STATISTICAL ANALYSIS OF SALES OF WINE

Group 5A

202130-S.R.V.Perera
202135-R.T.K.Prematilaka
202138-W.P.A.G.Rajapaksha
202142-R.M.B.N.A.Rathnamalala
202143-M.M.G.N.L.Rathnayake
192021-B.G.I.H.Dayasiri

STAT 3124 – Time Series Analysis
Final Report

Department of Mathematical Sciences
Faculty of Applied Sciences
Wayamba University of Sri Lanka
Kuliyapitiya

1. INTRODUCTION

The data set is used for the time series analysis of wine sales in Vino Vista Company from Italy. Data has been collected from January 2012 to December 2021. The quantity of wine sold is calculated in bottles each month. Over the past ten years, people have been choosing more expensive and craft wines. This has affected how many bottles of wine are sold each month. People are getting more into the culture of wine, seeing it as part of their lifestyle. This has led to a steady increase in sales. Also, it's now easier to buy wine online, which means more people can explore and buy different types of wine. This has helped sell even more bottles each month.

Throughout the period, seasonal variations played a role, with peak sales often observed during festive seasons and holidays. In recent years, sustainability and organic trends have gained traction, influenced consumer choices and subsequently impacted the wine market. The amount of wine sold each month between 2012 and 2021 was influenced by money, culture, and worldwide events. The wine industry showed it could handle changes and adapt to what people wanted, even when faced with challenges.

Table 01: Monthly sales of Wine, from 2012 to 2016

	2012	**2013**	**2014**	**2015**	**2016**
January	73821	73891	77234	74921	80342
February	60059	61234	67891	79761	82912
March	61743	63219	72542	69723	66014
April	50349	60583	63321	74896	70659
May	36214	50987	58923	67025	78521
June	41637	56032	62001	59783	65342
July	47892	60248	53214	64902	63341
August	41059	50632	48956	78943	60322
September	41472	47528	66342	92456	65761
October	54023	56789	78923	102132	73311
November	60628	55712	89125	92105	96312
December	69243	74325	84432	75521	110342

Table 02: Monthly sales of Wine, from 2017 to 2021

	2017	**2018**	**2019**	**2020**	**2021**
January	98623	93621	100399	104230	109601
February	88101	98102	86250	91585	96311
March	91783	88332	88826	95132	99441
April	81022	85214	78636	83986	88705
May	78013	90231	75563	81065	85719
June	83702	93346	80743	86630	91119
July	69022	76432	83380	89355	93807
August	71609	79678	66046	70999	75889
September	81571	90341	70994	72843	79065
October	102342	114102	79313	83844	88915
November	119432	130243	100903	106573	111155
December	106521	114794	116303	122670	126953

2. STATISTICAL ANALYSIS

2.1. Time Series Plot

This analysis focuses on the time series plot sales of Wine from 2012 to 2021

Figure 2.1.1: The time series plot of monthly sales of Wine from 2012 to 2021

This time series plot of sales of Wine shows an upward trend and increasing seasonal variation with seasonal length is 12. Therefore, there is a multiplicative seasonal component.

