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. Sankalapani

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
- ✓ 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.

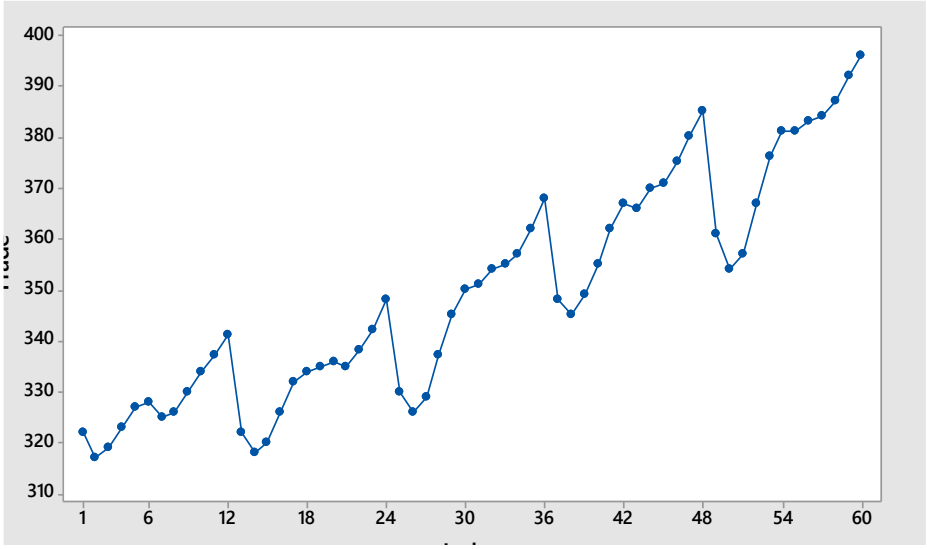

Figure 1 : Upward Trend

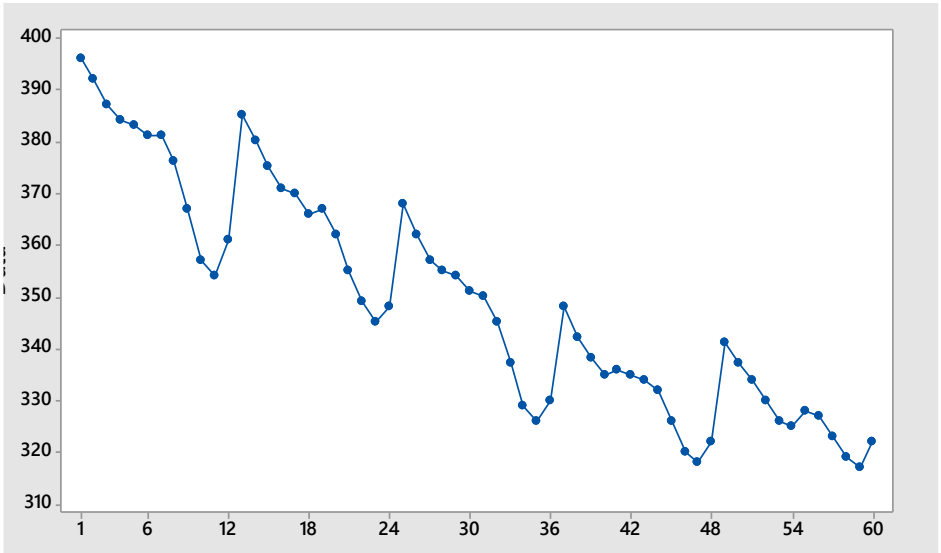

Figure 2 : Downward Trend

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.

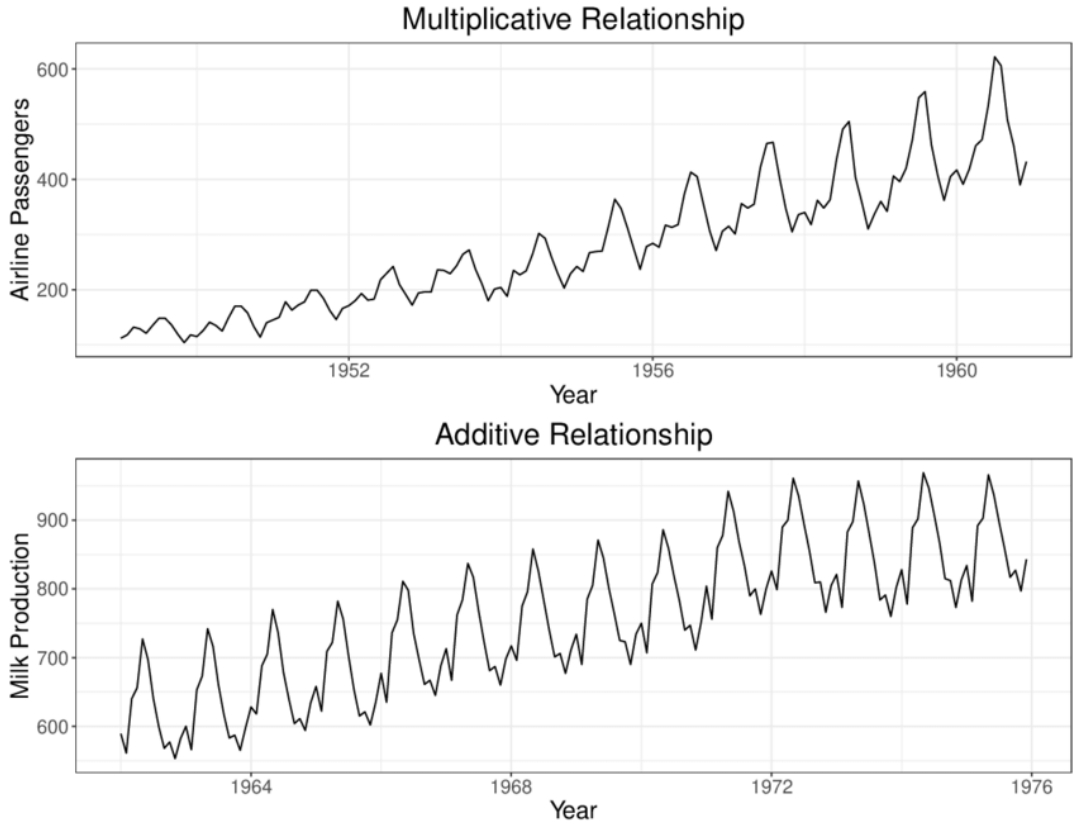

Figure 3 : Seasonal Variation

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.

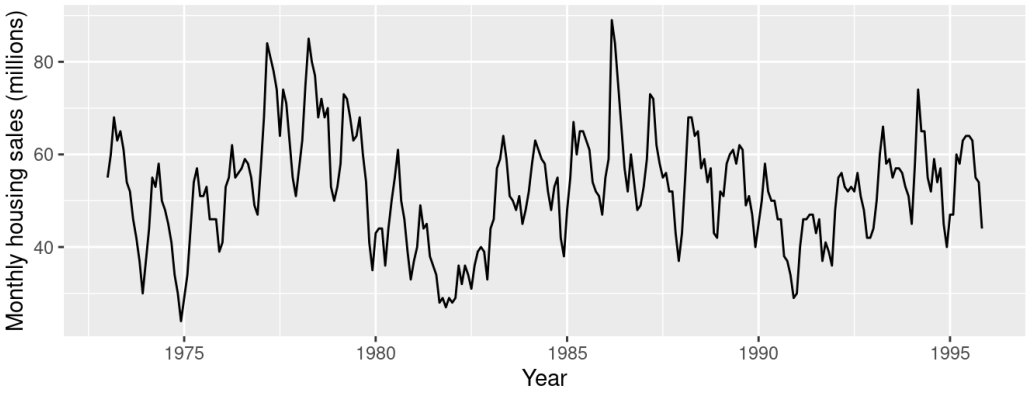

Figure 4 : Cyclic Variation

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]

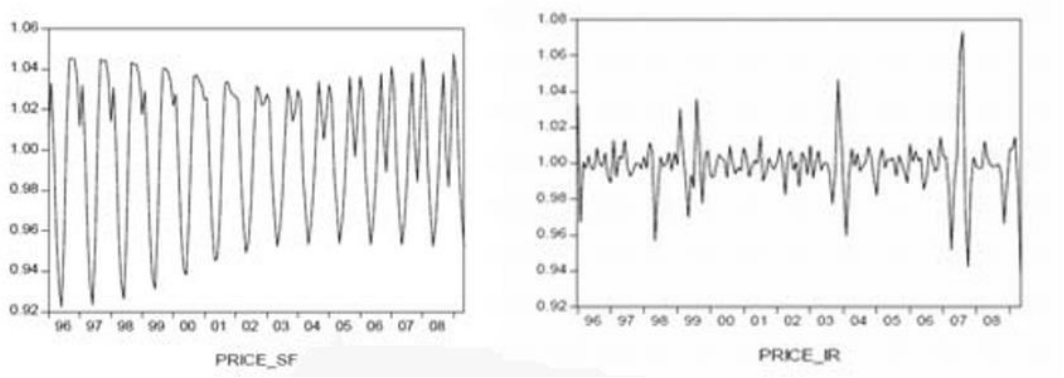

Figure 5 : Irregular Fluctuations

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;

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

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.

