	0.093	0.009	0.025
24	26.03	21	0.2053	0.003	-0.084	-0.005	-0.002	-0.032	-0.029
30	32.80	27	0.2038	0.025	0.028	0.024	0.027	-0.080	-0.009

Figure 11 Results of model ARIMA(1,1,1)x(0,1,1)$_{12}$

We now try to combine one regular AR and one regular MA component, to examine whether we could obtain an enhanced model fit. The resulting parameter estimates are φ_1=0.385, θ_1= 0.595 and Θ_1=0.769 and all of them are statistically significant. The white noise hypothesis for the model residuals is not rejected with very strong probabilities and the white noise probabilities are even higher compared to the ARIMA(0,1,1)x(0,1,1)$_{12}$ model (**B**) (See Figure 11). This model provides a great fit with information criteria **AIC=-12.387** and **SBC=1.283**. Unfortunately, it contains one major drawback. The AR and MA parameters are very highly correlated, with their coefficient being $\rho(\varphi 1,\theta 1)$=96.2%, suggesting that one of the two parameters must be dropped from the model.

Model selection

The top-2 performing models are the ARIMA(0,1,1)x(0,1,1)$_{12}$ (model **B**) and the ARIMA(1,1,1)x(0,1,1)$_{12}$ (model **D**). AIC prefers model (**D**), while SBC prefers model (**B**). Both fit the house sales time series data very well, with the 2nd providing an almost perfect fit, but ultimately not being chosen due to the high correlation between the AR-MA parameters ($\rho(\varphi_1,\theta_1)$=96.2%).

The final model that we chose to fit to the data and perform forecasts with, is the ARIMA(0,1,1)x(0,1,1)$_{12}$ given by:

$$(1-B)*(1-B^{12})*Z_t = (1-0.22628B)*(1-0.7725B^{12})*a_t$$

With $Z_t = \frac{Y_t^{0.25}-1}{0.25}$, where Y_t is the original series of the monthly house sales in thousands .

V. Forecasting the series in 12 months ahead $\widehat{Z_{t+n}}$

Using the above seasonal ARIMA model, we forecast the expected house sales for the next 12 months, starting from October 2022 until September 2023.

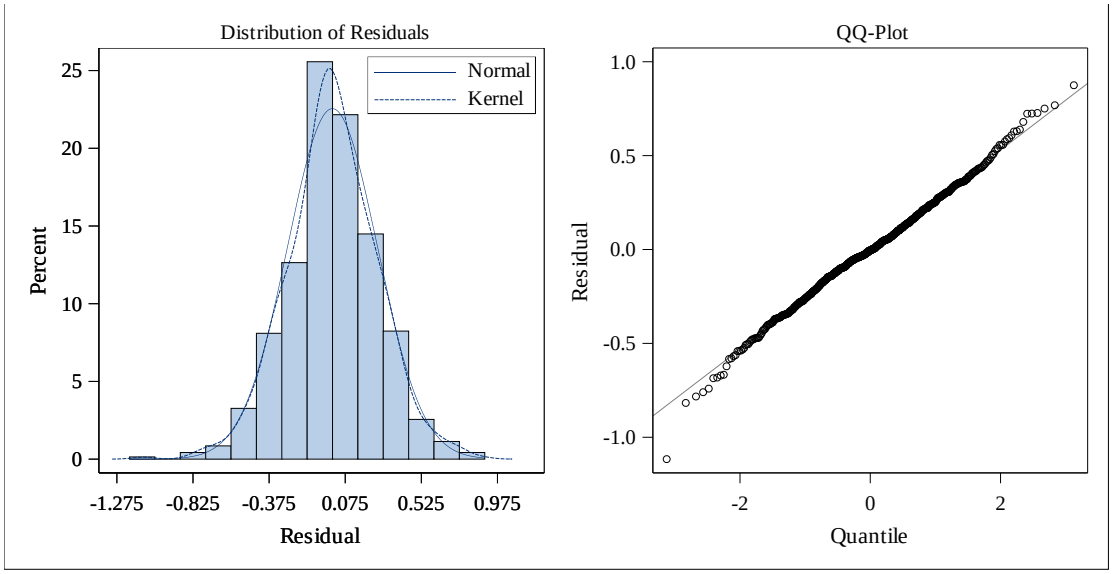


Figure 12 Residual Normality Diagnostics – Model ARIMA(0,1,1)x(0,1,1)$_{12}$

As the residuals of the model are approximately normally distributed (Figure 12), the unbiased forecasts and the 95% confidence intervals are as follows:

Forecasts for variable z				
Observation	**Forecast**	**Std Error**	**95% Confidence Limits**	
October 22	6.5327	0.2399	6.0624	7.0029
November 22	6.3319	0.3034	5.7373	6.9265
December 22	6.3684	0.3557	5.6713	7.0655
January 23	6.7544	0.4012	5.9681	7.5407
February 23	6.8952	0.4421	6.0288	7.7616
March 23	7.1196	0.4795	6.1799	8.0594
April 23	6.8599	0.5142	5.8522	7.8676
May 23	6.8472	0.5466	5.7759	7.9186

Forecasts for variable z				
Observation	**Forecast**	**Std Error**	**95% Confidence Limits**	
June 23	6.7842	0.5773	5.6527	7.9157
July 23	6.6540	0.6064	5.4655	7.8426
August 23	6.6396	0.6342	5.3966	7.8826
September 23	6.5629	0.6608	5.2677	7.8580