2.2. Multiplicative Decomposition Model

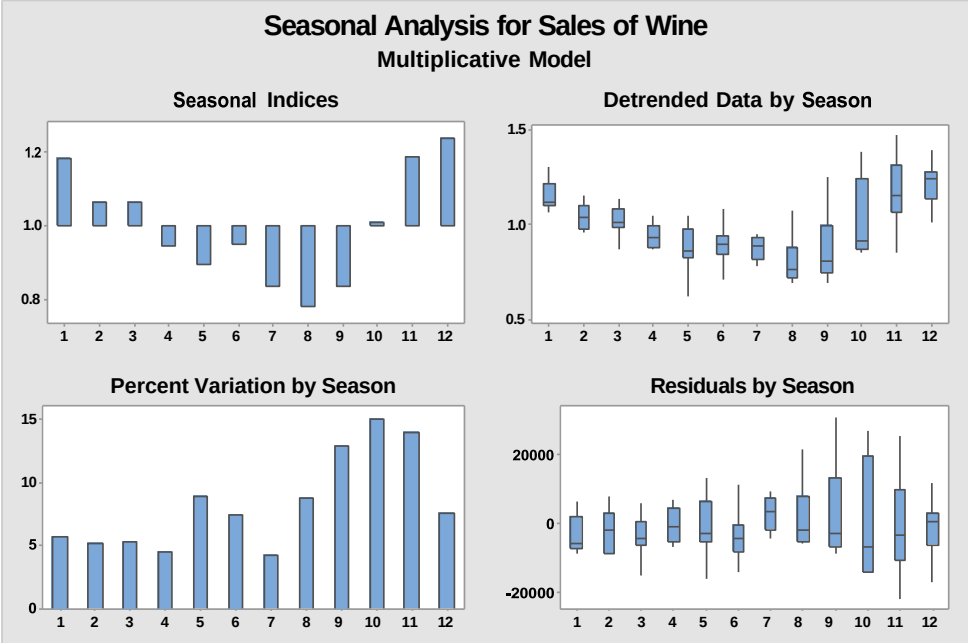


Figure 2.2.1: Seasonal Analysis for Sales of Wine

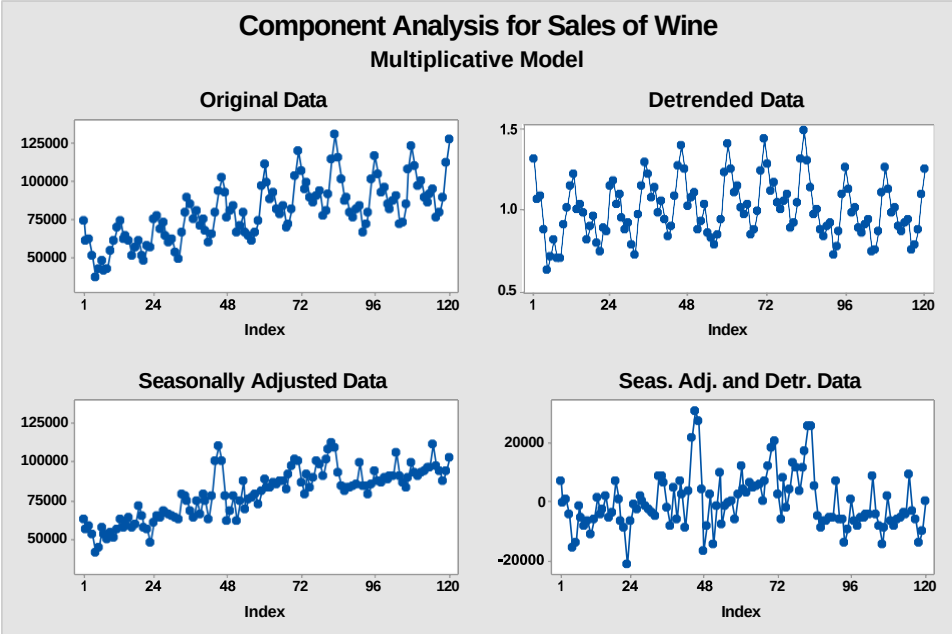


Figure 2.2.2: Components Analysis for Sales of Wine

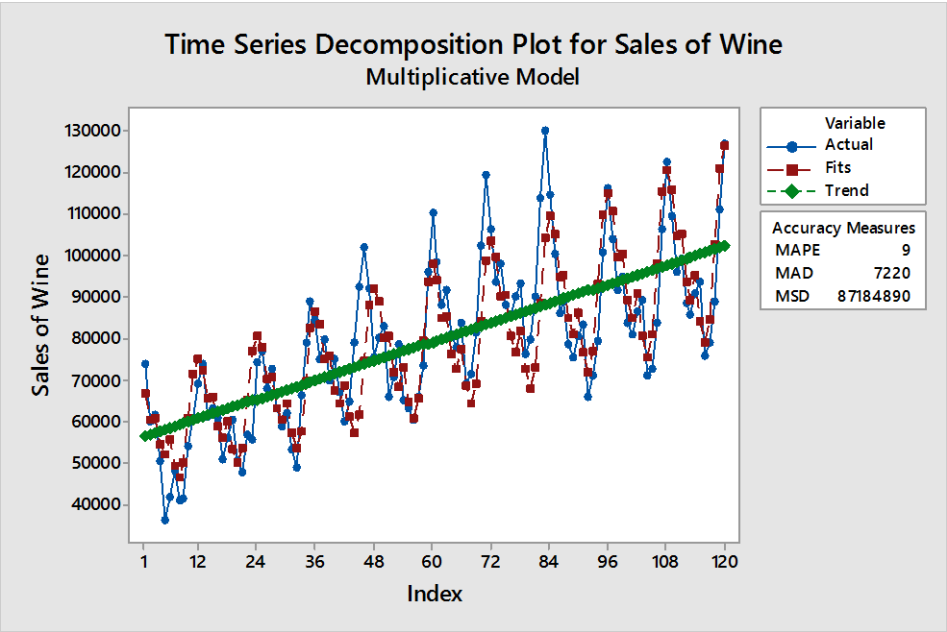


Figure 2.2.3. : Time Series Decomposition Plot

The magnitude of the seasonal variation is increasing over the time. Therefore, the multiplicative decomposition method is suitable. The seasonal lag is 12. We use decomposition models to forecast time series that exhibits trend and seasonal effects. The multiplicative decomposition model is useful when modeling time series that display increasing seasonal variation over time.

Multiplicative Decomposition Model:

$\mathbf{Y}_t = TR_t \mathbf{x}\, SN_t \mathbf{x}\, CL_t \mathbf{x}\, IR_t$ ---------- (1)

Here TR_t, SN_t, CL_t *and* IR_t *are* respectively, trend, seasonal, cyclical and irregular factors.

Fitted Trend Equation

$Y_t = 56084 + 385.7 \times t$ ---------- (2)

Seasonal Indices

Table 2.2.1: Seasonal indices for Monthly Data

Period	Index
1	1.18420
2	1.06348
3	1.06470
4	0.94527
5	0.89520
6	0.95127
7	0.83966
8	0.78304
9	0.83764
10	1.01149
11	1.18694
12	1.23710

Accuracy Measures

MAPE 9

MAD 7220

MSD 87184890

2.3 Autocorrelation Function (ACF)

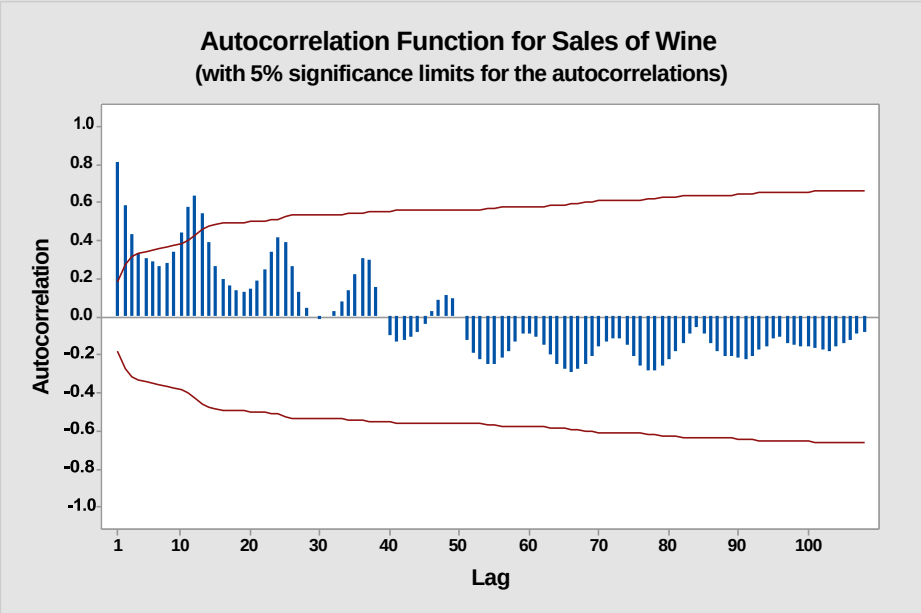


Figure 2.3: Autocorrelation function for monthly sales of Wine from 2012 to 2021

The Autocorrelation function (ACF) is exhibiting a seasonal variation and trend variation. That means the series is not stationary. We need to perform a non-seasonal (trend) difference of lag 1 to make the series stationary.

2.3.1 Autocorrelation function for monthly sales of Wine after trend difference

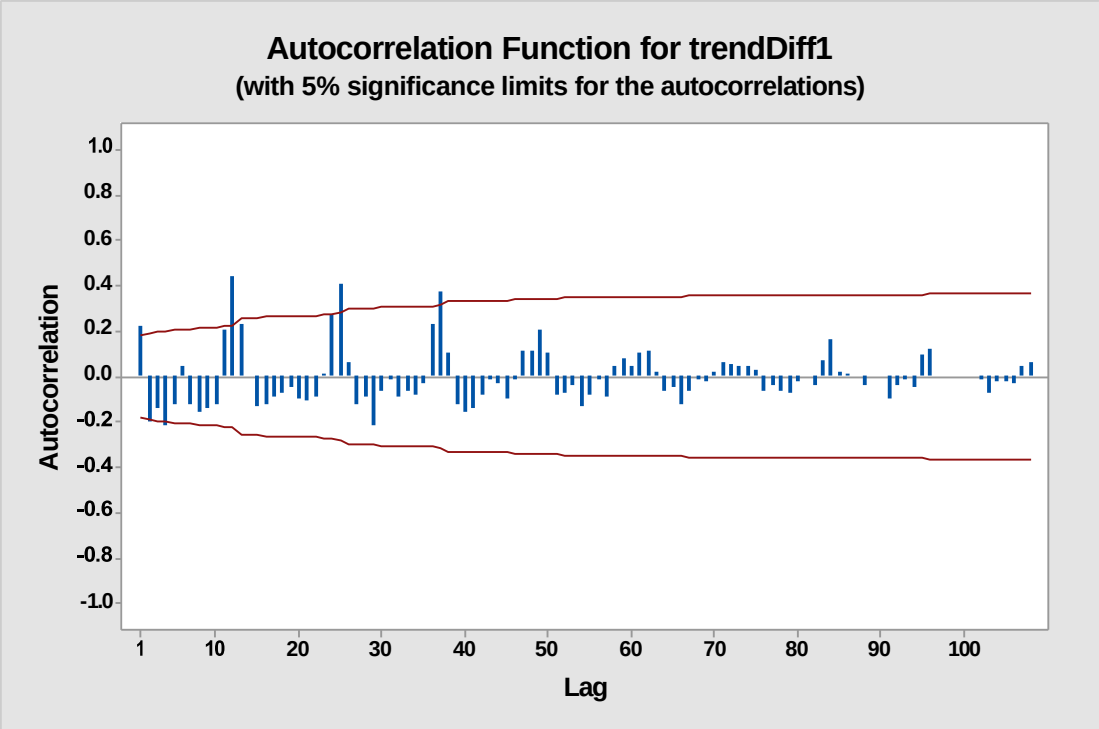


Figure 2.3.1: Autocorrelation function for trend differenced

Now plot is clearly seasonal. The series is not stationary. To make the series stationary we use seasonal difference with lag 12.