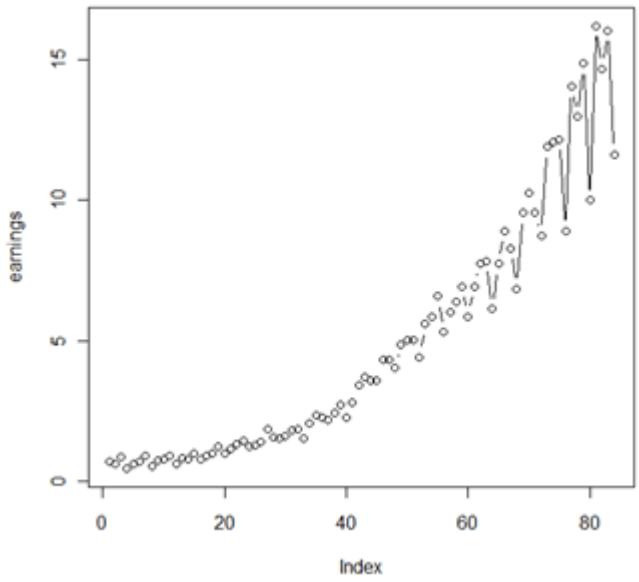

Figure 6 : Multiplicative Seasonality

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]

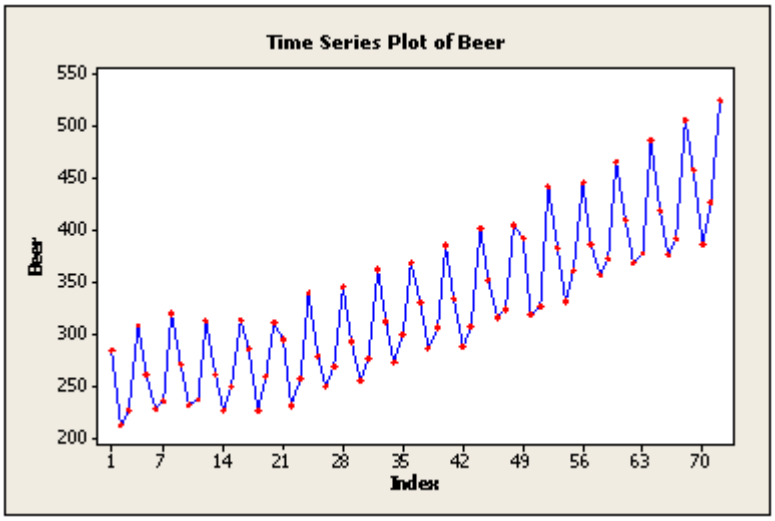

Figure 7 : Additive Seasonality

The additive decomposition model:

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

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

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 autocovarinace

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:

$$t_{\gamma_k} = \frac{\gamma_{\mathrm{k}}}{\frac{1}{\sqrt{n}}\sqrt{1 + 2\sum_{j=1}^{k-1} {\gamma_k}^2}}$$

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:

$$t = \frac{\gamma_{kk}}{\frac{1}{\sqrt{n}}}$$

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: Coefficient = 0

H_1: Not so

If p-value of the parameter is less than level of significance, H_0 can be rejected. i.e. coefficient of the parameter is 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:

$\mathrm{H}_0: \text{Coefficient} = 0$

$H_1: \text{Coefficient} \neq 0$

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 P-value of seasonal lags

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 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 this plot to look for the following:	Non-normality
Not a straight line	
Curve in the tails	Skewness
A point far away from the line	An outlier
Changing slope	An unidentified variable

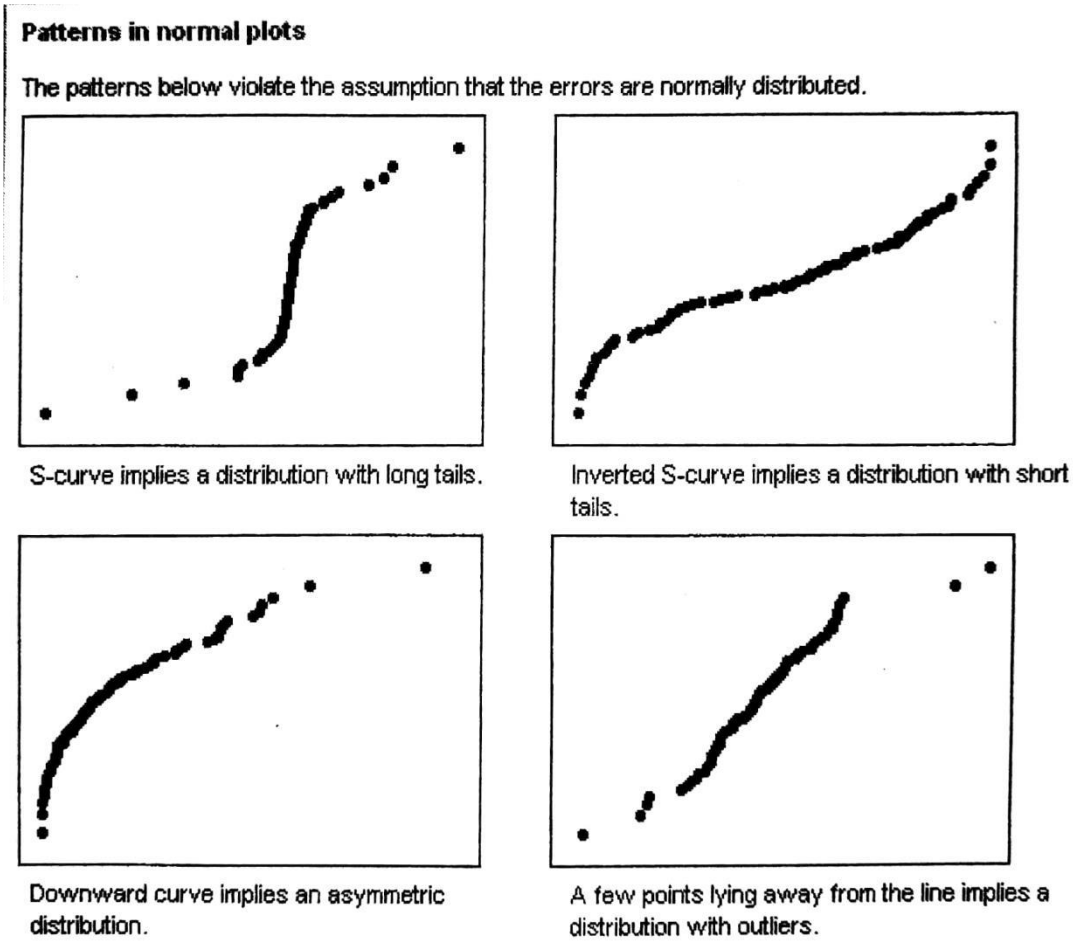

Figure 8 : Non Normal Patterns in Normal Probability Plot

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.

2.11.2 Exponential Smoothing

Exponential smoothing calculates the moving average by considering more past values and give them weightage as per their occurrence, as recent observation gets more weightage compared to past observation so that the prediction is accurate. hence the formula of exponential smoothing can be defined as.

$$Y_T = \alpha * X_T + \alpha(1-\alpha) * y_T-1$$

Alpha is a hyper parameter that defines the weightage to give. This is known as simple exponential smoothing, but we need to capture trend and seasonality components so there is double exponential smoothing which is used to capture the trend components. only a little bit of modification in the above equation is there.

$$Yt = \alpha * X_t + (1-\alpha)\ (y_t\text{-}1 + b_t\text{-}1)\ \#\text{trend component}$$

where, $b_t = beta * (Y_t - Yt\text{-}1) + (1\text{-}beta) * b_{t-1}$

hence here we are taking 2 past observations and what was in the previous cycle, which means we are taking two consecutive sequences, so this equation will give us the trend factor.

If we need to capture trend and seasonality for both components, then it is known as triple exponential smoothing which adds another layer on top of trend exponential smoothing where we need to calculate trend and seasonality for both.

$$Y = alpha * (X_t / C_t\text{-}1) + (1 - alpha) *(Y_{t-1} + b_{t-1})$$

where, $c_t = gamma * (x_t/y_t) + (1\text{-}alpha) * c_t\text{-}alpha$

here we are capturing trends as well as seasonality. Using smoothing we will be able to decompose our time series data and our time-series data will become easy to work with because in real-world scenarios working with time series is a complex task so you have to adopt such methods to make the process smooth. [9]

2.11.3 Accuracy of Forecasts

2.11.3.1 Forecasting Error

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

$$e_t=(X_t-\hat{X}_t) \quad \textit{Equation 8}$$

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.

2.11.3.2 Mean Absolute Error

$$\mathrm{MAE} = \frac{\sum_{t=1}^{n}|X_t-\hat{X}_t|}{n} \quad \textit{Equation 9}$$

Where,

X_t – Actual Value

$\hat{X}_t$ – Forecasted Value

n – Number of Data

2.11.3.3 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.

$$\text{MAPE} = \frac{\sum_{t=1}^{n} |X_t - \hat{X}_t| / X_t}{n} \times 100 \quad \textit{Equation 10}$$

Where,

X_t – Actual Value

$\hat{X}_t$ – Forecasted Value

n – Number of Data

2.11.3.4 Akaike Information Criterion (AIC)

The AIC is defined as

$$AIC = \log(residual\ sum\ of\ square) + (2/n)k \quad \textit{Equation 11}$$

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

2.11.4 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

	1955	1956	1957	1958	1959	1960
January	242	284	315	340	360	417
February	233	275	301	318	342	391
March	267	317	356	362	406	465
April	263	313	348	353	398	461
May	270	318	355	363	420	472
June	315	374	422	435	472	535
July	364	413	465	491	548	622
August	357	405	463	487	540	616
September	312	355	404	404	463	508
October	274	306	347	359	407	461
November	237	271	305	310	362	390
December	278	306	336	337	405	432

