Analyzing Economy and Finance with Time Series Techniques: Seasonal and Non-Seasonal Dataset Analysis

Han Chau - 20012654
Department of Mathematical Sciences
Stevens Institute of Technology
Hoboken, NJ, U.S.
hchau@stevens.edu
Supervisor: Hadi Safari Katesari

I. INTRODUCTION

Time series analysis is a powerful method for examining patterns and trends within sequential data points over time. Unlike cross-sectional data, which provides a snapshot at a single point in time, time series data allows us to understand how variables evolve and interact over successive intervals. By analyzing time series data, we can gain insights into the dynamics of various phenomena, such as asset prices, economic indicators, and sales figures.

This project focuses on exploring and analyzing two distinct types of time series datasets: seasonal and non-seasonal. Seasonal datasets exhibit recurring patterns or fluctuations at regular intervals, such as monthly or yearly cycles. In contrast, non-seasonal datasets lack these recurring patterns and may exhibit trends, cycles, or irregular fluctuations.

To effectively analyze these datasets, it is essential to employ appropriate techniques tailored to each type. For seasonal datasets, models such as Seasonal ARIMA (SARIMA) or Multiplicative Seasonal ARIMA are well-suited to capture and forecast the seasonal variations. For instance, in this project, we examine a seasonal dataset consisting of monthly truck sales over a span of 12 years, starting from 2002.

On the other hand, non-seasonal datasets require different approaches to address their unique characteristics. While the traditional ARIMA model can be utilized, it may lead to issues such as volatility clustering. To mitigate this challenge, incorporating additional models like Autoregressive Conditional Heteroskedasticity (ARCH) or Generalized ARCH (GARCH) can enhance the accuracy of predictions and better capture the volatility dynamics. For example, we apply these techniques to analyze the price of garlic, a non-seasonal variable, throughout the year 2023.

Through this project, we aim to demonstrate the practical application of time series analysis techniques in understanding and forecasting both seasonal and non-seasonal datasets, providing valuable insights for decision-making and future planning.

II. METHODOLOGY

To begin, the dataset's stationarity is assessed. Should non-stationarity be detected, techniques like differencing, transformation, or detrending are employed to achieve stationarity.

Subsequently, parameters for potential models including ARIMA(p,d,q), SARIMA(p,d,q)x(P,D,Q)s, and ARCH(q)/GARCH(p,q) are identified. The suggested models are then compared using information criteria like Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) to determine the most suitable model.

Following model selection, residual analysis is conducted to evaluate the adequacy of the chosen model. Diagnostic tests such as the Ljung-Box statistic, autocorrelation function (ACF) plot of residuals, and Shapiro-Wilk test are performed to assess the independence, autocorrelation, and normality of residuals. Utilizing the selected model, future time points are forecasted, offering insights into the anticipated behavior of the time series.

For datasets exhibiting changing variance over time (heteroskedasticity), autoregressive conditional heteroskedasticity (ARCH) or generalized ARCH (GARCH) models are applied. These models capture correlated changes in variance, particularly relevant in financial time series analysis. In nonseasonal dataset analysis, after identifying the best ARIMA model, the presence of volatility clustering is examined through ACF/PACF plots of squared residuals. If present, alternative models such as ARIMA-ARCH/GARCH or standalone ARCH/GARCH models are considered. Returns are transformed for ARCH/GARCH modeling, typically involving differencing and applying logarithmic transformations to the data. Following ARCH/GARCH modeling, residual analysis is performed, and forecasts are generated for future time points. The model's performance is evaluated, considering its ability to capture changing volatility dynamics.

It's acknowledged that ARCH/GARCH models may not explicitly account for trend or seasonality. In cases where these components are present in the data, additional considerations or adjustments may be required.

III. EXPERIMENT

1. Seasonal Dataset

a. Data Analysis

The dataset sourced from Kaggle encompasses monthly sales figures for trucks from a particular company spanning the years 2003 to 2014 including 144 observations recording monthly.

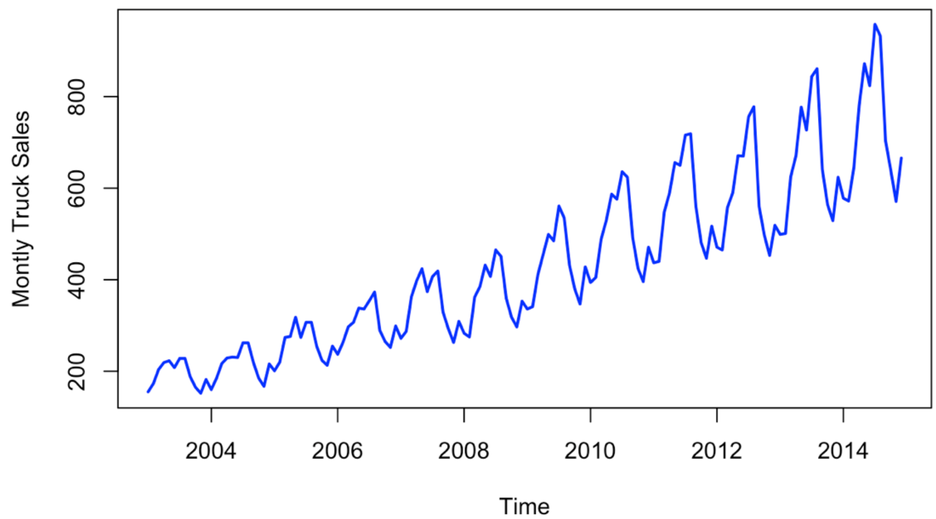

Figure 1: Plot the monthly sales for trucks from 2003 to 2014

Looking at the plot of this data, we observe a growth trend throughout each year, delineated by 12-month cycles. Notably, truck sales peak during the summer months, particularly in August, before declining during the winter season.

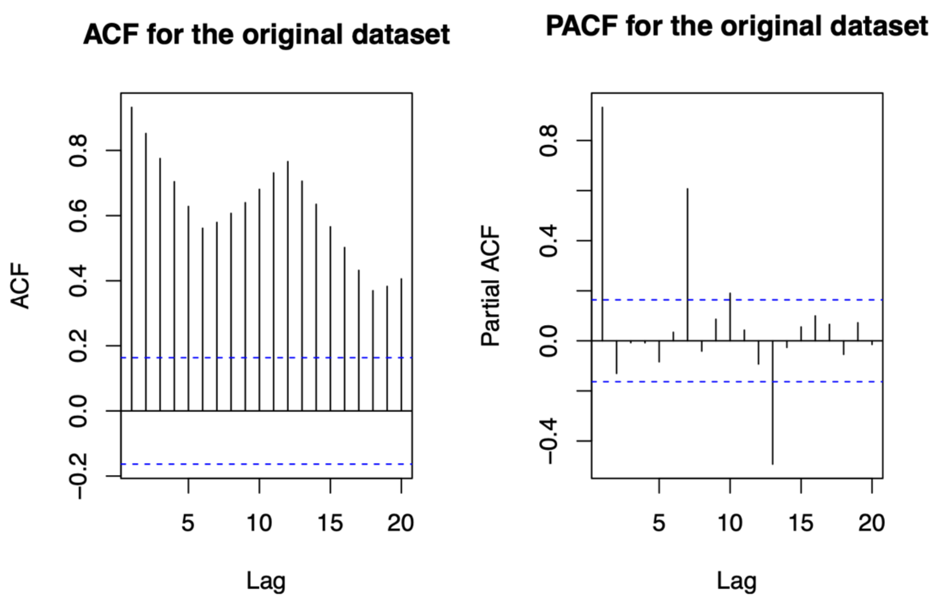

Figure 2: The ACF/PACF for the truck-sale dataset

The data exhibit a consistent upward trend, indicating that the mean is dependent on time (t). Additionally, the autocorrelation function (ACF) demonstrates a gradual decrease in lags, with an increase starting at lag 7. Consequently, this dataset is non-stationary, requiring the application of methods to achieve stationarity.

b. Models

Initially, we opted to achieve stationarity in the series by applying differencing. The Dickey-Fuller Test serves as a tool to ascertain whether the series attains stationarity. Considering the seasonal nature of the series, characterized by cycles occurring every 12 months, I specified the max.lag.y hyperparameter as 12 for the test. Consequently, the obtained p-value from this test is 0.09. Therefore, we fail to reject the null hypothesis, suggesting the presence of a unit root in an AR model and indicating that the data series remains non-stationary.

