FINAL REPORT **STAT 3124**

STATISTICAL ANALYSIS OF NUMBER OF MONTHLY AIR PASSENGERS

Group P

192055 – G.A.M.K. Jayathilaka

192059 – T.N.D. Kodippily

192126 – W.L.T. Sankalpani

192135 – M.A.S. Shavindi

ACKNOWLEDGEMENT

We would like to take this opportunity to express our gratitude to everyone who has contributed to the successful completion of this project. Firstly, we would like to extend our heartfelt appreciation to our Lecturer in charge, Mrs. P.M.O.P. Panahetipola and our Assistant Lecturer, Ms. Poornima Munasinghe for providing us with valuable guidance and support throughout the project. Your feedback and advice were instrumental in shaping our ideas and improving the overall quality of our work.

We also extend our sincere thanks to our teammates for their dedication and hard work. Each of us brought unique skills and perspectives to the project, and we worked collaboratively to meet our project goals successfully. Without their commitment and enthusiasm, this project would not have been able to complete.

Furthermore, we would like to thank the website Kaggle (https://www.kaggle.com/datasets/rakannimer/air-passengers) for providing the air passenger time series data set and giving us the opportunity to work on this project and we learned a lot from the experience.

Finally, we extend our deepest appreciation to everyone who contributed to this project, and we are proud of what we have accomplished as a team.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

- ✓ ACF – Autocorrelation Function
- ✓ PACF – Partial Autocorrelation Function
- ✓ SAR – Seasonal Auto Regressive
- ✓ SMA – Seasonal Moving Average
- ✓ AR – Autoregressive
- ✓ MA – Moving Average
- ✓ ARMA – Mixed Auto Regressive Moving Average
- ✓ ARIMA – Integrated Auto Regressive Moving Average
- ✓ SARIMA – Seasonal Integrated Auto Regressive Moving Average
- ✓ AD – Anderson Darling
- ✓ i.e. – That is

1. INTRODUCTION

Time Series Analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. The behaviour of time series variables such as monthly air passengers is not consistent and to forecast it is irrational. These decisions are made under the premise that patterns exist in the previous data and these patterns provide an indication of future movement of number of air passengers If such patterns exist, then it is possible in principle to apply modern mathematical tools and techniques such as Box-Jenkins ARIMA model to forecast the number of air passengers.

The goal of this study is to perform statistical analysis on the number of air passengers from 1955 and 1960. The properties of the data are described and basic time series techniques are applied to the data. Plots of the series, autocorrelation function and the partial autocorrelation function are some of the graphical tools used to analyse the series. We also aim to fit a model to the data in order to make credible forecasts from the model. The data was downloaded from the Kaggle website (https://www.kaggle.com/datasets/rakannimer/air-passengers). A year of data is considered to be 12 months which equals 12 data points per year. A 5% level of significance is used throughout the analysis.

2. THEORY

2.1 What Is Time Series?

Time series analysis is a specific way of analysing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time.

What sets time series data apart from other data is that the analysis can show how variables change over time. In other words, time is a crucial variable because it shows how the data adjusts over the course of the data points as well as the final results. It provides an additional source of information and a set order of dependencies between the data.

Time series analysis typically requires a large number of data points to ensure consistency and reliability. An extensive data set ensures you have a representative sample size and that analysis can cut through noisy data. It also ensures that any trends or patterns discovered are not outliers and can account for seasonal variance. Additionally, time series data can be used for forecasting—predicting future data based on historical data. [1]

2.2 Components of Time Series

The causes which changes the attributes of a time series are known as the components of a time series.

The following are the components of time series:

- Trend
- Seasonal Variation
- Cyclic Variation
- Irregular Fluctuations

2.2.1 Trend

Trend shows common tendency of data. Trend is the long term change in the mean level of data. It may move upward or downward over a certain long period of time. It is not mandatory for the data to move in the same direction. The direction or movement may change over the long-term period but the overall tendency should remain the same in a trend. A trend can be either linear or non-linear.

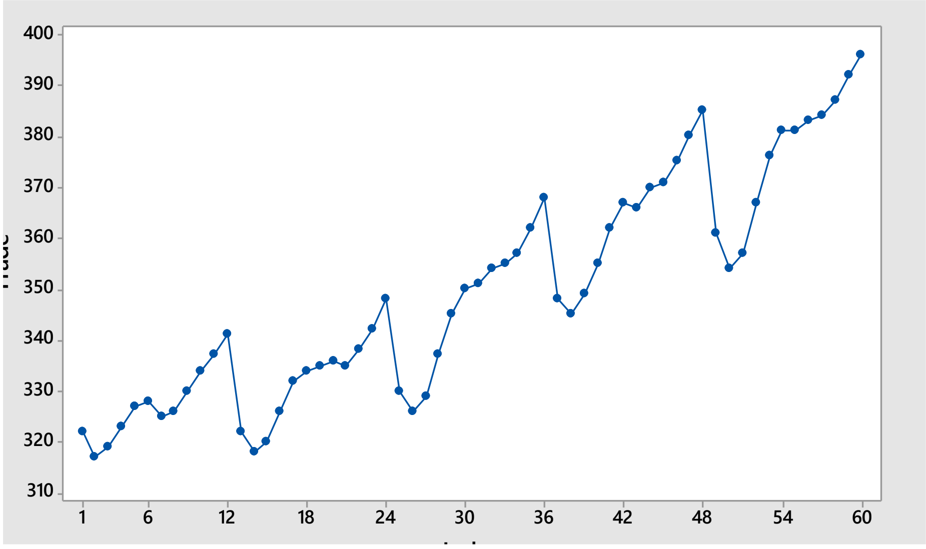

Figure 1 : Upward Trend

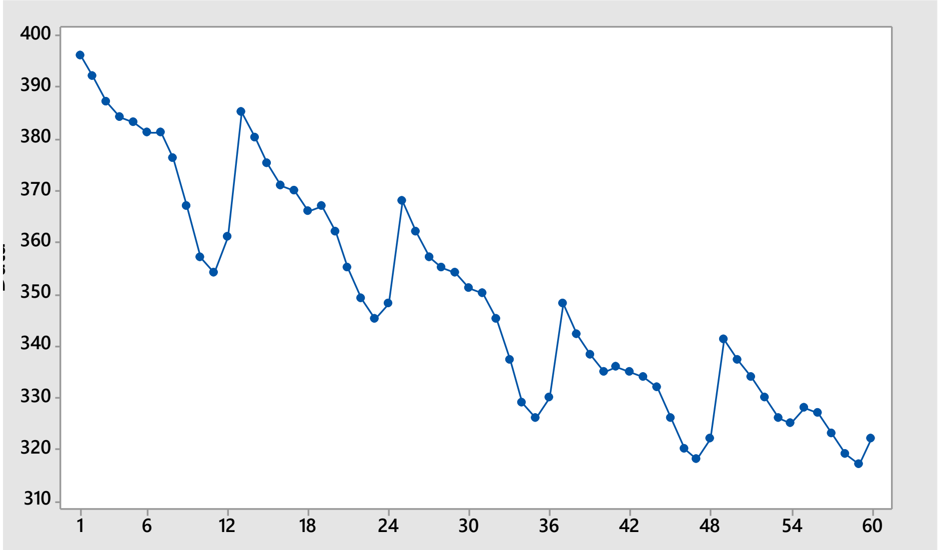

Figure 2 : Downward Trend

2.2.2 Seasonal Variation

Seasonal variations are changes in time series that occur in the short term, usually within less than 12 months. They usually show the same pattern of upward or downward growth in the 12-month period of the time series. These variations are often recorded as hourly, daily, weekly, quarterly, and monthly schedules.

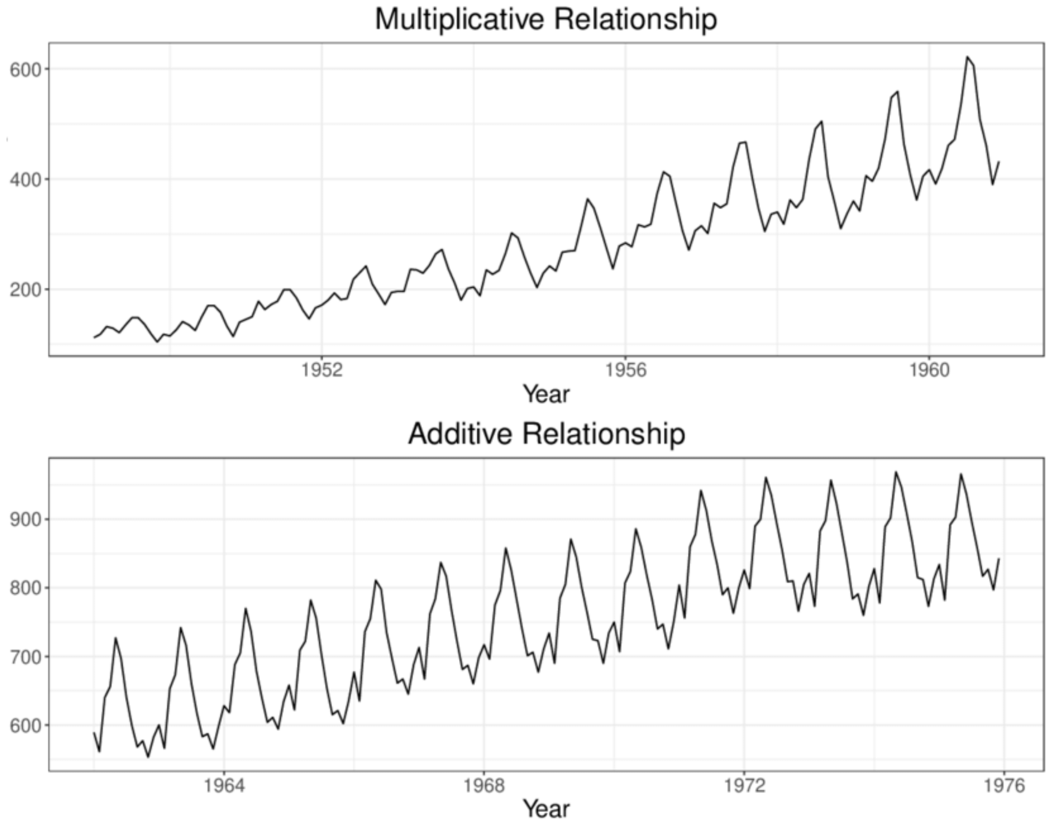

Figure 3 : Seasonal Variation

2.2.3 Cyclic Variation

Variations in time series that occur themselves for the span of more than a year are called Cyclical Variations. Such oscillatory movements of time serious often have a duration of more than a year. One complete period of operation is called a cycle.

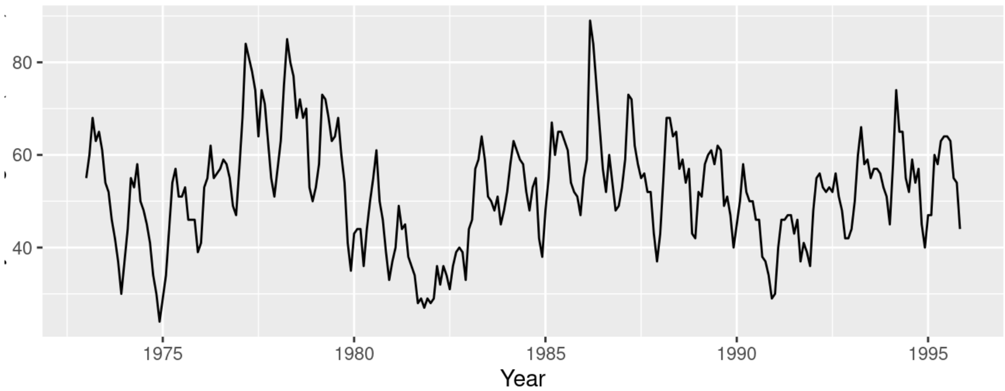

Figure 4 : Cyclic Variation

2.2.4 Irregular Fluctuations

There is another kind of movement that can be seen in the case of time series. It is pure Irregular and Random Movement. As the name suggests, no hypothesis or trend can be used to suggest irregular or random movements in a time series. These outcomes are unforeseen, erratic, unpredictable, and uncontrollable in nature. [2]

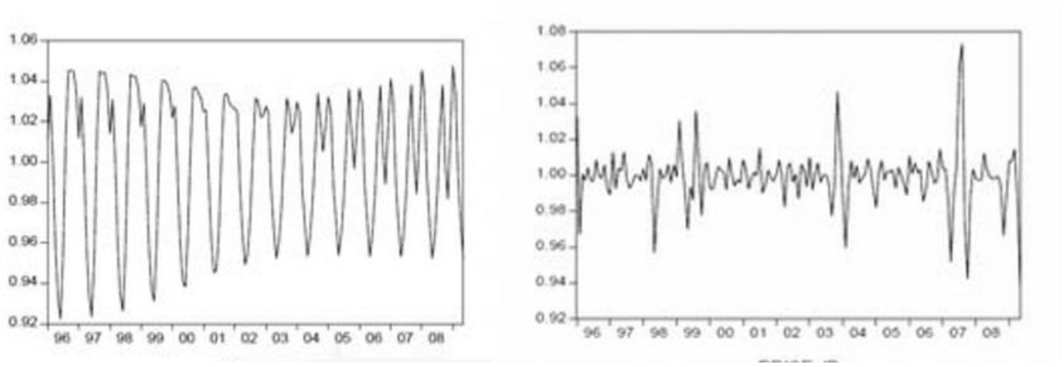

Figure 5 : Irregular Fluctuations

2.3 Traditional analysis

2.3.1 Regression models

If a time series shows a trend component only it can be modelled using regression model.

A time series y_t could be described by using a trend model.

The trend model is;

$$y_t = TR_t + \varepsilon_t \quad \text{Equation 1}$$

Where,

y_t – The value of the time series in period t

TR_t – The trend in time period t

ε_t – The error term in time period

2.3.2 Decomposition methods

Decomposition procedures are used in time series to describe the trend and seasonal factors in a time series. More extensive decompositions might also include long-run cycles, holiday effects, day of week effects and so on. Here, we'll only consider trend and seasonal decompositions.

One of the main objectives for a decomposition is to estimate seasonal effects that can be used to create and present seasonally adjusted values. A seasonally adjusted value removes the seasonal effect from a value so that trends can be seen more clearly. Decomposition is further classified into two as follows:

- Multiplicative
- Additive

2.3.2.1 Multiplicative Decomposition

The multiplicative model is useful when the seasonal variation is either increasing or decreasing over time.

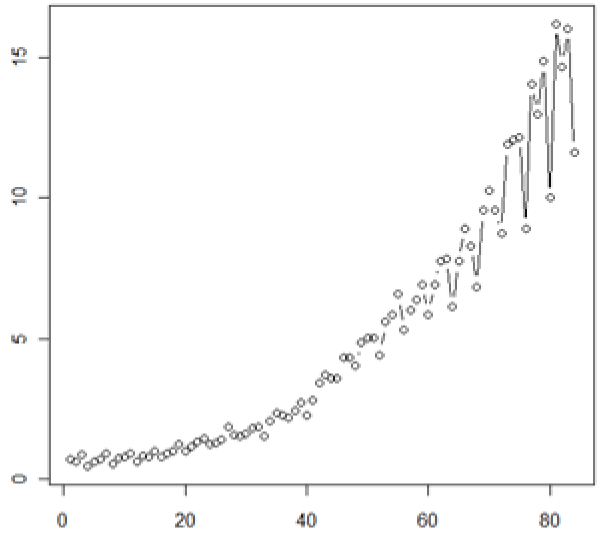

Figure 6 : Multiplicative Seasonality

The multiplicative decomposition model:

$$Y_t = TR_t \times SN_t \times CL_t \times IR_t \quad \textit{Equation 2}$$

Where,

Y_t – The observed value of the time series in time period t

TR_t – The trend component in time period t

SN_t – The seasonal component in time period t

CL_t – The cyclical component in time period t

IR_t – The irregular component in time period t

2.3.2.2 Additive Decomposition

The additive model is useful when the seasonal variation is constant over time. [3]

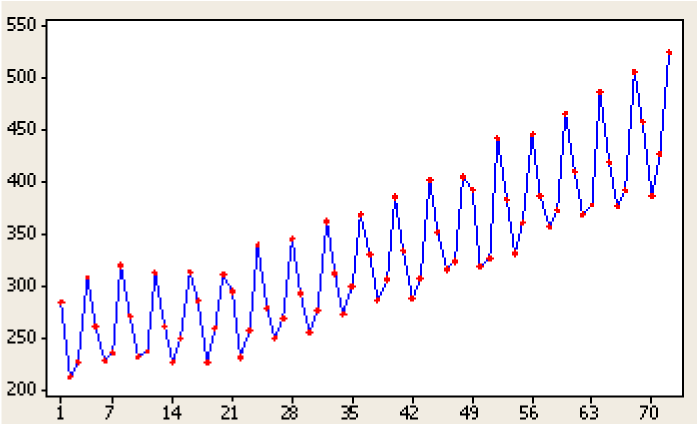

Figure 7 : Additive Seasonality

The additive decomposition model:

$$Y_t = TR_t + SN_t + CL_t + IR_t \quad \text{Equation 3}$$

Where,

Y_t – The observed value of the time series in time period t

TR_t – The trend component in time period t

SN_t – The seasonal component in time period t

CL_t – The cyclical component in time period t

IR_t – The irregular component in time period t

2.4 Probability Models

There are several number of probability models which can be modelled using time series data.