3.2 Time Series Plot

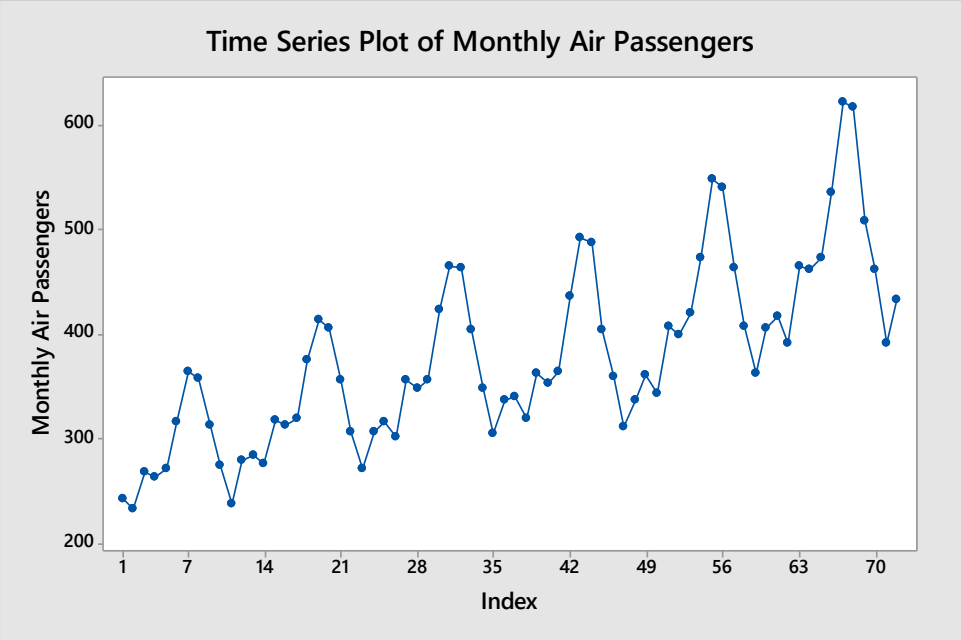

Figure 9 : Time Series Plot of Air Passenger Data

An upward trend and a 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

$$Y_t = 266.32 + 3.0603 \times t \qquad \textit{Equation 12}$$

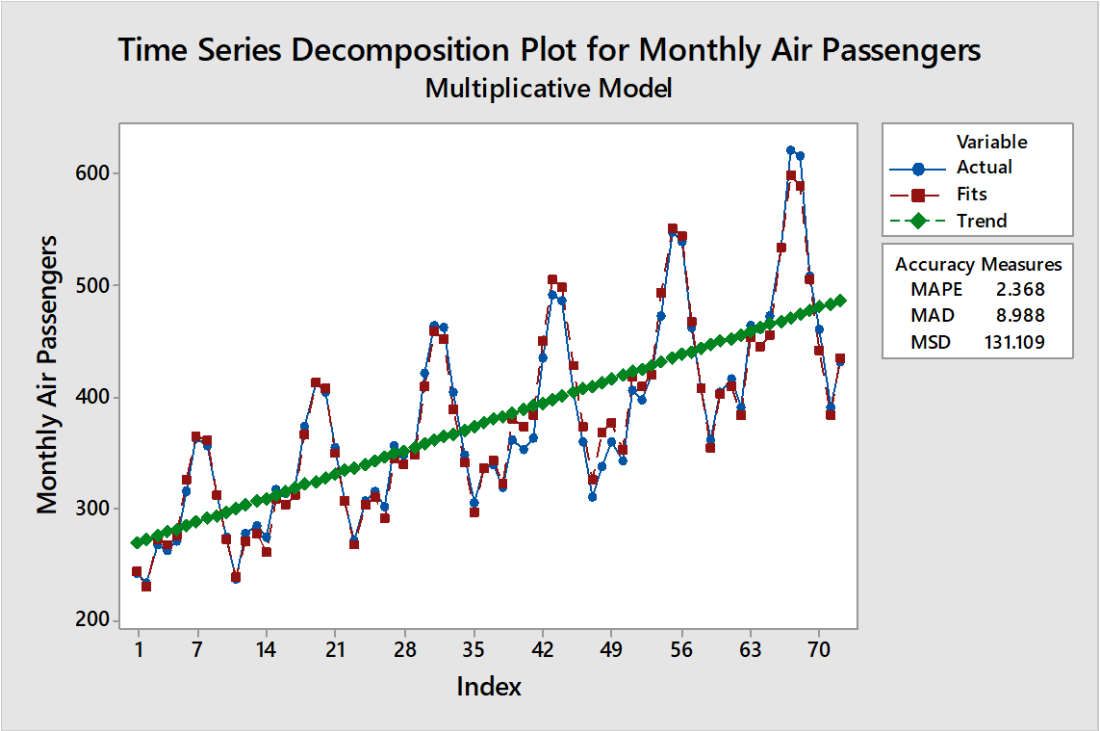

Figure 10 : Decomposition Plot of Air Passenger Data

Seasonal Indices:

Table 2 : Seasonal Indices of Multiplicative Decomposition

Period	Index
January	0.90463
February	0.841651
March	0.98828
April	0.962567
May	0.97914
June	1.142467
July	1.270627
August	1.243384
September	1.059975
October	0.919499
November	0.794709
December	0.893071

3.4 Autocorrelation Function

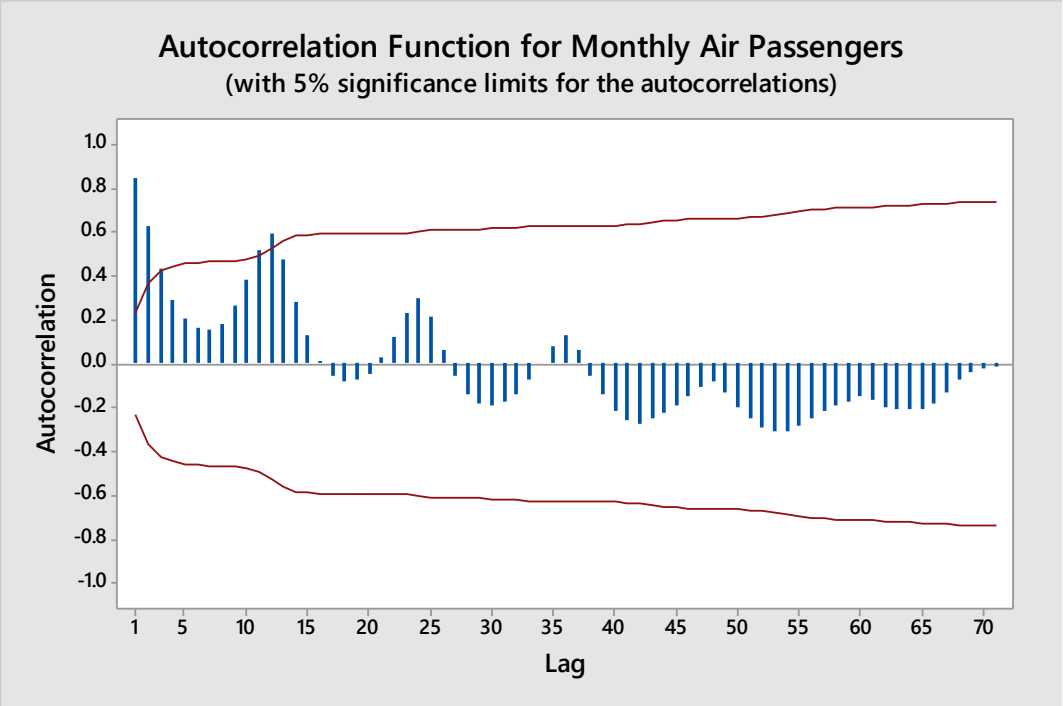

Figure 11 : Autocorrelation of Original Series

Table 3 : Autocorrelation of Original Data

	lag	ACF	T Statistic	LBQ
Non Seasonal Area	1	0.853397372	7.241316822	54.65230323
	2	0.630366375	3.412670422	84.89721024
	3	0.43288835	2.037106709	99.3671566
	4	0.287517666	1.281185316	105.8443128
	5	0.205471392	0.895399691	109.2016282
	6	0.160778437	0.692963646	111.2884044
	7	0.160527733	0.687314292	113.4006821
	8	0.184075712	0.783016409	116.2215141
	9	0.269659512	1.137425687	122.3712314
	10	0.382445561	1.584930451	134.9405512
	11	0.51674591	2.070478958	158.2637695
Seasonal	12	0.592953571	2.245866196	189.4853111
	24	0.29987687	0.994529958	238.2963495
	36	0.128481737	0.408188582	266.0631313

48	-0.083298233	-0.250535007	339.3757254
60	-0.149174653	-0.415643832	544.6177274

The absolute value of T statistic values of first three non-seasonal lags and the first seasonal lag were greater than 2.

i.e. Autocorrelations of first three non-seasonal lags and the first seasonal lag were significant from 0.

i.e. ACF was cut off at non seasonal lag 3 and seasonal lag 1.

Thus the original series was non-stationary.