2.3.2 Autocorrelation function for monthly sales of Wine after seasonal difference

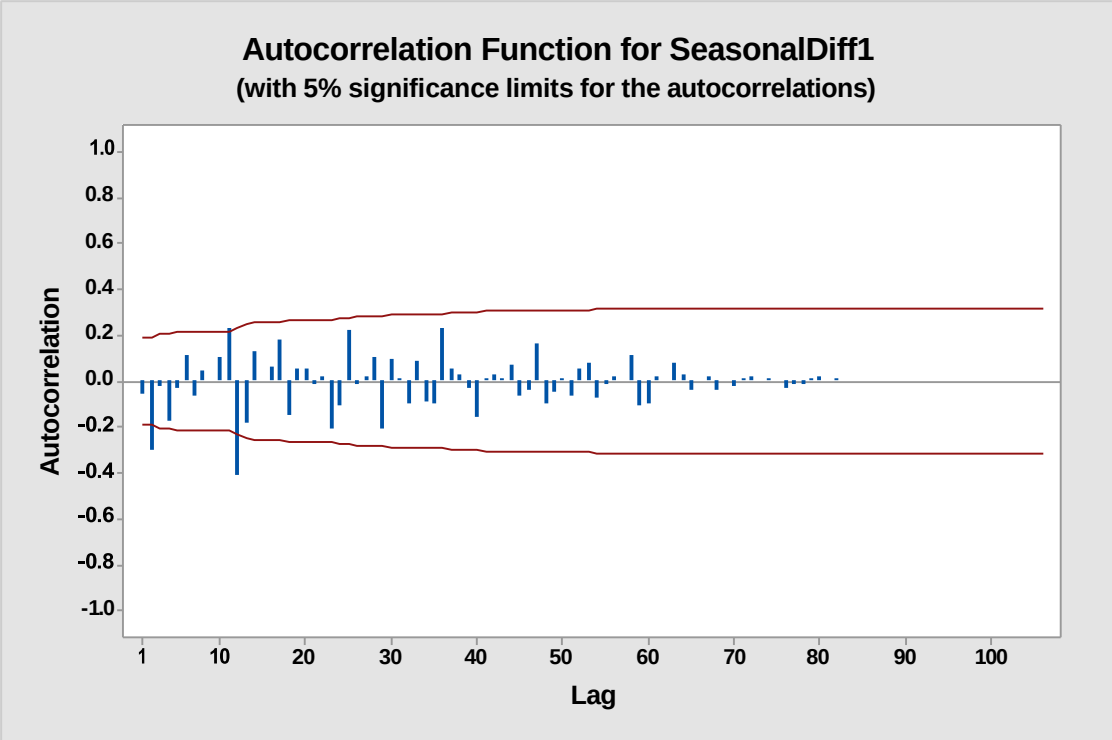


Figure 2.3.2: Autocorrelation function for Seasonal Differenced

Hypothesis:

H0: $\rho k = 0$

H1: $\rho k \neq 0$

ρk – Autocorrelation function of seasonal differenced monthly sales of wine of lag k

Checking T-values of seasonal differenced data:

For Non - Seasonal area

Lag	ACF	T	LBQ
1	-0.054590	-0.56	0.33
2	-0.299150	-3.09	10.27
3	-0.018730	-0.18	10.31
4	-0.175016	-1.66	13.78

5	-0.031318	-0.29	13.89
6	0.112251	1.04	15.34
7	-0.064018	-0.59	15.82
8	0.049428	0.45	16.11
9	-0.000200	-0.00	16.11
10	0.105260	0.96	17.44
11	0.233832	2.11	24.08

For Seasonal area

Lag	ACF	T	LBQ
12	-0.406706	-3.53	44.39
24	-0.106646	-0.77	66.52
36	0.228418	1.53	97.22
48	-0.099451	-0.64	111.35
60	-0.099795	-0.62	124.31
72	0.017409	0.11	127.63
84	0.001598	0.01	128.49
96	0.000087	0.00	128.52

In non- seasonal area 2nd lag is significant since the absolute value of t statistics are greater than 2(|T|>2).In other non-seasonal lags we don't have enough evidence reject H_o at 5% level of significance. It means ACF does not cuts off in non-seasonal area. And in the seasonal are ACF cuts off at lag 1 Therefore series is stationary.

2.4 Partial Autocorrelation Function

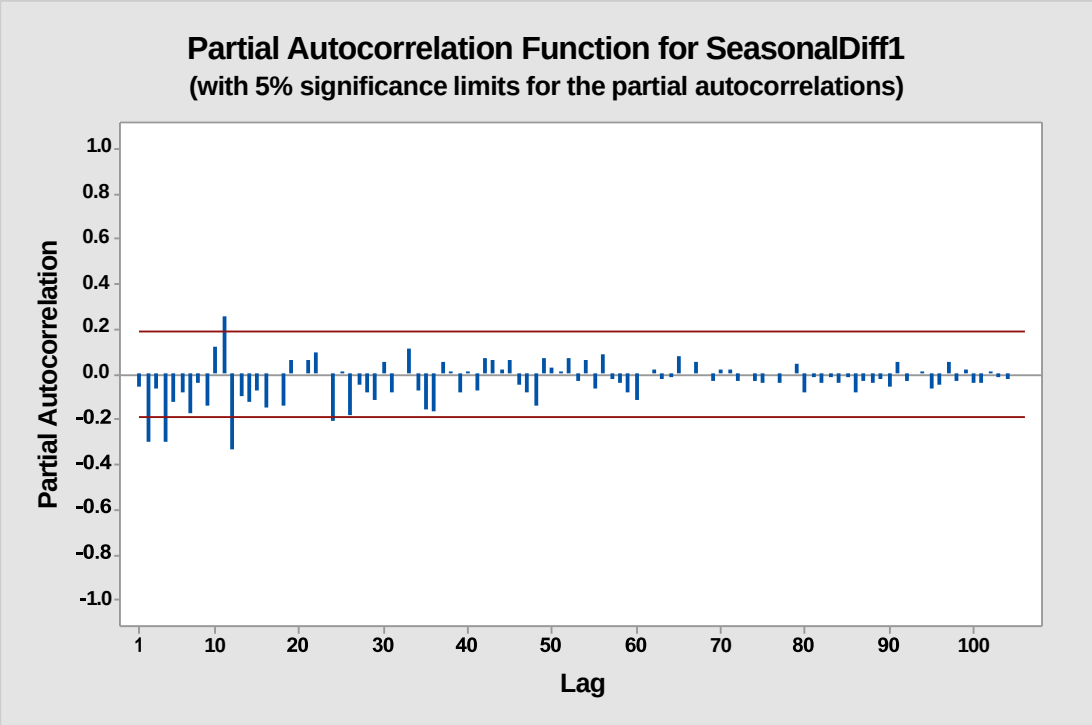


Figure 2.4: Partial Autocorrelation function for Seasonal Differenced monthly sales of wine

Hypothesis

H_0: $\rho k\text{k} = 0$

H_1: $\rho k\text{k} \neq 0$

For Non- Seasonal Area

Lag	PACF	T
1	-0.054590	-0.56
2	-0.303033	-3.13
3	-0.062456	-0.65
4	-0.301678	-3.12
5	-0.123380	-1.28
6	-0.084116	-0.87
7	-0.172474	-1.78
8	-0.036924	-0.38
9	-0.135286	-1.40
10	0.124721	1.29
11	0.259957	2.69\

For Seasonal Area

Lag	PACF	T
12	-0.333975	-3.45
24	-0.206552	-2.14
36	-0.162646	-1.68
48	-0.139843	-1.45
60	-0.116457	-1.20
72	-0.028985	-0.30
84	-0.034826	-0.36
96	-0.048415	-0.50

In 2nd and 4th Non – Seasonal lag, absolute values of t statistics are greater than 2(|T|>2).But in other Non -Seasonal lags, absolute values of t statistics are less than 2. (|T|<2). So, there is no enough evidence to reject H_0 at 5% level of significance. So PACF cutoff at lag zero at Non seasonal area. PACF Cuts off at seasonal area at lag 2.