- A purely random process
- Random walk
- Autoregressive process – AR(p)
- Moving average process – MA(q)

- Autoregressive moving average process – ARMA (p, q)
- Integrated autoregressive moving average process – ARIMA (p, d, q)
- Seasonal autoregressive process – SAR(P)
- Seasonal moving average process – SMA(Q)
- Seasonal integrated autoregressive process – SARIMA (p, d, q) (P, D, Q) $_s$

2.4.1 ARIMA Process

The ARIMA (p, d, q) model:

$$\phi_p(B)(1-B)^d X_t = \theta_q(B) Z_t \quad \textit{Equation 4}$$

Where,

$$\phi_P(B) = 1 - \alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_p B^p$$

$$\theta_q(B) = 1 + \beta_1 B + \beta_2 B^2 + \cdots + \beta_q B^q$$

2.4.2 SARIMA Process

The SARIMA (p, d, q) (P, D, Q) $_S$ model:

$$\phi_p(B)\Phi_P(B^S)(1-B^S)^D(1-B)^d X_t = \theta_q(B)\Theta_Q(B) Z_t \quad \textit{Equation 5}$$

Where,

$$\phi_P(B) = 1 - \alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_p B^p$$

$$\theta_q(B) = 1 + \beta_1 B + \beta_2 B^2 + \cdots + \beta_q B^q$$

$$\Phi_P(B) = 1 - \alpha'_1 B - \alpha'_2 B^2 - \cdots - \alpha'_P B^P$$

$$\theta_Q(B) = 1 + \beta'_1 B + \beta'_2 B^2 + \cdots + \beta'_Q B^Q$$

B – The backward shift operator

d – The number of non-seasonal differencing

D – The number of seasonal differencing

q – The number of non-seasonal moving average parameters

Q – The number of seasonal moving average parameters

p – The number of non-seasonal auto regression parameters

P – The number of seasonal auto regression parameters

$\alpha_1,\ldots,\alpha_p, \beta_1,\ldots \beta_q, \alpha'_1,\ldots,\alpha'_P, \beta'_1,\ldots,\beta'_Q$ - Coefficients of each parameter

2.5 Stationarity

A stationary time series is one whose properties do not depend on the time at which the series is observed. Thus, time series with trends, or with seasonality, are not stationary — the trend and seasonality will affect the value of the time series at different times. [4]

2.6 Autocorrelation

Autocorrelation is the correlation between two values in a time series. Correlation between observations a distance k apart is;

$$\gamma_k = \frac{\sum_{t=1}^{N-k}(x_t-\bar{x})(x_{t+k}-\bar{x})}{\sum_{t=1}^{N}(x_t-\bar{x})^2} \quad \textit{Equation 6}$$

$$\rho_k = \frac{\gamma_2}{\gamma_0} \quad \textit{Equation 7}$$

Where,

γ_k – Theoretical auto covariance

ρ_k – Theoretical autocorrelation

Using the autocorrelation function (ACF) we can identify which lags have significant correlations, understand the patterns and properties of the time series, and then use that information to model the time series data. From the ACF, you can assess the randomness and stationarity of a time series. You can also determine whether trends and seasonal patterns are present.

In an ACF plot, each bar represents the size and direction of the correlation. Bars that extend across the red line are statistically significant. For random data, autocorrelations should be near zero for all lags. The autocorrelation function declines to near zero rapidly for a stationary time series. In contrast, the ACF drops slowly for a non-stationary time series. When trends are present in a time series, shorter lags typically have large positive correlations because observations closer in time tend to have similar values. The correlations taper off slowly as the lags increase. When seasonal patterns are present, the autocorrelations are larger for lags at multiples of the seasonal frequency than for other lags. When a time series has both a trend and seasonality, the ACF plot displays a mixture of both effects.

2.7 Partial Autocorrelation

The partial autocorrelation function is similar to the ACF except that it displays only the correlation between two observations that the shorter lags between those observations do not explain. The partial autocorrelation function (PACF) is more useful during the specification process for an autoregressive model. [5]

2.8 Box-Jenkins Methodology

The Box-Jenkins Model is a mathematical model designed to forecast data ranges based on inputs from a specified time series. The Box-Jenkins Model can analyse several different types of time series data for forecasting purposes.

Its methodology uses differences between data points to determine outcomes. The methodology allows the model to identify trends using auto regression, moving averages, and seasonal differencing to generate forecasts.

Autoregressive integrated moving average (ARIMA) models are a form of Box-Jenkins model. The terms ARIMA and Box-Jenkins are sometimes used interchangeably. [6]

The Box-Jenkins methodology consists of five-step for identifying, selecting, and assessing conditional mean models (for discrete, univariate time series data).

- Determine whether the time series is stationarity. If the series is not stationary, successively difference it to attain stationarity. The sample autocorrelation function (ACF) and partial autocorrelation function (PACF) of a stationary series decay exponentially (or cut off completely after a few lags).
- Identify a stationary conditional mean model for the series. The sample ACF and PACF functions can help with this selection. For an autoregressive (AR) process, the sample ACF decays gradually, but the sample PACF cuts off after a few lags. Conversely, for a moving average (MA) process, the sample ACF cuts off after a few lags, but the sample PACF decays gradually. If both the ACF and PACF decay gradually, consider an ARMA model.
- Create a model template for estimation, and then fit the model to the series.
- Conduct goodness-of-fit checks to ensure the model describes the series adequately. Residuals should be uncorrelated, homoscedastic, and normally distributed with constant mean and variance.
- After choosing a model check its fit and forecasting ability and then you can use the model to forecast. [7]

2.8.1 Test Statistic for Autocorrelation

Hypothesis:

$H_0: \rho_k = 0$

$H_1: \rho_k \neq 0$

T distribution statistic:

$$t_{\gamma_k} = \frac{\gamma_k}{\frac{1}{\sqrt{n}}\sqrt{1+2\sum_{j=1}^{k-1}\gamma_k{}^2}} \quad \text{Equation 8}$$

Where,

γ_k – Sample autocorrelation at lag k

ρ_k – Autocorrelation at lag k

n – number of data in the series

If $|t_{\gamma_k}|$>2, the null hypothesis can be rejected. i.e. Autocorrelation is statistically significant from 0.

2.8.2Test Statistic for Partial Autocorrelation

Hypothesis:

$H_0: \rho_{kk} = 0$

$H_1: \rho_{kk} \neq 0$

T distribution statistic:

$$t = \frac{\gamma_{kk}}{\frac{1}{\sqrt{n}}} \quad \text{Equation 9}$$

Where,

γ_k – Sample autocorrelation at lag k

ρ_{kk} – Partial autocorrelation at lag k

n – number of data in the series

If |t|>2, the null hypothesis can be rejected. i.e. Partial autocorrelation is statistically significant from 0.

2.9 Parameter Estimation

Hypothesis:

H_0: Constant = 0 vs H_1: Not so

H'_0: Coefficient = 0 vs H'_1: Not so

If p-value of the parameter is less than level of significance, H_0 and H'_0 can be rejected. i.e. coefficient of the parameter and the constant are statistically significant from 0. Parameters of tentative model must be modified until all parameters are significant from 0.

2.10 Diagnostic Checking

Before forecasting with the fitted model it is necessary to perform a model adequacy tests to validate the good ness of fit of the fitted model. The best way to check the adequacy of box- Jenkins model is to analyse the residuals.

Characteristics of a good model:

- The residuals are random
- The residuals are approximately normally distributed
- All parameter estimates are significantly different from zero.

2.10.1 Significance of Parameters

Hypothesis:

H_0: Constant = 0 vs H_1: Not so

H'_0: Coefficient = 0 vs H'_1: Not so

If p-value < α the level of significance, the null hypothesis is rejected. i. e the parameters are significant from 0.

2.10.2 Randomness of Residuals

- Using ACF and PACF of residuals

  If residuals are random ACF and PACF statistically equals to zero.

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

  Hypothesis:

$$H_0: \rho_1 = \rho_2 = \cdots \rho_k = 0$$

$$H_1: \rho_1 \neq \rho_2 \neq \cdots \rho_k \neq 0$$

If p-value > α the level of significance, null hypothesis is not rejected. i. e the residuals are random.

2.10.3 Normality of Residuals

Bell shape in histogram or straight line pattern in normal probability plot indicates the normality of residuals. Use normal probability plot to look for the following:

Not a straight line	→	Non-normality
Curve in the tails	→	Skewness
A point far away from the line	→	An outlier
Changing slope	→	An unidentified variable

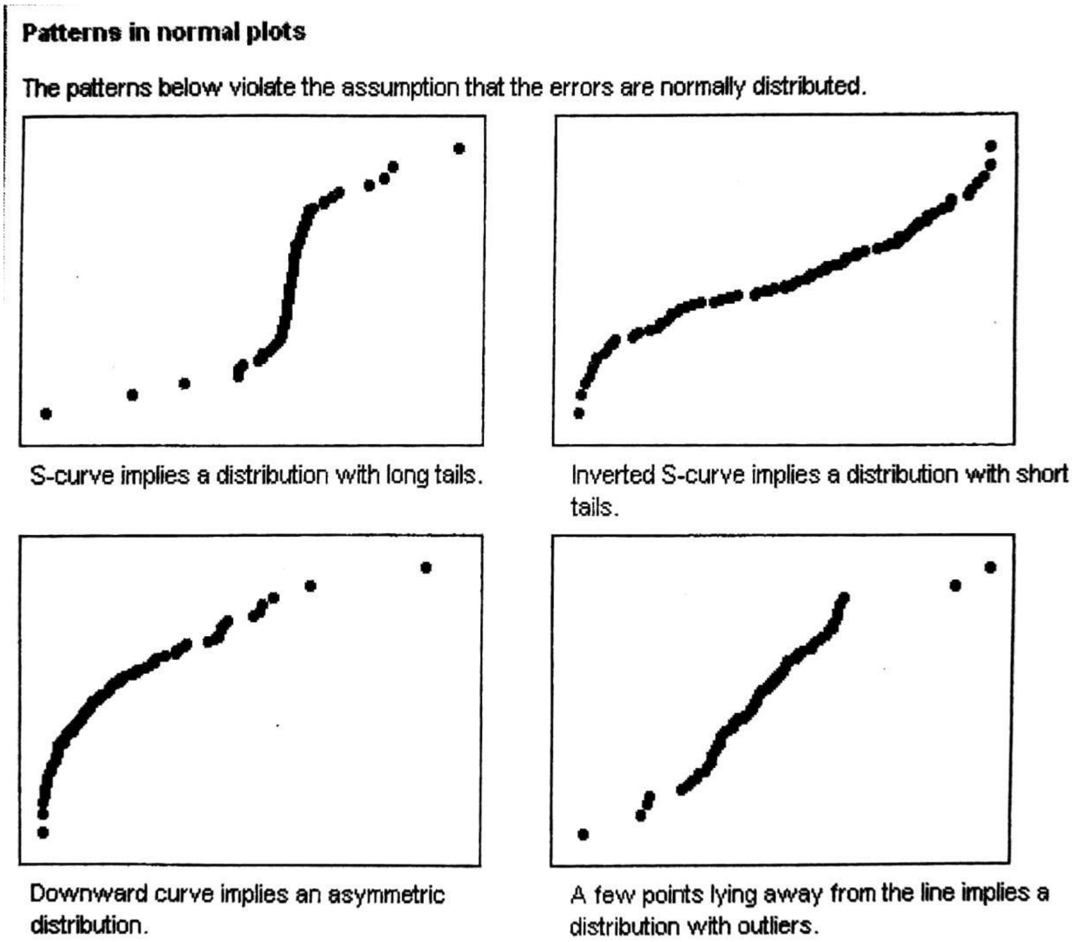

Figure 8 : Non Normal Patterns in Normal Probability Plot

If the dataset has fewer than 50 observations, the plot may display curvature in the tails even if the residuals are normally distributed. As the number of observations decreases, the probability plot may show even greater variation and nonlinearity. Using the normal probability plot and goodness-of-fit tests the normality of residuals in small data sets can be accessed.

2.10.4 Goodness of Fit Tests

Testing for normality is often a first step in analysing your data. Many statistical tools you might use have normality as an underlying assumption. If you fail that assumption, you may need to use a different statistical tool or approach.

2.10.4.1 Anderson Darling Test

The Anderson-Darling test is used to test if a sample of data comes from a population with a specific distribution. Its most common use is for testing whether your data comes from a normal distribution.

The normal distribution is a theoretical distribution. What you are really testing with the AD test is not whether your data is exactly consistent with a normal distribution, but whether your data is close enough to normal that you can use your statistical tool without concern.

In some cases, a statistical tool may be robust to the normality assumption, which means the statistical tool is not overly sensitive to some level of violation of the normality assumption. The normal distribution is popular because it describes many real-life situations, such as the distribution of people's heights, weights, and income.

The AD test is really a hypothesis test. The null hypothesis (H_o) is that your data is not different from normal. Your alternate or alternative hypothesis (H_1) is that your data is different from normal. You will make your decision about whether to reject or not reject the null based on your p-value.

Assuming you selected your alpha risk to be α, you will reject the null hypothesis if the p-value is less than α. That allows you to claim that your data is statistically different from a normal distribution. On the other hand, if your p-value is higher than α, you can state that your data is not statistically different from a normal distribution. [8]

2.10.5 Parameter Redundancy

The correlation matrix for estimated parameters provides a mean for recognizing the existence of parameter redundancy. A very high correlation (|correlation|>0.8/0.9) suggest parameter redundancy.

2.11 Forecasting

Time series forecasting occurs when you make scientific predictions based on historical time stamped data. It involves building models through historical analysis and using them to make observations and drive future strategic decision-making.
Forecasting methods may be broadly classified in to three categories.

- Subjective – Forecasting can be made using judgements, intuition, knowledge of the subject, previous experience and other relevant information
- Univariate – Forecasting is based entirely on the past observations of the time series. Usually fits a suitable model to the given data and extrapolate to the future.
- Multivariate – In this case we have to consider observations on other variables in to account in order to make forecast.

2.11.1 Point Forecasting

To obtain a point forecast, the final model (equation) must be written in terms of original data and then need to substitute respective past data in order to obtain the desired forecast value.

2.11.1 Forecasting Error

The accuracy of a forecasting model depends on how close the forecasted values ($\hat{X}t$) are to the actual values (Xt). In practice, we define the difference between the actual and the forecast values as the forecast error,

$$e_t = X_t - \hat{X}_t \quad \text{Equation 10}$$

Where,

e_t – Forecast Error

X_t – Actual Value

$\hat{X}_t$ – Forecasted Value

If the model is doing a good job in forecasting the actual data, the forecast error will be relatively small. In fact, if we have correctly modelled the data, what are left over are simply erratic fluctuations (errors) in a time series that have no definable pattern. Often, these fluctuations are caused by outside events that in themselves are not predictable. These fluctuations are caused by outside events that in themselves are not predictable. This means that et for each time period is purely random fluctuation around $\hat{X}t$. Thus, if we were to add them we should get a value equal to or near 0.

Define random forecast error as "the sum of the error terms equal to zero and the mean is equal to zero." The measure of this randomness (forecast accuracy) may be achieved by using either statistical or graphical methods.

2.11.2 Mean Absolute Error

$$MAE = \frac{\sum_{t=1}^{n} |X_t - \hat{X}_t|}{n} \quad \text{Equation 11}$$

Where,

X_t – Actual Value

$\hat{X}_t$ – Forecasted Value

n – Number of Data

2.11.3 Mean Absolute Percentage Error

Mean Absolute Percentage Error is the measure of how accurate a forecast system is. It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values.

$$MAPE = \frac{\sum_{t=1}^{n} |X_t - \hat{X}_t| / X_t}{n} \times 100 \quad \text{Equation 12}$$

Where,

X_t – Actual Value

$\hat{X}_t$ – Forecasted Value

n – Number of Data

2.11.4 Akaike Information Criterion (AIC)

The AIC is defined as

$$AIC = log(residual\ sum\ of\ square) + \frac{2}{n}k \quad \text{Equation 13}$$

Where n is the number of observations in the model and k is the number of parameters in the model.

