FINAL PROJECT REPORT

STAT 3124

TIME SERIES ANALYSIS

Submitted by

162037

162045

162048

162095

162112

Department of Mathematical sciences

Faculty of Applied Sciences

Wayamba University of Sri Lanka

1. INTRODUCTION

This dataset provides monthly totals of US Airline passengers from 1949 to 1960. This data set is taken from an inbuilt dataset of R called Air Passengers. We can use this data set to forecast the future values and help the business. Forecasting is a very important part for a business to satisfy their customers. US Airline passenger forecasting is useful for US for make some decisions about rules and regulations, decisions about visas and passports. For airline companies in US it is useful to decide service frequency, aircraft size, ticket prices, fight distance, and number of airports in the route. So airline companies can adopt with seasonal variations and trend variations. In this project Minitab software is used to forecast the future values for next twelve months.

Table 01 - Data Set of Airline Passengers

	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
Jan	112	115	145	171	196	204	242	284	315	340	360	417
Feb	118	126	150	180	196	188	233	277	301	318	342	391
Mar	132	141	178	193	236	235	267	317	356	362	406	419
Apr	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
Aug	148	170	199	242	272	293	347	405	467	505	559	606
Sep	136	158	184	209	237	259	312	355	404	404	463	508
Oct	119	133	162	191	211	229	274	306	347	359	407	461
Nov	104	114	146	172	180	203	237	271	305	310	362	390
Dec	118	140	166	194	201	229	278	306	336	337	405	432

2. STATISTICAL ANALYSIS

2.1 Time Series Plot

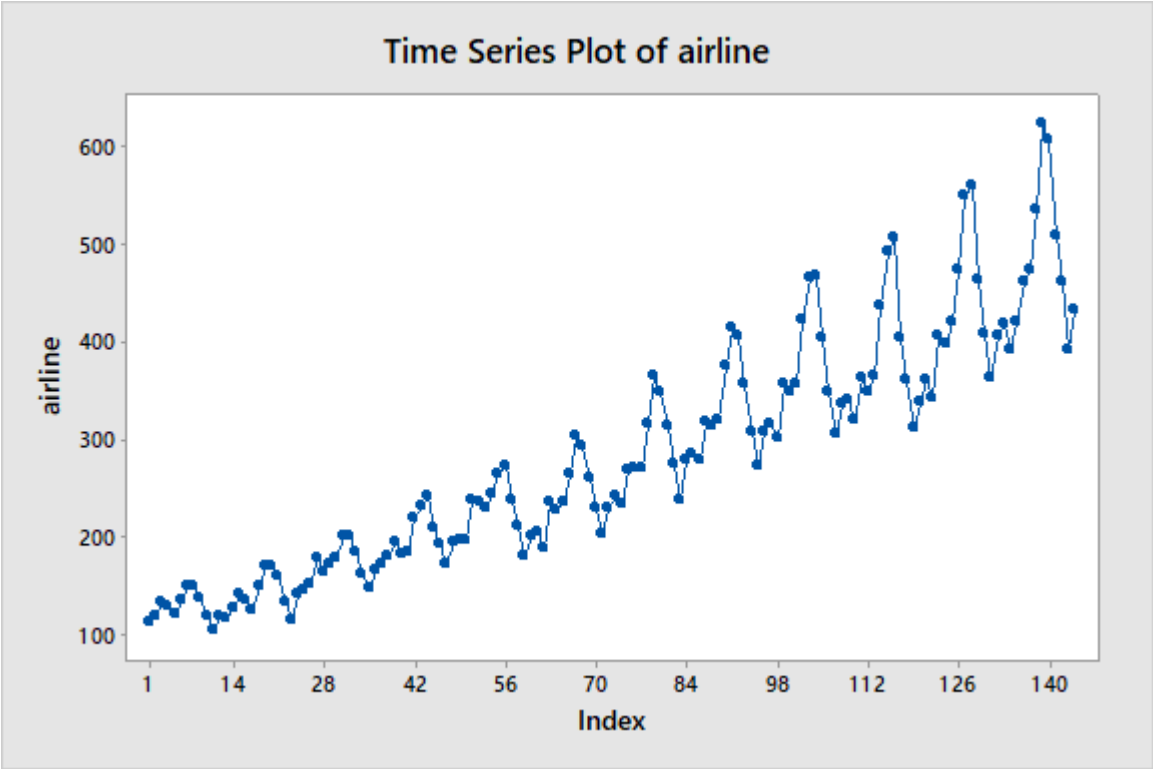


Figure 01- Time Series Plot of airline passengers

- There is an upward trend.
- There is an increasing seasonal variation.
- Seasonal length is 12.