2.5 Tentative Model

Non seasonal Differences	d= 1
Seasonal Differences	D=1
ACF Cuts off at Seasonal lag1	Q=1
ACF is zero at non-Seasonal lag	q=0
PACF cuts off at Seasonal Lag	P=2
PACF cuts off at non-seasonal lag	p=0

So, the tentative model is

SARIMA (0,1,0) $(2,1,1)_{12}$

2.6 Testing the Adequacy of the Tentative Model 1 and Diagnostic Checking

2.6.1. Significance of the Parameters

Hypothesis:

H_0: Coefficient = 0

H_1: Not so

```
Final Estimates of Parameters
Type          Coef  SE Coef      T      P
SAR  12    -0.7874   0.2385  -3.30  0.001
SAR  24    -0.4662   0.1160  -4.02  0.000
SMA  12    -0.2592   0.2689  -0.96  0.337
Constant        82     1142   0.07  0.943


Differencing: 1 regular, 1 seasonal of order 12
Number of observations:  Original series 120, after differencing 107
Residuals:    SS =  9064654010 (backforecasts excluded)
              MS =  88006350  DF = 103
```

Coefficient of SMA 12 is not significance from zero since p-value > 0.05. Do not reject H_0. Constant is not significant from zero since p-value > 0,05. H_0 is not rejected. That means parameters are not significant.

After removing constant term,

```
Final Estimates of Parameters

Type         Coef  SE Coef      T      P
SAR  12   -0.7880   0.2372  -3.32  0.001
SAR  24   -0.4664   0.1154  -4.04  0.000
SMA  12   -0.2599   0.2674  -0.97  0.333

Differencing: 1 regular, 1 seasonal of order 12
Number of observations:  Original series 120, after differencing 107
Residuals:    SS =  9064897018 (backforecasts excluded)
              MS =  87162471  DF = 104
```

Again P – Value of SMA 12 process is greater than 0.05. Therefore, H_0 is not rejected. That means parameters are not significant.

Remove SMA term,

```
Final Estimates of Parameters

Type         Coef  SE Coef      T      P
SAR  12   -0.5588   0.0904  -6.18  0.000
SAR  24   -0.3709   0.0902  -4.11  0.000


Differencing: 1 regular, 1 seasonal of order 12
Number of observations:  Original series 120, after differencing 107
Residuals:     SS =  9152649990 (backforecasts excluded)
               MS =  87168095  DF = 105
```

P values of all the processes are less than 0.05. So, there are enough evidences to reject H_0. Therefore, all the parameters are significant.

Hence, **SARIMA $(0,1,0)(2,1,0)_{12}$** is a significant model.

Diagnostic Checking

2.6.2 Randomness of Residuals

Hypothesis:

H_0: $\rho 1 = \rho 2 = \rho 3 \ldots = \rho k = 0$ (at Least one)

H_1: at least one $\rho_i \neq 0$ (i = 1, 2, 3… k)

ρk − residual autocorrelation of lag k

```
Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12     24     36     48
Chi-Square    19.4   29.2   54.3   70.5
DF              10     22     34     46
P-Value      0.035  0.139  0.015  0.012
```

Here, P-value should be greater than 0.05 to do not reject H_0.

Since the P-value at lag 1 < 0.05, we cannot decide whether the residuals are random or not by looking at Ljung-Box. We have to consider residual ACF and residual PACF now.

2.6.2.1 Autocorrelation Function for Residuals

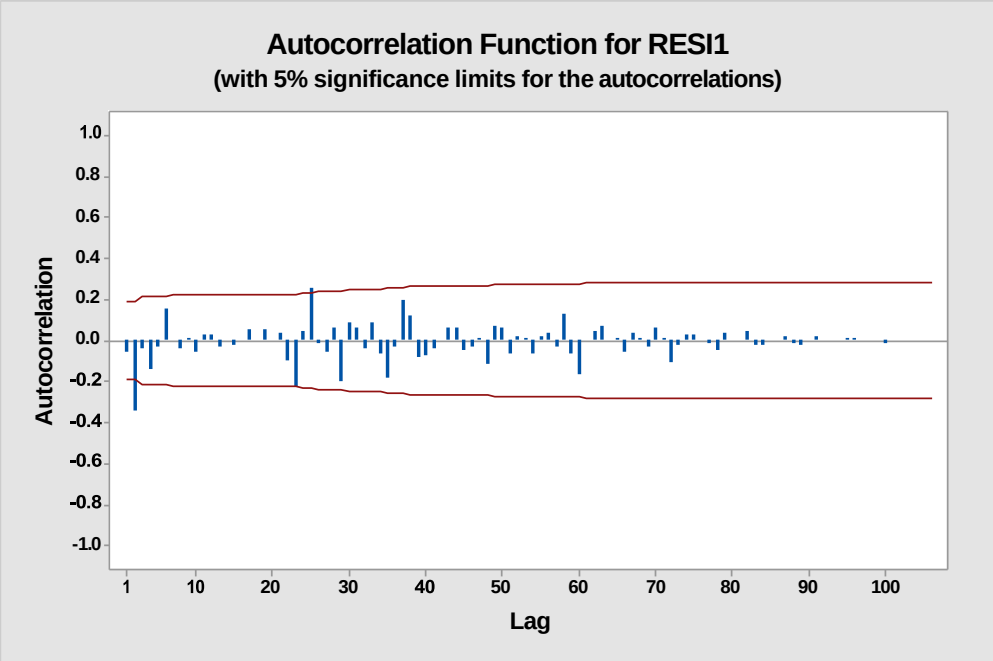


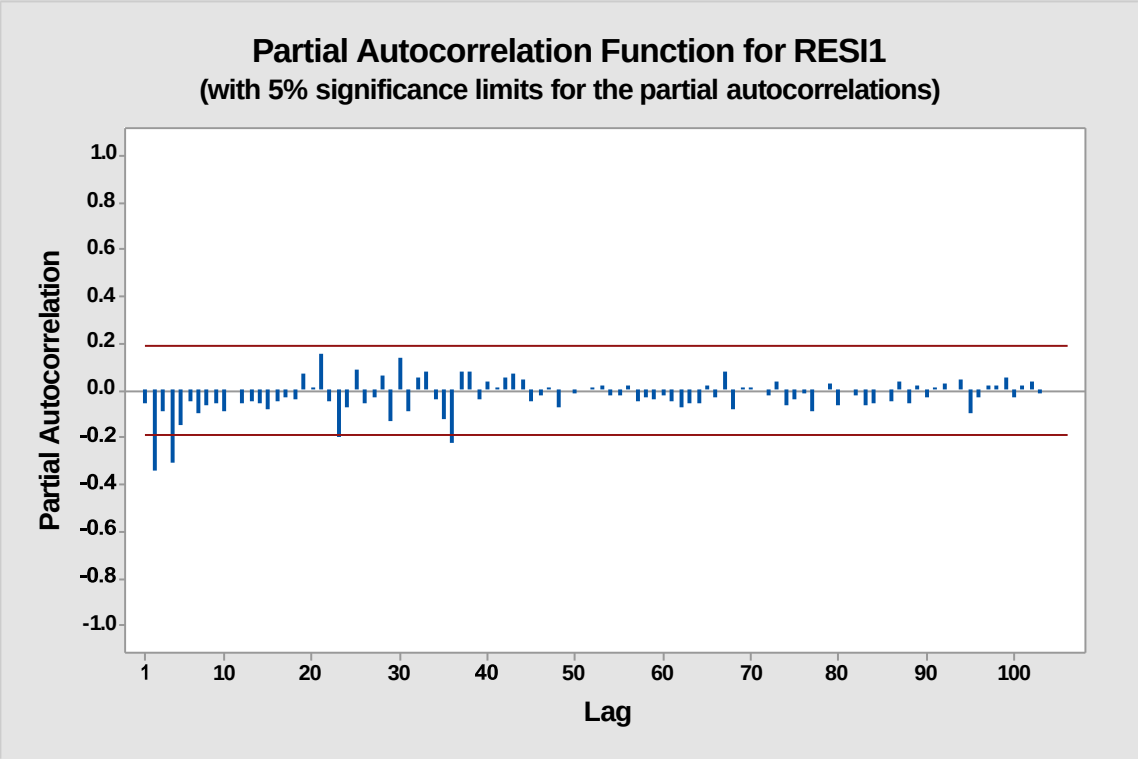
Figure 2.6.2.1: Autocorrelation function for Residuals of model

Lag	ACF	T	LBQ
1	-0.051044	-0.53	0.29
2	-0.342528	-3.53	13.32
3	-0.036499	-0.34	13.47
4	-0.136493	-1.27	15.58
5	-0.029723	-0.27	15.68
6	0.160556	1.47	18.66
7	0.006926	0.06	18.66
8	-0.041182	-0.37	18.86
9	0.010546	0.09	18.87
10	-0.057825	-0.52	19.28
11	0.026046	0.23	19.36
12	0.026290	0.23	19.44
24	0.050318	0.43	29.21

36	-0.029444	-0.23	54.31
48	-0.114129	-0.84	70.47
60	-0.163963	-1.17	86.62

In both seasonal and non – seasonal area, from 1st lag the absolute values of t statistics are not greater than 2 (|T|<2).