2.11.2 The Best Model.

The model with less mean absolute percentage error, less mean absolute error and lower AIC is the best model for forecast or the model with higher accuracy is chosen as the best fit model for forecasting. (Accuracy is obtained by, Accuracy = 100-MAPE Value)

3. STATISTICAL ANALYSIS

3.1 Dataset

Table 1 : Air Passenger Data Set

	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
January	112	115	145	171	196	204	242	284	315	340	360	417
February	118	126	150	180	196	188	233	277	301	318	342	391
March	132	141	178	193	236	235	267	317	356	362	406	419
April	129	135	163	181	235	227	269	313	348	348	396	461
May	121	125	172	183	229	234	270	318	355	363	420	472
June	135	149	178	218	243	264	315	374	422	435	472	535
July	148	170	199	230	264	302	364	413	465	491	548	622
August	148	170	199	242	272	293	347	405	467	505	559	606
September	136	158	184	209	237	259	312	355	404	404	463	508
October	119	133	162	191	211	229	274	306	347	359	407	461
November	104	114	146	172	180	203	237	271	305	310	362	390
December	118	140	166	194	201	229	278	306	336	337	405	432

3.2 Time Series Plot

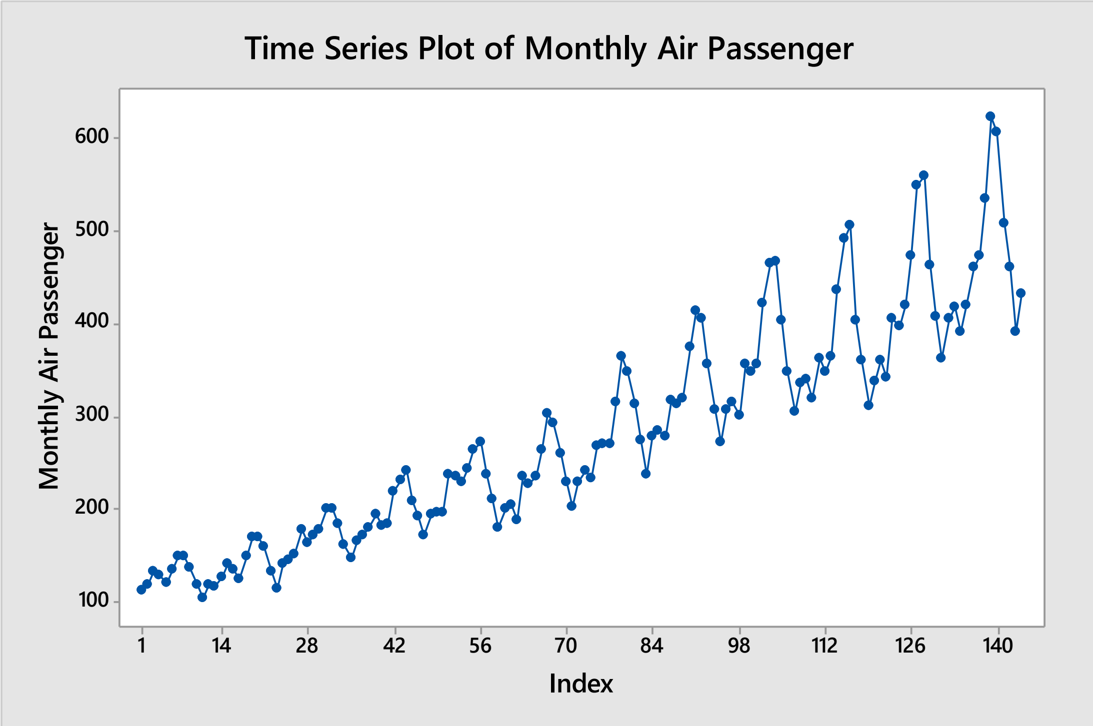

Figure 9 : Time Series Plot of Air Passenger Data

An upward trend and an increasing seasonal variation with lag 12 was indicated in time series plot.

3.3 Multiplicative Decomposition

Since seasonal variation was increasing over time multiplicative decomposition technique was used to analyse the trend.

Fitted Trend Equation,

$$Y_t = 88.42 + 2.6447 \times t \quad \textit{Equation 14}$$

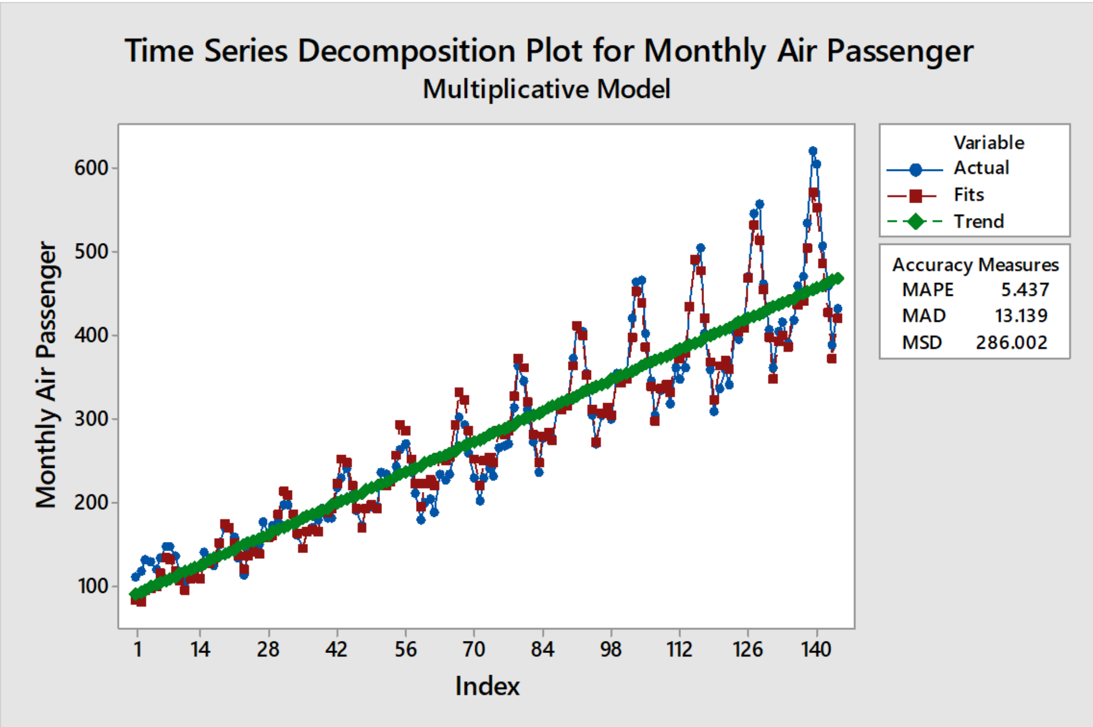

Figure 10 : Decomposition Plot of Air Passenger Data

<u>Seasonal Indices:</u>

Table 2 : Seasonal Indices of Multiplicative Decomposition

Period	Index
January	0.909268
February	0.874866
March	0.996728
April	0.974048
May	0.981221
June	1.114614
July	1.254857
August	1.208643
September	1.059133
October	0.92322
November	0.802955
December	0.900446

3.4 Autocorrelation Function

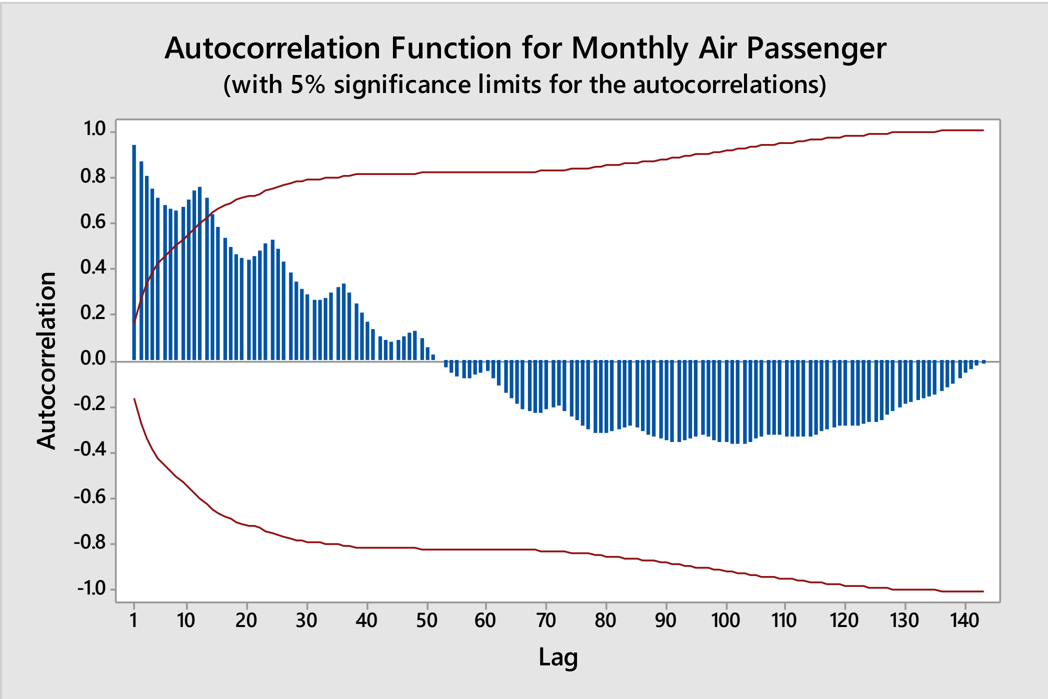

Figure 11 : Autocorrelation of Original Series

Table 3 : Autocorrelation of Original Data

	lag	ACF	T Statistic	LBQ
Non Seasonal Area	1	0.948047341	11.37656809	132.1415386
	2	0.875574835	6.281779328	245.6461603
	3	0.806681155	4.651538089	342.6748259
	4	0.752625417	3.805541848	427.7386836
	5	0.713769973	3.293054415	504.7965704
	6	0.681733603	2.932179205	575.6018536
	7	0.662904386	2.694831978	643.0385934
	8	0.655610484	2.540154295	709.4844982
	9	0.670948328	2.490384807	779.5912312
	10	0.702719921	2.502746835	857.0686386
	11	0.743240189	2.538924494	944.3903175
Seasonal Area	12	0.760395042	2.488515456	1036.481907
	24	0.53218983	1.397921623	1606.083817
	36	0.337023599	0.824754603	1866.625062
	48	0.132634564	0.318993692	1933.155822
	60	-0.046933623	-0.112605735	1943.671149

ACF died down slowly. There was a seasonal pattern too. Thus the original series was non-stationary.

Therefore, it was required to perform a suitable difference in order to make the original series stationary.

3.5 Tentative Model 01

Considering the trend component first, it was required to perform a non-seasonal differencing in order to make the series stationary.

3.5.1 Autocorrelation Function of Non-Seasonally Differenced Series

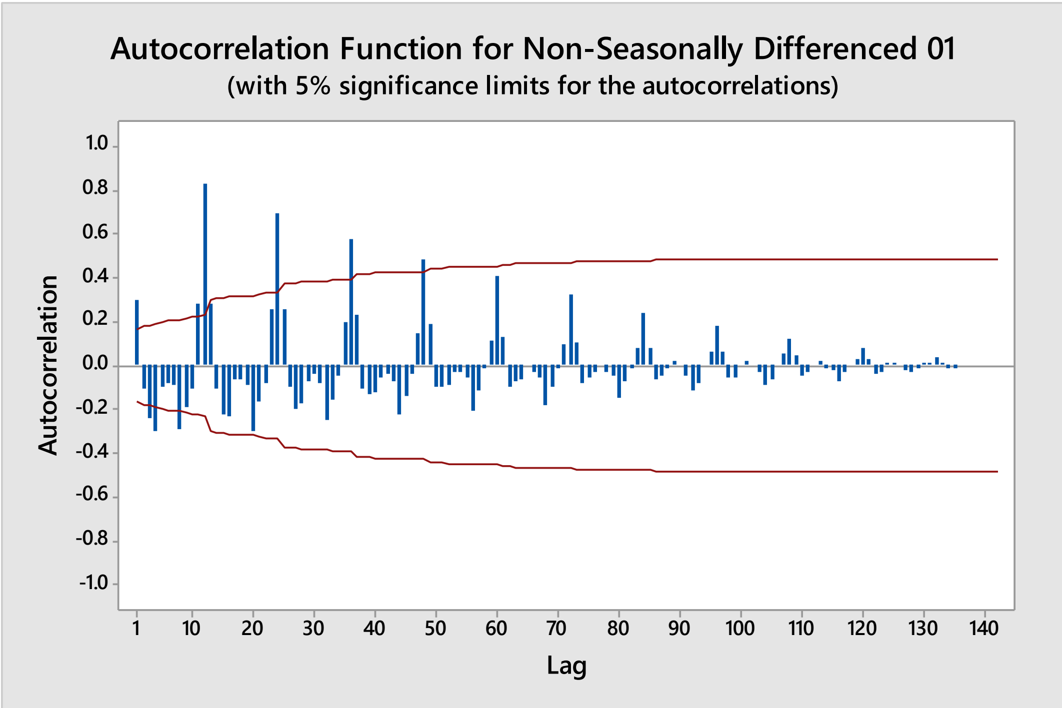

Figure 12 : Autocorrelation of Differenced Series

ACF showed a seasonal pattern. Thus the differenced series was non-stationary. Therefore, it was required to perform a seasonal difference in order to make the differenced series stationary.

3.5.2 Autocorrelation Function of Seasonally Differenced Series

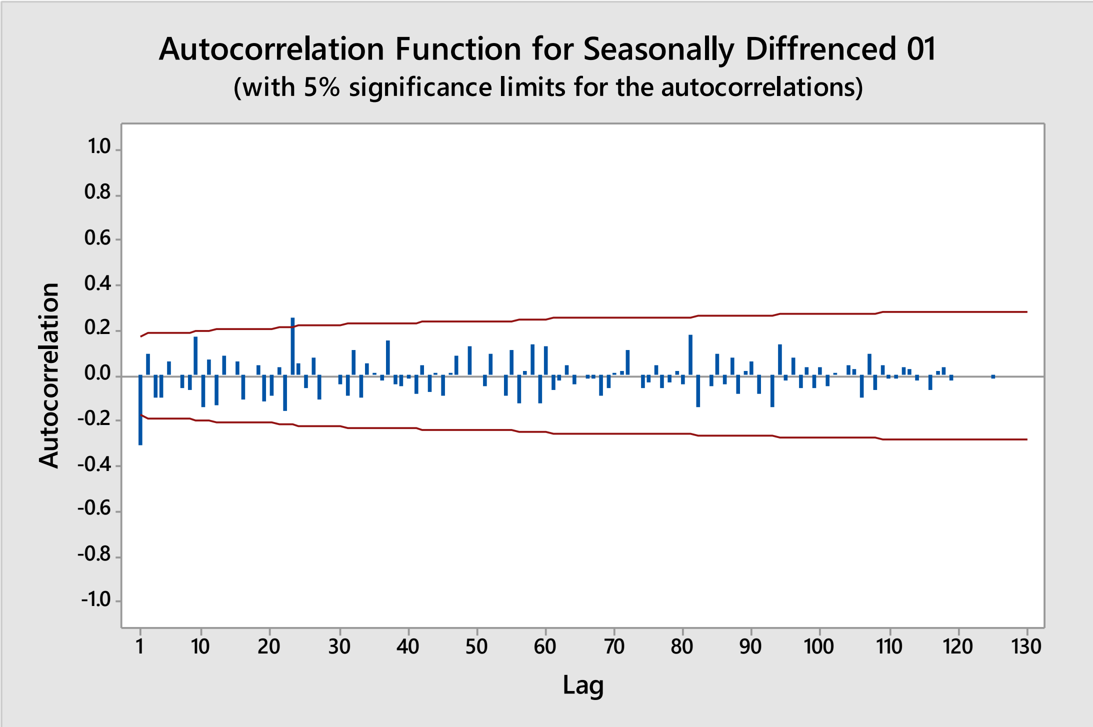

Figure 13 : Autocorrelation Function of Seasonally Differenced Series

Table 4 : Autocorrelation Function of Seasonally Differenced Series

	lag	ACF	T Statistic	LBQ
Non Seasonal Area	1	-0.309814643	-3.545990668	12.8642202
	2	0.095351458	0.999609289	14.09219062
	3	-0.096890891	-1.008087602	15.37003765
	4	-0.098995034	-1.022081511	16.7144919
	5	0.061000707	0.624843225	17.22903543
	6	-0.000287806	-0.002939312	17.22904698
	7	-0.056108484	-0.573025816	17.67138913
	8	-0.060965772	-0.621077605	18.19787895
	9	0.175916925	1.78686754	22.61743203
	10	-0.14027891	-1.391367506	25.45092626
	11	0.069735328	0.681675021	26.15699565
Seasonal	12	-0.133673434	-1.302070668	28.77316701
	24	0.052836049	0.464327546	51.36242016
	36	-0.018646436	-0.158426106	61.1863815

The absolute value of T statistic of non-seasonal lag 1 was greater than 2.

i.e. Autocorrelation of non-seasonal lag 1 was significant from 0.

i.e. ACF cut off at non-seasonal lag 1 and 0 at seasonal lags.

i.e. The differenced series is stationary.