2.2 The graph of Autocorrelation Function

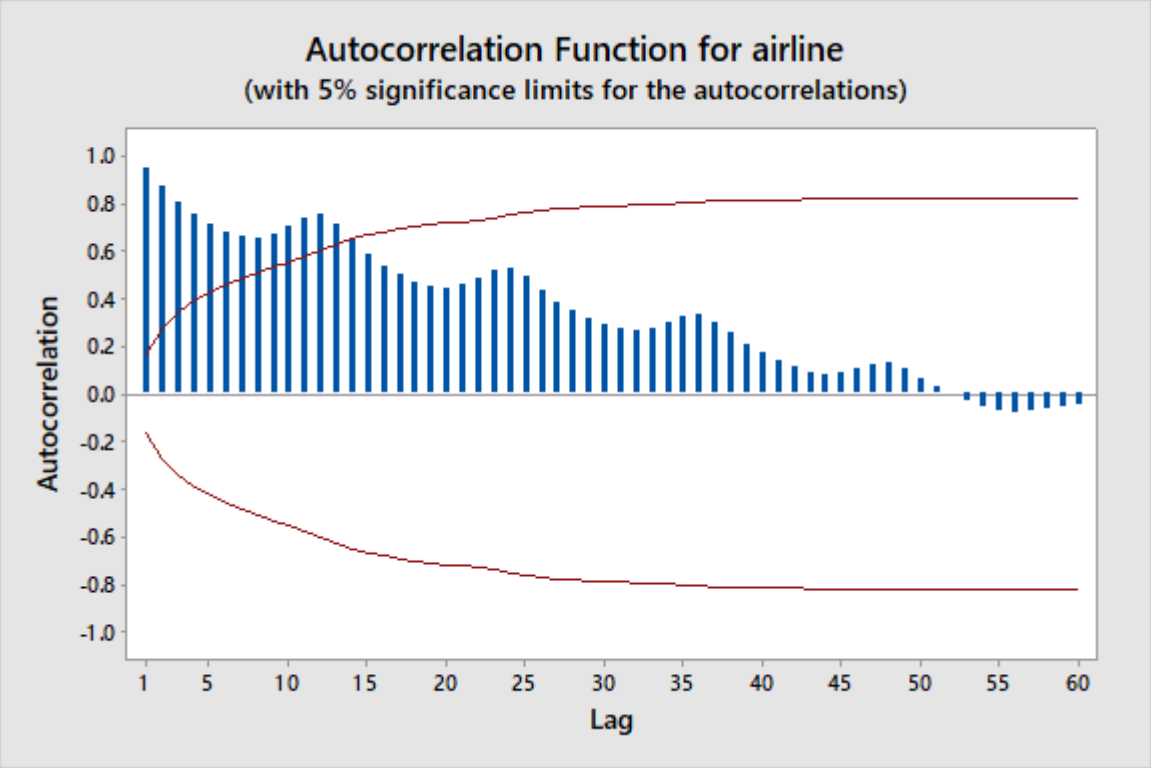


Figure 02- The graph of Autocorrelation Function.

- We can see a trend since the ACF slowly dies down.
- Therefore it is not stationary.
- Therefore it needs to perform a non-seasonal differencing.

(Appendix A: Page No 27)

2.3 The graph of non-seasonal differenced Autocorrelation Function

Hypothesis;

H_0: $\rho_k=0$

H_1: Not so

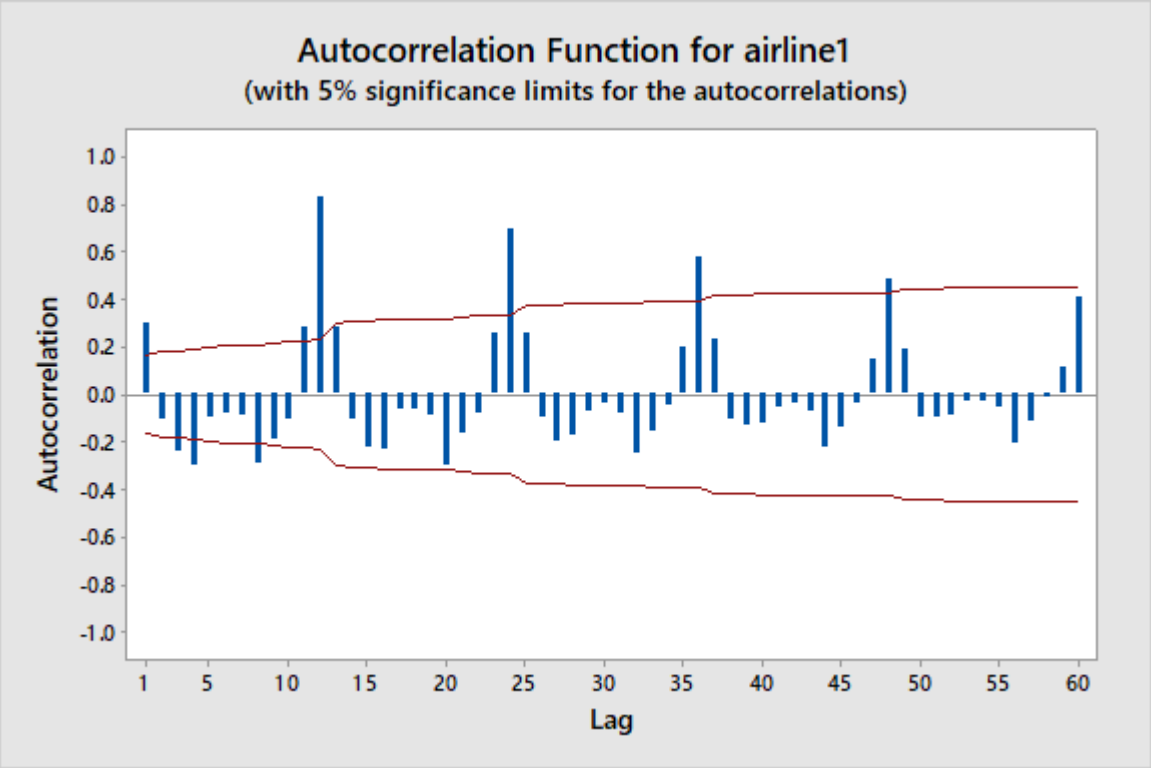


Figure 03- The graph of first non-seasonal differenced Autocorrelation Function.

- We can see a seasonal variation.
- Therefore it is not stationary.
- Therefore it needs a seasonal differencing.

(Appendix B: Page No 28)