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

3.5 Autocorrelation Function of Differenced Series

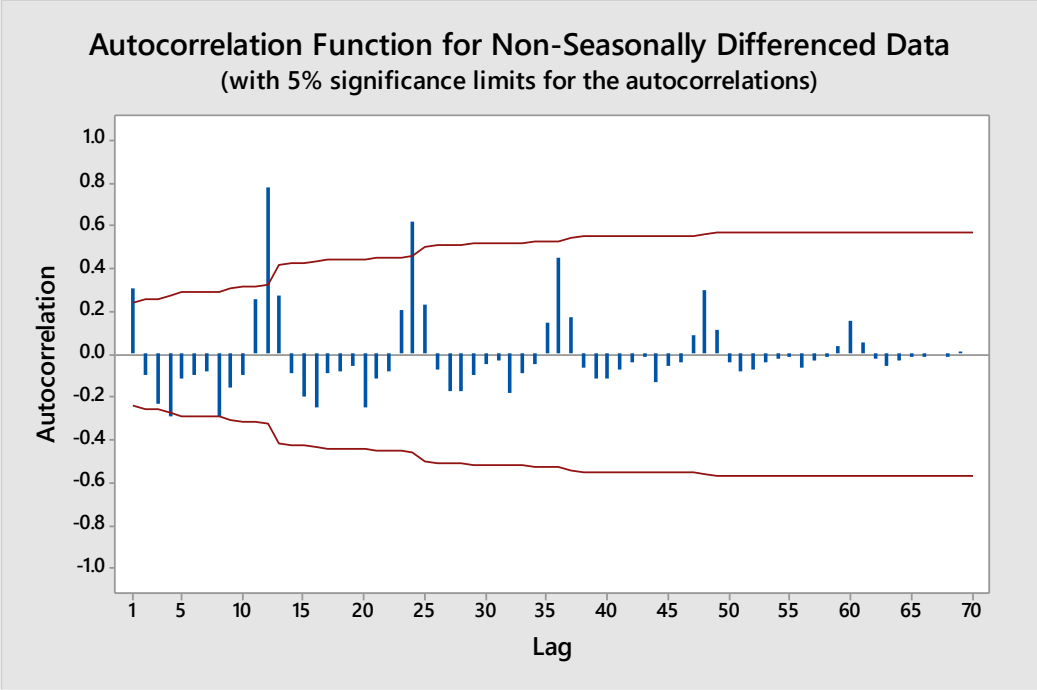

Figure 12 : Autocorrelation of Differenced Series

Table 4 : Autocorrelation of Differenced Series

	Lag	ACF	T Statistic	LBQ
Non Seasonal Area	1	0.308028	2.595493	7.025295
	2	-0.09951	-0.76868	7.769045
	3	-0.23499	-1.80034	11.97781
	4	-0.29503	-2.16376	18.71133
	5	-0.11133	-0.76748	19.68473
	6	-0.09518	-0.65076	20.40712
	7	-0.07884	-0.53587	20.91055
	8	-0.29105	-1.97023	27.87984
	9	-0.15259	-0.9807	29.8263
	10	-0.09681	-0.61392	30.62257
	11	0.258487	1.630612	36.39433
Seasonal Area	12	0.784302	4.772114	90.43185
	24	0.621869	2.690333	164.337
	36	0.453658	1.717111	218.8619
	48	0.296118	1.058167	255.1019
	60	0.156524	0.547501	275.9092

The absolute value of T statistic values of the first non-seasonal lag and first two seasonal lags were greater than 2.

i.e. Autocorrelations of the first non-seasonal lag and first two seasonal lags were significant from 0.

i.e. ACF was cut off at non seasonal lag 1 and seasonal lag 2.

Therefore, the one-time non-seasonally differenced series was stationary.

3.6 Partial Autocorrelation Function of Stationary Series

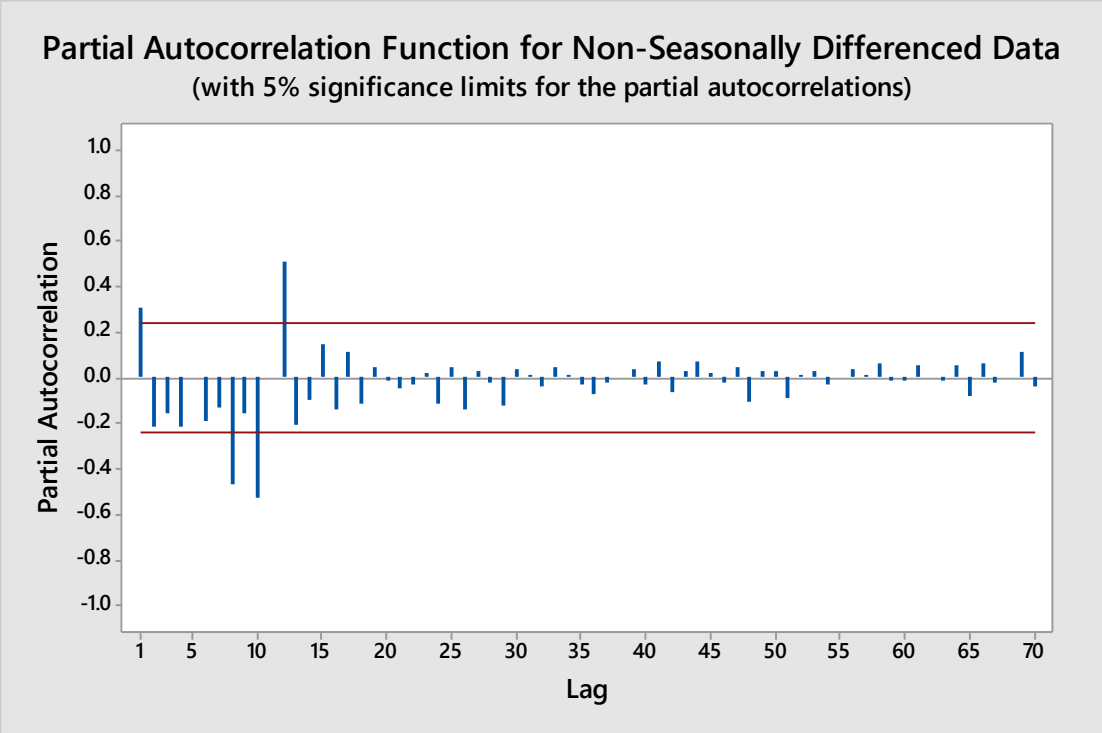

Figure 13 : Partial Autocorrelation of Stationary Series

Table 5 : Partial Autocorrelation of Stationary Series

	Lag	PACF	T Statistic
Non Seasonal Area	1	0.308028	2.595493
	2	-0.21476	-1.80964
	3	-0.15242	-1.28435
	4	-0.21764	-1.83388
	5	-0.00778	-0.06552
	6	-0.19372	-1.63233
	7	-0.12737	-1.07321
	8	-0.46532	-3.92089
	9	-0.15923	-1.34167
	10	-0.5324	-4.48606
	11	-0.00373	-0.03143
Seasona	12	0.510768	4.30381
	24	-0.11147	-0.93927

36	-0.06883	-0.58001
48	-0.10492	-0.88407
60	-0.01541	-0.12984

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

i.e. Autocorrelations of seasonal lag 1 and non-seasonal lag 1 were significant from 0.

i.e. PACF was cut off at non seasonal lag 1 and seasonal lag 1.

3.7 Tentative Model

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

Identified tentative model;

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

3.8 Parameter Estimation

3.8.1 Model 01

Final Estimates of Parameters:

Table 6 : Significant Parameter Estimates of Model One

Type	Coef SE	Coef	T	P
SAR 12	1.0162	0.0294	34.53	0.000

Identified model;

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

3.8.2 Model 02

Final Estimates of Parameters:

Table 7 : Significant Parameter Estimates of Model Two

Type	Coef SE	Coef	T	P
AR 1	-0.2399	-2.04	0.1176	0.045
SAR 12	1.0168	0.0262	38.74	0.000

Identified model;

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

3.8.3 Model 03

Final Estimates of Parameters

Table 8 : Significant Parameter Estimates of Model Three

Type	Coef SE	Coef	T	P
MA 1	-0.3083	0.1173	-2.63	0.011
SMA 12	-0.8057	0.1612	-5.00	0.000

Identified model;

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

3.9 Diagnostic Checking

3.9.1 Model 01

```
Final Estimates of Parameters

Type       Coef SE Coef      T      P
SAR  12  1.0162   0.0294  34.53  0.000
```

Since p-values of estimated parameters were less than 0.05 parameters were significant from zero.