3.5.3 Partial Autocorrelation Function of Stationary Series

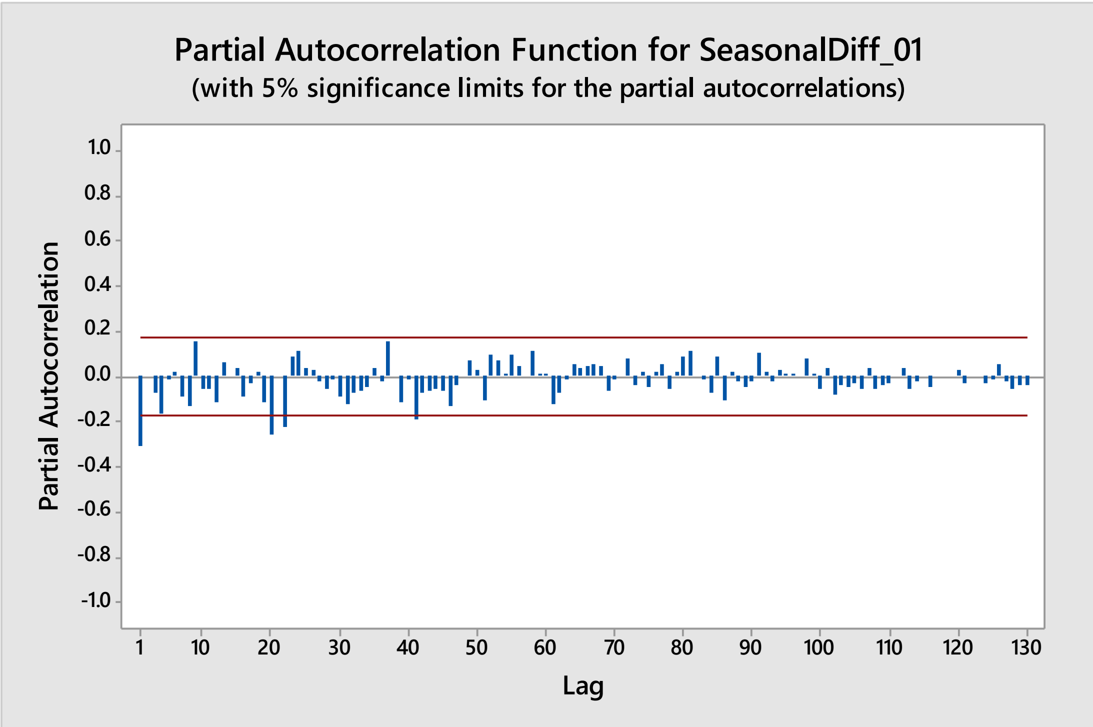

Figure 14 : Partial Autocorrelation of Stationary Series

Table 5 : Partial Autocorrelation of Stationary Series

	lag	PACF	T Statistic
Non Seasonal Area	1	-0.30982	-3.54599
	2	-0.0007	-0.00802
	3	-0.07472	-0.85519
	4	-0.16676	-1.90866
	5	-0.01515	-0.17336
	6	0.018288	0.20931
	7	-0.08809	-1.00819
	8	-0.13389	-1.53241
	9	0.156405	1.79014
	10	-0.05875	-0.67242
	11	-0.05243	-0.60007
Seasonal Area	12	-0.11501	-1.31638
	24	0.117491	1.34475
	36	-0.0212	-0.24268
	48	0.001826	0.0209
	60	0.012838	0.14694

The absolute value of T statistic of non-seasonal lag 1 was greater than 2.
i.e. Partial Autocorrelation of non-seasonal lag 1 was significant from 0.
i.e. PACF was cut off at non-seasonal lag 1 and 0 at seasonal lags.

3.5.4 Tentative Model

- Number of non-seasonal differences: 1 → d=1
- Number of seasonal differences: 1 → D=1
- ACF at non-seasonal lag: 1→ q=1
- ACF at seasonal lag: 0 → Q=0
- PACF at non-seasonal lag: 1 → p=1
- PACF at seasonal lag: 0 → P=0

Identified tentative model;

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

3.5.3 Diagnostic Checking

Considering the order of removing non-significant parameters in this model we could obtain two different adequate models for forecast.

3.5.3.1 Model 01

```
Final Estimates of Parameters

Type          Coef  SE Coef      T      P
AR   1    -0.3012   0.2765  -1.09  0.278
MA   1     0.0100   0.2907   0.03  0.973
Constant    0.230    1.024   0.22  0.822
```

Since P-values of constant and MA parameter are not less than 0.05, they are not significant.

After Removing Constant;

```
Final Estimates of Parameters

Type        Coef  SE Coef      T      P
AR   1  -0.3053   0.2752  -1.11  0.269
MA   1   0.0052   0.2897   0.02  0.986
```

Since P-value of AR and MA parameters are not less than 0.05, they are not significant.

After Removing MA Parameter;

```
Final Estimates of Parameters

Type        Coef  SE Coef      T      P
AR   1  -0.3099   0.0834  -3.72  0.000


Differencing: 1 regular, 1 seasonal of order 12
Number of observations:  Original series 144, after differencing 131
Residuals:    SS =  17946.8 (backforecasts excluded)
              MS =  138.1  DF = 130
```

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

Lag            12     24     36     48
Chi-Square   11.7   38.3   45.0   60.1
DF             11     23     35     47
P-Value     0.385  0.024  0.120  0.096
```

Since p-values of estimated parameters of modified model SARIMA (1,1,0) $(0,1,0)_{12}$ were less than 0.05 parameters were significant from zero.

All p-values of Modified Box-Pierce (Ljung-Box) Chi-Square statistic at seasonal lags were not greater than 0.05. Therefore, residuals were not random. ACF and PACF of residuals need to be evaluated to verify the randomness of residuals.

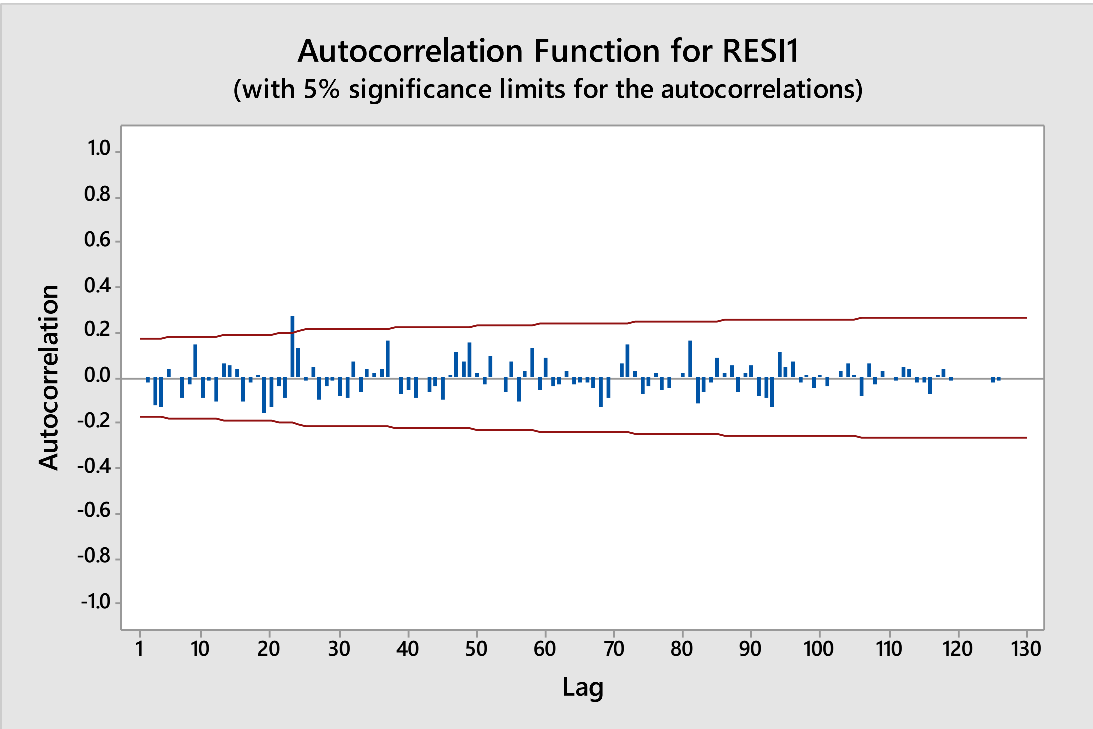

Figure 15 : Autocorrelation of Residuals of Model 01

Table 6 : Autocorrelation of Residuals of Model 01

	lag	ACF	T Statistic	LBQ
Non Seasonal Area	1	-0.000341099	-0.003904054	1.55934E-05
	2	-0.024230695	-0.277332949	0.079314093
	3	-0.118494564	-1.355436538	1.990530311
	4	-0.132856434	-1.498844048	4.412029875
	5	0.039122435	0.433987122	4.623672987
	6	0.001553131	0.017204241	4.624009212
	7	-0.088908858	-0.984853219	5.73469525
	8	-0.032543003	-0.357842666	5.884709692
	9	0.14536129	1.596831763	8.902302555
	10	-0.086985134	-0.937480894	9.991803575
	11	-0.009959305	-0.10662331	10.00620481
Seasonal Area	12	-0.108317396	-1.159534392	11.72400402
	24	0.135142997	1.278155538	38.2624426
	36	0.034225322	0.311960072	45.01276866
	48	0.071731917	0.625448869	60.06344427
	60	0.091607733	0.762937952	79.66236551
	72	0.14502264	1.178888359	96.48949766
	84	-0.021657439	-0.169504245	116.2118189
	96	0.068709039	0.522038872	147.0553094
	108	-0.027031542	-0.203198134	160.6003723
	120	-0.006140898	-0.045902912	174.1917475

ACF is 0 at both seasonal and non-seasonal lags.

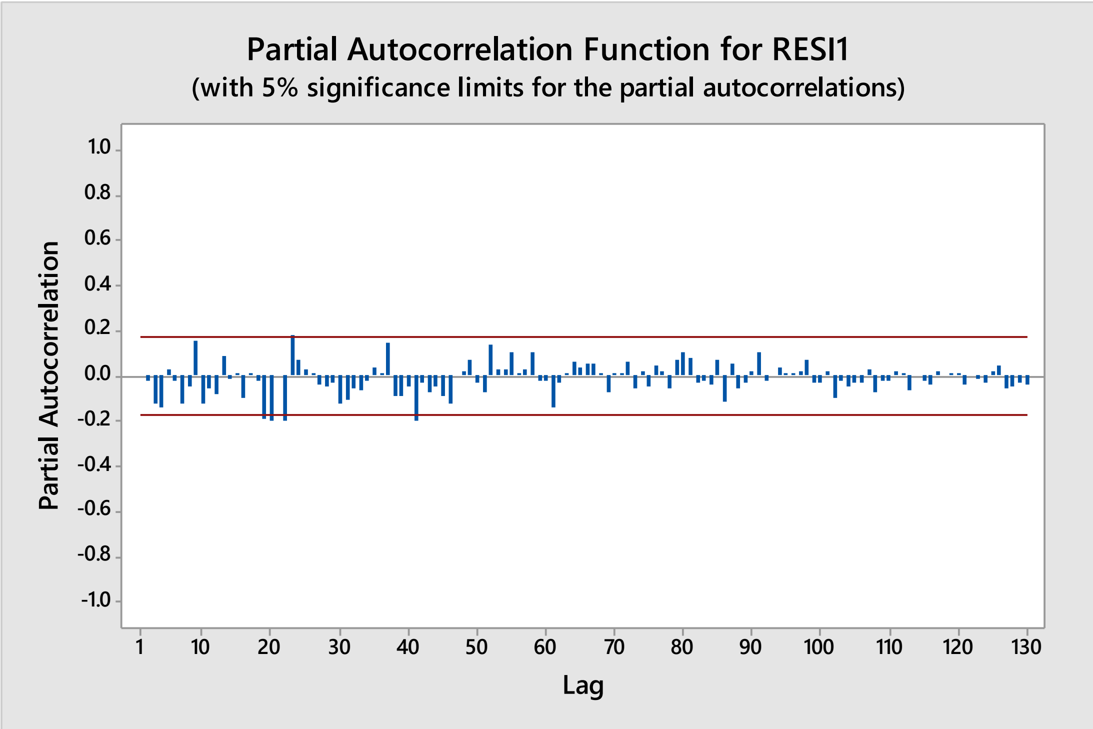

Figure 16 : Partial Autocorrelation of Residuals of Model 01

Table 7 : Partial Autocorrelation of Residuals of Model 01

	lag	PACF	T Statistic
Non Seasonal Area	1	-0.000341099	-0.003904054
	2	-0.024230814	-0.277334345
	3	-0.118580931	-1.35722079
	4	-0.135855872	-1.554941532
	5	0.031387209	0.35924303
	6	-0.018930342	-0.216667672
	7	-0.123391823	-1.412283964
	8	-0.048239988	-0.552131899
	9	0.153982246	1.762407356
	10	-0.122665319	-1.403968748
	11	-0.050960044	-0.583264363
Seasonal Area	12	-0.081471109	-0.932479466
	24	0.074327352	0.850715425
	36	0.016374903	0.187419327
	48	0.022846568	0.261490917
	60	-0.019554426	-0.223810632

PACF was 0 at both seasonal and non-seasonal lags.

Therefore, residuals were random.

Since one parameter was in the model parameter redundancy did not exist in this model.

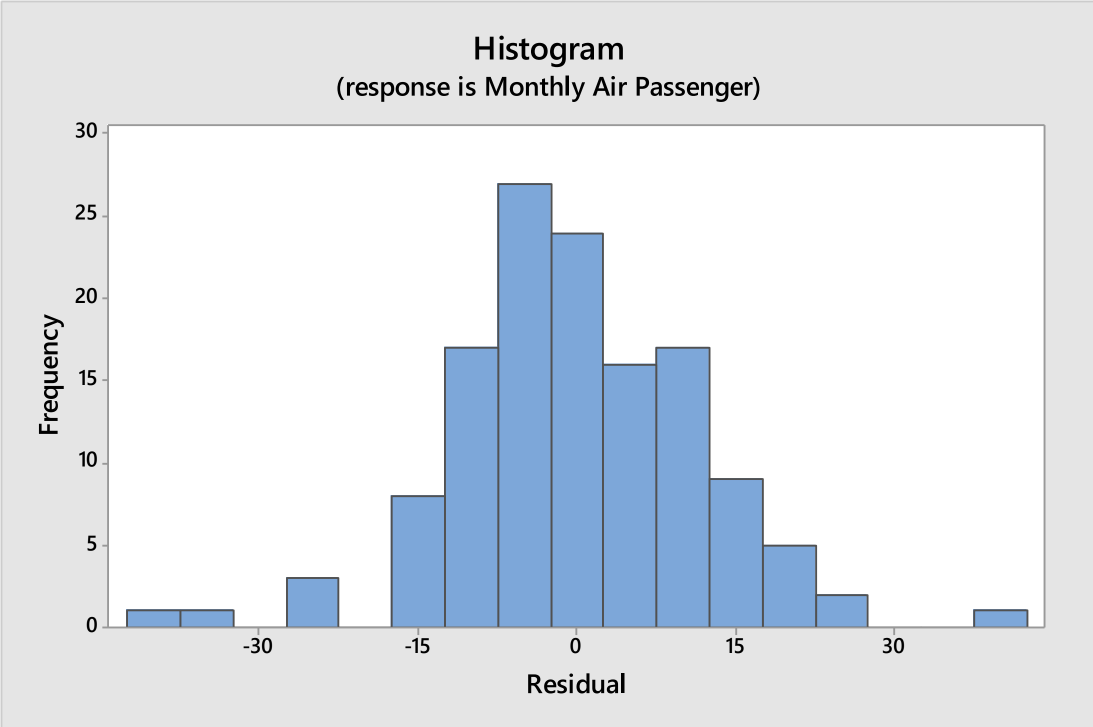

Figure 17 : Histogram of residuals of model 01

A bell shape was shown in the histogram of the residuals.

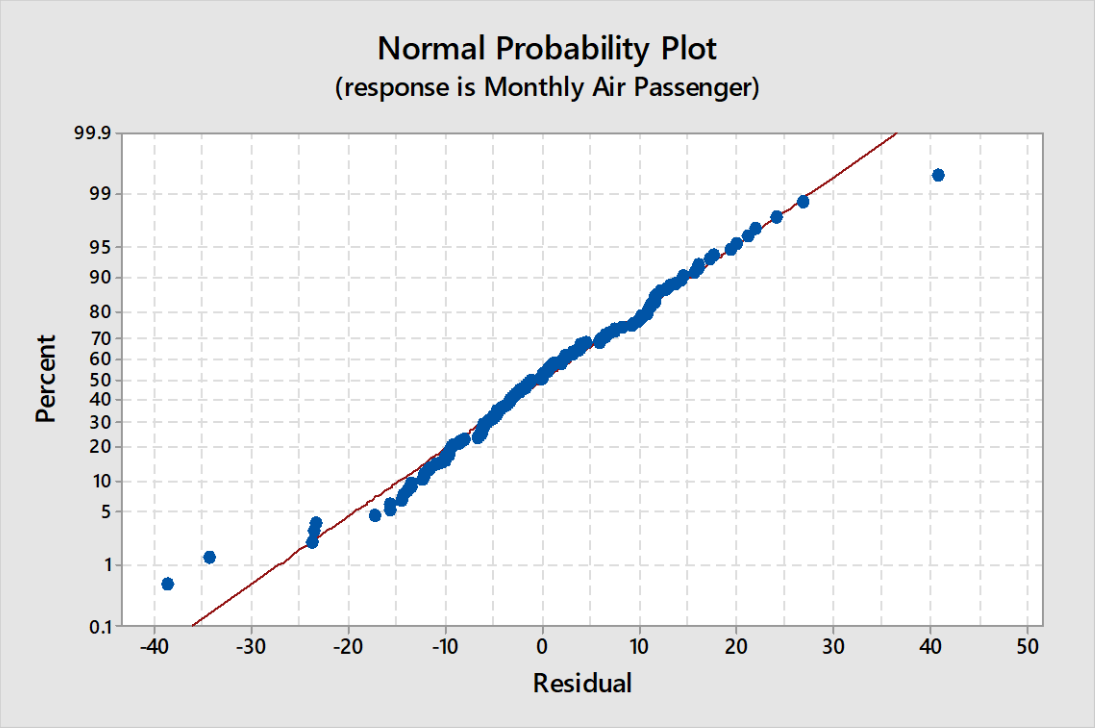

Figure 18 : Normal probability plot of residuals in model 01

Normal probabilities of residuals were approximately scattered in a straight line and some outliers were visible in the plot.

Therefore, we could conclude that the residuals are normally distributed by inspecting histogram alone.

Hence, the identified model SARIMA (1,1,0) $(0,1,0)_{12}$ was adequate.

$$(1-\alpha_1 B)(1-B^{12})(1-B)X_t = Z_t$$ *Equation 15*

$$X_t = 0.6901X_{t-1} + 0.3099X_{t-2} + X_{t-12} - 0.6901X_{t-13} - 3.099X_{t-14} + Z_t$$

Equation 16

3.5.3.2 Model 02

Changing the order of removing parameters provided another adequate model for this tentative model as thus.