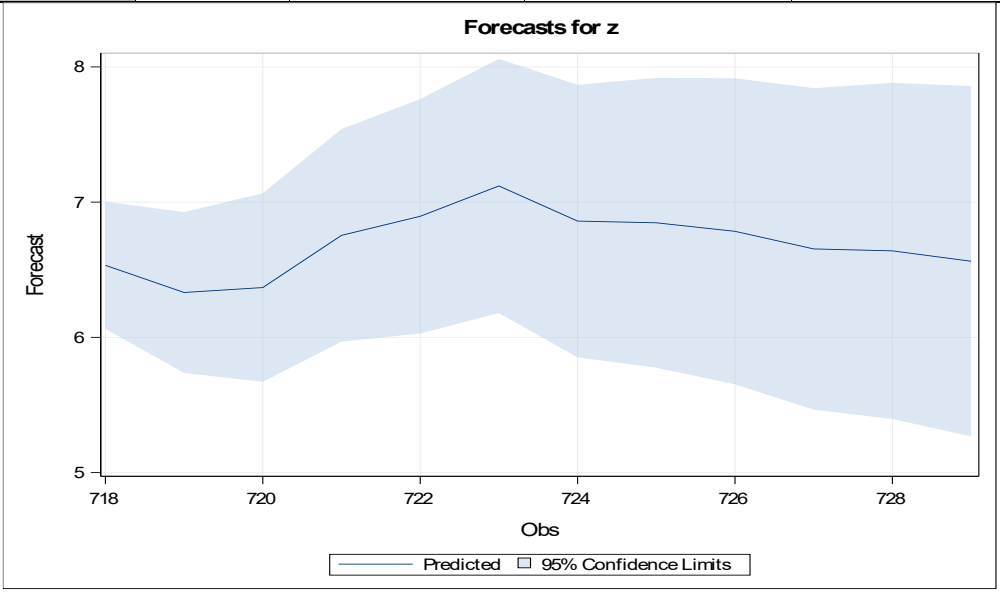


Figure 13 Forecasting results

Limitations of the Forecasts

It is important to note that the forecasts of the model are performed for the transformed times series $Z_t = \frac{Y_t^{0.25}-1}{0.25}$. We can obtain the forecasts for the original series by solving the above equation for Y_t and returning to the original scale of thousands of houses sold per month, but then the corresponding 95% confidence intervals will be biased. Transforming back to the original series, the forecasts for Y_t can be seen in Figure 14. These forecasts give us estimates of the number of residential houses in thousands that will be sold in the U.S. in the next 12 months starting in October 22.

Observation	Forecasts for Original Series Y
October 22	48.075
November 22	44.512
December 22	45.145
January 23	52.252
February 23	55.043
March 23	59.720
April 23	54.333
May 23	54.079
June 23	52.834
July 23	50.328
August 23	50.057
September 23	48.629

.

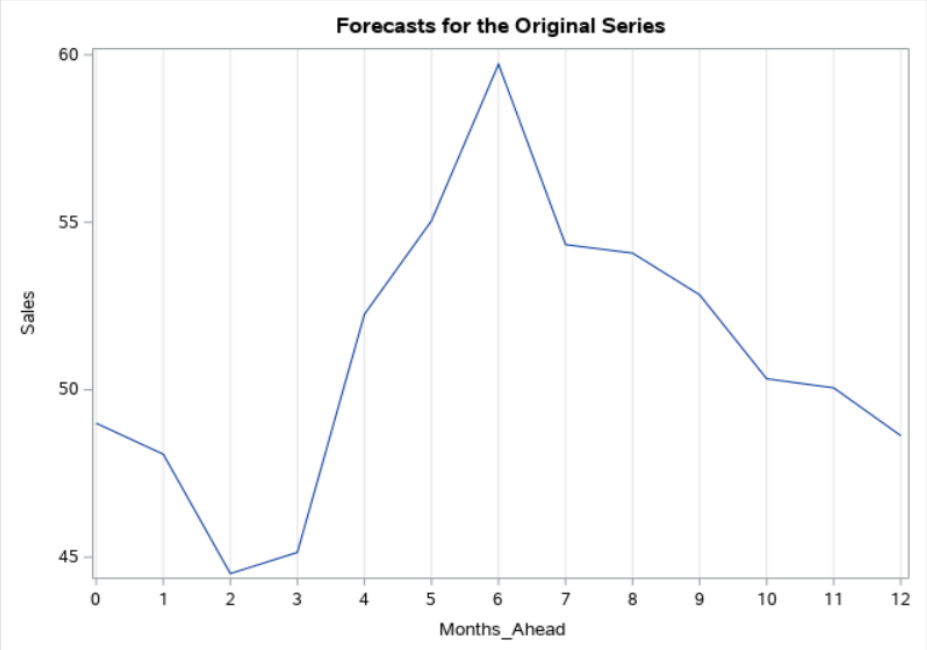


Figure 14 Results of forecasting for the number of houses sold

VI. Conclusions

Using the time series data of the monthly house sales , we performed time-domain time series analysis to find a good model for the data that can be used for forecasting. In the analysis, we performed a variance stabilizing transformation to the series and were able to identify non-stationarity and seasonality effects. After accounting for these features by differencing, we fitted 4 different models and finally chose an ARIMA$(0,1,1)\times(0,1,1)_{12}$ for forecasting. The forecasts show that housing sales is expected to increase from November 2022 to March 2023, and then decrease.

**Appended to this document are all the statistical computations and results in SAS as well as the code used to produce the above results.*