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

Lag            12     24     36     48
Chi-Square   13.6   30.4   38.9   43.0
DF             11     23     35     47
P-Value     0.258  0.139  0.299  0.639
```

All p-values of Modified Box-Pierce (Ljung-Box) Chi-Square statistic at seasonal lags were greater than 0.05. Therefore, residuals were random.

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

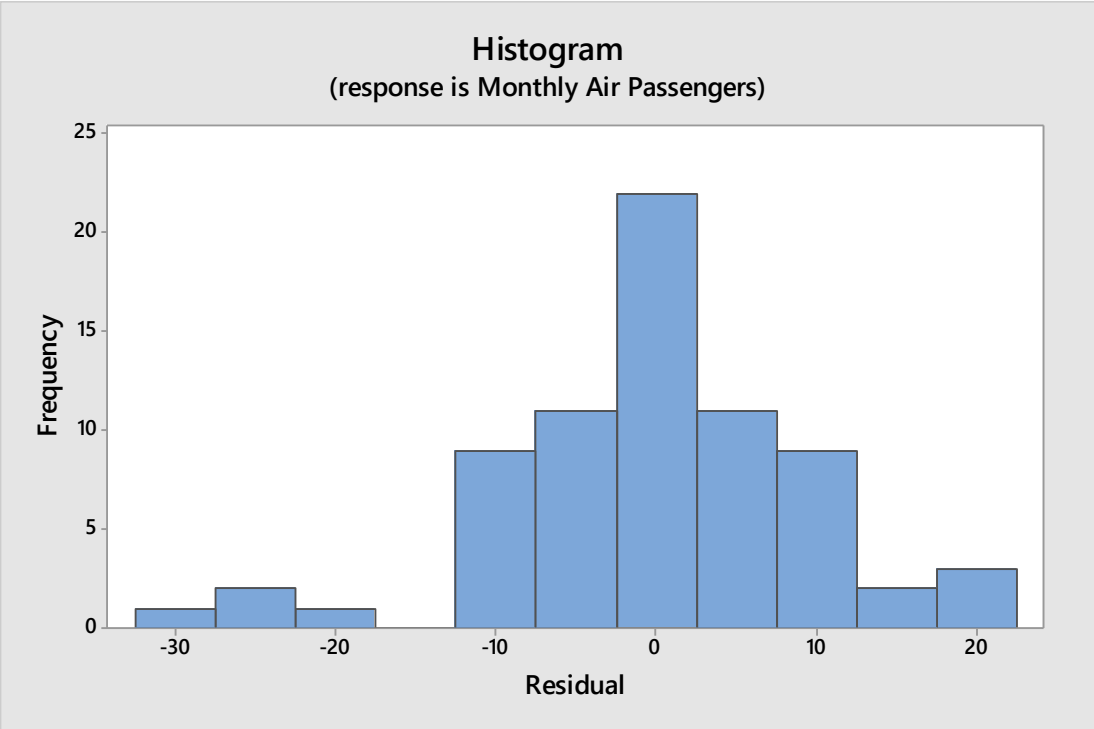

Figure 14 : Histogram of residuals of model 01

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

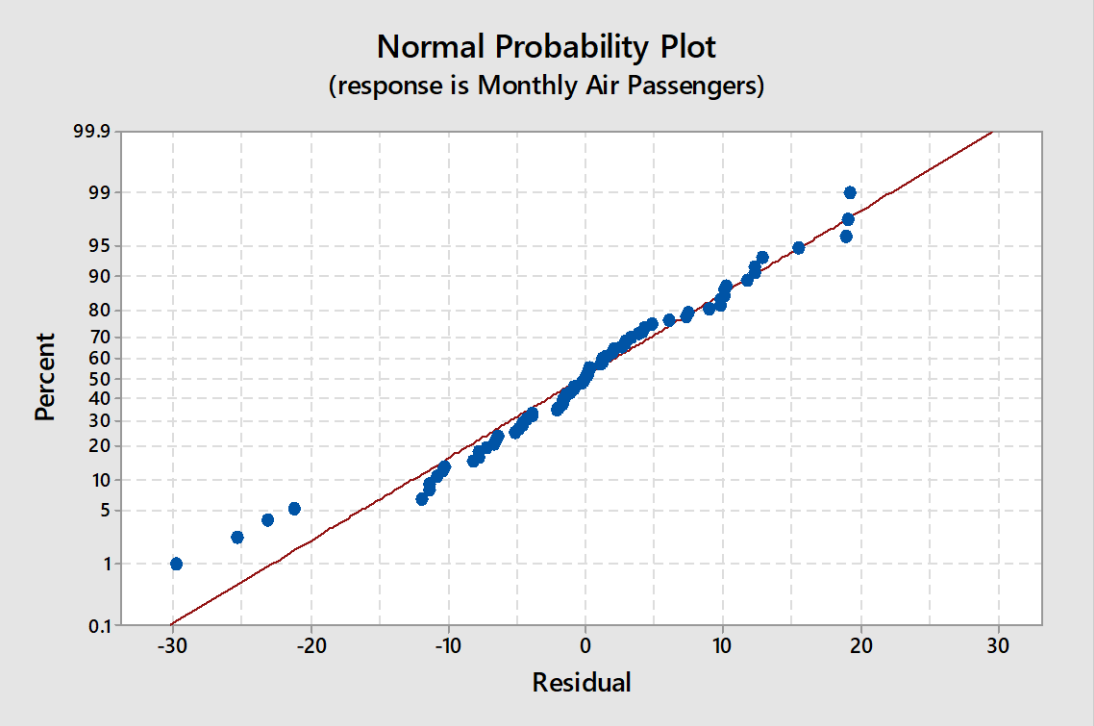

Figure 15 : Normal probability plot of residuals in model 01

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

Therefore, we could not conclude that the residuals are normally distributed. Thus it was required to conduct Anderson darling goodness of fit test.

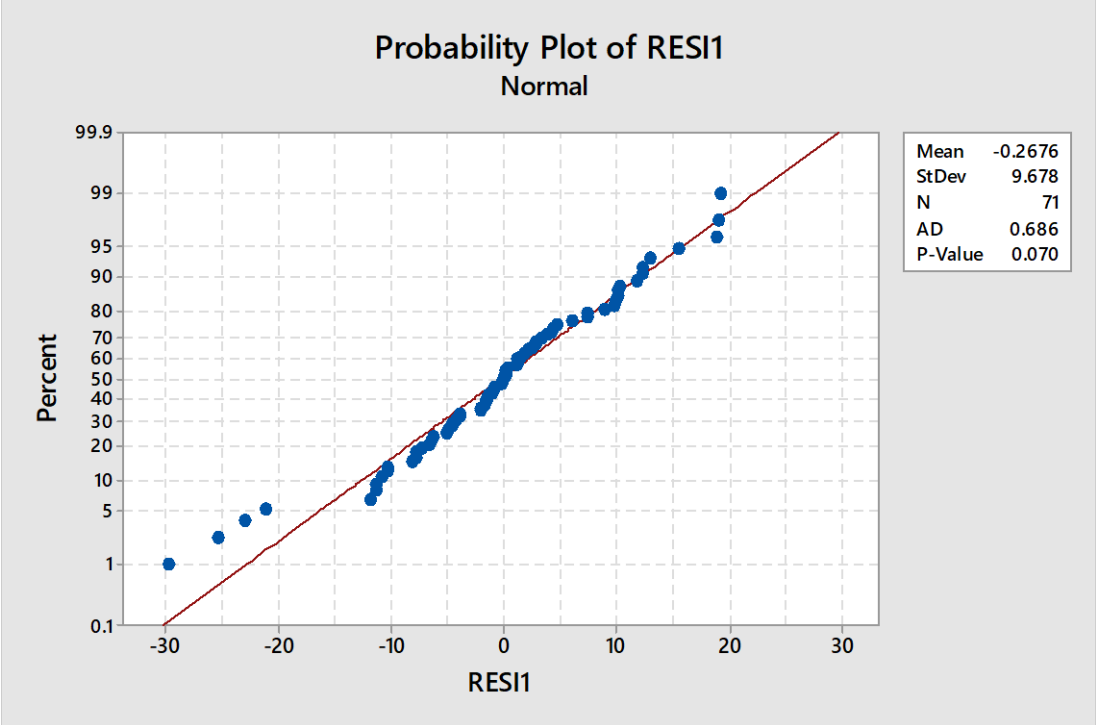

Figure 16 : Output of Anderson Darling Test for Model 01

Since the p value of Anderson darling test is greater than 0.05 the residuals are normally distributed.

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

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

$$X_t = X_{t-1} + 1.0162X_{t-12} - 1.0162X_{t-13} + Z_t \quad \textit{Equation 13}$$

3.9.2 Model 02

```
Final Estimates of Parameters

Type         Coef  SE Coef      T      P
AR   1   -0.2399   0.1176  -2.04  0.045
SAR  12   1.0168   0.0262  38.74  0.000
```

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

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

Lag	12	24	36	48
Chi-Square	7.1	19.1	27.9	31.4
DF	10	22	34	46
P-Value	0.720	0.637	0.758	0.951

All p-values of Modified Box-Pierce (Ljung-Box) Chi-Square statistic at seasonal lags were greater than 0.05. Therefore, residuals were random.