```
Final Estimates of Parameters


Type            Coef  SE Coef       T      P
AR    1      -0.3012   0.2765   -1.09  0.278
MA    1       0.0100   0.2907    0.03  0.973
Constant       0.230    1.024    0.22  0.822
```

Since P-values of constant and MA parameter are not less than 0.05, they are not significant.

After Removing Constant;

```
Final Estimates of Parameters

Type       Coef  SE Coef      T      P
AR   1  -0.3053   0.2752  -1.11  0.269
MA   1   0.0052   0.2897   0.02  0.986
```

Since P-value of AR and MA parameters are not less than 0.05, they are not significant.

After Removing AR Parameter;

```
Final Estimates of Parameters

Type      Coef  SE Coef     T      P
MA   1  0.3212   0.0837  3.84  0.000


Differencing: 1 regular, 1 seasonal of order 12
Number of observations:  Original series 144, after differencing 131
Residuals:    SS =  17978.9 (backforecasts excluded)
              MS =  138.3  DF = 130


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

Lag            12     24     36     48
Chi-Square   12.0   38.6   46.1   62.1
DF             11     23     35     47
P-Value     0.367  0.022  0.099  0.069
```

All p values were less than 0.05. Therefore, parameters of modified model SARIMA (0,1,1) $(0,1,0)_{12}$ were significant.

All p-values of Modified Box-Pierce (Ljung-Box) Chi-Square statistic at seasonal lags were not greater than 0.05. Therefore, residuals were not random. It was required to inspect ACF and PACF of residuals to verify the randomness of residuals.

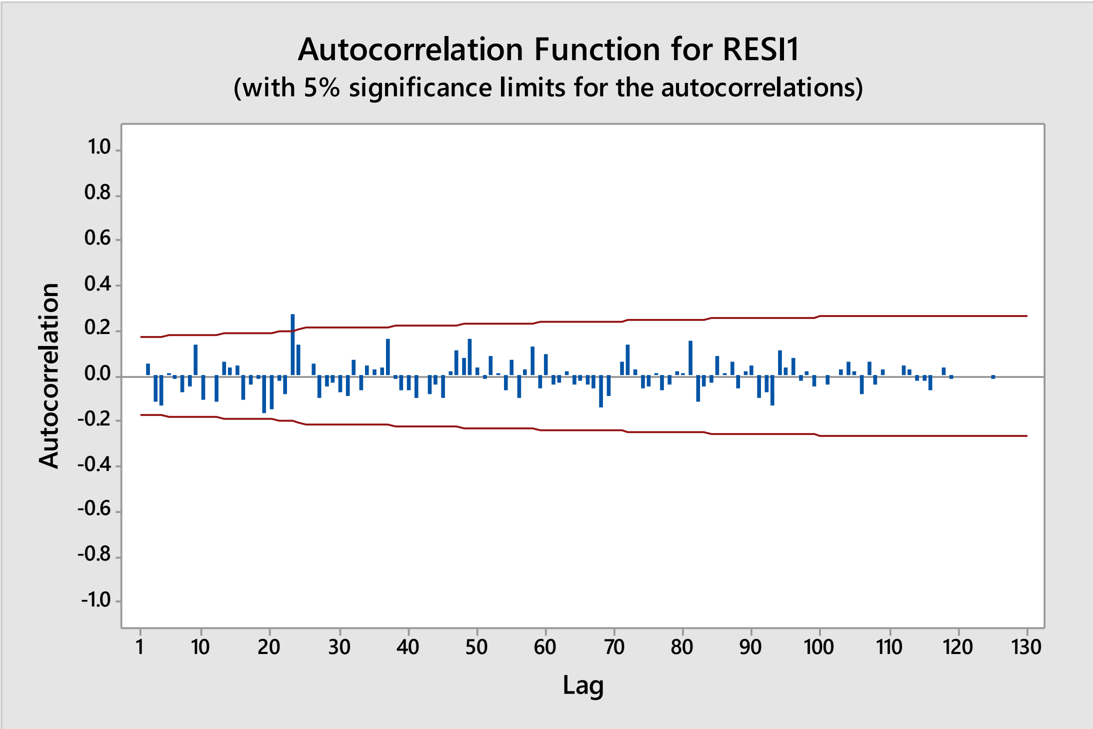

Figure 19 : Autocorrelation of Residuals of Model 02

Table 8 : Autocorrelation of Residuals of Model 02

	lag	ACF	T Statistic	LBQ
Non Seasonal Area	1	-0.002739538	-0.031355441	0.001005852
	2	0.056933352	0.651627113	0.438796764
	3	-0.116819802	-1.332740801	2.296369801
	4	-0.129513673	-1.457923634	4.597549069
	5	0.014700298	0.162858549	4.627430724
	6	-0.015217894	-0.168558657	4.659709868
	7	-0.071372399	-0.790373991	5.375460663
	8	-0.042354819	-0.466814141	5.629571679
	9	0.141408172	1.555945699	8.485268645
	10	-0.103580981	-1.119228265	10.03015881
	11	0.00482985	0.051696199	10.03354576
Seasonal Area	12	-0.11470606	-1.227728891	11.95995566
	24	0.136765604	1.291664489	38.64346969
	36	0.037743538	0.342648389	46.10628881
	48	0.079046137	0.685182902	62.12406699
	60	0.099705753	0.82550291	82.02575054

ACF was 0 at both seasonal and non-seasonal lags.

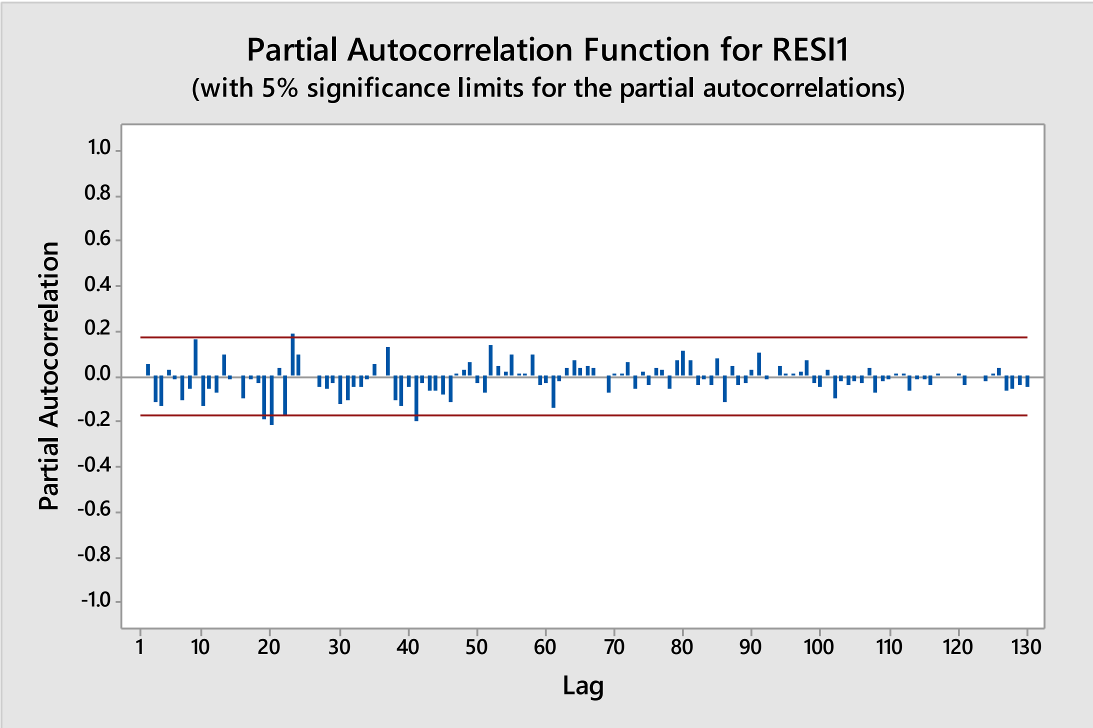

Figure 20 : Partial Autocorrelation of Residuals of Model 02

Table 9 : Partial Autocorrelation of Residuals of Model 02

	lag	PACF	T Statistic
Non Seasonal Area	1	-0.002739538	-0.031355441
	2	0.056926275	0.651550994
	3	-0.116896447	-1.337940993
	4	-0.134856995	-1.543508856
	5	0.027917416	0.319529427
	6	-0.012880839	-0.147427944
	7	-0.109075355	-1.248424495
	8	-0.05709836	-0.653520603
	9	0.162532079	1.860264677
	10	-0.13048315	-1.49344791
	11	-0.059004162	-0.675333497
Seasonal	12	-0.072187924	-0.826228556
	24	0.093705174	1.072504742
	36	0.006225119	0.071249739

PACF was 0 at both seasonal and non-seasonal lags.

Therefore, residuals are random.

parameter redundancy did not exist in this model.

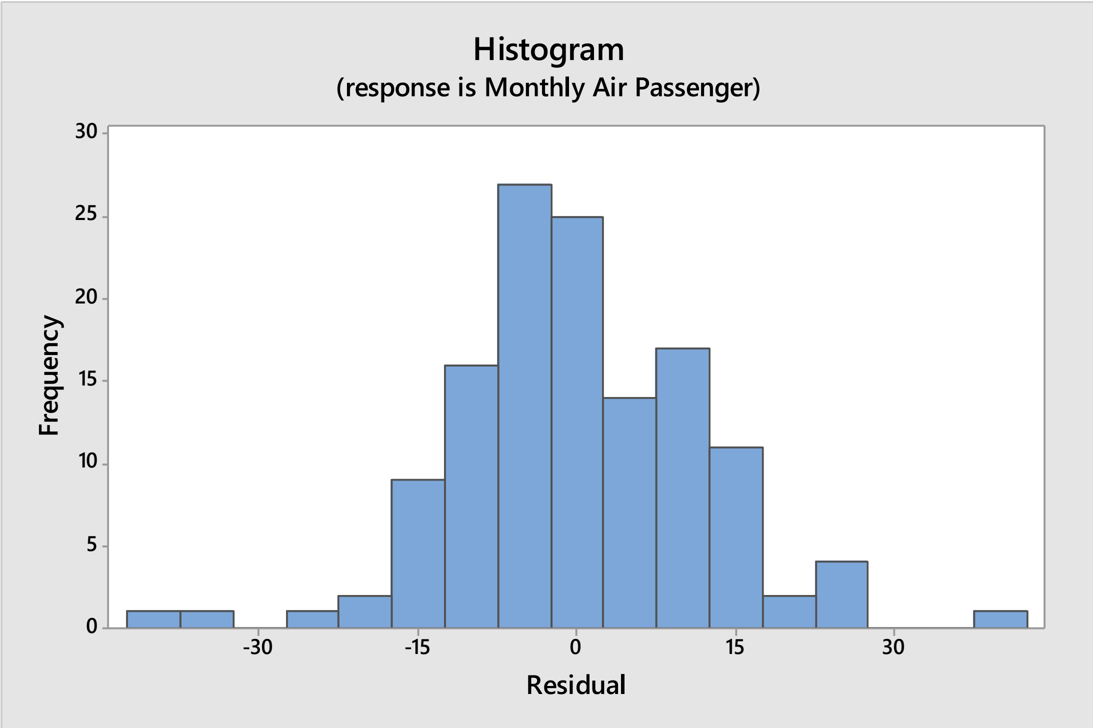

Figure 21 : Histogram of residuals of Model 02

A bell shape was shown in the histogram of residuals.

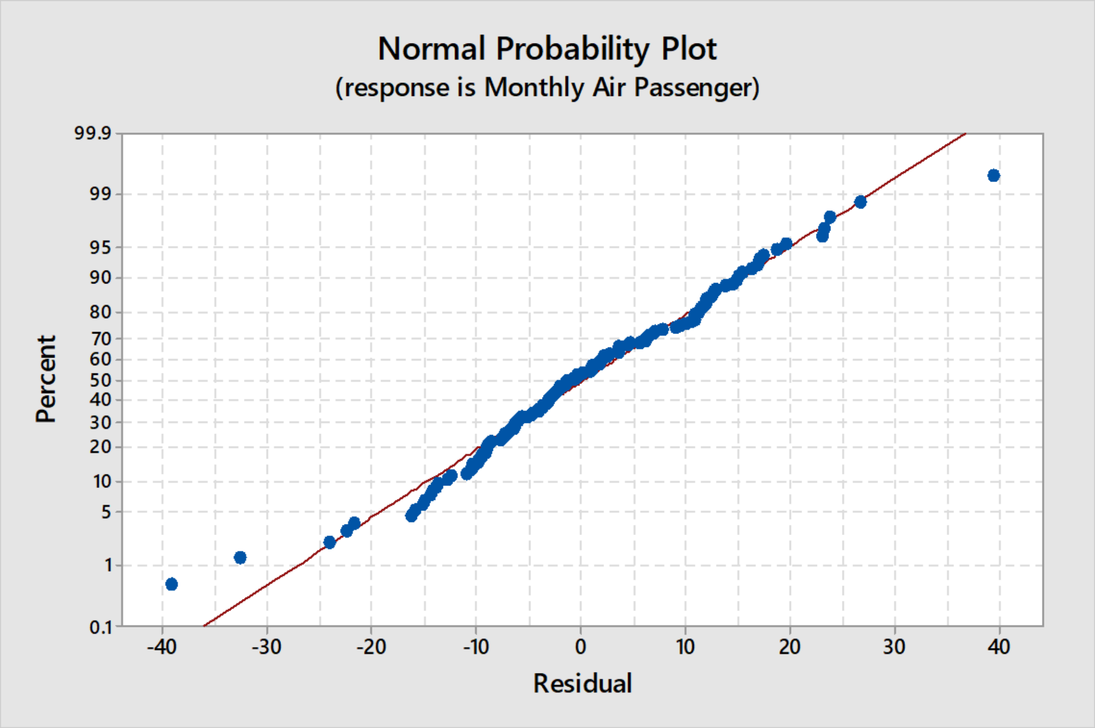

Figure 22 : Normal probability Plot of residuals of Model 02

Normal probabilities of residuals were approximately scattered in a straight line and some outliers were visible in the plot.

Therefore, we could conclude that the residuals are normally distributed by inspecting histogram alone.

Hence, the identified model SARIMA (0,1,1) $(0,1,0)_{12}$ was adequate.

$$(1-B^{12})(1-B)X_t = (1+\beta_1 B)Z_t \quad \textit{Equation 17}$$

$$X_t = X_{t-1} + X_{t-12} - X_{t-13} + Z_t + 0.3212Z_{t-1} \quad \textit{Equation 18}$$

3.6 Tentative Model 02

Considering the seasonal component first, it was required to perform a seasonal differencing in order to make the series stationary.

3.6.1 Auto Correlation Function of Differenced Series

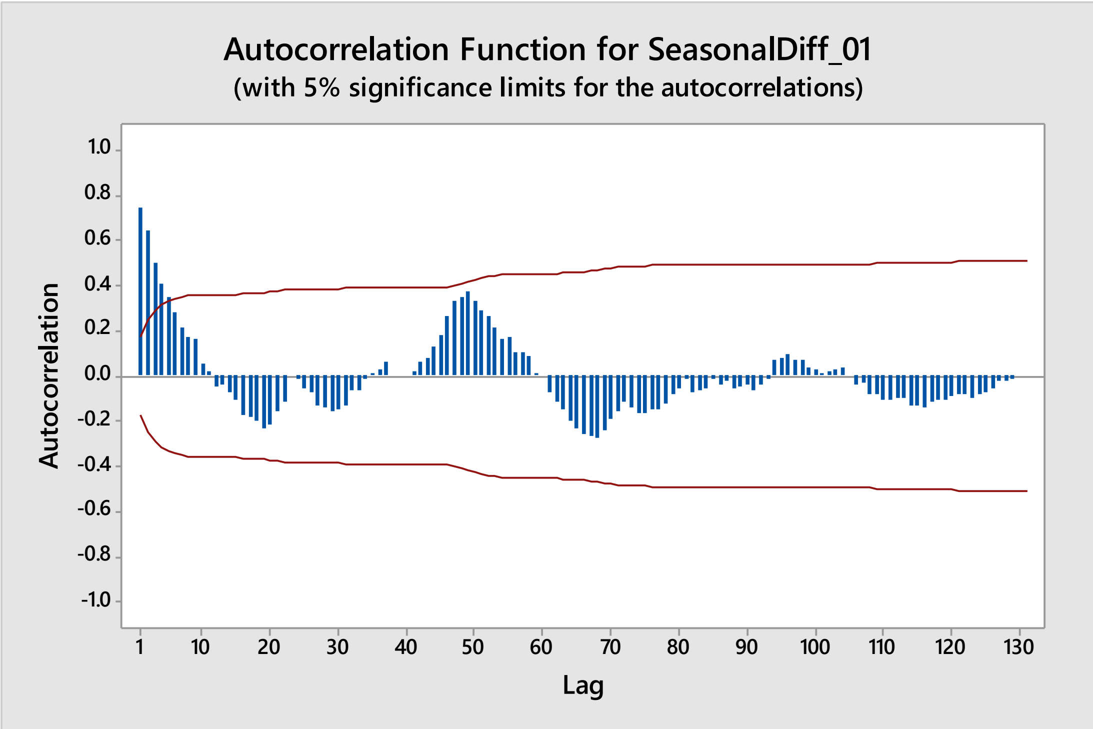

Figure 23 : Autocorrelation of Differenced Series

Table 10 : Autocorrelation of Differenced Series

	lag	ACF	T Statistic	LBQ
Non Seasonal Area	1	0.74646	8.57618	75.235
	2	0.647083	5.11273	132.206
	3	0.504892	3.37629	167.16
	4	0.40646	2.50993	189.989
	5	0.354782	2.09319	207.52
	6	0.283378	1.61903	218.793
	7	0.216276	1.21183	225.412
	8	0.175025	0.96996	229.782
	9	0.164732	0.90648	233.684
	10	0.057201	0.31282	234.159
	11	0.01907	0.10421	234.212
Seasonal Area	12	-0.043736	-0.23899	234.494
	24	-0.013841	-0.07151	275.036
	36	0.033021	0.1667	295.398
	48	0.353707	1.70702	371.638

ACF died down quickly at non-seasonal area and ACF is 0 at seasonal lags. Thus the differenced series is stationary.

3.6.2 Partial Autocorrelation Function of Stationary Series

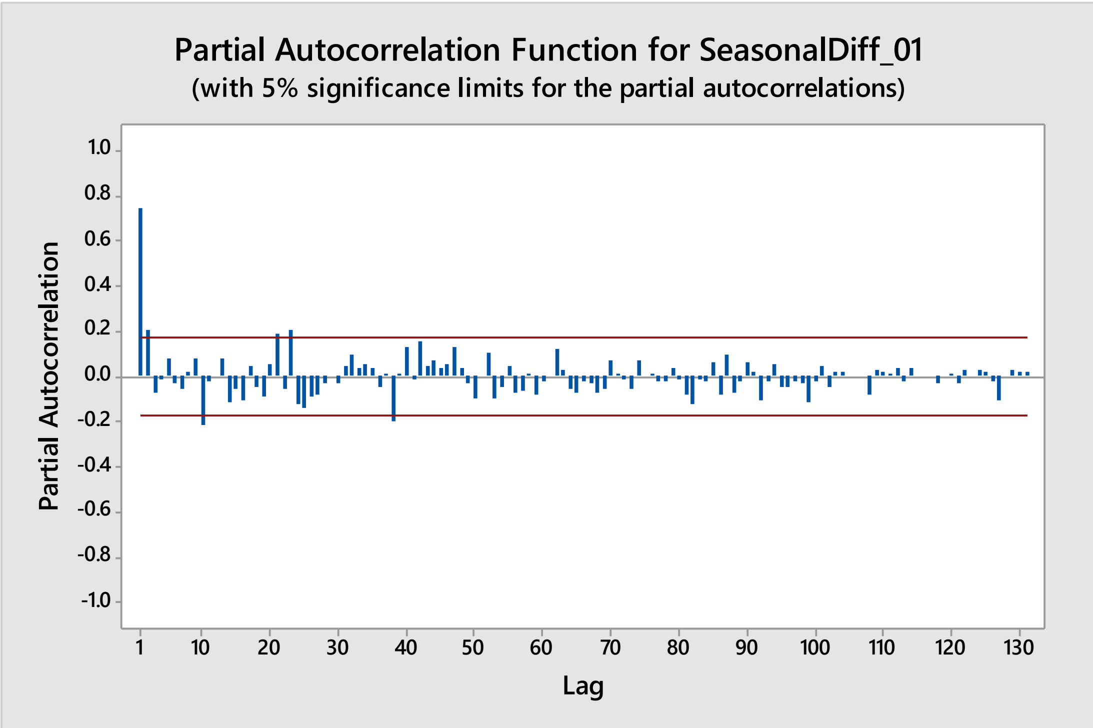

Figure 24 : Partial Autocorrelation of Stationary Series

Table 11 : Partial Autocorrelation of Stationary Series

	lag	PACF	T Statistic
Non Seasonal Area	1	0.74646	8.57618
	2	0.202983	2.33209
	3	-0.074439	-0.85523
	4	-0.014184	-0.16296
	5	0.08005	0.91971
	6	-0.030608	-0.35166
	7	-0.059	-0.67786
	8	0.022455	0.25799
	9	0.077707	0.89279
	10	-0.218328	-2.5084
	11	-0.020258	-0.23275
Seasonal Area	12	-0.007385	-0.08484
	24	-0.123419	-1.41798
	36	-0.046152	-0.53024
	48	0.040117	0.46091

The absolute value of T statistic of non-seasonal lag 1,2 were greater than 2.i.e. Partial Autocorrelations of non-seasonal lag 1,2 was significant from 0.i.e. PACF was cut off at non-seasonal lag 2 and 0 at seasonal lags.

3.6.3 Tentative Model

- Number of non-seasonal differences: 0→ d=0
- Number of seasonal differences: 1 → D=1
- ACF at non-seasonal lag: 0→ q=0
- ACF at seasonal lag: 0 → Q=0
- PACF at non-seasonal lag: 2 → p=2
- PACF at seasonal lag: 0 → P=0

Identified tentative model;

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

3.6.4 Diagnostic Checking