2.4 The graph of seasonal differenced Autocorrelation Function

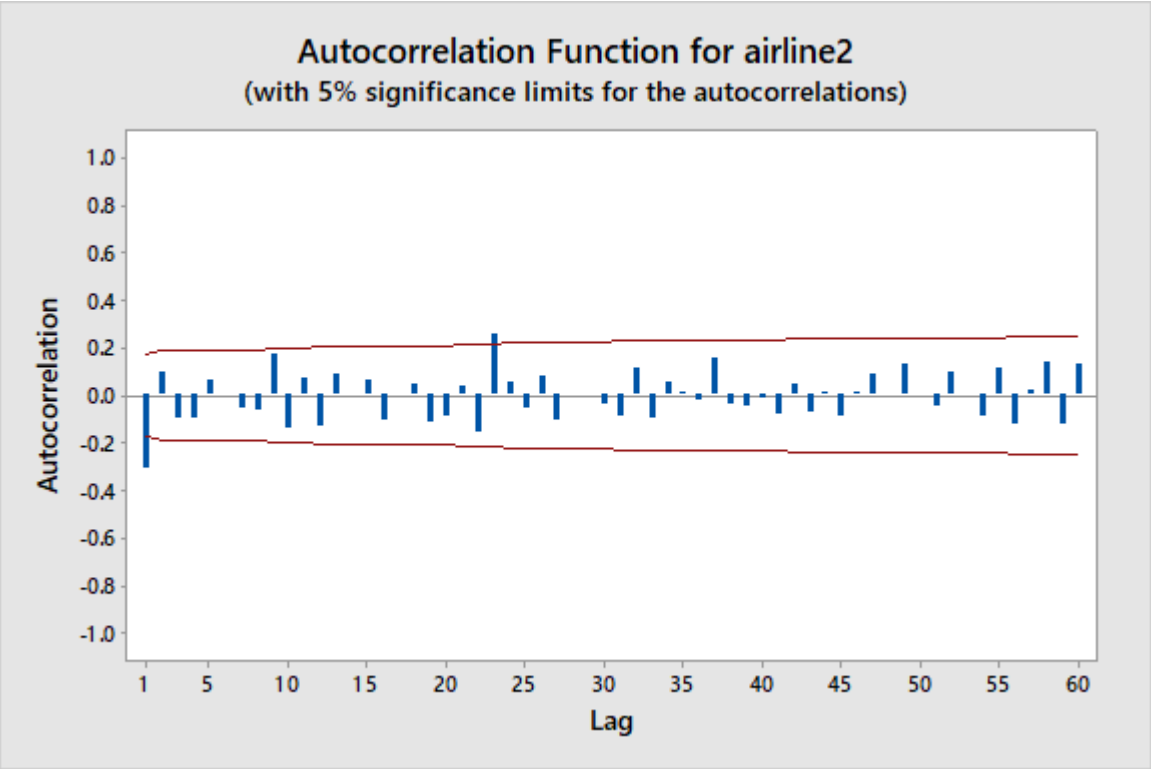


Figure 04- The graph of seasonal differenced Autocorrelation Function.

Autocorrelation Function: airline2

Lag	ACF	T	LBQ
1	-0.309815	-3.55	12.86
2	0.095351	1.00	14.09
3	-0.096891	-1.01	15.37
4	-0.098995	-1.02	16.71
5	0.061001	0.62	17.23
6	-0.000288	-0.00	17.23
7	-0.056108	-0.57	17.67

8	-0.060966	-0.62	18.20
9	0.175917	1.79	22.62
10	-0.140279	-1.39	25.45
11	0.069735	0.68	26.16
12	-0.133673	-1.30	28.77
24	0.052836	0.46	51.36
36	-0.018646	-0.16	61.19
48	-0.003843	-0.03	72.28
60	0.128968	1.01	99.67

In non-seasonal area

|T|>2 in lag 1.

The ACF is cut off at lag 1.

In seasonal area

In all seasonal lags |T|< 2.

Do not reject H_0.

ACF is zero.

Therefore the series is stationary.

2.5 The graph of Partial Autocorrelation Function

Hypothesis;

H_0: $\rho_k=0$

H_1: Not so

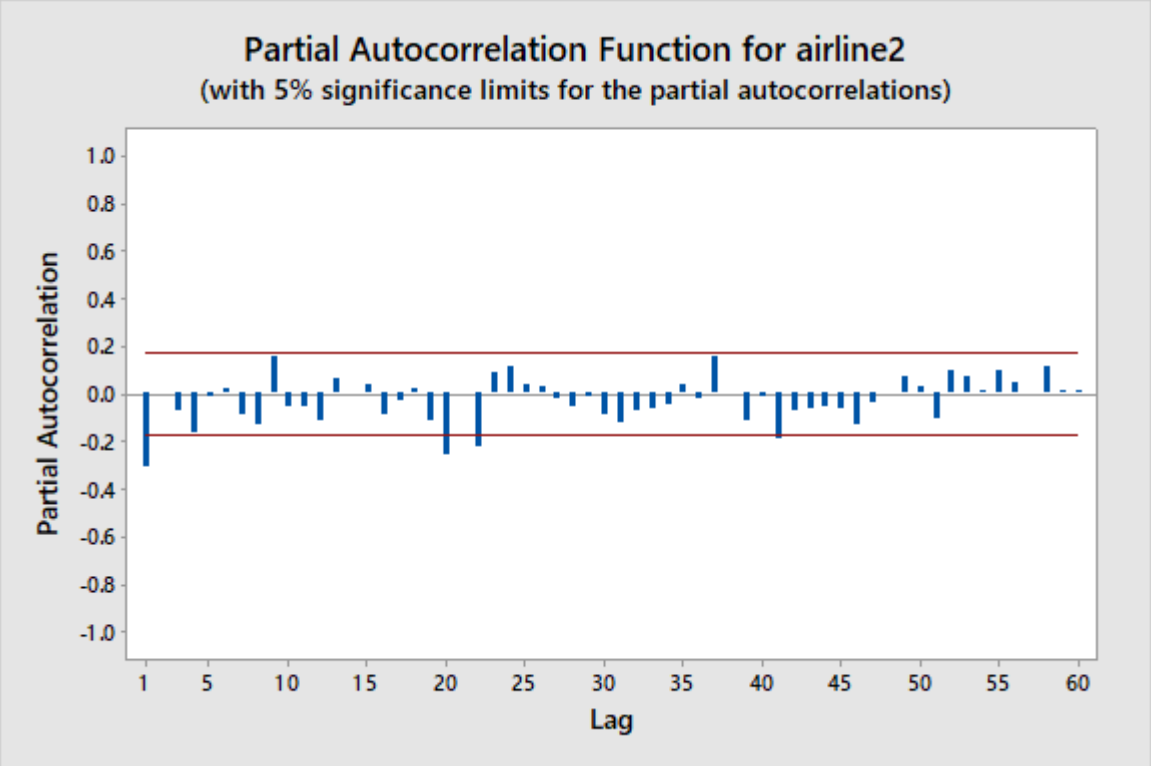


Figure05- The graph of Partial Autocorrelation Function.