```
Correlation matrix of the estimated parameters

        1
2 -0.131
```

Since less correlation was recorded parameter redundancy did not exist in this model.

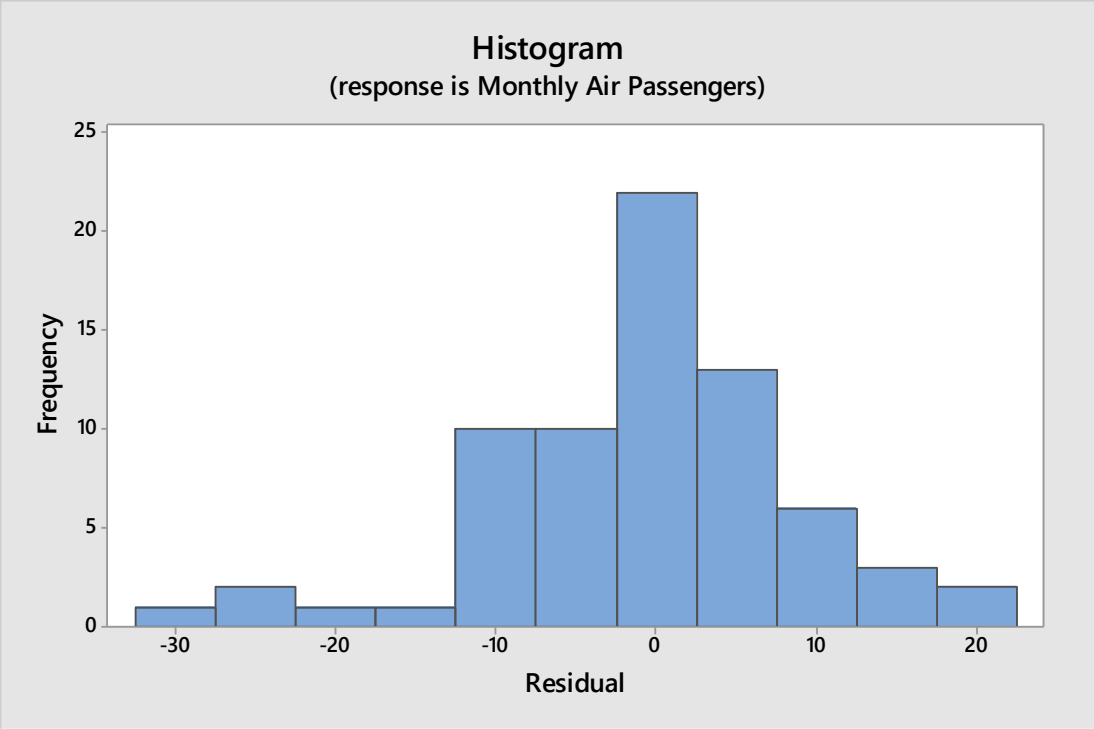

Figure 17 : Histogram of residuals of Model 02

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

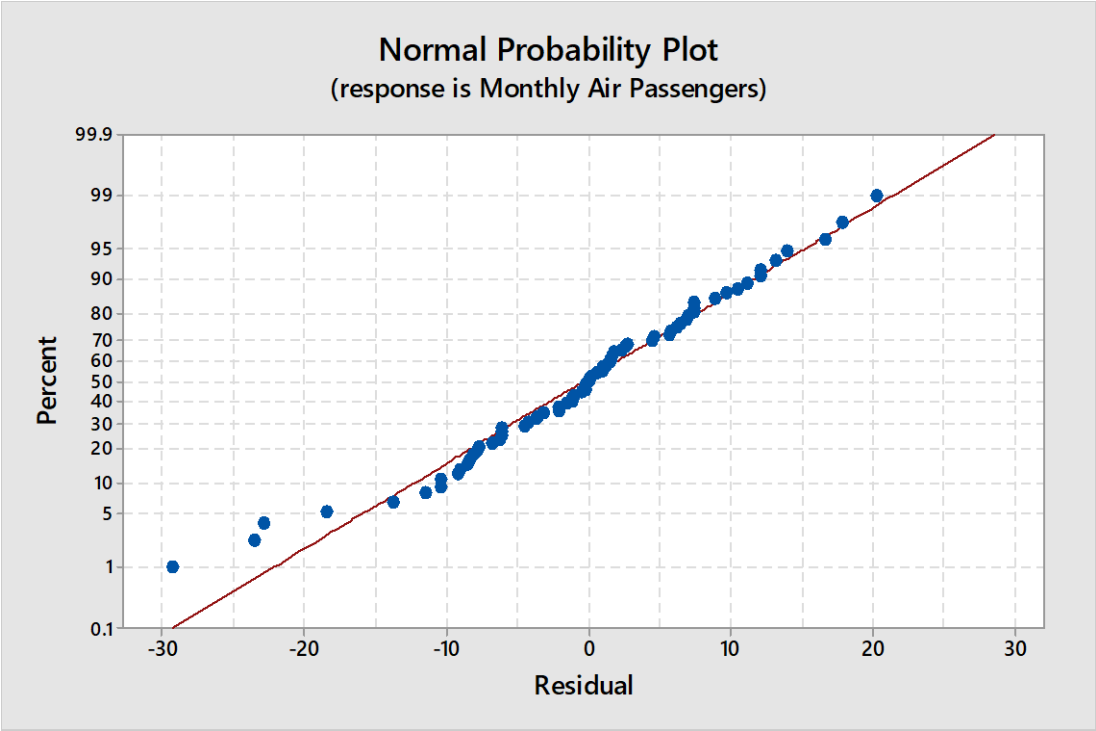

Figure 18 : Normal probability Plot of residuals of Model 02

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

Therefore, we could not conclude that the residuals are normally distributed. Thus it was required to conduct Anderson darling goodness of fit test.

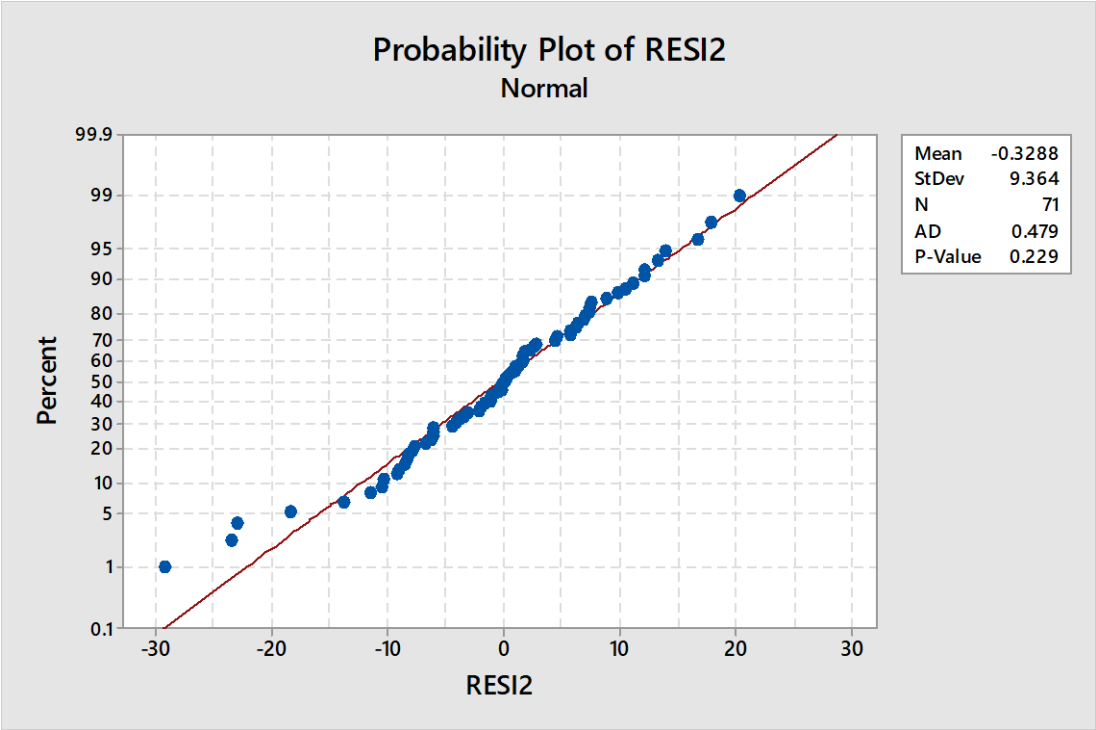

Figure 19 : Output of Anderson Darling Test of Model 02

Since the p value of Anderson darling test is greater than 0.05 the residuals are normally distributed.

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

$$(1 - \propto_1 B)(1 - \propto_1' B^{12})(1 - B)X_t = Z_t$$

$$X_t = 0.7601X_{t-1}0.2399X_{t-2}1.0168X_{t-12} - 0.7729X_{t-13} - 0.0040X_{t-14} + Z_t$$

Equation 14

3.9.3 Model 03

```
Final Estimates of Parameters

Type         Coef  SE Coef      T      P
MA   1   -0.3083   0.1173  -2.63  0.011
SMA  12  -0.8057   0.1612  -5.00  0.000
```

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

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

Lag             12     24     36     48
Chi-Square    46.8   96.5  129.9  150.2
DF              10     22     34     46
P-Value      0.000  0.000  0.000  0.000
```

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 modify the model again.

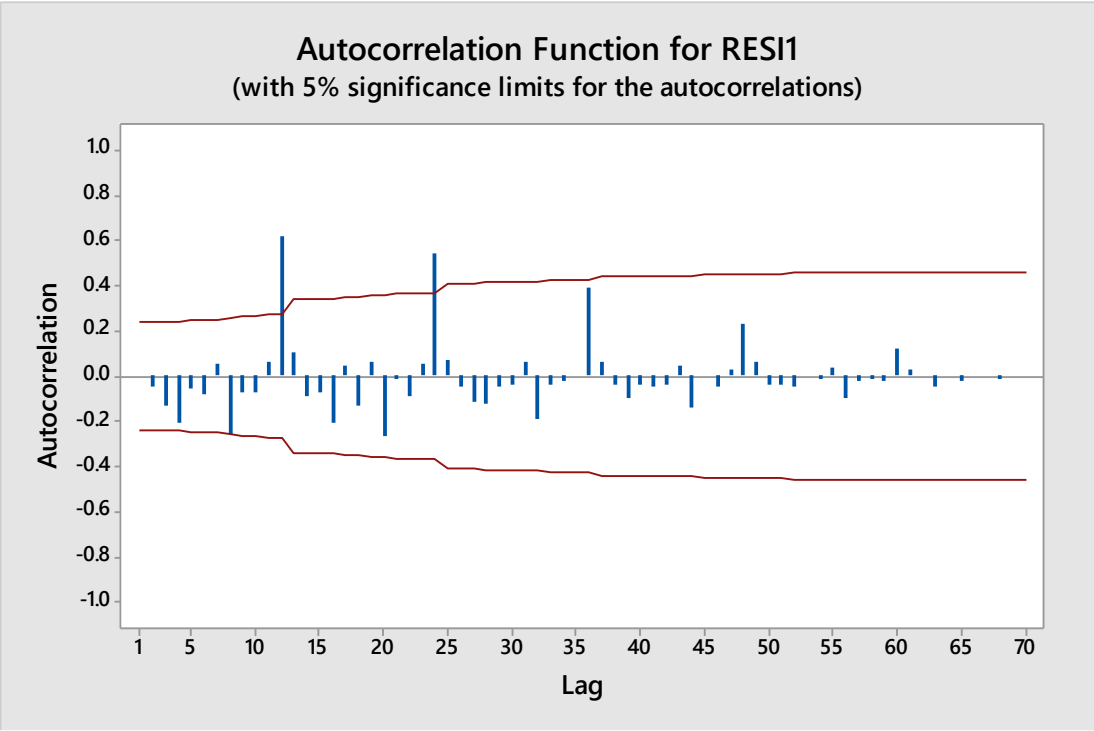

Figure 20 : Autocorrelation function for residuals of model 03

Table 9 : Autocorrelation for residuals of model 03

	lag	ACF	T Statistic	LBQ
Non Seasonal Area	1	-0.00333	-0.02807	0.000822
	2	-0.04565	-0.38469	0.157389
	3	-0.13418	-1.12823	1.529611
	4	-0.20959	-1.73157	4.927789
	5	-0.05252	-0.41665	5.144387
	6	-0.07928	-0.62747	5.645625
	7	0.057058	0.449082	5.909279
	8	-0.25711	-2.01786	11.34757
	9	-0.0732	-0.54416	11.79553
	10	-0.07245	-0.53636	12.24156
	11	0.062108	0.457926	12.57478
Seasonal Area	12	0.623819	4.585902	46.76062
	24	0.543558	2.930597	96.51448
	36	0.396976	1.872017	129.8936

48	0.235725	1.047556	150.2314
60	0.12303	0.53548	163.7822

ACF was cut off at seasonal lag 2. Therefore, two SMA parameter was added to the model.