```
Final Estimates of Parameters

Type         Coef  SE Coef     T      P
AR   1     0.6013   0.0859  7.00  0.000
AR   2     0.2164   0.0855  2.53  0.013
Constant  5.5586   0.9895  5.62  0.000


Differencing: 0 regular, 1 seasonal of order 12
Number of observations:  Original series 144, after differencing 132
Residuals:    SS =  16599.2 (backforecasts excluded)
              MS =  128.7  DF = 129


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

Lag            12     24     36     48
Chi-Square    9.7   33.9   41.6   54.6
DF              9     21     33     45
P-Value     0.379  0.037  0.146  0.155


Correlation matrix of the estimated parameters

        1       2
2  -0.760
3  -0.032  -0.013
```

All p values were less than 0.05. therefore, parameters were significant.

All p-values of Modified Box-Pierce (Ljung-Box) Chi-Square statistic at seasonal lags were not greater than 0.05. Therefore, residuals were not random.

It was required to check ACF and PACF of residuals to verify the randomness of residuals.

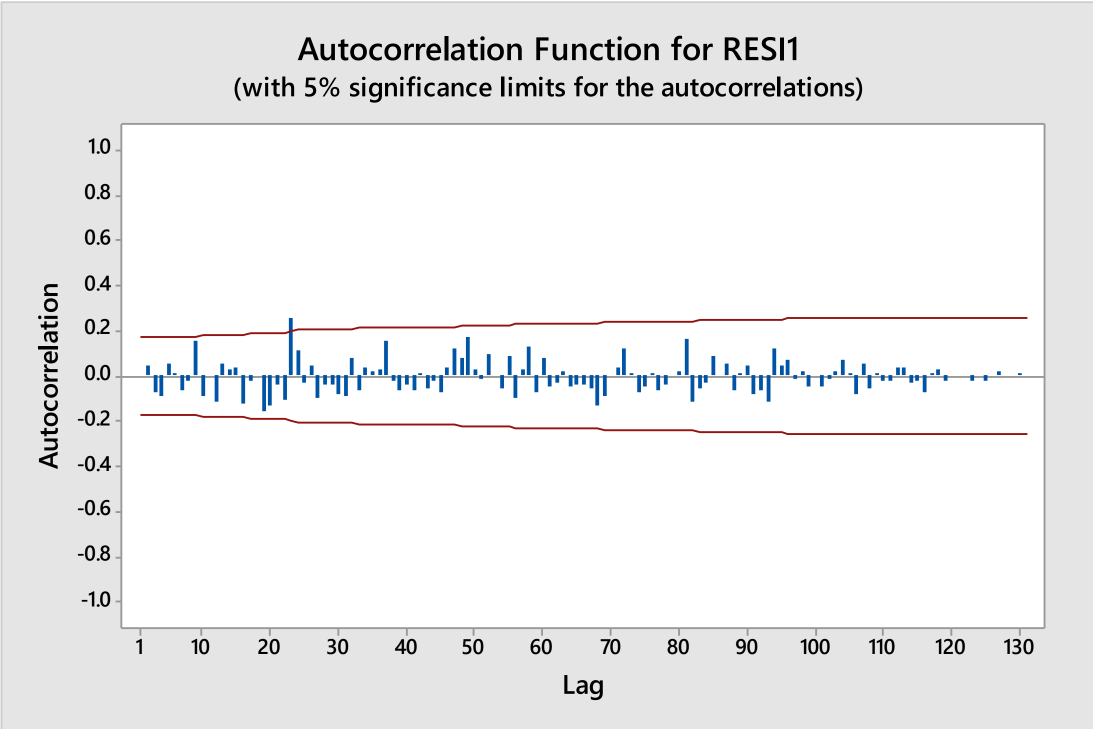

Figure 25 : Autocorrelation function for residuals of model 03

Table 12 : Autocorrelation function for residuals of model 03

Area	lag	ACF	T Statistic	LBQ
Non Seasonal Area	1	0.00116742	0.013412637	0.000184019
	2	0.04269258	0.490499729	0.248177448
	3	-0.069508515	-0.797139363	0.910645651
	4	-0.090339704	-1.031084529	2.03843006
	5	0.056026967	0.634370179	2.475619064
	6	0.01413253	0.159531171	2.503657127
	7	-0.06341878	-0.715747449	3.072777956
	8	-0.023406872	-0.263151774	3.150930621
	9	0.153847687	1.728726084	6.554664228
	10	-0.087680202	-0.963652364	7.669271787
	11	-0.005401024	-0.058946902	7.673536066
Seasonal Area	12	-0.115870896	-1.264584332	9.652537993
	24	0.110555169	1.067654793	33.93916644
	36	0.025771553	0.23966532	41.55713935
	48	0.077178276	0.690642468	54.59757666

ACF was 0 at both seasonal and non-seasonal lags.

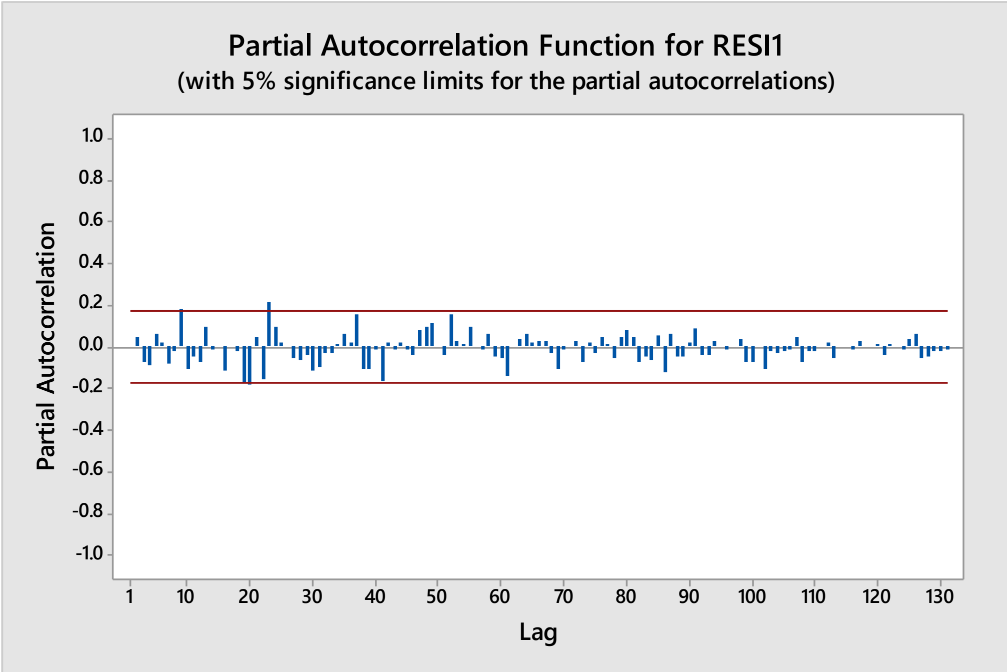

Figure 26 : Partial autocorrelation function for residuals in Model 03

Table 13 : Partial autocorrelation function for residuals in Model 03

	lag	PACF	T Statistic
Non Seasonal Area	1	0.00116742	0.013412637
	2	0.042691275	0.490485408
	3	-0.069733253	-0.801174079
	4	-0.092416753	-1.061787649
	5	0.063087571	0.724821007
	6	0.017731886	0.20372386
	7	-0.083632066	-0.960859279
	8	-0.025161997	-0.289089334
	9	0.180072265	2.068872812
	10	-0.103847761	-1.193119934
	11	-0.047023492	-0.540258792
Seasonal Area	12	-0.075999824	-0.8731715
	24	0.094213358	1.082429074
	36	0.023879841	0.274358481
	48	0.095909093	1.101911583
	60	-0.051959789	-0.596972523

PACF was o at both seasonal and non-seasonal lags.

Therefore, residuals were random.

Since correlation between first parameter and second parameter was high parameter redundancy existed in this model. Therefore, one of them need to be removed in order to omit parameter redundancy.

After removing one AR parameter;

```
Final Estimates of Parameters

Type         Coef  SE Coef      T      P
AR   1     0.7625   0.0569  13.39  0.000
Constant   7.362    1.008   7.31  0.000


Differencing: 0 regular, 1 seasonal of order 12
Number of observations:  Original series 144, after differencing 132
Residuals:    SS =  17400.4 (backforecasts excluded)
              MS =  133.8  DF = 130


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

Lag            12     24     36     48
Chi-Square   18.9   39.6   49.2   59.5
DF             10     22     34     46
P-Value     0.041  0.012  0.044  0.087