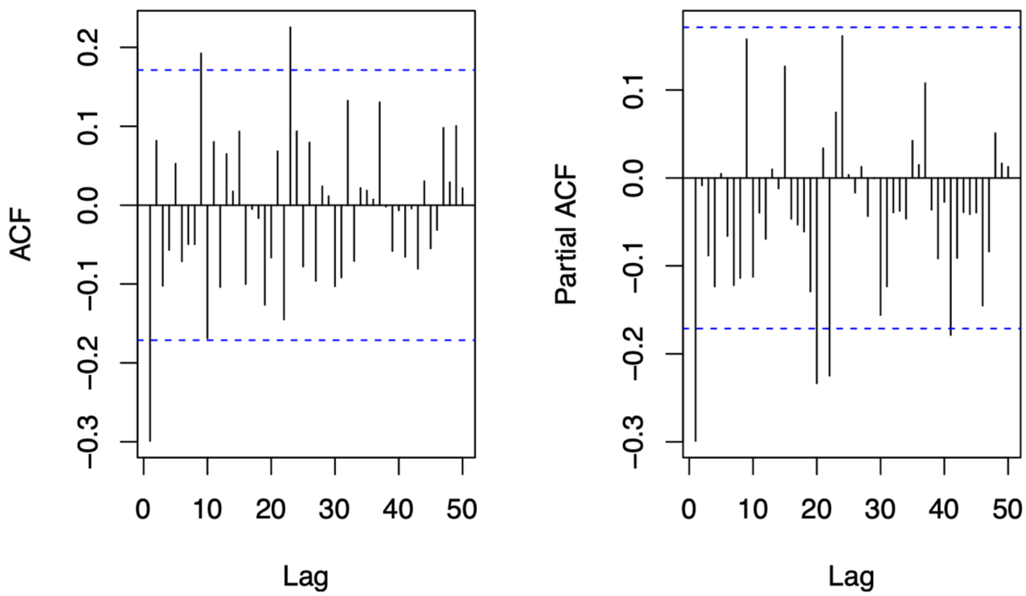

Figure 3: The ACF/PACF for the $\nabla_{12}\nabla Y_t$

When conducting the Dickey-Fuller Test for diff(diff(data), lag=12), we reject the null hypothesis. However, despite this rejection, the ACF and PACF plots do not provide clear indications, making it difficult to determine the exact number of parameters for the models. In light of this ambiguity, we opt to apply Variance stabilization through the Box-Cox transformation to the dataset. This transformation aims to identify the appropriate lambda value for adjustment, facilitating a more accurate modeling process.

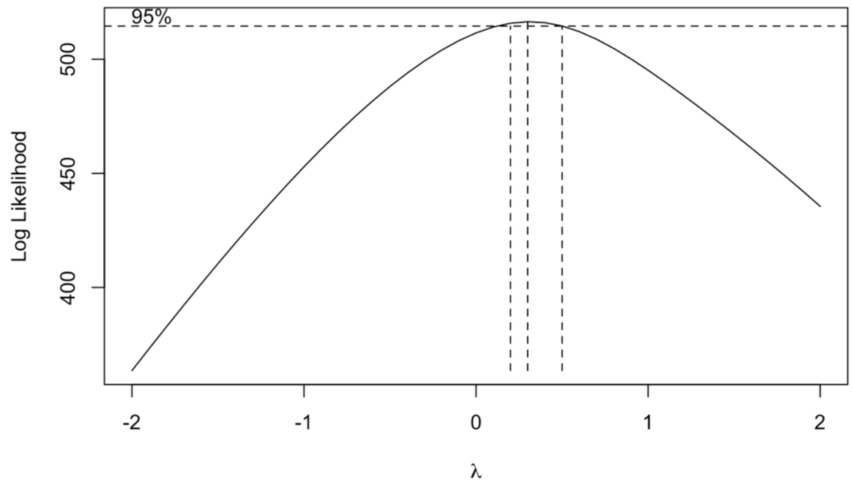

Figure 4: The plot of Box-Cox transformation

The Box-Cox transformation indicated $\lambda = 0.3$ as the optimal choice. While using the cube root of the data is an alternative, we must weigh the interpretability of the transformed variable and its suitability for subsequent analysis.

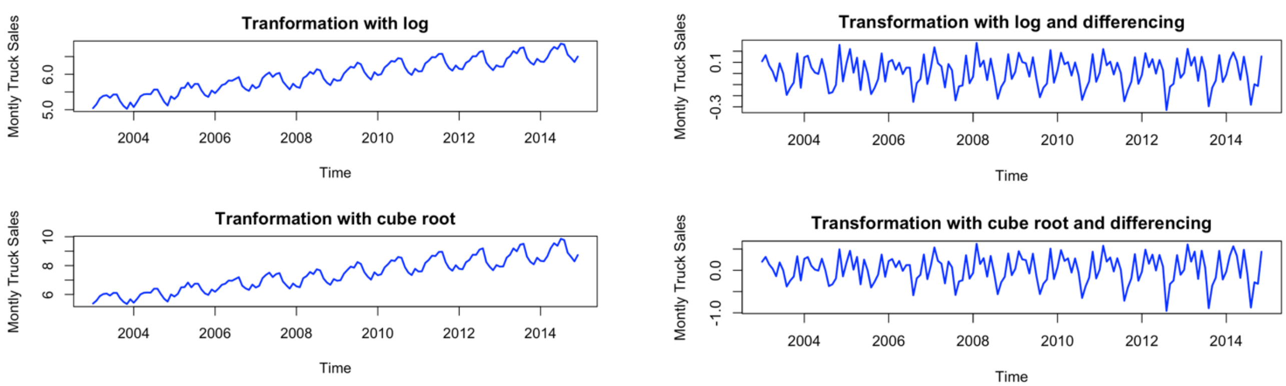

Figure 5: The plot comparing between log and cube root.

Based on the combined plot of transformation and differencing, it is evident that even with the cube root transformation, there is a variance dependence on t. Consequently, we opt for $\lambda = 0$ proceed with a log-transformation for each observation.

```
 ADF test

data:  seasonal_data
ADF(12) = -4.4032, p-value = 0.003176
alternative hypothesis: true delta is less than 0
sample estimates:
    delta
-2.220249
```

Based on the results of the ADF test, with a p-value of 0.05, we can reject the null hypothesis, indicating that the series is stationary. Furthermore, considering this dataset as a seasonal time series with a cycle of 12 months, we perform differencing with a lag of 12. The resulting p-value of 0.003 leads us to accept that this series is also stationary.

To determine the number of parameters, we go through the ACF and PACF plot.

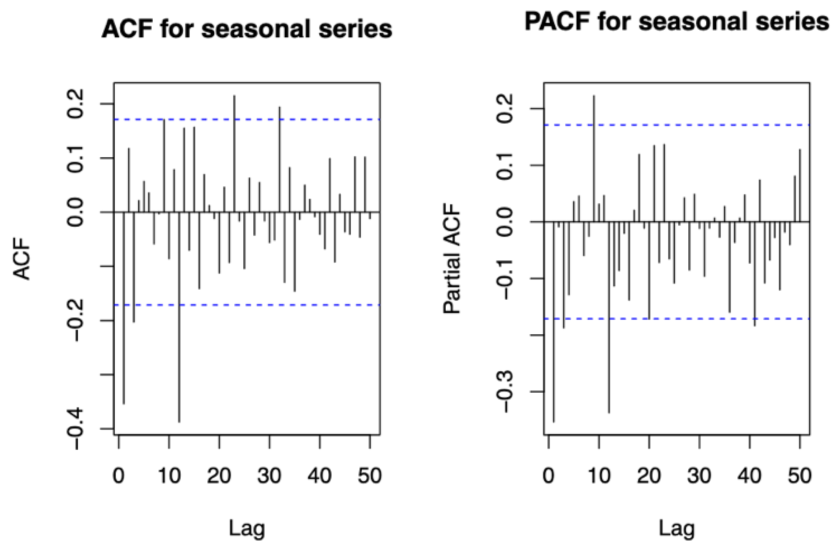

Figure 6: The ACF/PACF for the $\nabla_{12}\nabla\log(Y_t)$

Looking at these plots, we can discern the following:

- In the nonseasonal component, there is one significant lag in both the ACF and PACF plots, suggesting potential values for q (for MA) and p (for AR) models, including 0 and 1.
- In the seasonal component, a significant lag appears at lag 12 in both ACF and PACF plots, indicating possible values for Q (for seasonal MA) and P (for seasonal AR) models, also including 0 and 1.

Since the series exhibits both nonseasonal and seasonal components, we opt for a Multiplicative SARMA(p,d,q) × (P,D,Q)s model. Therefore, Consequently, we explore several potential models, such as SARIMA(0,1,1)x(1,1,1)[12], SARIMA(1,1,0)x(1,1,1)[12], SARIMA(1,1,1)x(1,1,1)[12], SARIMA(0,1,1)x(0,1,1)[12], SARIMA(1,1,0)x(0,1,1)[12], SARIMA(1,1,1)x(0,1,1)[12], SARIMA(0,1,1)x(1,1,0)[12], SARIMA(1,1,0)x(1,1,0)[12], and SARIMA(1,1,1)x(1,1,0)[12].

From the table of results, we observe the Akaike Information Criterion (AIC), log-likelihood, and Bayesian Information Criterion (BIC) values for different models. Lower values of AIC, log-likelihood, and BIC indicate better model fit and parsimony.