Partial Autocorrelation Function: airline2

Lag	PACF	T
1	-0.309815	-3.55
2	-0.000701	-0.01
3	-0.074718	-0.86
4	-0.166761	-1.91
5	-0.015146	-0.17

6	0.018288	0.21
7	-0.088086	-1.01
8	-0.133888	-1.53
9	0.156405	1.79
10	-0.058750	-0.67
11	-0.052428	-0.60
12	-0.115013	-1.32
24	0.117491	1.34
36	-0.021203	-0.24
48	0.001826	0.02
60	0.012838	0.15

In non-seasonal area

At non seasonal lag 1 $|T|> 2$

The PACF is cuts off at non seasonal lag 1.

In seasonal area

At All seasonal lags $|T|< 2$.

Do not reject H_0.

PACF is zero.

ACF

The ACF is cuts off at non-seasonal lag 1 ⟶ . q=1

The ACF is zero at seasonal lags. ⟶ Q=0

PACF

The PACF is cuts off in non-seasonal lag 1 ⟶. p=1

The PACF is zero in seasonal lags. ⟶ P=0

Number of seasonal differencing=1 ⟶ D=1

Number of non-seasonal differencing=1 ⟶ d=1

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

2.6 Parameter estimation and diagnostic checking

- Parameter significance

Hypothesis;

H_0: Coefficient = 0 and constant = 0

H_1: Not so

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

- P values of AR1 and MA1 terms are greater than 0.05

  Hence the coefficient of AR1 and MA1 are not significant from zero.

  Do not reject null hypothesis at 5% level of significance.

- Hence need to remove constant term, AR1 and MA1 terms.

After removing constant term:

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

- P values of AR1 and MA1 terms are greater than 0.05

Hence the coefficient of AR1 and MA1 are not significant from zero.

Do not reject null hypothesis at 5% level of significance.

- Remove MA1 and AR1 terms separately.
- Then the tentative models are,
  SARIMA(1, 1, 0) $(0, 1, 0)_{12}$
  SARIMA(0, 1, 1) $(0, 1, 0)_{12}$

(Appendix C: Page No 30)

2.6.1 Checking for the first tentative model

- Parameter significance

After removing one non-seasonal MA value;

Final Estimates of Parameters

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

- P values of AR1 term is less than 0.05
- Hence the coefficient of AR1 is significant from zero.
- Reject null hypothesis at 5% level of significance.
- Therefore the parameters are significance
- Hence the tentative model is SARIMA(1, 1, 0) $(0, 1, 0)_{12}$

- Randomness of residuals

Hypothesis

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

$H_0 : \rho_k \neq 0$ (at least one)

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

- P value at lag 24 is less than 0.05
- Therefore the residuals are not random
- Need to check the ACF and PACF of residuals

Hypothesis;

H_0: $\rho_k=0$

H_1: Not so

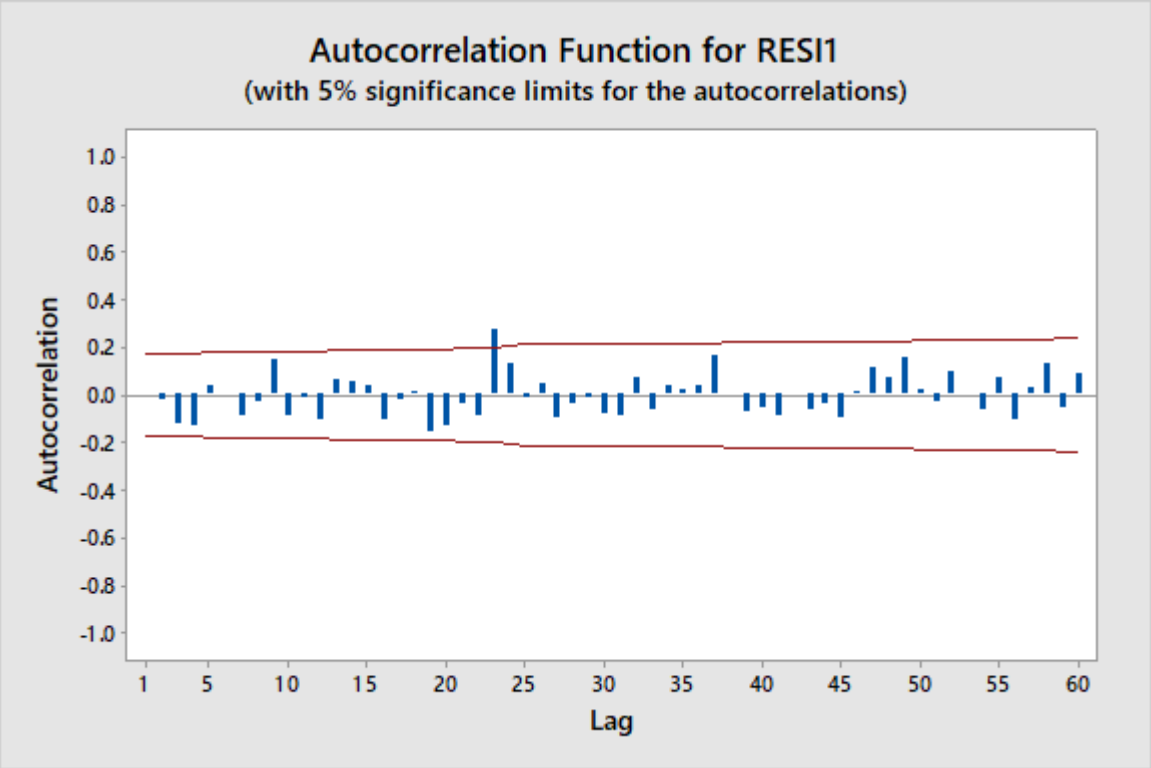


Figure 06-The graph of Autocorrelation Function for Residuals

Autocorrelation Function: RESI1

Lag	ACF	T	LBQ
1	-0.000341	-0.00	0.00
2	-0.024231	-0.28	0.08
3	-0.118495	-1.36	1.99
4	-0.132856	-1.50	4.41
5	0.039122	0.43	4.62
6	0.001553	0.02	4.62
7	-0.088909	-0.98	5.73
8	-0.032543	-0.36	5.88
9	0.145361	1.60	8.90
10	-0.086985	-0.94	9.99
11	-0.009959	-0.11	10.01
12	-0.108317	-1.16	11.72
24	0.135143	1.28	38.26
36	0.034225	0.31	45.01
48	0.071732	0.63	60.06
60	0.091608	0.76	79.66

In non-seasonal area; |T|<2 in all lags

In seasonal area; |T|<2 in all lags

Do not reject H_0.