2.6.2.2. Partial Autocorrelation Function for Residuals



Partial Autocorrelation function for Residuals of model

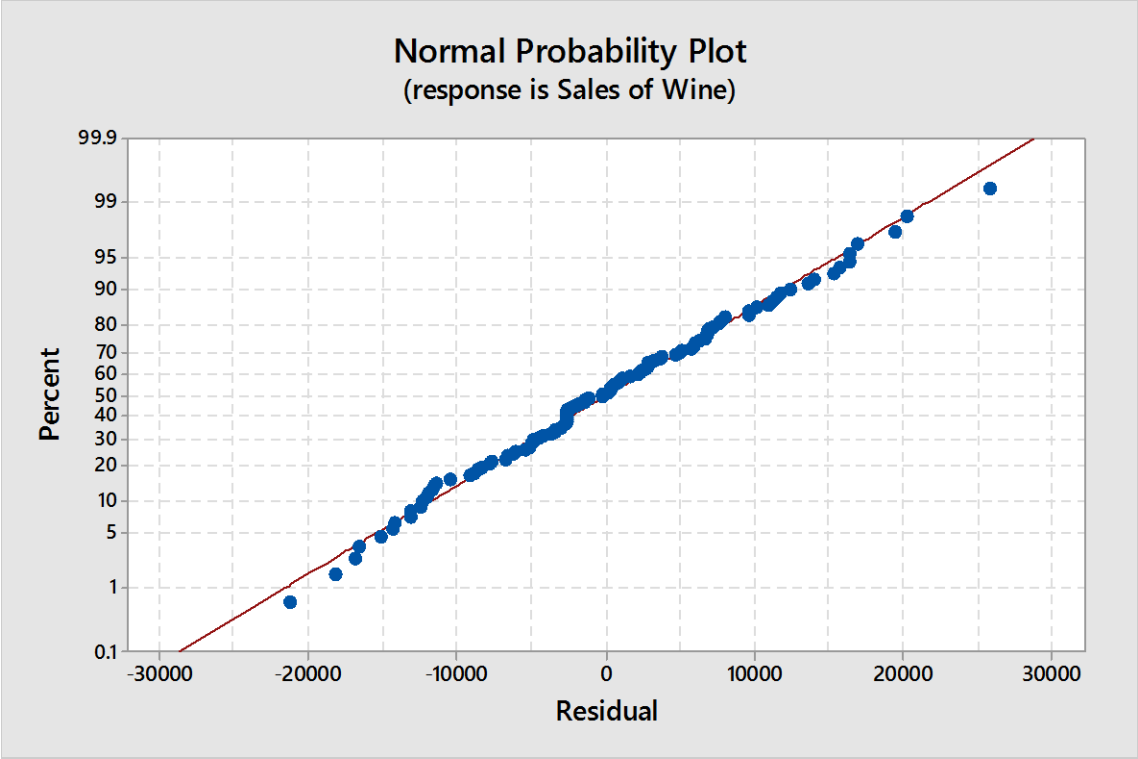
Lag	PACF	T
1	-0.051044	-0.53
2	-0.346035	-3.58
3	-0.088496	-0.92
4	-0.304498	-3.15
5	-0.151759	-1.57
6	-0.045607	-0.47
7	-0.097861	-1.01
8	-0.064530	-0.67
9	-0.053738	-0.56
10	-0.087888	-0.91
11	-0.005862	-0.06

12	-0.056783	-0.59
24	-0.076005	-0.79
36	-0.221133	-2.29
48	-0.071106	-0.74
60	-0.023077	-0.24

In both seasonal and non – seasonal area, from 1st lag the absolute values of t statistics are not greater than 2(|T|<2).

Therefore, Residuals are randomly distributed even though the Ljung box Chi – Square Statistic is not indicated as the residuals are random.

2.6.3. Normality of the Residuals



Normal Probability plot for Model

Almost all the data points lie on the straight line of Normal Probability Plot. So, it is concluded that residuals are normally distributed.

2.6.4 Parameter Redundancy

```
Correlation matrix of the estimated parameters

       1
2  0.408
```

There are no correlation values close to 0.8. So, there are no parameter redundancy does not exist.

Therefore, the model **SARIMA (0,1,0)(2,1,0)$_{12}$** is adequate with obeying the following Conditions;

- ✓ The parameters are significant.
- ✓ Residuals are random.
- ✓ Residuals are normally distributed.
- ✓ No parameter redundancy.

2.7 Testing the Adequacy of the Tentative Model 2 and Diagnostic Checking

SARIMA (0,1,0) (1,1,0)$_{12}$

2.7.1. Significance of the Parameters

Hypothesis:

H_0: Coefficient = 0

$H_{1:}$ Not so

```
Final Estimates of Parameters

Type          Coef  SE Coef      T      P
SAR  12    -0.4166    0.0886  -4.70  0.000
Constant     54.8     965.8   0.06  0.955
```

```
Differencing: 1 regular, 1 seasonal of order 12
Number of observations:  Original series 120, after differencing 107
Residuals:    SS =  10479352410 (backforecasts excluded)
              MS =  99803356  DF = 105
```

Constant is not significant from zero since p-value > 0,05. H_0 is not rejected. That means parameters are not significant.

After removing constant term,

```
Final Estimates of Parameters

Type         Coef  SE Coef       T      P
SAR  12  -0.4166   0.0882  -4.72  0.000

Differencing: 1 regular, 1 seasonal of order 12
Number of observations:  Original series 120, after differencing 107
Residuals:    SS =  10479620764 (backforecasts excluded)
              MS =  98864347  DF = 106
```

P values of all the processes are less than 0.05. So, there are enough evidences to reject H_0. Therefore, all the parameters are significant.

Hence, **SARIMA $(0,1,0)(1,1,0)_{12}$** is a significant model.

Diagnostic Checking

2.7.2 Randomness of Residuals

Hypothesis:

H_0: $\rho 1 = \rho 2 = \rho 3 \ldots = \rho k = 0$ (at Least one)

H_1: at least one $\rho_i \neq 0$ (i = 1, 2, 3… k)