Index	Models	AIC	log.likelihood	BIC
<int>	<chr>	<dbl>	<dbl>	<dbl>
1	SARIMA(0,1,1)x(1,1,1)[12]	-482.3055	245.1527	-470.8047
2	SARIMA(1,1,0)x(1,1,1)[12]	-480.2373	244.1186	-468.7365
3	SARIMA(1,1,1)x(1,1,1)[12]	-480.5769	245.2884	-466.2009
4	SARIMA(0,1,1)x(0,1,1)[12]	-483.8448	244.9224	-475.2192
5	SARIMA(1,1,0)x(0,1,1)[12]	-482.0477	244.0239	-473.4221
6	SARIMA(1,1,1)x(0,1,1)[12]	-482.1830	245.0915	-470.6822
7	SARIMA(0,1,1)x(1,1,0)[12]	-477.9483	241.9742	-469.3227
8	SARIMA(1,1,0)x(1,1,0)[12]	-475.4768	241.9742	-466.8512
9	SARIMA(1,1,1)x(1,1,0)[12]	-475.9761	241.9881	-464.4753

Figure 7: The table for AIC, log likelihood, and BIC from 9 seasonal models

After analyzing the model results, two models stand out as the most promising candidates, these are model 4 and model 1. The first model, SARIMA(0,1,1)x(0,1,1)[12], exhibits a relatively low AIC of -483.8448 and BIC of -475.2192, suggesting a good balance between model fit and complexity. The second model, SARIMA(0,1,1)x(1,1,1)[12], also demonstrates strong performance with an AIC of -482.3055 and BIC of -470.8047. However, it's worth noting that the value of log-likelihood does not consistently align with the trend of AIC and BIC. Thus, we can disregard this value in our evaluation. These two models appear as top contenders based on their performance metrics, warranting further evaluation to determine the optimal choice for forecasting purposes.

c. Residual Analysis:

Model diagnostics involve assessing the goodness of fit of a model and suggesting appropriate modifications if the fit is inadequate. Residuals play a crucial role in evaluating whether a model effectively captures the information present in the data. These residuals represent the discrepancy between the observed values and the corresponding predicted values, calculated as residuals = observed - predicted. An essential step in diagnostics is to ensure that the standardized residuals adhere to the characteristics of standard Normal White Noise, typically verified using the Shapiro-Wilk normality test. Subsequently, plotting the residuals allows for the assessment of significant residual autocorrelation (via Sample ACF and PACF), normality (via Q-Q plots and histograms), and consistent variance. Furthermore, the Ljung-Box test serves as a common method for testing against residual autocorrelation, evaluating the null hypothesis that a series of residuals demonstrates no autocorrelation.

From the aforementioned models, we have selected two candidates for residual analysis: Model 4, SARIMA(0,1,1)x(0,1,1)[12], and Model 1, SARIMA(0,1,1)x(1,1,1)[12].

- Initially, we plot the residuals against the processed seasonal data for Model 4.

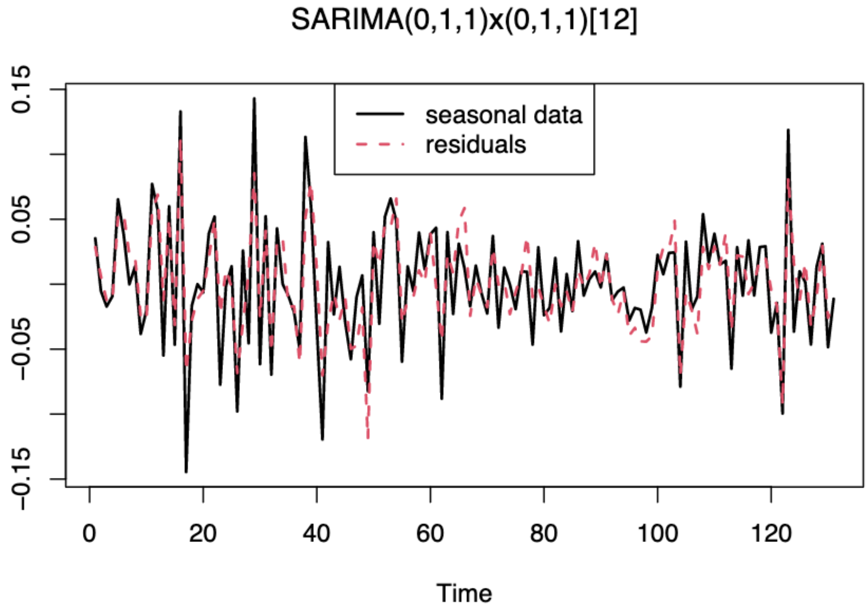

Figure 8: The plot for actual $\nabla_{12}\nabla\log(Y_t)$ and the residual of Model 4

The resulting plot indicates that the residuals exhibit a mean of 0 and display consistent variance. Moreover, the shape of the residuals closely resembles that of the processed seasonal data.

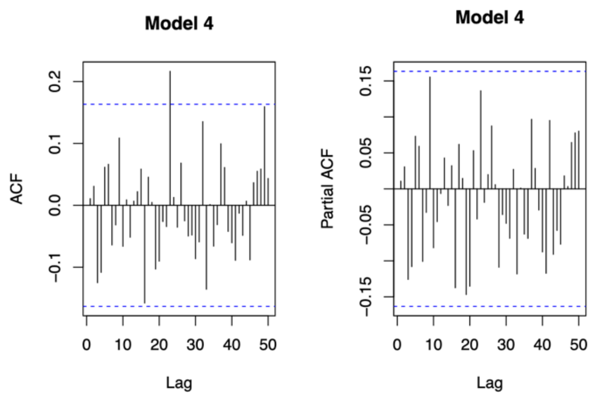

Figure 9: The ACF/PACF plot of the residual of Model 4

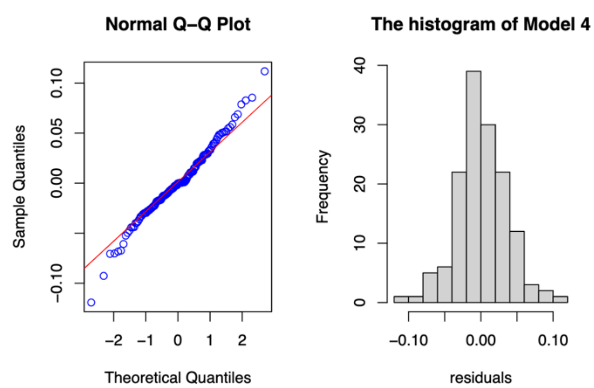

Figure 10: The plot of Normal Q-Q and the histogram in Model 4

Upon examination of the ACF plot spanning 50 lags, the majority of the autocorrelation coefficients fall within the confidence interval, indicating no significant autocorrelation between the residuals. However, a notable exception occurs at lag 23, where a deviation from the confidence interval is observed. Despite this isolated occurrence, the overall pattern suggests that the residuals conform well to the assumptions of the model. As a result, we can confidently accept this model as a suitable representation of the data. The normality of residuals can be inferred from Q-Q plots and histograms. Upon conducting the Shapiro-Wilk normality test, we obtained a p-value of 0.1565, leading us to accept the null hypothesis, indicating that the residuals adhere to a normal distribution.

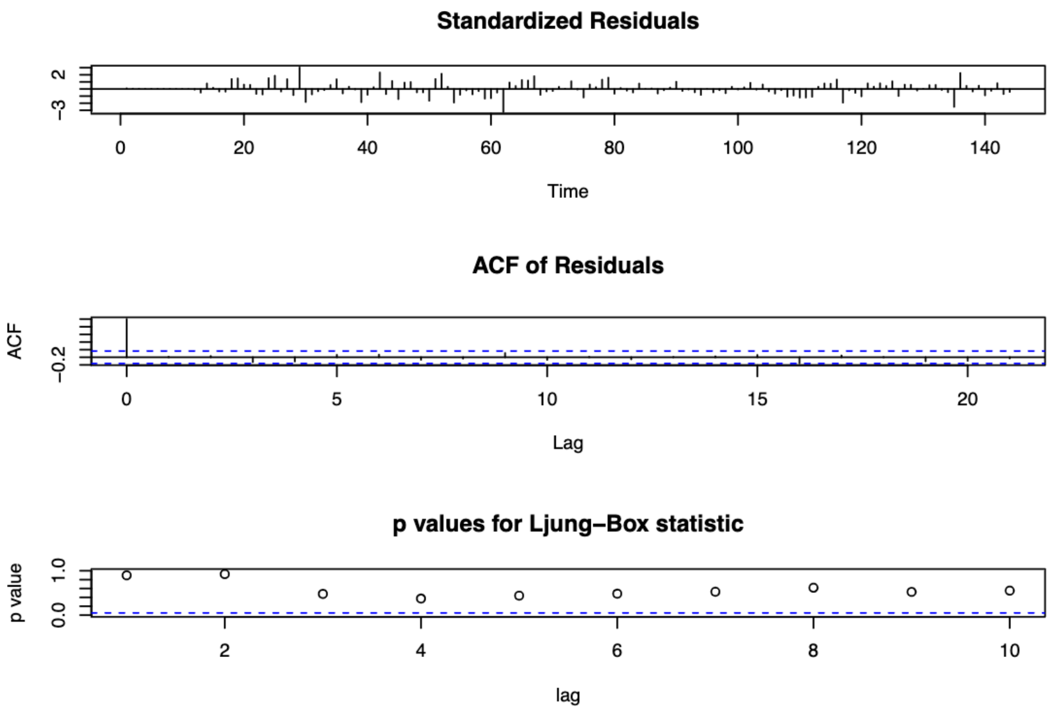

Figure 11: Residual Analysis for Model 4

Based on the Ljung-Box statistic derived from the plot above, we observe that all lags exceed 0.05. Therefore, we accept the null hypothesis of no autocorrelation.