ACF is zero.

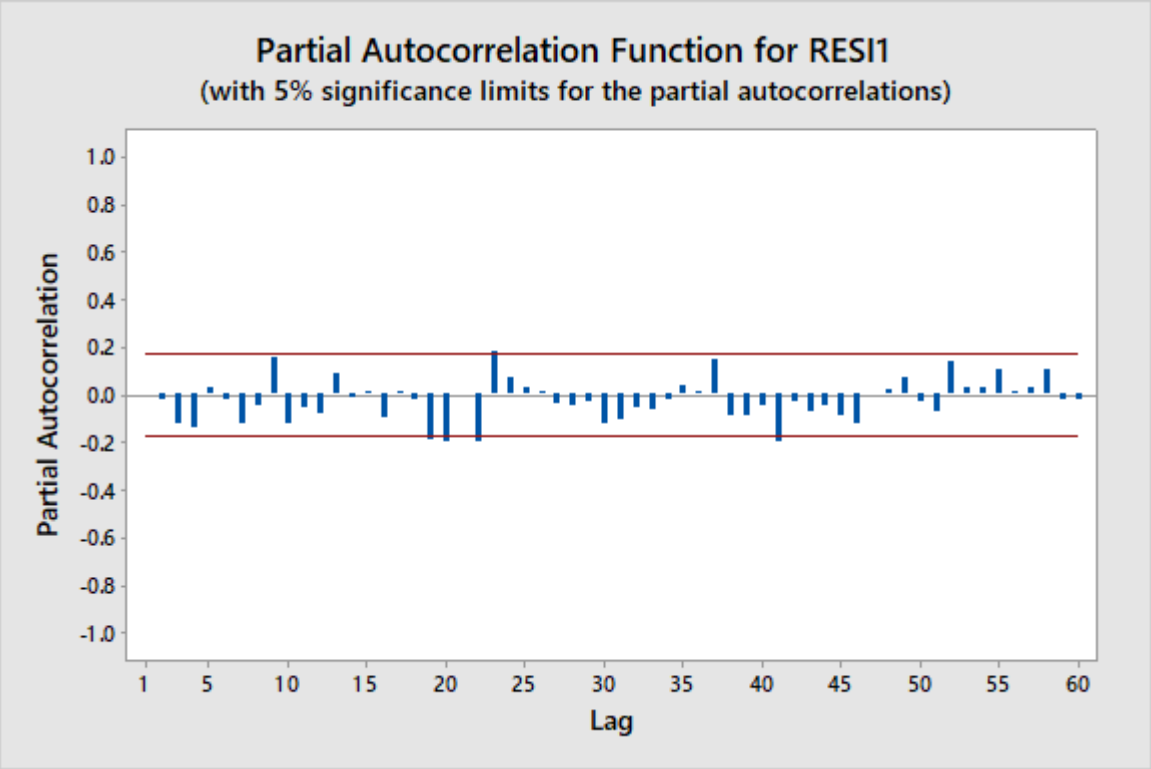


Figure 07-The graph of Partial Autocorrelation Function for Residuals

Hypothesis;

H_0: $\rho_k=0$

H_1: Not so

Partial Autocorrelation Function: RESI1

Lag	PACF	T
1	-0.000341	-0.00
2	-0.024231	-0.28
3	-0.118581	-1.36
4	-0.135856	-1.55
5	0.031387	0.36
6	-0.018930	-0.22
7	-0.123392	-1.41
8	-0.048240	-0.55
9	0.153982	1.76
10	-0.122665	-1.40
11	-0.050960	-0.58
12	-0.081471	-0.93
24	0.074327	0.85
36	0.016375	0.19
48	0.022847	0.26
60	-0.019554	-0.22

In non-seasonal area; |T|<2 in all lags

In seasonal area; |T|<2 in all lags

Do not reject H_0.

PACF is zero.