ρk – residual autocorrelation of lag k

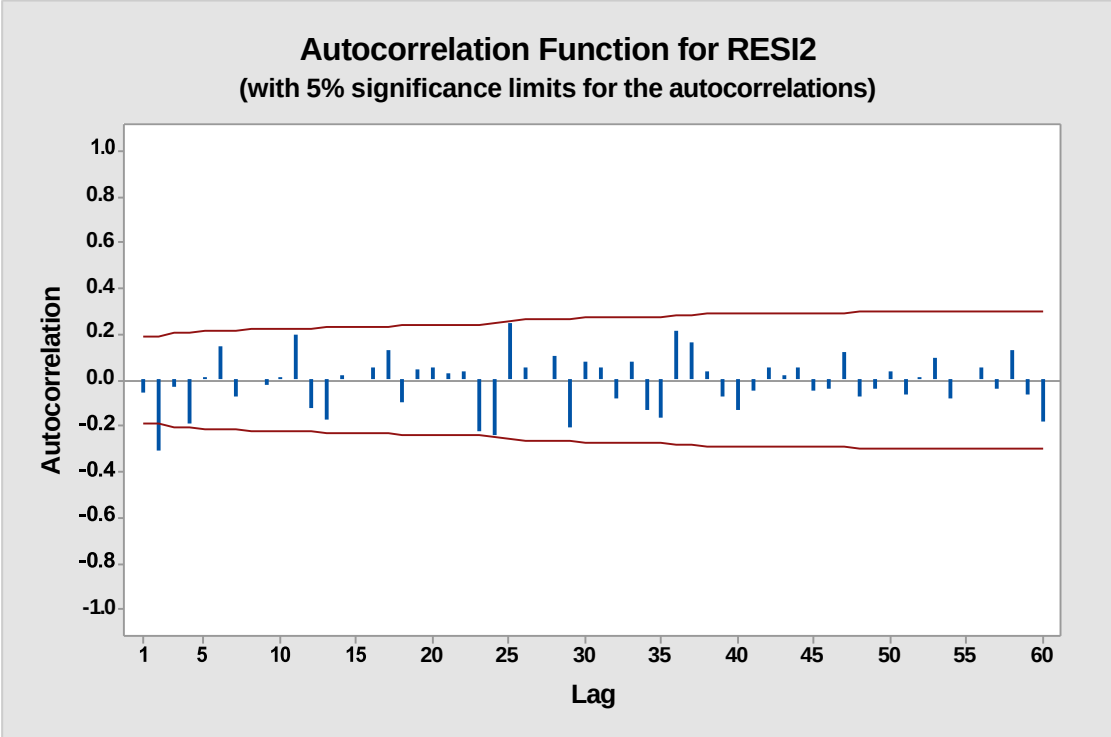
```
Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12     24     36     48
Chi-Square    24.7   48.4   83.4   98.1
DF              11     23     35     47
```

```
P-Value      0.010   0.001   0.000   0.000
```

Here, P-value should be greater than 0.05 to do not reject H_0.

Since all the P-values are < 0.05, we cannot decide whether the residuals are random or not by looking at Ljung-Box. We have to consider residual ACF and residual PACF now.

2.7.2.1 Autocorrelation Function for Residuals

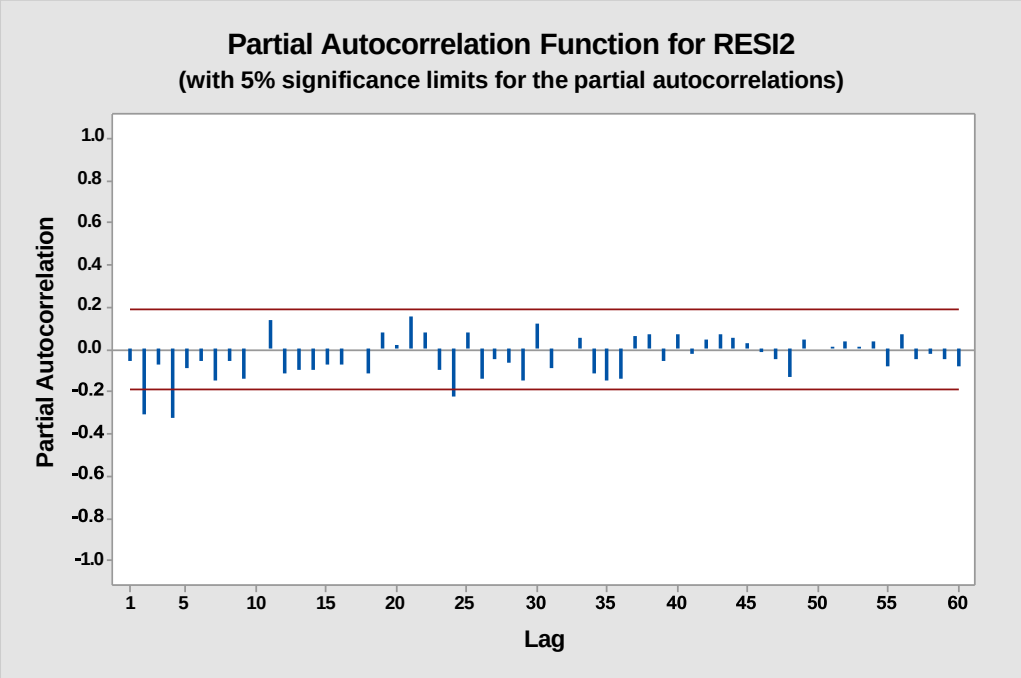


Lag	ACF	T	LBQ
1	-0.052014	-0.54	0.30
2	-0.308079	-3.18	10.84
3	-0.030048	-0.28	10.94
4	-0.187455	-1.77	14.92
5	0.013497	0.12	14.94
6	0.145988	1.34	17.40
7	-0.075517	-0.68	18.07
8	0.005363	0.05	18.07
9	-0.018777	-0.17	18.11
10	0.010232	0.09	18.13
11	0.198369	1.78	22.91

12	-0.122531	-1.07	24.75
24	-0.239088	-1.92	48.37
36	0.215893	1.52	83.43
48	-0.070345	-0.47	98.13
60	-0.180856	-1.18	116.98

In both seasonal and non – seasonal area, from 1st lag the absolute values of t statistics are not greater than 2 (|T|<2).

2.7.2.2 Partial Autocorrelation Function for Residuals



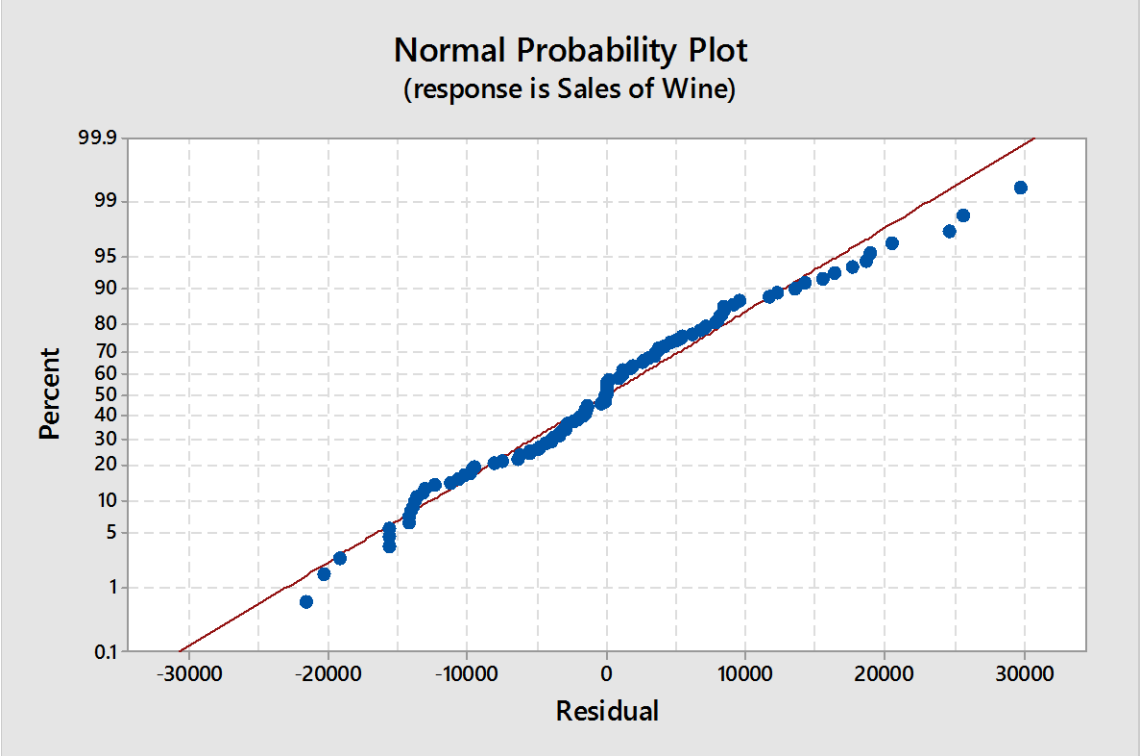
Lag	PACF	T
1	-0.052014	-0.54
2	-0.311627	-3.22
3	-0.074712	-0.77
4	-0.325732	-3.37
5	-0.087427	-0.90
6	-0.056339	-0.58
7	-0.150088	-1.55
8	-0.058089	-0.60
9	-0.135630	-1.40
10	-0.007569	-0.08
11	0.142958	1.48

12	-0.111208	-1.15
24	-0.227623	-2.35
36	-0.138695	-1.43
48	-0.128384	-1.33
60	-0.078373	-0.81

In both seasonal and non – seasonal area, from 1st lag the absolute values of t statistics are not greater than 2(|T|<2).