- We will perform a similar check for Model 1:

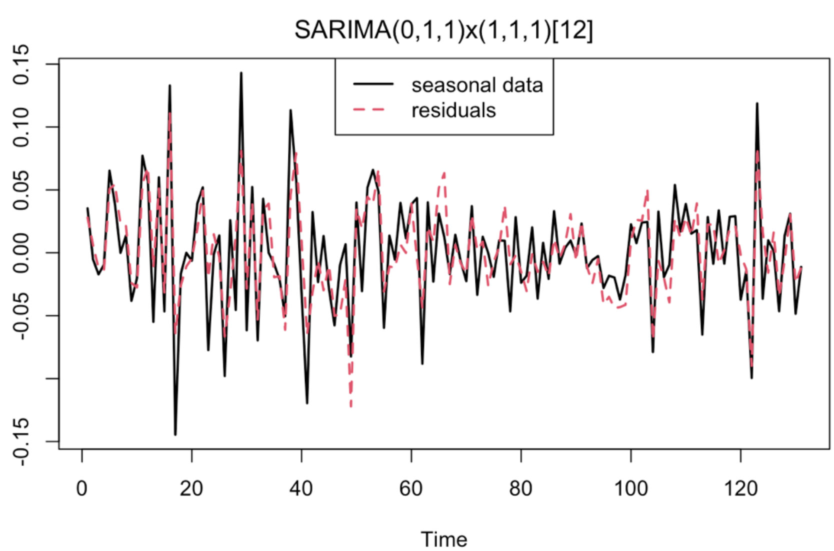

Figure 12: The plot for actual $\nabla_{12}\nabla\log(Y_t)$ and the residual of Model 1

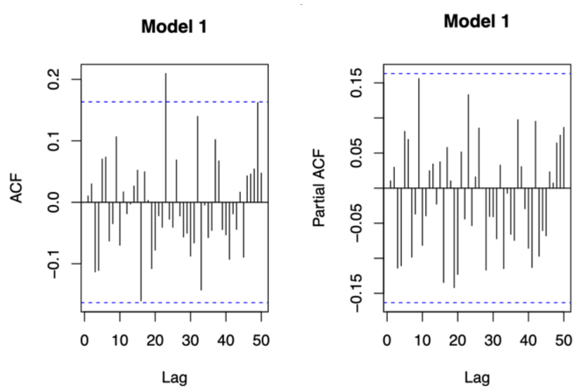

Figure 13: The ACF/PACF plot of the residual of Model 4

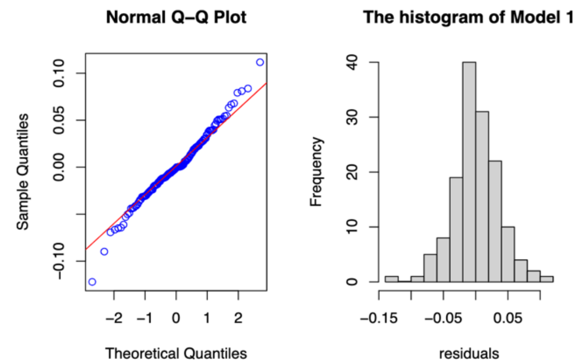

Figure 14: The plot of Normal Q-Q and the histogram in Model 4

The residuals plot in this model also exhibits a mean of 0 and consistent variance. However, there is a slight difference observed in the ACF plot. In addition to lag 23 extending beyond the confidence interval, lag 49 touches the interval as well. Regarding the Shapiro-Wilk normality test, this model also yields a p-value higher than the confidence interval, leading us to accept the null hypothesis. However, in the histogram, the shape appears to slightly deviate from normality compared to Model 4.

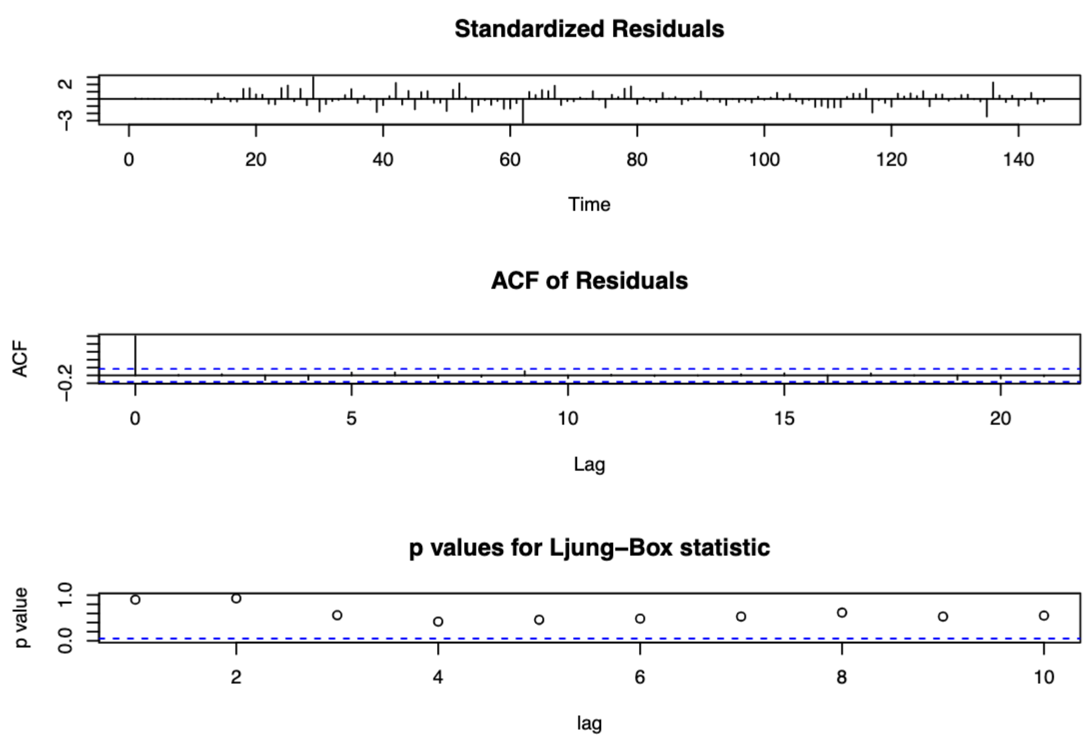

Figure 15: Residual Analysis for Model 1

Finally, based on the Ljung-Box statistic derived from the plot of Model 1, we observe that all lags exceed 0.05. Therefore, we also accept the null hypothesis of no autocorrelation. Considering the characteristics observed and the number of parameter, we suggest that Model 4 is the most suitable choice for forecasting purposes.

d. Forecasting

One of the primary objectives of constructing a time series model is to forecast future values of the series accurately. Equally important is the evaluation of the precision of these forecasts. To accomplish this step, we employ the forecast() function to generate forecasts from the model. We will set the hyperparameter with a 95% confidence interval and predict for the next 2 cycles based on the previous 12 cycles.

	Point Forecast <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
Jan 2015	625.0779	581.3695	672.0723
Feb 2015	623.5220	573.1458	678.3260
Mar 2015	738.4597	671.8389	811.6868
Apr 2015	835.2940	752.9127	926.6893
May 2015	941.9772	841.8695	1053.9889
Jun 2015	899.4354	797.5079	1014.3900
Jul 2015	1033.2005	909.3335	1173.9404
Aug 2015	1028.5213	898.8828	1176.8566
Sep 2015	774.1628	672.0854	891.7438
Oct 2015	689.9296	595.1549	799.7966
Nov 2015	629.5360	539.7509	734.2565
Dec 2015	736.0349	627.3647	863.5286
Jan 2016	688.5002	577.0194	821.5192
Feb 2016	686.7865	569.5080	828.2161
Mar 2016	813.3861	667.7502	990.7851
Apr 2016	920.0455	748.1303	1131.4656
May 2016	1037.5531	836.0116	1287.6812
Jun 2016	990.6948	791.2922	1240.3461
Jul 2016	1138.0322	901.3417	1436.8771
Aug 2016	1132.8782	889.9876	1442.0573
Sep 2016	852.7117	664.6363	1094.0076
Oct 2016	759.9319	587.8175	982.4420
Nov 2016	693.4107	532.4023	903.1110
Dec 2016	810.7152	617.9981	1063.5294

24 rows

Figure 16: Forecasting the next 2 years by Model 4

From the above table, it's evident that the points forecasted fall within the range defined by the low 95% and high 95% interval. This suggests that the forecasts have captured the uncertainty in the data and provide a reasonable range of potential outcomes. The narrow spread between the low and high intervals indicates a relatively high level of confidence in the forecasted values.