- Normality of residuals

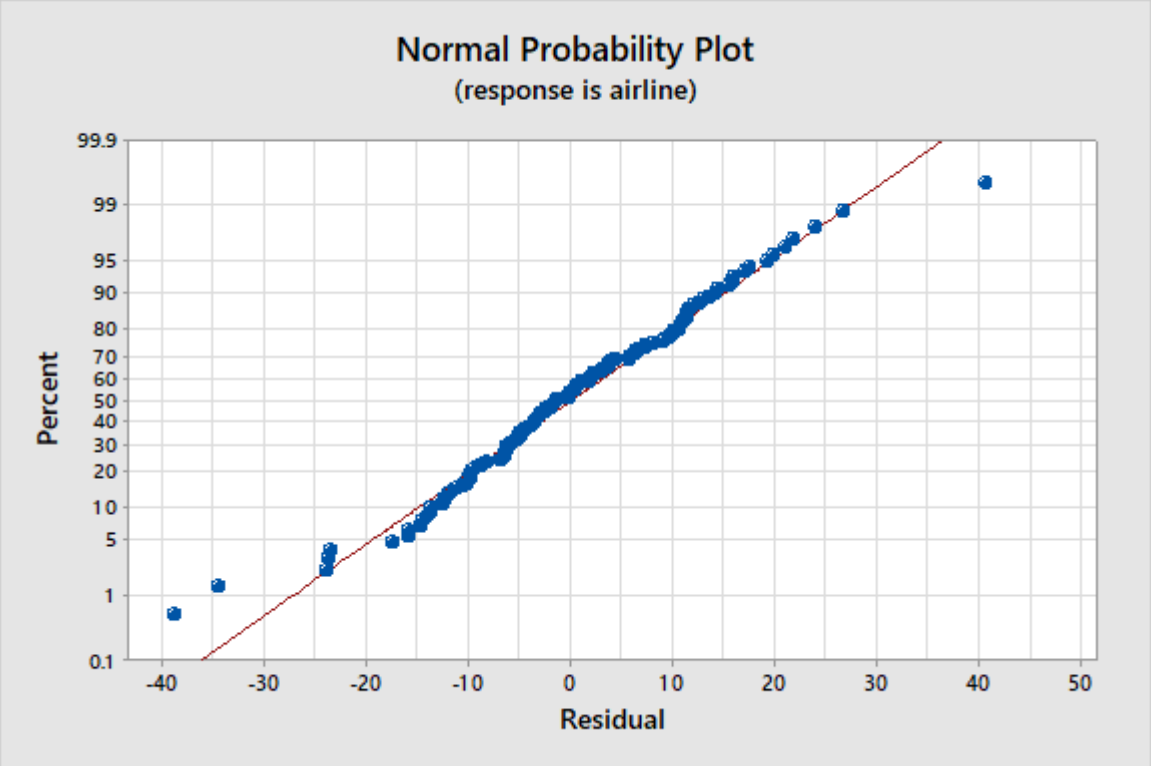


Figure 08- The Normal Probability Plot

- Residuals lie on a straight line
- Therefore the residuals are normally distributed
- Hence SARIMA$(1,1,0)(0,1,0)_{12}$ is an adequate model

2.6.2 Checking for the second tentative model

- Parameter significance

After removing MA1 term:

Final Estimates of Parameters

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

- P values of MA1 terms is less than 0.05
- Hence the coefficient of MA1 term is significant from zero.
- Therefore the parameters are significance 5% level of significance.
- Hence the tentative model is SARIMA(0, 1, 1) $(0, 1, 0)_{12}$

- Randomness of residuals

Hypothesis

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

$H_0 : \rho_k \neq 0$ (at least one)

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

- P value at lag 24 is less than 0.05
- Therefore the residuals are not random
- Need to check the ACF and PACF of residuals

Hypothesis;

H_0: $\rho_k=0$ H_1: Not so

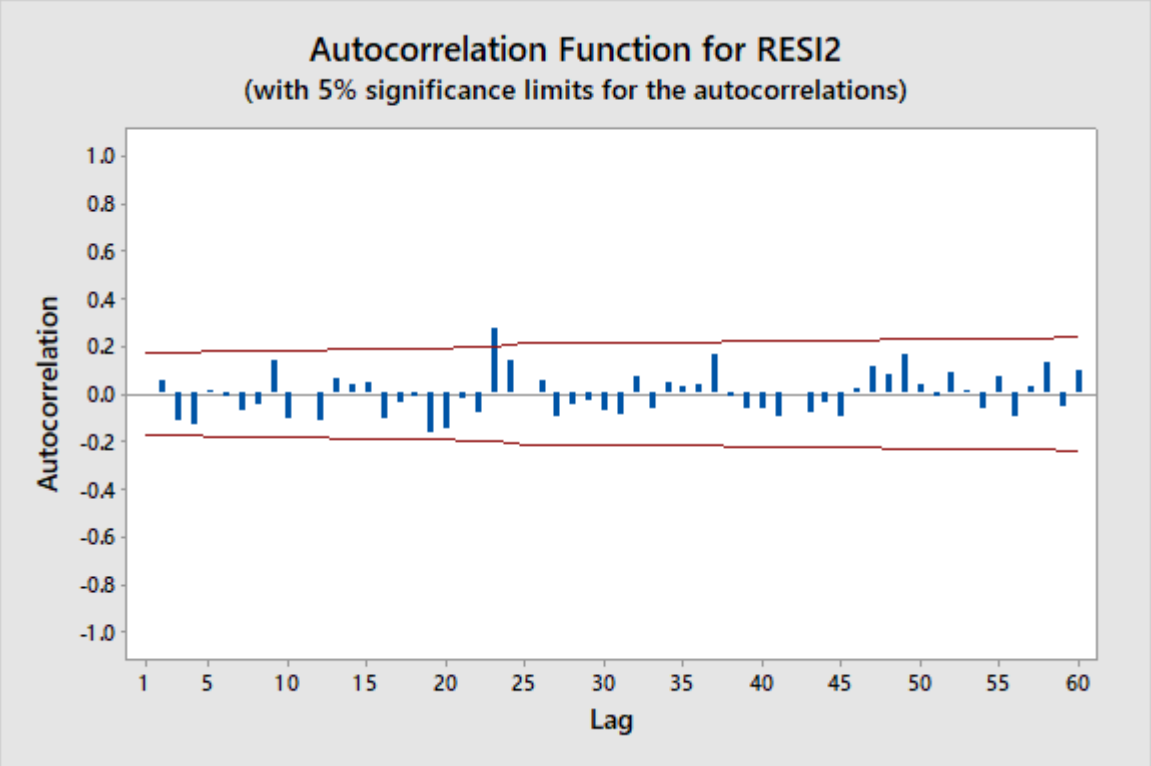


Figure 09-The graph of Autocorrelation Function for Residuals

Autocorrelation Function: RESI2

Lag	ACF	T	LBQ
1	-0.002740	-0.03	0.00
2	0.056933	0.65	0.44
3	-0.116820	-1.33	2.30

4	-0.129514	-1.46	4.60
5	0.014700	0.16	4.63
6	-0.015218	-0.17	4.66
7	-0.071372	-0.79	5.38
8	-0.042355	-0.47	5.63
9	0.141408	1.56	8.49
10	-0.103581	-1.12	10.0
11	0.004830	0.05	10.03
12	-0.114706	-1.23	11.96
24	0.136766	1.29	38.64
36	0.037744	0.34	46.11
48	0.079046	0.69	62.12
60	0.099706	0.83	82.03

In non-seasonal area; $|T|<2$ in all lags

In seasonal area; $|T|<2$ in all lags

Do not reject H_0.