Therefore, Residuals are randomly distributed even though the Ljung box Chi – Square Statistic is not indicated as the residuals are random.

2.7.3. Normality of the Residuals



Almost all the data points lie on the straight line of Normal Probability Plot. So, it is concluded that residuals are normally distributed.

The correlation matrix of the estimated parameters is empty

There is no high correlation between terms because all the correlation values < 0.8. Therefore, parameter redundancy does not exist.

Therefore, the model **SARIMA (0,1,0)(1,1,0)$_{12}$** is adequate with obeying the following

Conditions;

- ✓ The parameters are significant.
- ✓ Residuals are random.
- ✓ Residuals are normally distributed.
- ✓ No parameter redundancy.

2.8 Testing the Adequacy of the Tentative Model 3 and Diagnostic Checking

SARIMA (0,1,0) (1,1,1)$_{12}$

2.8.1. Significance of the Parameters

Hypothesis:

H_0: Coefficient = 0

$H_{1:}$ Not so

```
Final Estimates of Parameters

Type         Coef  SE Coef      T      P
SAR  12    0.2124   0.1182   1.80  0.075
SMA  12    0.9044   0.0771  11.73  0.000
Constant     46.7    117.9   0.40  0.693
```

Constant is not significant from zero since p-value > 0,05. H_0 is not rejected. That means parameters are not significant.

After removing constant term,

```
Final Estimates of Parameters

Type        Coef  SE Coef       T      P
SAR  12   0.2126   0.1177    1.81  0.074
SMA  12   0.9044   0.0768   11.78  0.000
```

SAR12 is not significant from zero since p-value > 0,05. H_0 is not rejected. That means parameters are not significant.

After removing SAR12 term,

```
Final Estimates of Parameters

Type       Coef    SECoef       T      P
SMA  12   0.8734   0.0655   13.33  0.000
```

P values of all the processes are less than 0.05. So, there are enough evidences to reject H_0. Therefore, all the parameters are significant.

Hence, **SARIMA (0,1,0)(0,1,1)$_{12}$** is a significant model.

Diagnostic Checking

2.8.2 Randomness of Residuals

Hypothesis:

H_0: $\rho 1 = \rho 2 = \rho 3 \ldots = \rho k = 0$ (at Least one)

H_1: at least one $\rho_i \neq 0$ (i = 1, 2, 3… k)

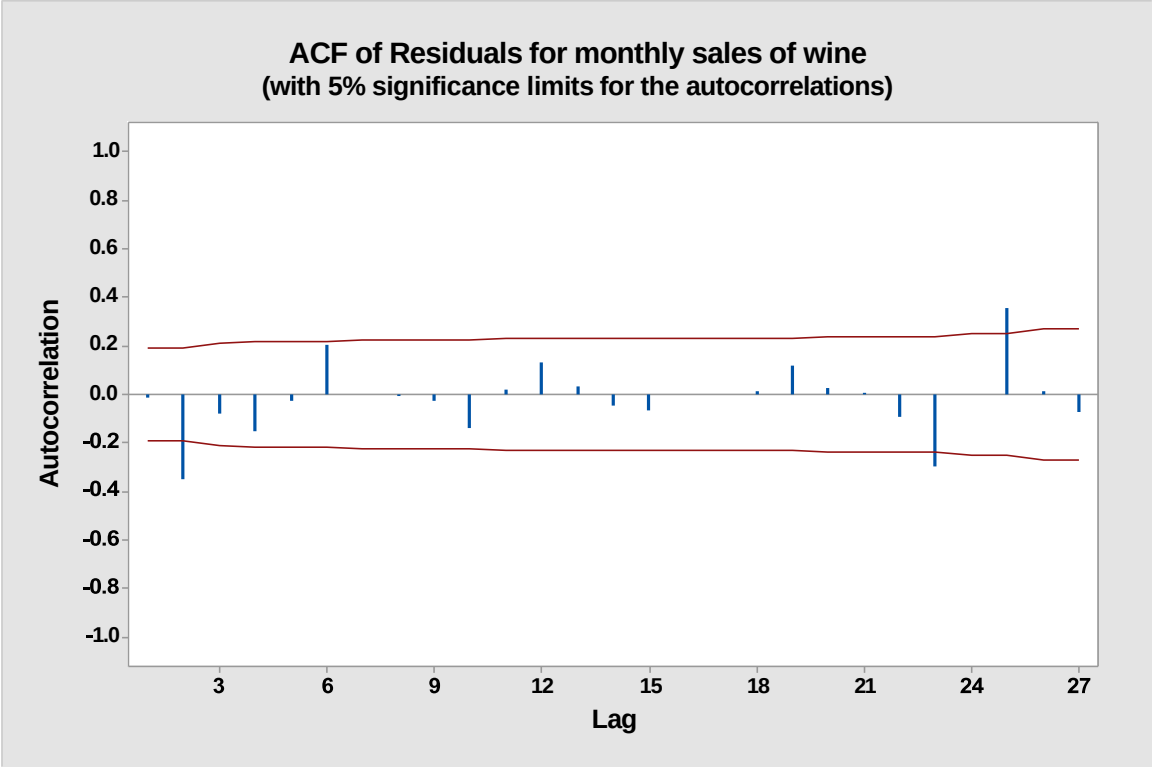
ρk − residual autocorrelation of lag k

```
Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12     24     36     48
Chi-Square    26.6   43.1   83.6  105.5
DF              11     23     35     47
P-Value      0.005  0.007  0.000  0.000
```

Here, P-value should be greater than 0.05 to do not reject H_0.

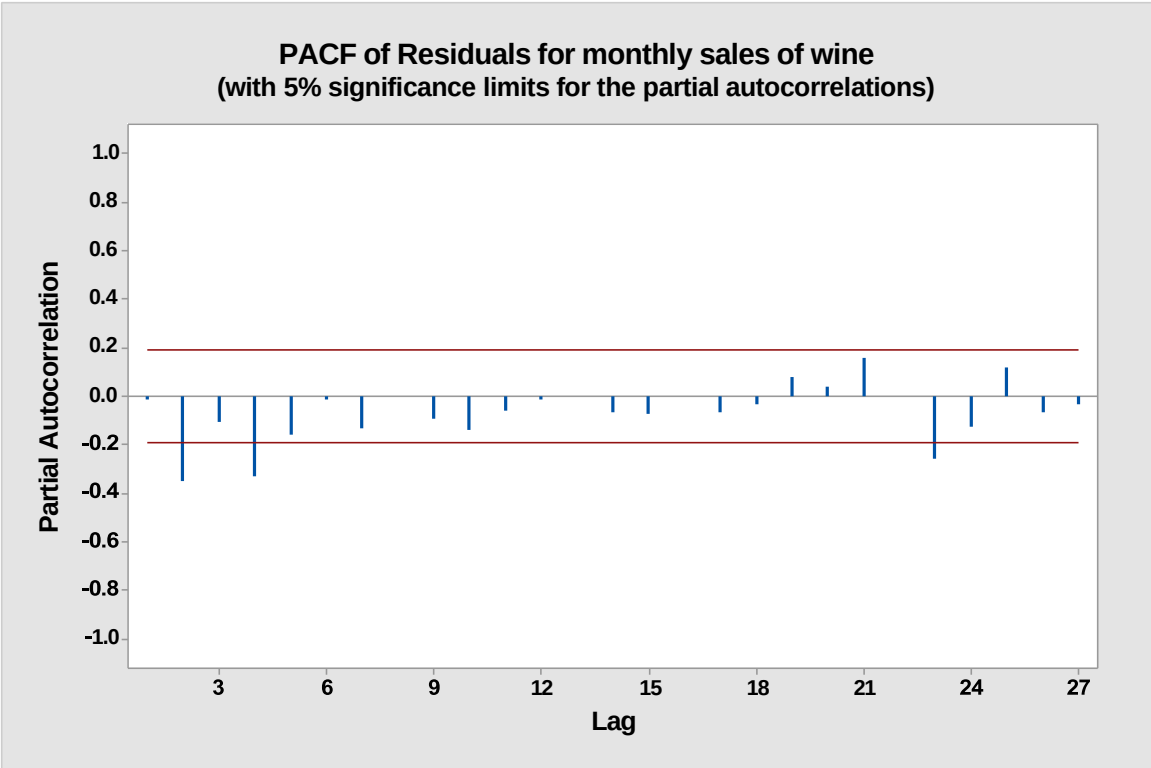
Since all the P-values are < 0.05, we cannot decide whether the residuals are random or not by looking at Ljung-Box. We have to consider residual ACF and residual PACF now.