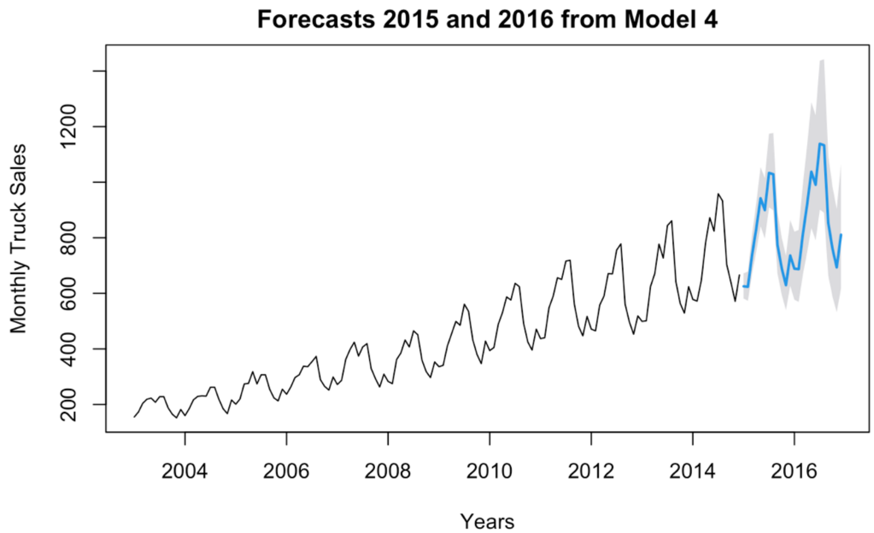

Figure 16: The plot of the original dataset and forecasting the next 2 years

Upon examining the forecast plot for 2015 and 2016 generated by Model 4, it becomes evident that the shape of each cycle closely resembles that of the original data. This indicates that the model has effectively captured the underlying patterns and seasonal fluctuations present in the data. Moreover, the consistency between the forecasted and observed shapes enhances the confidence in the model's predictive capability. It suggests that the model has successfully replicated the temporal dynamics of the series and provides reliable forecasts for future time periods.

2. Nonseasonal dataset

a. Data Analysis

This dataset, obtained from Kaggle and collected from an authorized source (https://agmarknet.gov.in/), provides the prices of a wide range of vegetables in the year 2023. For this project, the price of garlic has been selected for analysis to determine its future trend.

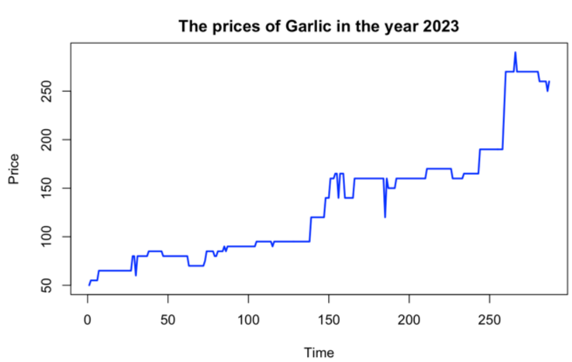

Figure 17: The plot of the price of Garlic in the year 2023

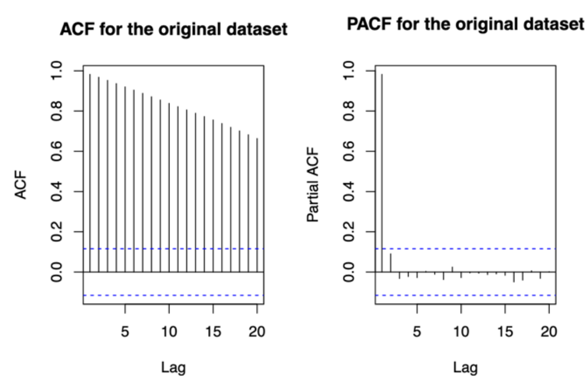

Figure 18: The ACF/PACF for the price of garlic dataset

Based on the ACF plot, it appears to exhibit slow decay or persistence over many lags, indicating the presence of a unit root or a trend. However, the PACF plot either cuts off after 1 lag or decreases to zero relatively quickly following the exponential sin, suggesting non-stationarity. To ensure accuracy, further verification with the Dickey-Fuller Test is recommended. We obtained a test statistic of -1.8182 with a corresponding p-value of 0.5243. With a significance level typically set at 0.05, the p-value exceeds this threshold, indicating insufficient evidence to reject the null hypothesis. The null hypothesis states that the series has a unit root or is non-stationary. Therefore, based on these results, we fail to reject the null hypothesis, suggesting that the data likely possesses a unit root or exhibits non-stationarity.

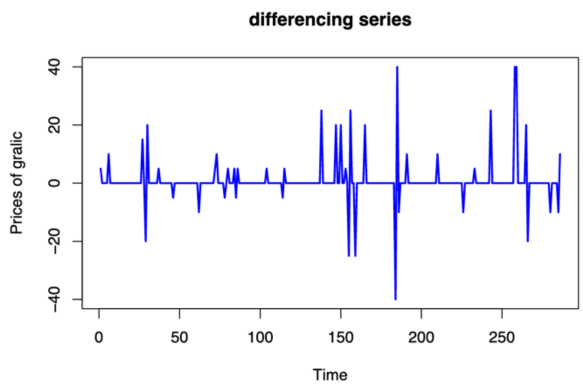

Figure 19: The plot of the differencing price of Garlic in the year 2023

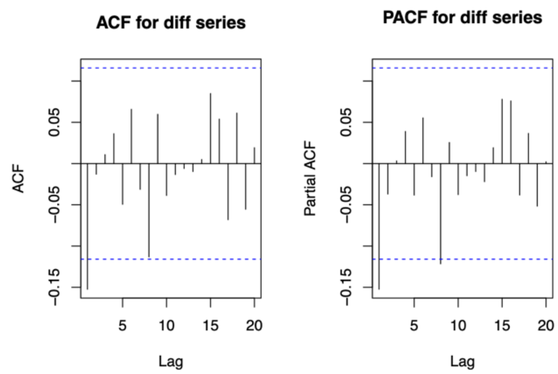

Figure 20: The ACF/PACF for the differencing series

After applying first differencing method to the price series, we conducted another Augmented Dickey-Fuller (ADF) test. The results show a significant improvement in stationarity, with a p-value much lower than the commonly used significance level of 0.05 (p-value < 2.2e-16). Consequently, we reject the null hypothesis in favor of the alternative hypothesis, suggesting that the differenced series is stationary. The estimated delta value is -1.2, indicating a strong indication of stationarity. This transformation indicates that the original series had a unit root or non-stationarity, which was successfully addressed through differencing, ensuring that the data is now suitable for further time series analysis. In both the ACF and PACF plots, we observe that they both cut off after lag 1. Hence, we have decided to consider values of 0 or 1 for both p and q in the ARIMA model selection.

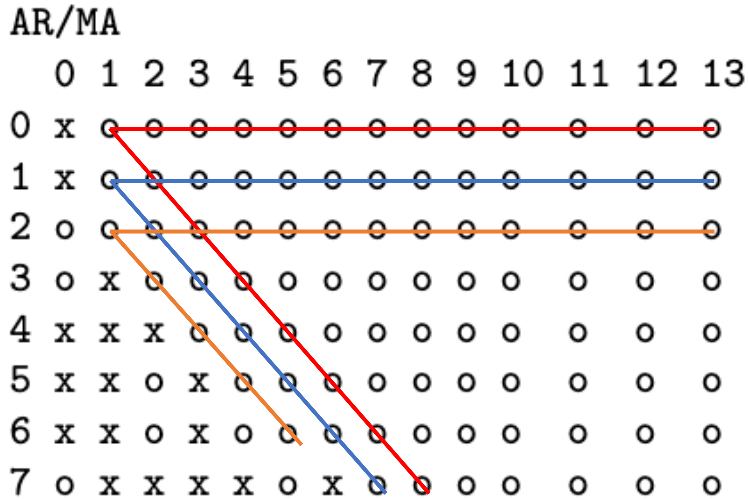

Figure 21: EACF of the differencing series

Additionally, in the Extended Autocorrelation Function (EACF), several triangles emerge, notably with a vertex at (0,1), (1,1), and (2,1). As a result, potential models include ARIMA(1,1,0), ARIMA(1,1,1), ARIMA(0,1,1), and ARIMA(2,1,1). These model specifications are derived from the observed patterns in autocorrelation and partial autocorrelation functions, providing a solid foundation for further time series analysis and forecasting.

Index <int>	Models <chr>	AIC <dbl>	log.likelihood <dbl>	BIC <dbl>
1	ARIMA(1,1,0)	1918.641	−958.3203	1927.953
2	ARIMA(0,1,1)	1918.559	−958.2793	1927.871
3	ARIMA(1,1,1)	1920.547	−958.2734	1933.515
4	ARIMA(2,1,1)	1922.497	−958.2485	1939.121

Figure 22: The table for AIC, log likelihood, and BIC from ARIMA models

Comparing the models, we see that ARIMA(0,1,1) has the lowest AIC (1918.559) and BIC (1927.871) values, indicating a good balance between model fit and complexity. Following this, ARIMA(1,1,0) possesses the next lowest AIC (1918.641) and BIC (1927.953), suggesting its competitiveness as well. Hence, ARIMA(0,1,1) and ARIMA(1,1,0) come out as the two best candidate models due to their lower AIC and BIC values. Further scrutiny through residual analysis and model validation would be prudent to confirm their suitability for forecasting.

b. Addressing Heteroskedasticity:

While the ARIMA model accounts for autocorrelation in the data, it may not adequately address volatility clustering or heteroscedasticity in the residuals. In this case, we consider employing ARCH (Autoregressive Conditional Heteroscedasticity) or GARCH (Generalized Autoregressive Conditional Heteroscedasticity) models to capture any remaining volatility patterns in the residuals. To determine whether ARCH or GARCH models are appropriate, we examine the ACF and PACF plots of the squared residuals to detect significant autocorrelations. If these plots display significant autocorrelations, it suggests that garlic prices are not independently and identically distributed, providing evidence for the need to employ ARCH or GARCH models.