ACF is zero.

Hypothesis;

H_0: $\rho_k=0$ H_1: Not so

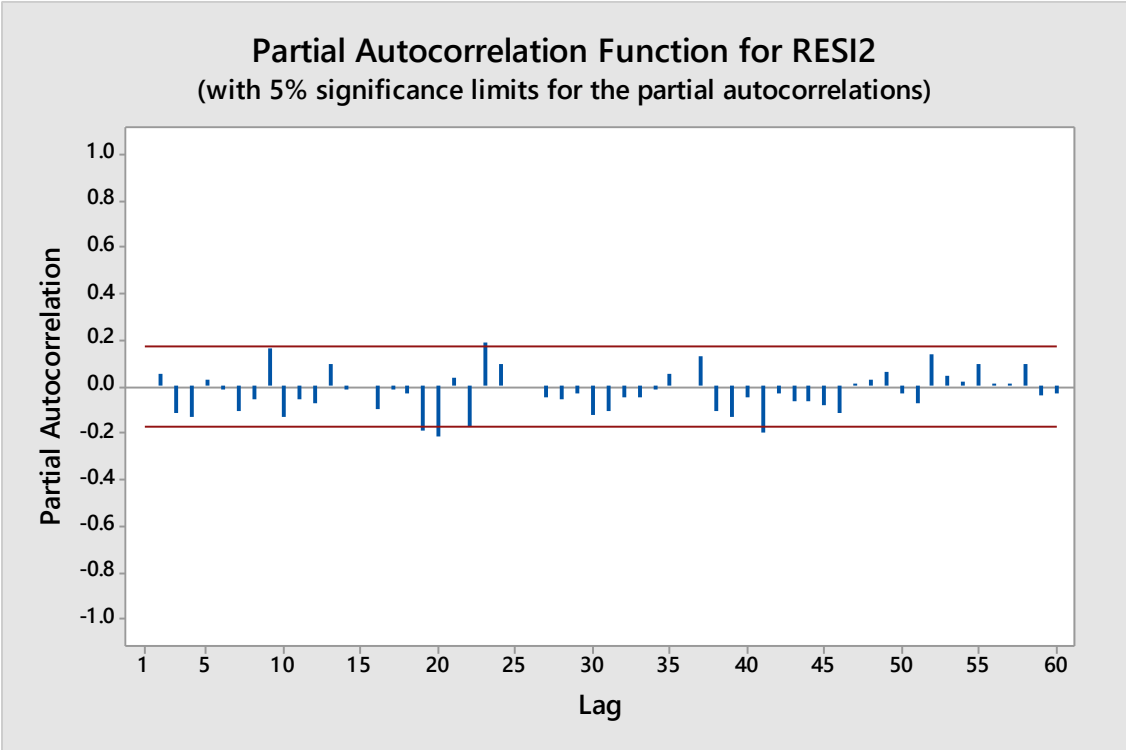


Figure 10-The graph of Partial Autocorrelation Function for Residuals

Partial Autocorrelation Function: RESI2

Lag	PACF	T
1	-0.002740	-0.03
2	0.056926	0.65
3	-0.116896	-1.34
4	-0.134857	-1.54
5	0.027917	0.32
6	-0.012881	-0.15

7	-0.109075	-1.25
8	-0.057098	-0.65
9	0.162532	1.86
10	-0.130483	-1.49
11	-0.059004	-0.68
12	-0.072188	-0.83
24	0.093705	1.07
36	0.006225	0.07
48	0.025570	0.29
60	-0.032501	-0.37

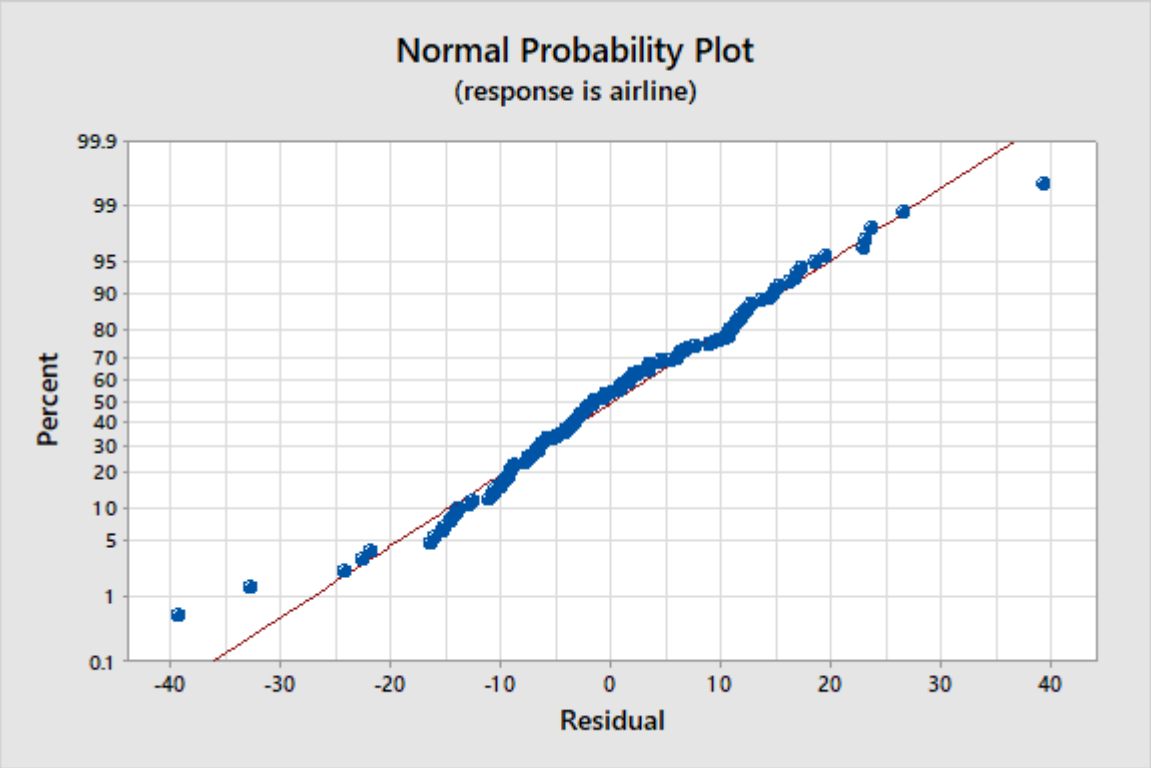
In non-seasonal area; $|T|<2$ in all lags

In seasonal area; $|T|<2$ in all lags

Do not reject H_0.