2.8.2.1 Autocorrelation Function for Residuals



In both seasonal and non – seasonal area, from 1st lag the absolute values of t statistics are not greater than 2 (|T|<2).

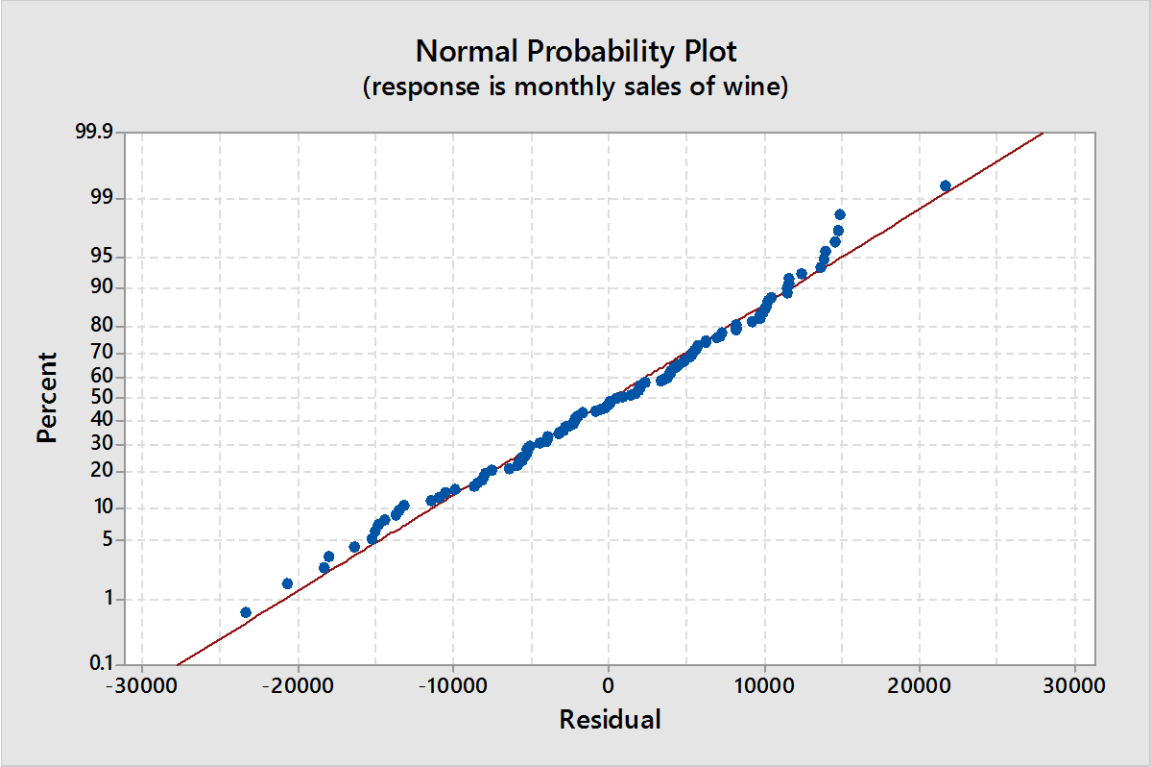
2.8.2.2 Partial Autocorrelation Function for Residuals



In both seasonal and non – seasonal area, from 1st lag the absolute values of t statistics are not greater than 2(|T|<2).

Therefore, Residuals are randomly distributed even though the Ljung box Chi – Square Statistic is not indicated as the residuals are random.

2.8.3. Normality of the Residuals



Almost all the data points lie on the straight line of Normal Probability Plot. So, it is concluded that residuals are normally distributed.

The correlation matrix of the estimated parameters is empty

There is no high correlation between terms because all the correlation values < 0.8. Therefore, parameter redundancy does not exist.

Therefore, the model **SARIMA $(0,1,0)(0,1,1)_{12}$** is adequate with obeying the following

Conditions;

- ✓ The parameters are significant.
- ✓ Residuals are random.
- ✓ Residuals are normally distributed.
- ✓ No parameter redundancy.

3 CONCLUSION

3.1 Accuracy Checking

	Fitted Model	MAPE	Accuracy
1	SARIMA $(0,1,0)(2,1,0)_{12}$	7.23051	92.7695
2	SARIMA $(0,1,0)(1,1,0)_{12}$	0.0466514	99.9533
3	SARIMA $(0,1,0)(1,1,0)_{12}$	11.4995	88.5004

By considering the accuracy of the models SARIMA $(0,1,0)(2,1,0)_{12}$, SARIMA $(0,1,0)(1,1,0)_{12}$ and SARIMA $(0,1,0)(1,1,0)_{12}$ the model SARIMA $(0,1,0)(1,1,0)_{12}$ has the highest accuracy and the lowest MAPE value.

According to the statistical theory, the model SARIMA $(0,1,0)(1,1,0)_{12}$ is considered the Best-fitted model for this series of data.

$$\mathbf{X}_t - 0.5834\,\mathbf{X}_{t-12} + \mathbf{X}_{t-1} - 0.4166\,\mathbf{X}_{t-13} - 0.4166\mathbf{X}_{t-24} + 0.4166\mathbf{X}_{t-25} = \mathbf{Z}_t$$

3.2 Forecasting

Period	Forecast	Lower	Upper	Actual
109	109630	90137	129122	109601
110	96358	68792	123924	96311
111	99501	65739	133262	99441
112	88753	49768	127738	88705
113	85769	42182	129355	85719
114	91173	43427	138919	91119
115	93862	42290	145433	93807
116	75931	20799	131064	75889
117	79068	20591	137545	79065
118	88952	27312	150592	88915
119	111207	46558	175855	111155
120	127013	59490	194537	126953

3.2.1. Forecasting Future Values

Period	Forecast	Lower	Upper
121	114299	94807	133791
122	101278	73711	128844
123	104581	70820	138343
124	93675	54690	132659
125	90716	47129	134302
126	96184	48438	143931
127	98888	47316	150460
128	80787	25655	135920
129	83408	24931	141885
130	93738	32098	155378
131	116182	51533	180830
132	132104	64581	199628

131 DISCUSSION

In this data set of monthly sales of wine in USA wine company, the time series plot exhibits an upward trend with increasing seasonal variation. Therefore, we performed multiplicative decomposition model. After identification of the preliminary specifications of the model, estimation of parameters and diagnostic checking of the model adequacy, the best fit of a time series model was obtained.

Among the three models obtained , SARIMA $(0,1,0)(1,1,0)_{12}$ has the Highest accuracy. The model equation is,

$$X_t - 0.5834 X_{t-12} - X_{t-1} + 0.5834 X_{t-13} - 0.4166 X_{t-24} + 0.4166 X_{t-25} = Z_t$$

After forecasting last 12 values using this model, we calculated MAPE value. Then, we used our final model to forecast future values for next 12 months.