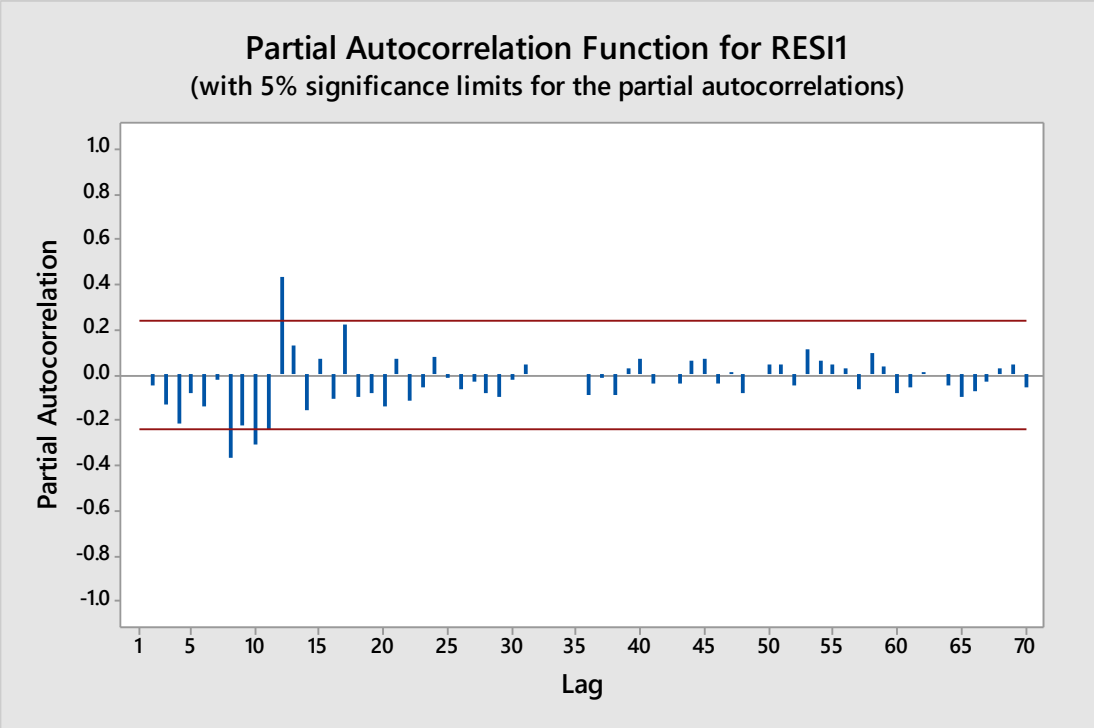

Figure 21 : Partial autocorrelation function for residuals in Model 03

Table 10 : Partial autocorrelation for residuals in Model 03

	lag	PACF	T Statistic
Non Seasonal Area	1	-0.00333	-0.02807
	2	-0.04567	-0.38479
	3	-0.13477	-1.13559
	4	-0.21784	-1.83558
	5	-0.08058	-0.67898
	6	-0.13554	-1.14204
	7	-0.02461	-0.20737
	8	-0.37073	-3.12384
	9	-0.2252	-1.89756
	10	-0.30618	-2.57993

	11	-0.2448	-2.06273
Seasonal Area	12	0.439519	3.70345
	24	0.078188	0.658824
	36	-0.08728	-0.73545
	48	-0.082	-0.69098
	60	-0.07694	-0.6483

PACF was cut off at seasonal lag 1. Therefore, one SAR parameter was added.

The modified model is: SARIMA $(0,1,1)\ (1,0,3)_{12}$

After removing non-significant parameters;

```
Final Estimates of Parameters

Type       Coef  SE Coef      T      P
SAR  12  1.0162   0.0294  34.53  0.000
```