Correlation matrix of the estimated parameters

        1
2  -0.036
```

All p values were less than 0.05. therefore, parameters of modified model SARIMA (1,0,0) $(0,1,0)_{12}$ and constant were significant.

All p-values of Modified Box-Pierce (Ljung-Box) Chi-Square statistic at seasonal lags were not greater than 0.05. Therefore, residuals were not random.

It was required to check ACF and PACF of residuals to verify the randomness of residuals.

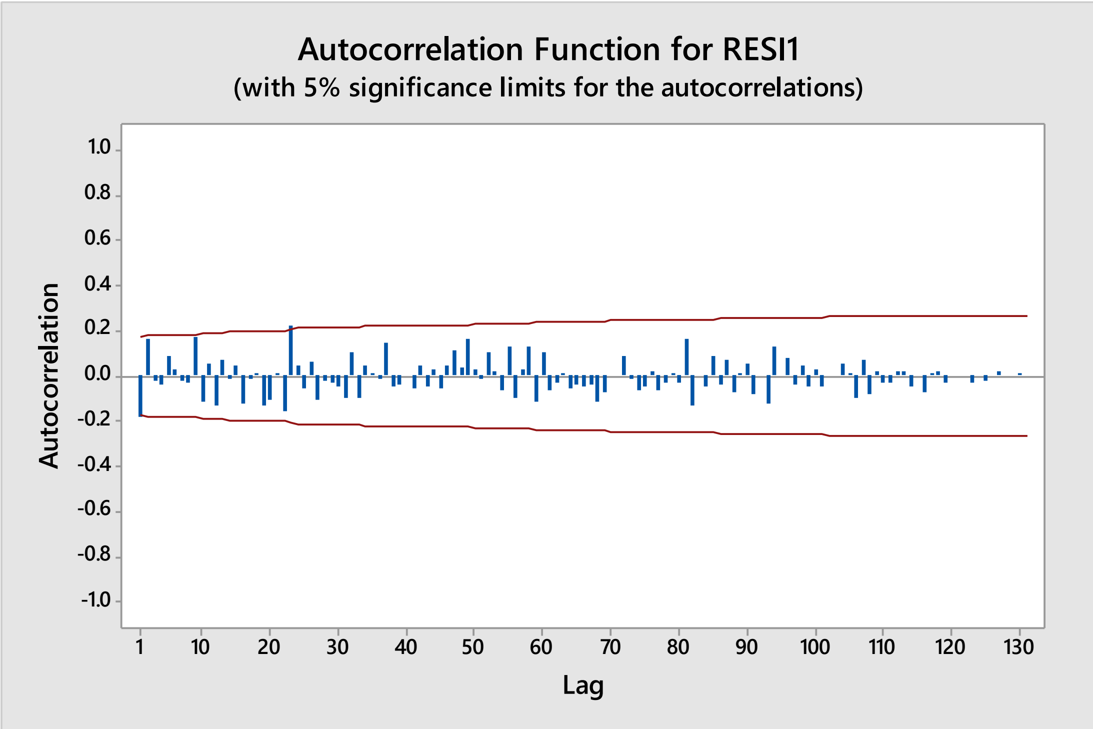

Figure 27 : Autocorrelation Function of Residuals of Modified Model of Model 03

Table 14 : Autocorrelation Function of Residuals of Modified Model of Model 03

	lag	ACF	T Statistic	LBQ
Non Seasonal Area	1	-0.178650945	-2.05254309	4.309412523
	2	0.161911463	1.803549109	7.876309544
	3	-0.022399751	-0.243582955	7.945107478
	4	-0.038358819	-0.416940244	8.14843643
	5	0.088735995	0.96324599	9.245103128
	6	0.029343549	0.316314118	9.365977284
	7	-0.023380133	-0.251839575	9.443327703
	8	-0.033655845	-0.362350551	9.604904144
	9	0.175972743	1.89270216	14.05802441
	10	-0.118187895	-1.23803346	16.0832103
	11	0.053468713	0.553699262	16.50112998
Seasonal Area	12	-0.128449835	-1.327093143	18.93313566
	24	0.044471591	0.41369559	39.56844404
	36	-0.01190338	-0.106786505	49.20831395
	48	0.037681048	0.327933586	59.52244883

Since ACF cut off at non seasonal lag 1 and 0 at seasonal lags, one MA parameter need to be added to the model.

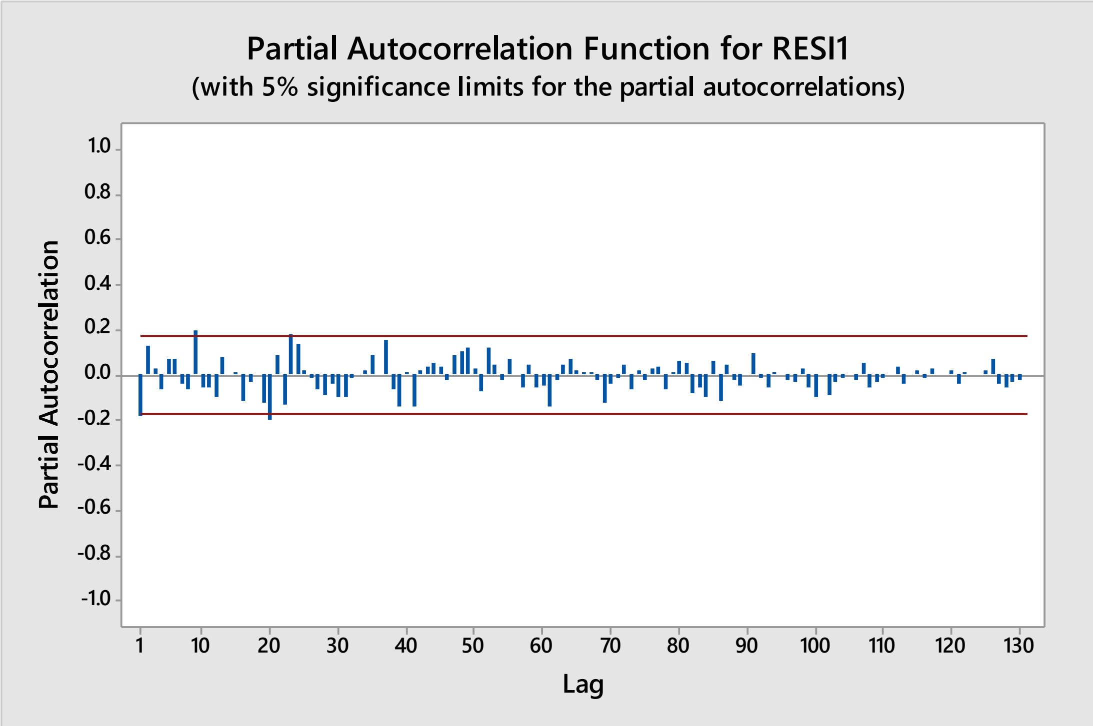

Figure 28 : Partial Autocorrelation Function of Residuals of Modified Model of Model 03

Table 15 : Partial Autocorrelation Function of Residuals of Modified Model of Model 03

	lag	PACF	T Statistic
Non Seasonal Area	1	-0.178650945	-2.05254309
	2	0.134281038	1.54277167
	3	0.028014278	0.321859546
	4	-0.062477076	-0.717806954
	5	0.076022213	0.873428736
	6	0.074778732	0.859142226
	7	-0.03630101	-0.417066852
	8	-0.064299221	-0.738741811
	9	0.194749912	2.237506136
	10	-0.058897961	-0.676686056
	11	-0.052928661	-0.608104022
Seasonal Area	12	-0.10020713	-1.151292276
	24	0.140495946	1.614175529
	36	0.007328341	0.084196227
	48	0.104545153	1.201132364
	60	-0.047935037	-0.550731647

Since PACF cut off at non seasonal lag 1 and 0 at seasonal lags, one AR parameter need to be added to the model.

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

```
Final Estimates of Parameters

Type          Coef   SE Coef       T      P
AR   1      0.3713    0.3499    1.06  0.291
AR   2      0.3935    0.2657    1.48  0.141
MA   1     -0.2392    0.3741   -0.64  0.524
Constant     7.185     1.228    5.85  0.000
```

Since p values of AR and MA parameters were not less than 0.05, parameters were not significant.

Removing AR parameter;

```
Final Estimates of Parameters

Type         Coef  SE Coef      T      P
AR   1     0.8592   0.0577  14.88  0.000
MA   1     0.2267   0.1090   2.08  0.040
Constant  4.2979   0.7686   5.59  0.000


Differencing: 0 regular, 1 seasonal of order 12
Number of observations:  Original series 144, after differencing 132
Residuals:    SS =  16755.1 (backforecasts excluded)
              MS =  129.9  DF = 129


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

Lag            12     24     36     48
Chi-Square   11.1   34.6   42.8   55.5
DF              9     21     33     45
P-Value     0.272  0.031  0.119  0.136