- ❖ *ARIMA-ARCH/GARCH*

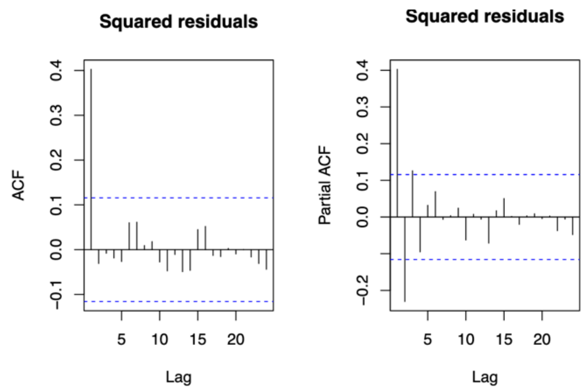

Figure 23: The ACF/PACF for the squared residual of ARIMA(0,1,1)

AR/MA

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	o	o	o	o	o	o	o	o	o	o	o	o	o
1	x	x	o	o	o	o	o	o	o	o	o	o	o	o
2	x	x	o	o	o	o	o	o	o	o	o	o	o	o
3	x	x	o	o	o	o	o	o	o	o	o	o	o	o
4	x	x	x	o	x	o	o	o	o	o	o	o	o	o
5	x	x	x	o	x	x	o	o	o	o	o	o	o	o
6	o	o	x	o	x	o	x	o	o	o	o	o	o	o
7	x	o	o	x	x	o	x	o	o	o	o	o	o	o

Figure 24: EACF for the squared residual of ARIMA(0,1,1)

We observe that the plots of ACF and PACF for the squared residuals of the ARIMA(0,1,1) model exhibit some significant lags outside the confidence interval. This suggests the presence of autocorrelation in the squared residuals, indicating potential volatility clustering or heteroscedasticity. Consequently, considering ARCH or GARCH models for modeling the volatility dynamics may be appropriate. Based on the analysis of the plots, potential models such as GARCH(1,1), GARCH(2,1), and GARCH(3,1) are suggested. However, it's noted that the lag 3 of the PACF lies just outside the boundary without significant deviation. To further explore the models, focus will be on evaluating the GARCH(1,1) and GARCH(2,1) specifications. Furthermore, the Extended Autocorrelation Function (EACF) reveals several distinctive triangles pattern at the various vertexes such as (0,1), (1,2), and (3,2). Therefore, we design 10 models by combined ARIMA and ARCH or GARCH such as ARIMA(0,1,1)-GARCH(1,1), ARIMA(0,1,1)-GARCH(2,1), ARIMA(0,1,1)-GARCH(1,2), ARIMA(0,1,1)-GARCH(3,2), ARIMA(0,1,1)-ARCH(1), ARIMA(1,1,0)-GARCH(1,1), ARIMA(1,1,0)-GARCH(2,1), ARIMA(1,1,0)-GARCH(1,2), ARIMA(1,1,0)-GARCH(3,2), and ARIMA(1,1,0)-ARCH(1).

Index <int>	Models <chr>	AIC <dbl>
1	ARIMA(0,1,1)–GARCH(1,1)	1863.354
2	ARIMA(0,1,1)–GARCH(2,1)	1857.060
3	ARIMA(0,1,1)–GARCH(1,2)	1858.371
4	ARIMA(0,1,1)–GARCH(3,2)	1854.945
5	ARIMA(0,1,1)–ARCH(1)	1855.659
6	ARIMA(1,1,0)–GARCH(1,1)	1863.712
7	ARIMA(1,1,0)–GARCH(2,1)	1857.422
8	ARIMA(1,1,0)–GARCH(1,2)	1858.705
9	ARIMA(1,1,0)–GARCH(3,2)	1855.276
10	ARIMA(1,1,0)–ARCH(1)	1856.057

Figure 25: The table for AIC from ARIMA-ARCH/GARCH models

From analyzing AIC values, the ARIMA(0,1,1)-GARCH(3,2) model emerges as the optimal choice with an AIC score of 1854.945. Following closely are the ARIMA(1,1,0)-GARCH(3,2) and ARIMA(0,1,1)-ARCH(1) models, with AIC scores of 1855.276 and 1855.659, respectively.

```
Coefficient(s):
    Estimate  Std. Error  t value Pr(>|t|)
a0 3.518e+01   9.338e+01    0.377 0.706406
a1 1.461e-01   4.065e-02    3.593 0.000326 ***
a2 9.726e-13   3.819e-01    0.000 1.000000
b1 2.709e-02   2.570e+00    0.011 0.991592
b2 3.162e-02   1.378e-01    0.229 0.818518
b3 2.897e-02   1.533e-01    0.189 0.850119
```

```
Coefficient(s):
    Estimate  Std. Error  t value Pr(>|t|)
a0 3.521e+01   9.398e+01    0.375 0.707930
a1 1.462e-01   4.074e-02    3.588 0.000334 ***
a2 5.930e-13   3.838e-01    0.000 1.000000
b1 2.686e-02   2.585e+00    0.010 0.991707
b2 3.140e-02   1.379e-01    0.228 0.819910
b3 2.877e-02   1.520e-01    0.189 0.849890
```

```
Coefficient(s):
    Estimate  Std. Error  t value Pr(>|t|)
a0  34.17805     0.86663   39.438  < 2e-16 ***
a1   0.17715     0.05424    3.266  0.00109 **
```

Figure 26: The coefficients for ARIMA(0,1,1)-GARCH(3,2), ARIMA(1,1,0)-GARCH(3,2), and ARIMA(0,1,1)-ARCH(1)models

However, a deeper examination of the model coefficients reveals noteworthy insights. While the first two models exhibit only one significant coefficient, all coefficients of the ARIMA(0,1,1)-ARCH(1) model are deemed significant. This signifies a robustness and reliability in the parameter estimates of the ARIMA(0,1,1)-ARCH(1) model compared to its counterparts. Moreover, although the difference in AIC values among the models is relatively small, the ARIMA(0,1,1)-ARCH(1) model stands out due to its parsimony, possessing fewer parameters than the other models under consideration.

Therefore, considering both the significance of coefficients and the model's parsimony, we conclude that the ARIMA(0,1,1)-ARCH(1) model is the most suitable candidate for further analysis, including residual analysis and forecasting. This selection ensures a balance between model complexity and explanatory power, facilitating meaningful insights into the underlying dynamics of the time series data.

- *ARCH/GARCH*

To tackle volatility variance, we exclusively employ ARCH/GARCH models. This process begins by transforming the original dataset through natural logarithm application, followed by computing the differences between consecutive observations, called return. Next, we employ ACF, PACF, and EACF analyses to find out the suitable number of parameters for the ARCH or GARCH model.

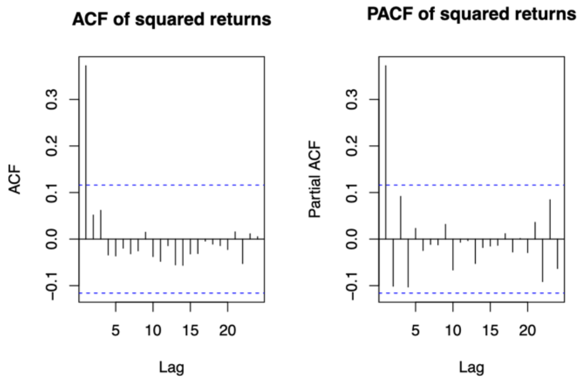

Figure 23: The ACF/PACF of the squared return

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o o o  o  o  o
1 x x o o o o o o o o o  o  o  o
2 x o x o o o o o o o o  o  o  o
3 x x x x o o o o o o o  o  o  o
4 x x x o o o o o o o o  o  o  o
5 x x o o o o o o o o o  o  o  o
6 x o o o o o o o o o o  o  o  o
7 x o x o o x o o o o o  o  o  o
```

Figure 24: EACF the squared return

From the ACF and PACF plots, we observe only one significant lag. This suggests potential models such as GARCH(1,1), ARCH(1), and GARCH(1,0). Additionally, the EACF reveals a triangle at vertex (0,1), (1,2), (2,3), (3,4), and (4,3). Based on these findings, we consider seven candidate models: GARCH(1,1), GARCH(1,0), GARCH(1,2), GARCH(2,3), GARCH(3,4), GARCH(4,3), and ARCH(1).

Index <int>	Models <chr>	AIC <dbl>
1	GARCH(1,1)	-930.0416
2	GARCH(1,0)	-870.9756
3	GARCH(1,2)	-926.0122
4	GARCH(2,3)	-915.0952
5	GARCH(3,4)	-904.8684
6	GARCH(4,3)	-904.1128
7	ARCH(1)	-920.3315

Figure 25: The table for AIC from ARCH/GARCH models