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

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

Lag            12     24     36     48
Chi-Square   13.6   30.4   38.9   43.0
DF             11     23     35     47
P-Value     0.258  0.139  0.299  0.639
```

All p-values of Modified Box-Pierce (Ljung-Box) Chi-Square statistic at seasonal lags were greater than 0.05. Therefore, residuals were random.

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

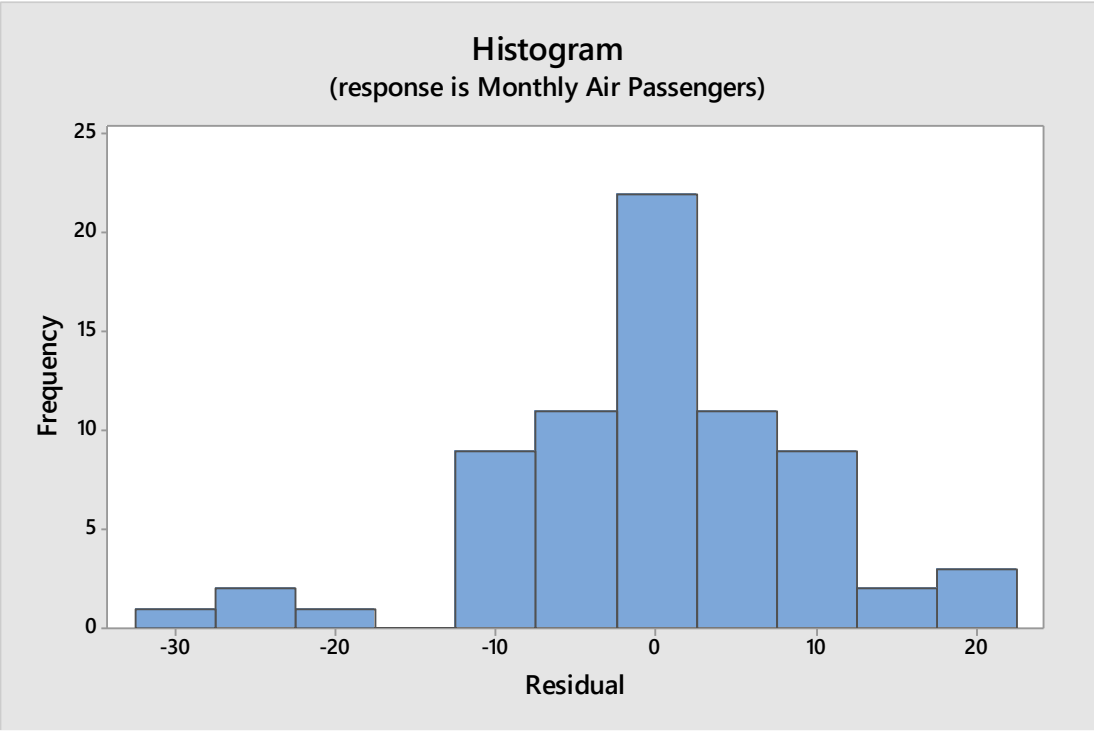

Figure 22 : Histogram of Residuals of Modified Model

A negatively skewed bell shape was shown in histogram of residuals.

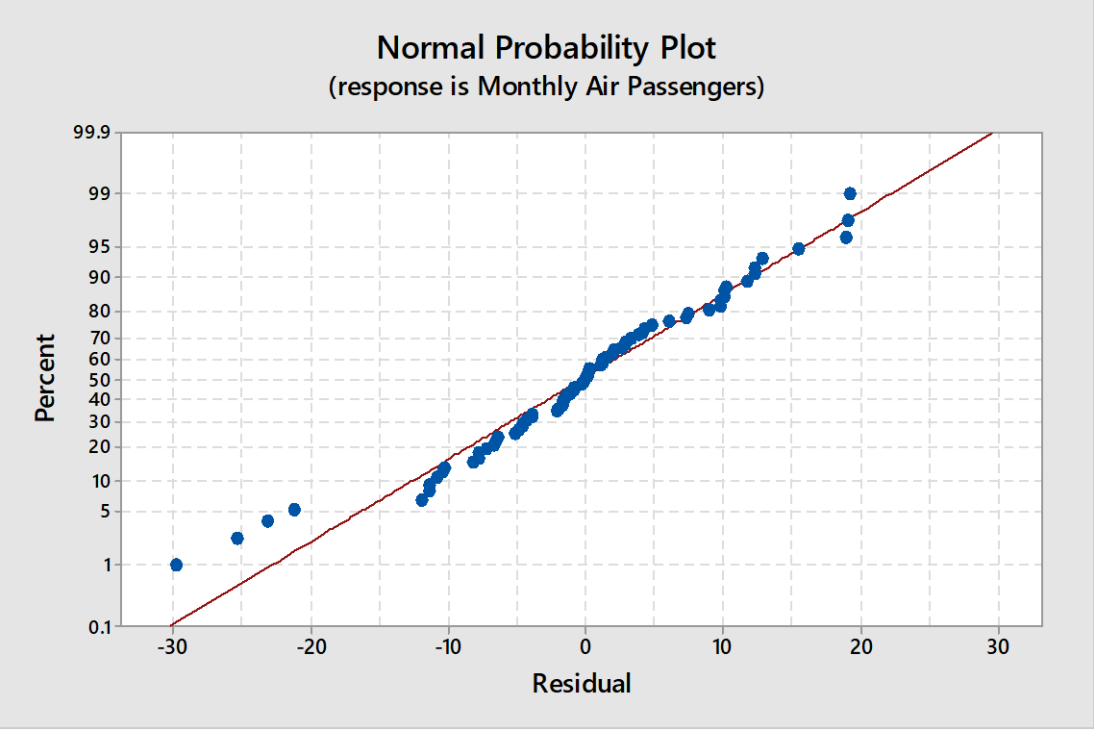

Figure 23 : Normal Probability plot of Residuals of Modified Model

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

Therefore, we could not conclude that the residuals are normally distributed. Thus it was required to conduct Anderson darling goodness of fit test.

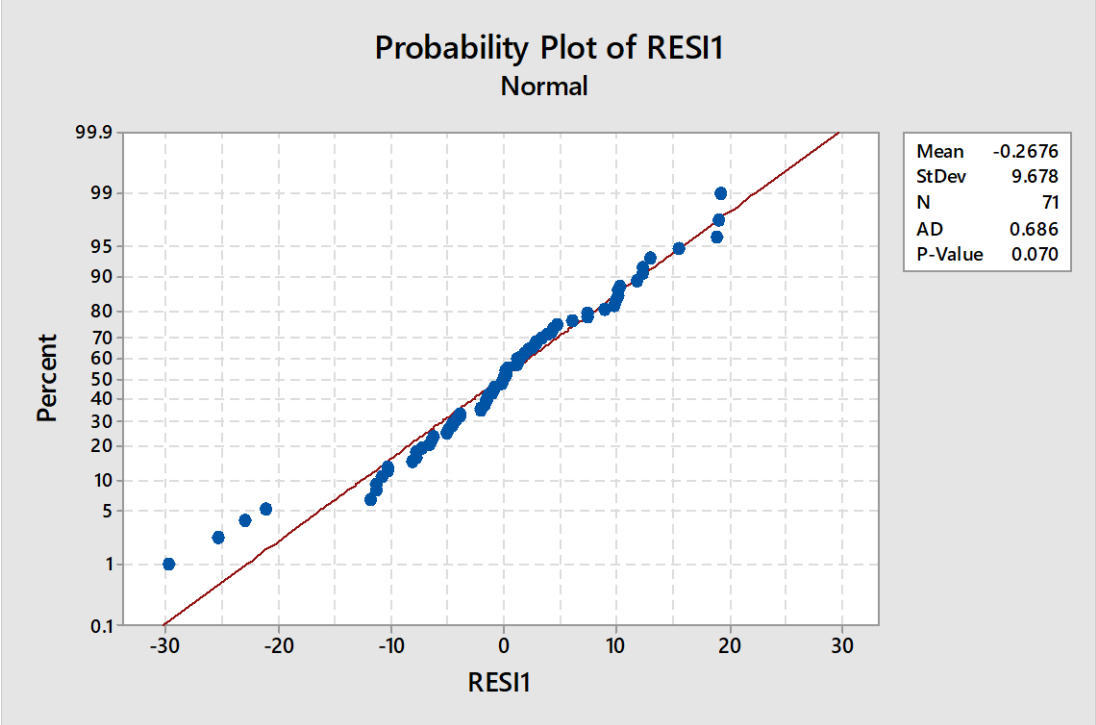

Figure 24 : Output of Anderson Darling Test of Model 03

Since the p value of Anderson darling test is greater than 0.05 the residuals are normally distributed.

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

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

$$X_t = X_{t-1} + 1.0162X_{t-12} - 1.0162X_{t-13} + Z_t \qquad \textit{Equation 15}$$

3.10 Forecasting

3.10.1 Forecasts for Last Observations of Adequate Model 01: SARIMA (0,1,0) $(1,0,0)_{12}$

Table 11 : Forecasted Values for Last Observations of Model 01

Period 1960 Jan-Dec	Actual Values	Forecasted Values
Jan	417	428.3715209
Feb	391	410.0807654
Mar	465	475.1145628
Apr	461	466.9853381
May	472	489.340706
Jun	535	542.1806664
Jul	622	619.4083008
Aug	616	611.2790761
Sep	508	533.0352886
Oct	461	476.1307159
Nov	390	430.4038271
Dec	432	474.0984097

3.10.2 Forecasts for Last Observations of Adequate Model 02: SARIMA (1,1,0) $(1,0,0)_{12}$

Table 12 : Forecasted Values for Last Observations of Model 02

Period 1960 Jan-Dec	Actual Values	Forecasted Values
Jan	417	424.6550266
Feb	391	407.2484526
Mar	465	472.1060728
Apr	461	464.0235512
May	472	486.3798146
Jun	535	539.2540777
Jul	622	616.5268002
Aug	616	608.3929253
Sep	508	530.1026945
Oct	461	473.1643844

Nov	390	427.410375
Dec	432	471.1308712

3.10.3 Future Forecasts for Adequate Model 01: SARIMA $(0,1,0)\ (1,0,0)_{12}$

Table 13 : Forecasted Values for Next Year in Model 01

Period 1961 Jan-Dec	Forecasted Values
Jan	444.193837
Feb	417.7738568
Mar	492.9691851
Apr	488.9045727
May	500.0822566
Jun	564.0999009
Jul	652.5052193
Aug	646.4083008
Sep	536.6637677
Oct	488.9045727
Nov	416.7577037
Dec	459.4361333

3.10.4 Future Forecasts for Adequate Model 02: SARIMA $(1,1,0)\ (1,0,0)_{12}$

Table 14 : Forecasted Values of Next Year in Model 02

Period 1961 Jan-Dec	Forecasted Values
Jan	444.6138693
Feb	418.079177
Mar	493.3428634
Apr	489.2701389
May	500.4558197
Jun	564.5111013
Jul	652.968927
Aug	646.8683739
Sep	537.0587615
Oct	489.2712419

Nov	417.0815866
Dec	459.7853265

3.11 Accuracy Measurements

Table 15 : Accuracy Measurements of Adequate Models

Model	MAPE Value	Forecasting Accuracy
SARIMA (0,1,0) $(1,0,0)_{12}$	3.799576445	96.20042355
SARIMA (1,1,0) $(1,0,0)_{12}$	3.314339915	96.68566008

4. RESULTS

- Time Series Plot:
  - An upward trend
  - A seasonal variation of lag 12
- Tentative Model: SARIMA (1,1,1) $(1,0,2)_{12}$
- Adequate Models:
  - SARIMA (0,1,0) $(1,0,0)_{12}$

$$X_t = X_{t-1} + 1.0162X_{t-12} - 1.0162X_{t-13} + Z_t$$

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

$$X_t = 0.7601X_{t-1}0.2399X_{t-2}1.0168X_{t-12} - 0.7729X_{t-13} - 0.0040X_{t-14} + Z_t$$

- Forecasting Accuracy of Adequate Models:
  - SARIMA (0,1,0) $(1,0,0)_{12}$ = 96.2004
  - SARIMA (1,1,0) $(1,0,0)_{12}$ = 96.6856

5. CONCLUSION

- The Best Fitted Model:

  Since forecasting accuracy of SARIMA $(0,1,0)\ (1,0,0)_{12}$ < forecasting accuracy of SARIMA $(1,1,0)\ (1,0,0)_{12}$, the best model is SARIMA $(1,1,0)\ (1,0,0)_{12}$

- Final Model in Usual Notation:

$$X_t = 0.7601X_{t-1}0.2399X_{t-2}1.0168X_{t-12} - 0.7729X_{t-13} - 0.0040X_{t-14} + Z_t$$

- Future Forecasts of Final Model:

Table 16 : Forecasted Values of Final Model

Period 1961 Jan-Dec	Forecasted Values
Jan	444.6138693
Feb	418.079177
Mar	493.3428634
Apr	489.2701389
May	500.4558197
Jun	564.5111013
Jul	652.968927
Aug	646.8683739
Sep	537.0587615
Oct	489.2712419
Nov	417.0815866
Dec	459.7853265

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 three significant spikes in non-seasonal area. Therefore, original series was not stationary. In order to make the series stationary a non-seasonal difference was carried out. After the difference the autocorrelation function was stationary and partial autocorrelation function was also stationary. By inspecting autocorrelation function and partial autocorrelation function of stationary series and the number of differences we obtained a tentative model for the data as SARIMA $(1,1,1)\ (1,0,2)_{12}$. While checking the parameter significance three different models with significant parameters was found. They were SARIMA $(0,1,0)\ (1,0,0)_{12}$, SARIMA $(1,1,0)\ (1,0,0)_{12}$, and SARIMA $(0,1,1)\ (0,0,1)_{12}$ respectively. However, when diagnostic checking number of adequate models were reduced into two as it was required to modified the third model in order to make it adequate. Final model modified for the model SARIMA $(0,1,1)\ (0,0,1)_{12}$ was the model SARIMA $(0,1,0)\ (1,0,0)_{12}$ which was an existing model.in diagnostic checking process normality of residuals of all models could not be proved using normal probability plot and histogram of residuals as they were violating the rules of normality. Therefore, we had to perform Anderson darling goodness of fit test to verify the normality of residuals. Anderson darling test's output was that residuals are normally distributed. After diagnostic checking we were left with two 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 $(0,1,0)\ (1,0,0)_{12}$ and SARIMA $(1,1,0)\ (1,0,0)_{12}$ were 96.2004 and 96.6856 respectively. That concluded that the model SARIMA $(1,1,0)\ (1,0,0)_{12}$ was the best model to forecast as the accuracy is higher than the other adequate model. Finally using the best fit model forecasts for the next year (1961) were calculated.

As we mentioned before normality checking step was a bit exaggerated due to the number of data points in the data set was less. That lead us to do a goodness of fit test to verify the normality in advance. If the data set contained higher number of data points we could have avoid that issue. However, from this analyse we learned and understood the Box- Jenkins methodology of identifying an adequate model for time series analysis.

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/