Correlation matrix of the estimated parameters

        1      2
2   0.613
3  -0.046  0.006
```

All p values were less than 0.05. therefore, parameters of modified model SARIMA (1,0,1) $(0,1,0)_{12}$ and constant were significant.

All p-values of Modified Box-Pierce (Ljung-Box) Chi-Square statistic at seasonal lags were not greater than 0.05. Therefore, residuals were not random.

It was required to check ACF and PACF of residuals to verify the randomness of residuals.

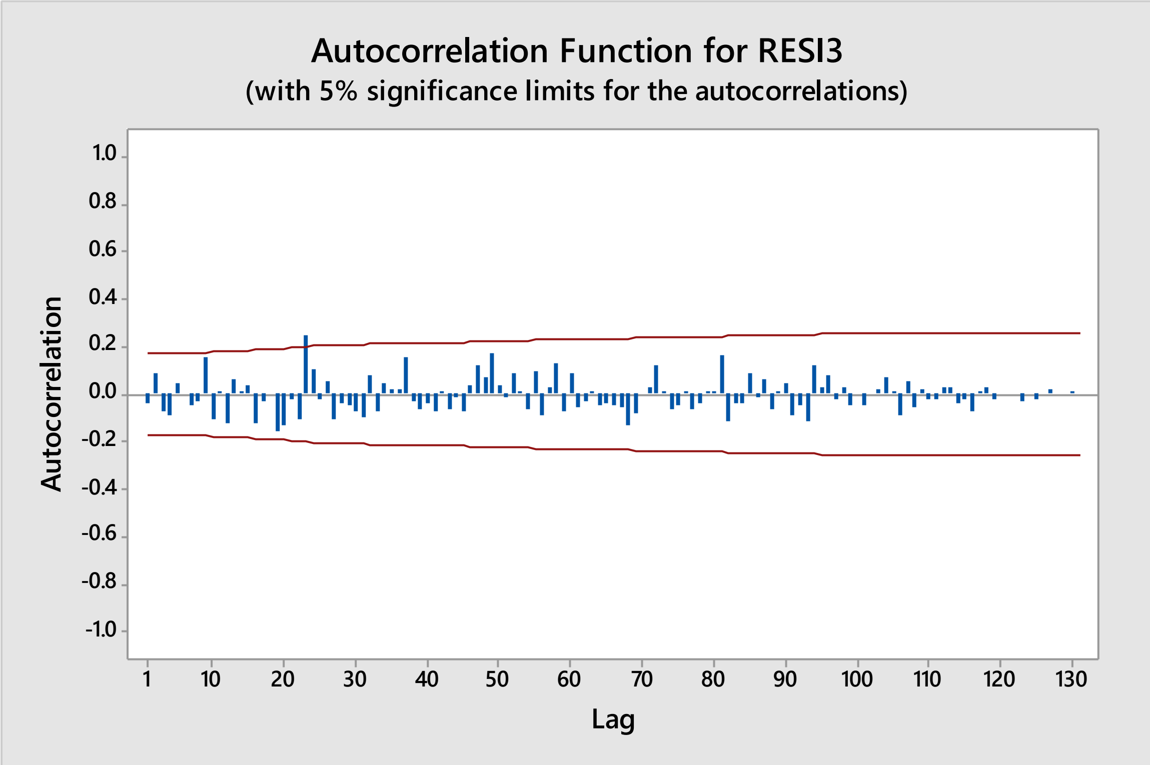

Figure 29 : Autocorrelation Function of Residuals of Modified Model of Model 03

ACF was 0 at both non seasonal and seasonal lags.

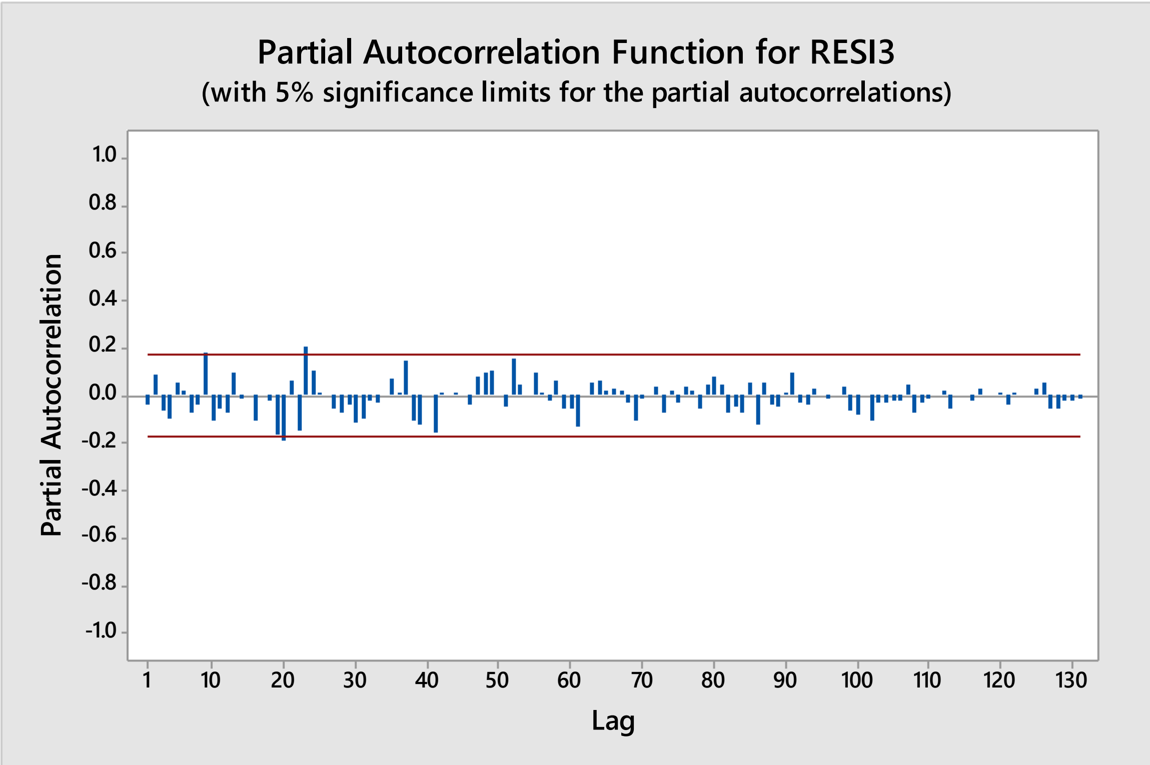

Figure 30 : Partial Autocorrelation Function of Residuals of Modified Model of Model 03

PACF was 0 at both non seasonal and seasonal lags.

Therefore, residuals were random.

Since no high correlation was there parameter redundancy did not exist in this model.

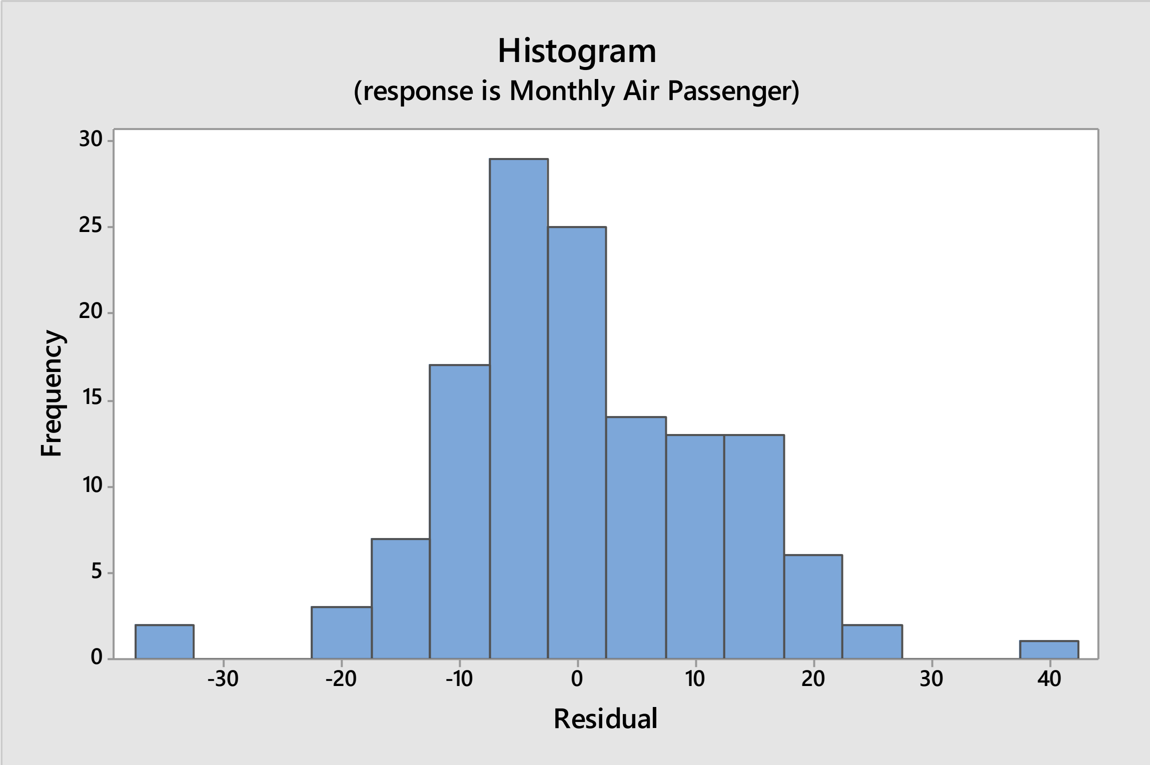

Figure 31 : Histogram of Residuals of Modified Model of Model 03

A bell shape was shown in the histogram of residuals

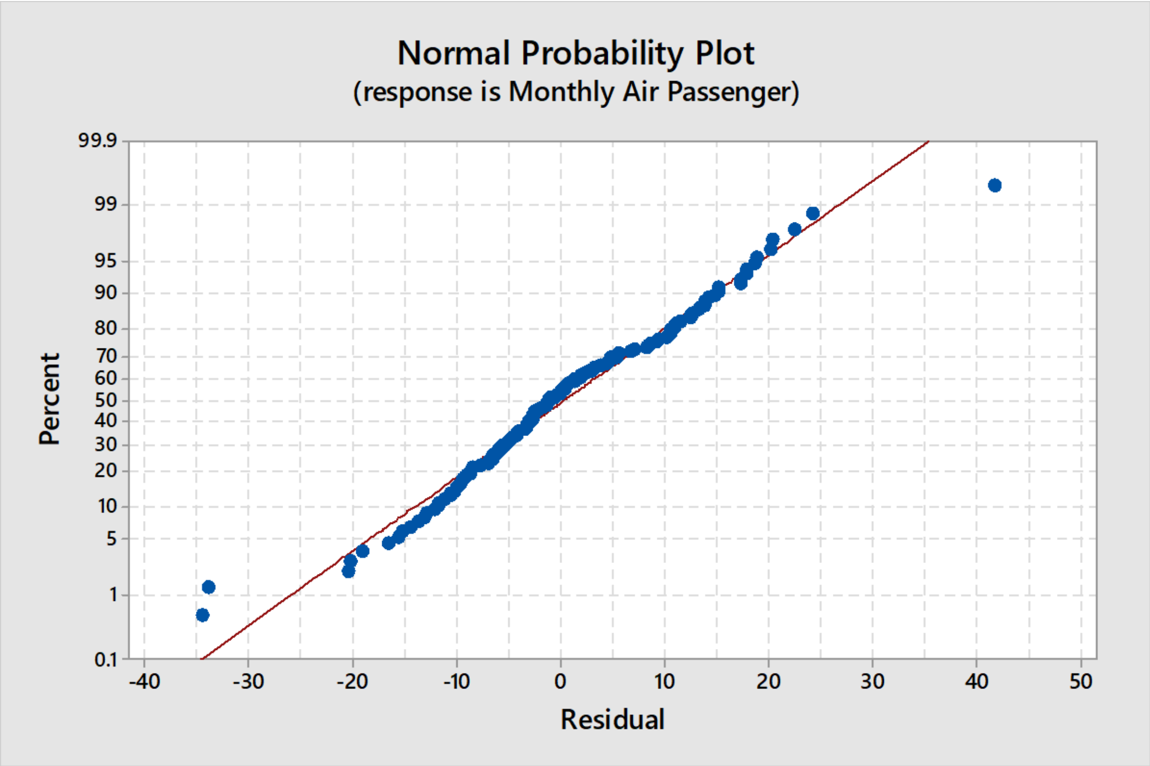

Figure 32 : Normal Probability Plot of Residuals of Modified Model for Model 03

Normal probabilities of residuals were approximately scattered in a straight line and some outliers were visible in the plot.

Therefore, we could conclude that the residuals are normally distributed by inspecting histogram alone.

Hence, the modified model SARIMA (1,0,1) $(0,1,0)_{12}$ for the tentative model SARIMA (2,0,0) $(0,1,0)_{12}$ was adequate.

$$(1-\alpha_1 B)(1-B^{12})X_t = (1-\beta_1 B)Z_t \qquad \textit{Equation 19}$$

$$X_t = 0.8592\, X_{t-1} + X_{t-12} - 0.8592X_{t-13} + Z_t - 0.2267Z_{t-1} + 4.2979 \qquad \textit{Equation 20}$$

3.7 Forecasting

3.7.1 Forecasts for Last Observations of Adequate Models

Table 16 : Forecasts for Last Observations of Adequate Models

Period 1960 Jan-Dec	Actual Values	Forecasted Values SARIMA (1,1,0) $(0,1,0)_{12}$	Forecasted Values SARIMA (0,1,1) $(0,1,0)_{12}$	Forecasted Values SARIMA (1,0,1) $(0,1,0)_{12}$
Jan	417	423.0408044	422.7537083	418.1345737
Feb	391	406.5779057	404.7537083	396.2469456
Mar	419	470.1014816	468.7537083	456.90671
Apr	461	460.2491491	458.7537083	444.036792
May	472	484.2033796	482.7537083	465.5709691
Jun	535	536.2175658	534.7537083	515.4523433
Jul	622	612.2131688	610.7537083	589.6320278
Aug	606	623.2145317	621.7537083	599.0680196
Sep	508	527.2141092	525.7537083	501.7242295
Oct	461	471.2142402	469.7537083	444.5696501
Nov	390	426.2141996	424.7537083	398.5776397
Dec	432	469.2142122	467.7537083	440.7253081

3.7.2 Future Forecasts for Adequate Models

Table 17 : Future Forecasts for Adequate Models

Period 1961 Jan-Dec	Forecasted Values SARIMA (1,1,0) $(0,1,0)_{12}$	Forecasted Values SARIMA (0,1,1) $(0,1,0)_{12}$	Forecasted Values SARIMA (1,0,1) $(0,1,0)_{12}$
Jan	444.3099497	446.8026567	445.8506368
Feb	418.2138809	420.8026567	420.0862959
Mar	446.2436574	448.8026567	448.2887734
Apr	488.2344282	490.8026567	490.4627413
May	499.2372888	501.8026567	501.6122138
Jun	562.2364021	564.8026567	564.7406401
Jul	649.2366769	651.8026567	651.8509834
Aug	633.2365918	635.8026567	635.9457901
Sep	535.2366182	537.8026567	538.0272475
Oct	488.23661	490.8026567	491.0972355
Nov	417.2366125	419.8026567	420.1573689
Dec	459.2366117	461.8026567	462.2090353

3.8 Accuracy Measurements

Table 18 : Accuracy Measurements of Adequate Models

Model	MAPE Value	Forecasting Accuracy
SARIMA (1,1,0) $(0,1,0)_{12}$	4.076410436	95.92358956
SARIMA (0,1,1) $(0,1,0)_{12}$	3.881024454	96.11897555
SARIMA (1,0,1) $(0,1,0)_{12}$	2.893567068	97.10643293

4. RESULTS

- Time Series Plot:
  - An upward trend
  - A seasonal variation of lag 12

- Tentative Models:
  - SARIMA (1,1,1) $(0,1,0)_{12}$
  - SARIMA (2,0,0) $(0,1,0)_{12}$

- Adequate Models:
  - SARIMA (1,1,0) $(0,1,0)_{12}$

    $X_t = 0.6901X_{t-1} + 0.3099X_{t-2} + X_{t-12} - 0.6901X_{t-13} - 3.099X_{t-14} + Z_t$ *Equation 21*

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

    $X_t = X_{t-1} + X_{t-12} - X_{t-13} + Z_t + 0.3212Z_{t-1}$ *Equation 22*

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

    $X_t = 0.8592\, X_{t-1} + X_{t-12} - 0.8592X_{t-13} + Z_t - 0.2267Z_{t-1} + 4.2979$ *Equation 23*

- Forecasting Accuracy of Adequate Models:
  - SARIMA (1,1,0) $(0,1,0)_{12}$ = 95.92358956
  - SARIMA (0,1,1) $(0,1,0)_{12}$ = 96.11897555
  - SARIMA (1,0,1) $(0,1,0)_{12}$ = 97.10643293

5. CONCLUSION

- The Best Fitted Model:

Forecasting accuracy of SARIMA (1,1,0) $(0,1,0)_{12}$ < forecasting accuracy of SARIMA (0,1,1) $(0,1,0)_{12}$ < forecasting accuracy of SARIMA (1,0,1) $(0,1,0)_{12}$

Therefore, the best model is SARIMA (1,0,1) $(0,1,0)_{12}$

- Final Model in Usual Notation:

$$X_t = 0.8592\, X_{t-1} + X_{t-12} - 0.8592 X_{t-13} + Z_t - 0.2267 Z_{t-1} + 4.2979$$

Equation 24

- Future Forecasts of Final Model:

Table 19 : Forecasted Values of Final Model

Period 1961 Jan-Dec	Forecasted Values
Jan	445.8506368
Feb	420.0862959
Mar	448.2887734
Apr	490.4627413
May	501.6122138
Jun	564.7406401
Jul	651.8509834
Aug	635.9457901
Sep	538.0272475
Oct	491.0972355
Nov	420.1573689
Dec	462.2090353

6. DISCUSSION

Time series analysis is a powerful statistical tool for analysing data over time, and it has many applications in various fields such as finance, economics, and engineering. Mainly the Box-Jenkins approach of modelling time series model for forecasting was studied throughout this project.

Number of monthly air passengers which was downloaded from Kaggle website was analysed using time series analysis principles to obtain an adequate model to forecasting. The time series plot indicated a seasonal variation with lag 12 and an upward trend in the original dataset. Autocorrelation function of original data showed a slowly dies down pattern with a slight seasonal pattern. Therefore, original series was not stationary. In order to make the series stationary we used two approaches by considering trend component and by considering seasonal component. In two approaches after doing suitable differences the series obtained were stationary. By inspecting autocorrelation function and partial autocorrelation function of stationary series and the number of differences we obtained two tentative models for the data as SARIMA (1,1,1) $(0,1,0)_{12}$ and SARIMA (2,0,0) $(0,1,0)_{12}$. While checking the parameter significance two different models with significant parameters was found for SARIMA (1,1,1) $(0,1,0)_{12}$. They were SARIMA (1,1,0) $(0,1,0)_{12}$, SARIMA (0,1,1) $(0,1,0)_{12}$ respectively and SARIMA (2,0,0) $(0,1,0)_{12}$ was modified into SARIMA (1,0,1) $(0,1,0)_{12}$. In diagnostic checking process normality of residuals of all models could not be concluded using normal probability plot as they were violating the rules of normality due to the presence of outliers. Therefore, we had to consider the histogram of the residuals as well to verify the normality of residuals. After diagnostic checking we were left with three adequate models to forecast. Then by forecasting last 12 data and considering their actual values the forecast errors and accuracy measurements were calculated. The accuracy values of two adequate models SARIMA (1,1,0) $(0,1,0)_{12}$, SARIMA (0,1,1) $(0,1,0)_{12}$and SARIMA (1,0,1) $(0,1,0)_{12}$ were 95.92358956, 96.11897555 and 97.10643293 respectively. That concluded that the model SARIMA (1,0,1) $(0,1,0)_{12}$ was the best model to forecast as the accuracy is higher than the other adequate models. Finally using the best fit model forecasts for the next year (1961) were calculated.

As we mentioned before normality checking step was a bit challenging due to the existence of few outliers. That could have lead us to do a goodness of fit test to verify the normality in advance. However, since histogram verified the normality goodness of fit was not necessary for the analyse. If the data set contained higher number of data points we could have avoided that issue. And also the parameter redundancy step and parameter significance step lead us to modify the model. However, from this analyse we learned and understood the Box- Jenkins approach for identifying an adequate model for time series analysis and forecast.

7. REFERRENCES

[1] Tableau, “Time Series Analysis: Definition, Types, Techniques, and When It’s Used,” *Tableau*, 2022. https://www.tableau.com/learn/articles/time-series-analysis

[2] Tutorials Point, “Time Series Analysis: Definition and Components,” *www.tutorialspoint.com*. https://www.tutorialspoint.com/time-series-analysis-definition-and-components#

[3] Online Stat Psu, “5.1 Decomposition Models | STAT 510,” *PennState: Statistics Online Courses*. https://online.stat.psu.edu/stat510/lesson/5/5.1

[4] *8.1 Stationarity and differencing | Forecasting: Principles and Practice*. Available: https://otexts.com/fpp2/stationarity.html

[5] J. Frost, “Autocorrelation and Partial Autocorrelation in Time Series Data,” *Statistics By Jim*, May 17, 2021. https://statisticsbyjim.com/time-series/autocorrelation-partial-autocorrelation/

[6] Investopedia, “Box-Jenkins Model,” *Investopedia*, 2019. https://www.investopedia.com/terms/b/box-jenkins-model.asp

[7] MathWorks, “Box-Jenkins Model Selection,” *Mathworks.com*, 2019. https://www.mathworks.com/help/econ/box-jenkins-model-selection.html (accessed Jun. 16, 2019).

[8] “Anderson-Darling Normality Test – iSixSigma,” *Isixsigma.com*, 2017. https://www.isixsigma.com/dictionary/anderson-darling-normality-test/

[9] “Time-series Forecasting -Complete Tutorial | Part-1,” Analytics Vidhya, Jul. 16, 2021. https://www.analyticsvidhya.com/blog/2021/07/time-series-forecasting-complete-tutorial-part-1/

8. APPENDIX

Table 20 : Autocorrelation Function of Non-Seasonally Differenced Series

	lag	ACF	T Statistic	LBQ
Non Seasonal Area	1	0.302855	3.62162	13.393
	2	-0.102148	-1.12285	14.928
	3	-0.241273	-2.62911	23.549
	4	-0.300402	-3.12581	37.011
	5	-0.094073	-0.91814	38.341
	6	-0.078443	-0.76112	39.272
	7	-0.092362	-0.89257	40.573
	8	-0.294802	-2.83317	53.921
	9	-0.191778	-1.74758	59.612
	10	-0.104917	-0.93627	61.328
	11	0.282931	2.50952	73.903
Seasonal Area	12	0.829178	7.05062	182.728
	24	0.701086	4.11578	334.365
	36	0.579577	2.90341	451.192
	48	0.485694	2.23124	544.154
	60	0.40961	1.78559	618.236
	72	0.322778	1.36104	673.179
	84	0.240946	0.99601	712.379
	96	0.181751	0.74272	742.517
	108	0.122571	0.49749	764.454
	120	0.083328	0.33708	780.628
	132	0.035053	0.1416	789.044

Table 21 : Autocorrelation Function of Residuals of Modified Model of Model 03

	lag	ACF	T Statistic	LBQ
Non Seasonal Area	1	-0.03633074	-0.417408424	0.178219788
	2	0.092796534	1.064746549	1.349872461
	3	-0.070683134	-0.80414063	2.034919843
	4	-0.085560189	-0.968658687	3.046527794
	5	0.047557771	0.534632264	3.361533192
	6	0.006733223	0.075529788	3.367897544
	7	-0.05061674	-0.5677677	3.730438479
	8	-0.031405151	-0.351414259	3.871127048
	9	0.155808183	1.74181825	7.362161573
	10	-0.101464283	-1.109089084	8.854767906
	11	0.011420793	0.123691448	8.873835062
Seasonal Area	12	-0.12163679	-1.317219099	11.05469306
	24	0.10185207	0.978486974	34.6175159
	36	0.021545111	0.199214575	42.76635879
	48	0.072040242	0.641642917	55.49896322
	60	0.091696529	0.776442065	76.38816828
	72	0.119233774	0.985934871	90.48875439
	84	-0.036003179	-0.287882212	108.9824505

Table 22 : Autocorrelation Function of Residuals of Modified Model of Model 03

	lag	PACF	T Statistic
Non Seasonal Area	1	-0.03633074	-0.417408424
	2	0.091597513	1.052375303
	3	-0.06492245	-0.745902165
	4	-0.099512933	-1.14331655
	5	0.055615956	0.638978685
	6	0.023339529	0.26815077
	7	-0.074974349	-0.861389694
	8	-0.040265761	-0.462618375
	9	0.185064201	2.126225788
	10	-0.101444807	-1.165512096
	11	-0.053667285	-0.616590158
Seasonal Area	12	-0.075681596	-0.869515335
	24	0.109102309	1.253490098
	36	0.014833722	0.170426486
	48	0.096187323	1.105108201
	60	-0.051822543	-0.595395691
	72	0.035462523	0.407433368
	84	-0.071423102	-0.820588964