```
Call:
garch(x = return_value_garch, order = c(1, 1))

Model:
GARCH(1,1)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8719 -0.1350 -0.1348 -0.1209  5.3358

Coefficient(s):
    Estimate  Std. Error  t value Pr(>|t|)
a0 0.0012877   0.0001733    7.430 1.09e-13 ***
a1 0.1960170   0.0618172    3.171  0.00152 **
b1 0.2902501   0.0908966    3.193  0.00141 **
```

Figure 26: The coefficients of GARCH(1,1)

The table presented above indicates that the GARCH(1,1) model stands out as the preferred option, showcasing the lowest AIC value. This finding suggests that the GARCH(1,1) model strikes the optimal balance between model complexity and goodness of fit. Furthermore, all parameters associated with this model demonstrate statistical significance, further bolstering its suitability for the analysis.

c. Residual Analysis:

❖ *ARIMA(0,1,1)-ARCH(1):*

Upon analyzing the ACF and PACF plot of the squared residuals obtained from the ARIMA-ARCH model, several noteworthy observations emerge. Firstly, the absence of significant lags extending beyond the confidence interval in both plots suggests that there is no systematic pattern of autocorrelation present in the squared residuals. This indicates that the squared residuals exhibit a random fluctuation pattern over time, without any discernible trend or cyclic behavior.

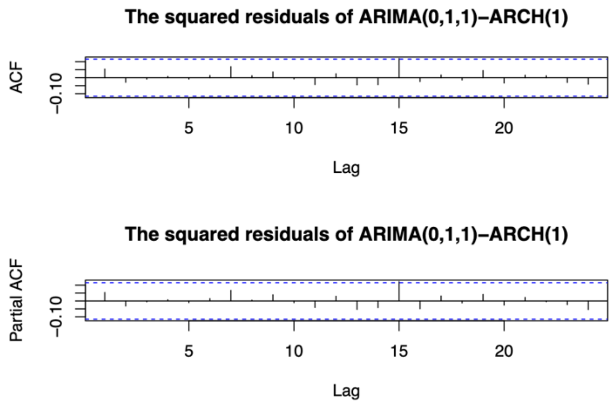

Figure 27: The ACF/PACF of the squared residuals from ARIMA-ARCH model

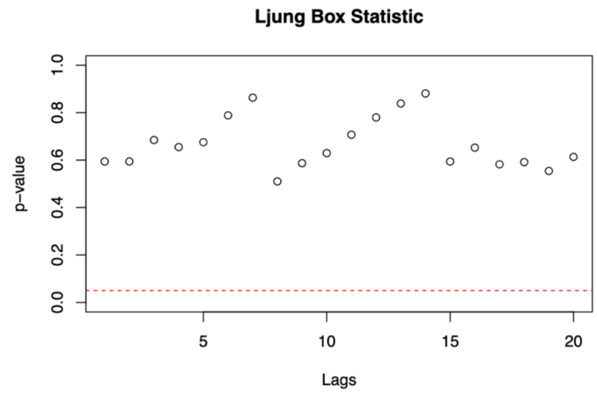

Figure 28: The plot of p-value from Ljung Box statistic for ARIMA-ARCH model

Furthermore, the Ljung-Box statistic, which assesses the overall autocorrelation in the squared residuals, yields p-values exceeding the conventional significance level of 5%. This implies that the squared residuals are statistically uncorrelated across different time periods. In other words, the fluctuations observed in the squared residuals do not depend on previous observations, reinforcing the notion of randomness and lack of serial correlation.

The absence of autocorrelation in the squared residuals suggests that the ARCH component adequately captures the volatility clustering present in the data. This indicates that the model effectively accounts for the changing variance over time, providing a robust framework for forecasting future volatility patterns. Additionally, the independence of the standardized residuals implies that the model's predictions are unbiased and reliable, as they are not influenced by past observations.

- *GARCH(1,1):*

Simillarly to ARIMA(0,1,1)-ARCH(1) model, the ACF and PACF plots for the squared residuals of the GARCH(1,1) model suggests that the model effectively captures the volatility patterns present in the data. This absence signifies that the GARCH(1,1) process adequately accounts for any systematic patterns or dependencies, resulting in squared residuals devoid of significant autocorrelation.

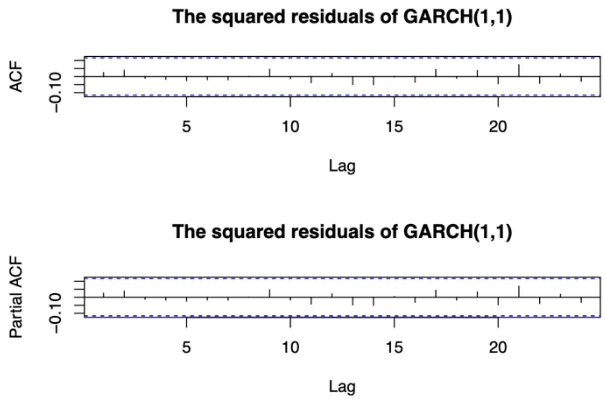

Figure 29: The ACF/PACF of the squared residuals from GARCH model

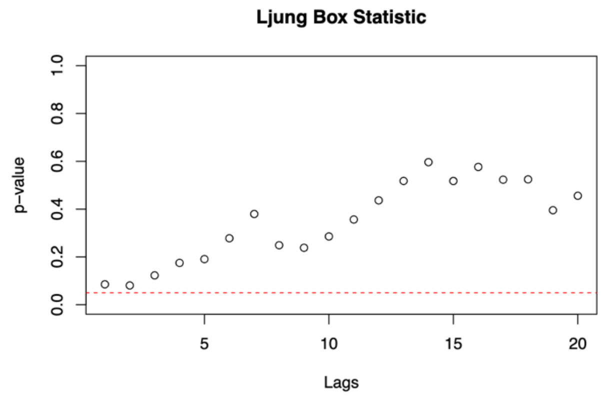

Figure 30: The plot of p-value from Ljung Box statistic for GARCH model

Additionally, the p-values obtained from the Ljung-Box statistic exceed the conventional significance level of 5%. This further corroborates the lack of significant autocorrelation in the squared residuals, providing additional evidence of the model's adequacy in capturing the underlying volatility dynamics.

These observations highlight the robustness of the GARCH(1,1) model in modeling volatility patterns and generating reliable forecasts. By effectively addressing volatility clustering and capturing the conditional heteroskedasticity in the data, this model offers valuable insights for risk management and decision-making in various fields, particularly in finance and economics.

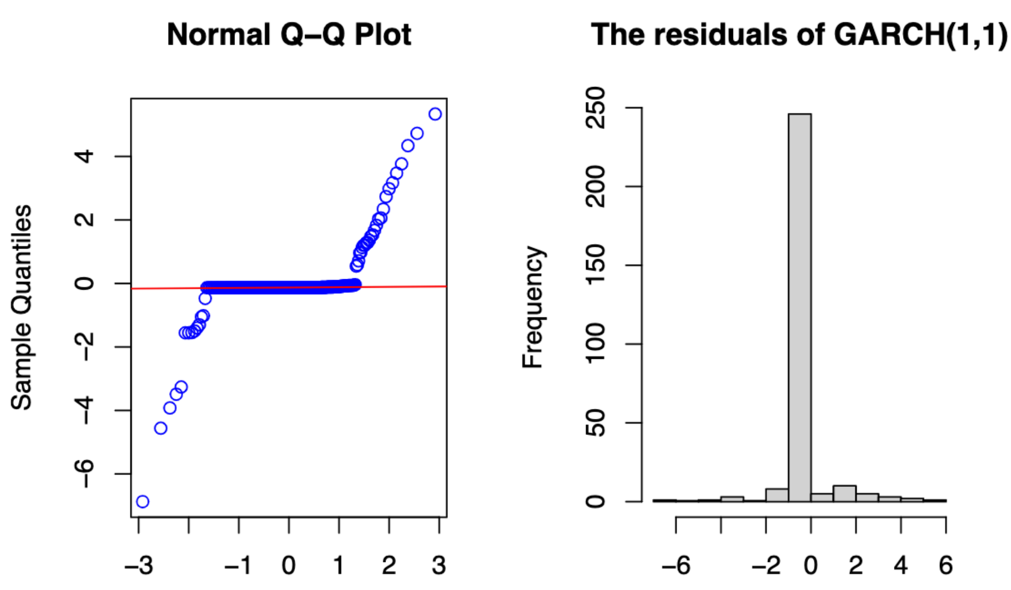

Figure 31: The normal Q-Q plot and histogram of the residuals from GARCH(1,1) model

Furthermore, upon closer examination of the squared residuals distribution through the Q-Q plot and histogram for both models, deviations from the expected normal distribution are evident. The observed shape of the distribution displays noticeable deviations from symmetry, particularly at the beginning and end of the distribution. These deviations raise concerns regarding the assumption of normality, prompting further scrutiny through additional model diagnostics.

```
 Shapiro-Wilk normality test

data:  na.omit(residuals(garch_model_1))
W = 0.46719, p-value < 2.2e-16


 Jarque Bera Test

data:  na.omit(residuals(garch_model_1))
X-squared = 3251.4, df = 2, p-value < 2.2e-16

The Skewness is  0.09  and the Kurtosis is 16.55
```

Regarding the Shapiro-Wilk normality test and Jarque Bera Test, the obtained p-values approach zero, indicating a rejection of the null hypothesis and suggesting that the data does not conform to a normal distribution. Additionally, while the skewness value falls within the range indicative of nearly symmetrical distribution, the kurtosis values exceeding zero imply a light-tailed distribution with a steeper peak, commonly known as positive kurtosis or leptokurtic distribution. Therefore, it's important to acknowledge that not all time series datasets follow a normal distribution when processing them.

d. Forecasting

❖ *ARIMA(0,1,1)-ARCH(1):*

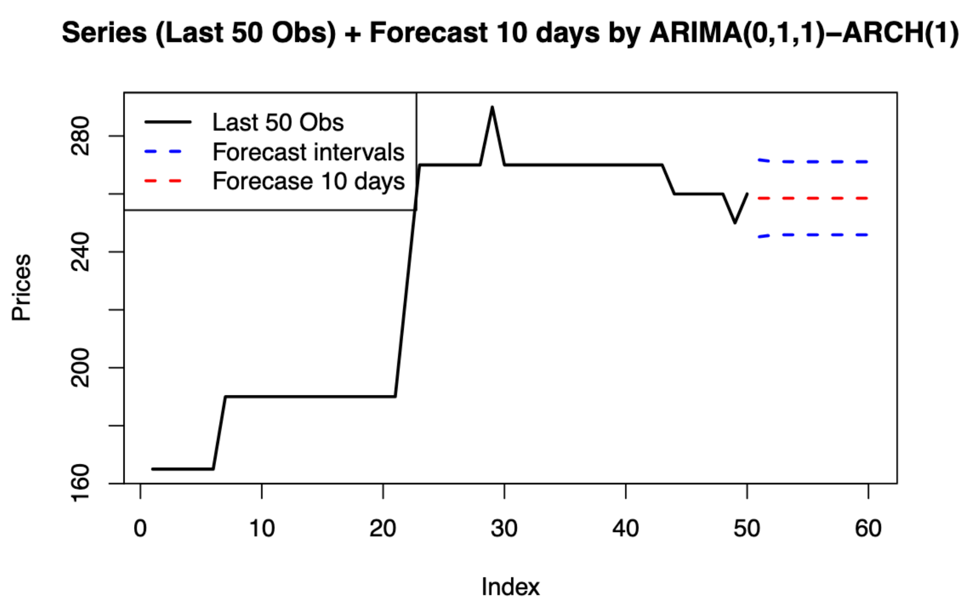

Figure 32: The plot of the last 50 observations and forecasting the next 10 days from ARIMA(0,1,1)-ARCH(1) model

Observing the plot of the ARIMA-ARCH model, it's evident that the forecasted values represented by the red dashed line closely align with the last observations, indicating a strong predictive capability. Additionally, the variance depicted by the blue dashed line remains relatively stable over time, suggesting that the ARCH component effectively captures the volatility clustering inherent in the ARIMA model.

❖ *GARCH(1,1):*

Conversely, in the GARCH(1,1) model plot, the forecasted values coincide exactly with the last observation, implying a prediction of no change in the price of Garlic for the next 10 days. Furthermore, similar to the ARIMA-ARCH model, the variance remains unchanged over time, indicating a consistent level of volatility captured by the model.

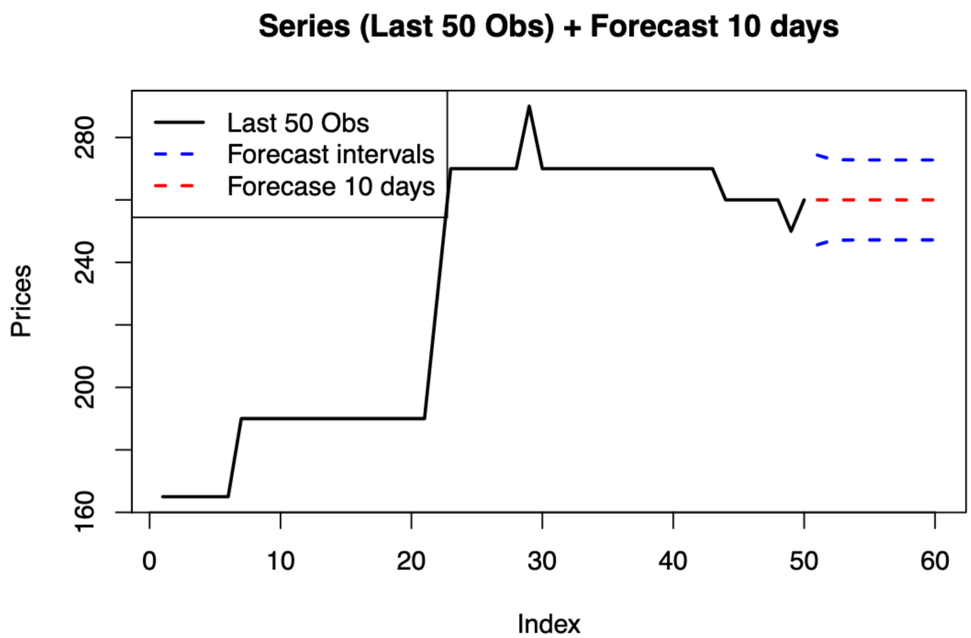

Figure 33: The plot of the last 50 observations and forecasting the next 10 days from GARCH(1,1) model

Upon analysis, it appears that both models provide a good fit for the dataset, as evidenced by their ability to accurately forecast future values and capture volatility patterns. However, when comparing the two, the GARCH(1,1) model exhibits slightly superior performance, as it accurately predicts no change in price and maintains stable variance over time.

IV. CONCLUSION

In conclusion, this project examined the application of various time series models to analyze seasonal and nonseasonal datasets. For the seasonal dataset concerning monthly truck sales, Multiplicative Seasonal ARIMA proved to be a suitable choice for modeling, while SARIMA(0,1,1)x(0,1,1)[12] emerged as the optimal model for forecasting.

On the other hand, for the nonseasonal dataset, specifically the price of Garlic, it was observed that ARIMA alone was insufficient in addressing volatility clustering. Consequently, two effective models, ARIMA(0,1,1)-ARCH(1) and GARCH(1,1), were identified to tackle this issue. Further analysis indicated that the GARCH(1,1) model exhibited slightly superior performance compared to ARIMA(0,1,1)-ARCH(1) in capturing volatility patterns and forecasting future values.

However, it's important to acknowledge the need for additional validation and scrutiny to confirm the selected models' suitability and robustness for forecasting purposes. Further research and analysis could involve assessing model performance under different scenarios and conducting rigorous validation tests to ensure reliable predictions.

In essence, this project underscores the importance of selecting appropriate time series models tailored to the specific characteristics of the dataset at hand. By leveraging the strengths of various modeling techniques, accurate forecasts can be generated, providing valuable insights for decision-making in diverse fields such as finance, economics, and business.

V. APPENDIX

For further code in this project see the url: https://github.com/hanchau94/Time-Series-Project

VI. REFFERENCES

1. https://www.kaggle.com/datasets/ddosad/dummy-truck-sales-for-time-series
2. https://www.kaggle.com/datasets/ksamiksha19/vegetable-prices
3. https://www.sciencedirect.com/topics/economics-econometrics-and-finance/dickey-fullertest#:~:text=The%20null%20hypothesis%20of%20DF,the%20test%20is%20being%20used.
4. https://www.sciencedirect.com/topics/mathematics/wilk-test#:~:text=Form%20null%20and%20alternative%20hypotheses,are%2095%25%20and%2099%25.
5. https://www.idrisstsafack.com/post/garch-models-with-r-programming-a-practical-example-with-tesla-stock
6. https://huskiecommons.lib.niu.edu/cgi/viewcontent.cgi?article=1176&context=studentengagement-honorscapstones
7. https://online.stat.psu.edu/stat510/lesson/11/11.1
8. https://www.rdocumentation.org/packages/rugarch/versions/1.5-1/topics/ugarchspec-methods
9. https://www.mathworks.com/help/econ/archtest.html
10. https://stackoverflow.com/questions/62480098/time-series-forecasting-with-garch
11. https://otexts.com/fpp2/residuals.html
12. https://simplilearn.com/tutorials/statistics-tutorial/skewness-and-kurtosis#:~:text=Hence%2C%20you%20can%20say%20that,pull%20or%20the%20peak's%20height.
13. https://www.researchgate.net/publication/265468228_Component_GARCH_Models_to_Account_for_Seasonal_Patterns_and_Uncertainties_in_Travel_Time_Prediction#:~:text=However%2C%20a%20GARCH%20model%20arguably,GARCH%20model%20may%20be%20affected.
14. https://machinelearningmastery.com/develop-arch-and-garch-models-for-time-series-forecasting-in-python/