PACF is zero.

- Normality of residuals



Residuals lie on a straight line

Therefore the residuals are normally distributed

The adequate model is SARIMA(0, 1, 1) $(0, 1, 0)_{12}$

There are two adequate models

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

2.7 Forecasting

Table 02 – Forecasting future values for two adequate models

Month(1961)	SRIMA(1,1,0)(0,1,0)	SRIMA(0,1,1)(0,1,0)
Jan	444.310	446.803
Feb	418.214	420.803
Mar	446.244	448.803
Apr	488.234	490.803
May	499.237	501.803
June	562.236	564.803
July	649.237	651.803
Aug	633.237	635.803
Sep	535.237	537.803
Oct	488.237	490.803
Nov	417.237	419.803
Dec	459.237	461.803

Table 03 – Forecasting values for last 12 month for calculating MAPE

Actual	SRIMA(1,1,0)(0,1,0)	SRIMA(0,1,1)(0,1,0)
417	423.041	422.754
391	406.578	404.754
419	470.101	468.754
461	460.249	458.754
472	484.203	482.754
535	536.218	534.754
622	612.213	610.754
606	623.215	621.754

508	527.214	525.754
461	471.214	469.754
390	426.214	424.754
432	469.214	467.754

MAPE = (SUM(ABS('Actual'-'Forecast')/'Actual')/12)*100 ⟶ Equation 01

Accuracy = 100-(SUM(ABS('Actual'-'Forecast')/'Actual')/12)*100 ⟶ Equation 02

Table 04 – MAPE value and accuracy of the models

	$SRIMA(1,1,0)(0,1,0)_{12}$	$SRIMA(0,1,1)(0,1,0)_{12}$
MAPE	4.07641	3.88102
Accuracy	95.9236	96.1190

$SRIMA(0,1,1)(0,1,0)_{12}$ has the highest accuracy value.

Hence $SRIMA(0,1,1)(0,1,0)_{12}$ is the most suitable model for forecasting the monthly totals of US Airline passengers.

3. APPENDIX

Appendix A- Autocorrelation Function values of Airline Passengers

Autocorrelation Function: airline

Lag	ACF	T	LBQ
1	0.948047	11.38	132.14
2	0.875575	6.28	245.65
3	0.806681	4.65	342.67
4	0.752625	3.81	427.74
5	0.713770	3.29	504.80
6	0.681734	2.93	575.60
7	0.662904	2.69	643.04
8	0.655610	2.54	709.48
9	0.670948	2.49	779.59
10	0.702720	2.50	857.07
11	0.743240	2.54	944.39
12	0.760395	2.49	1036.48
24	0.532190	1.40	1606.08
36	0.337024	0.82	1866.63
48	0.132635	0.32	1933.16
60	-0.046934	-0.11	1943.67

$$T_{rk} = \frac{r_k}{\frac{1}{\sqrt{n}}\sqrt{1+2\sum Y_j^2}} \longrightarrow$$ Equation 03

Where;

n = number of data

$T_{rk} = T\ statistic$

$r_k = sample\ autocorrelation\ at\ lag\ k$

Y_j = data value

Non seasonal area

| T | > 2 at lag 1-11. Therefore, reject H_0. Hence ACF is not zero.

ACF slowly dies down in non-seasonal area, therefore non seasonal area is not stationary.

Seasonal area

| T | > 2 at seasonal lag 1. ACF cuts off at seasonal lag 1.

| T | < 2 in lag 24, 36, 48, 60. Therefore, do not reject H_0. Hence ACF is zero.

Appendix B

Autocorrelation Function: airline1

Lag	ACF	T	LBQ
1	0.302855	3.62	13.39
2	-0.102148	-1.12	14.93
3	-0.241273	-2.63	23.55

4	-0.300402	-3.13	37.01
5	-0.094073	-0.92	38.34
6	-0.078443	-0.76	39.27
7	-0.092362	-0.89	40.57
8	-0.294802	-2.83	53.92
9	-0.191778	-1.75	59.61
10	-0.104917	-0.94	61.33
11	0.282931	2.51	73.90
12	0.829178	7.05	182.73
24	0.701086	4.12	334.36
36	0.579577	2.90	451.19
48	0.485694	2.23	544.15
60	0.409610	1.79	618.24

Non seasonal area

$|T| > 2$ at lag 1, 3, 4, 8, 11. Therefore, reject H_0.

Therefore non seasonal area is not stationary.

Seasonal area

| T | > 2 at seasonal lag 12, 24,36, 48.

| T | > 2 in lag 12, 24, 36, 48.

Appendix C

In Final Estimates of Parameters after removing the constant term ,P values of both AR1 and MA1 are greater than 0.05.First we have removed the MA1 term and checked the adequacy of the model and got the first fitted model.

And then we have removed the AR1 term and checked the adequacy of the model and got the second fitted model.

4. CONCLUSION AND DISCUSSION

- Seasonal variation and trend variation are identified in this dataset from 1949 to 1960.
- From the statistical analysis two adequate models can be identified for the US airline passenger dataset. They are,

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

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

- After measuring the forecast error using MAPE, the lowest value is obtained through $SRIMA(0,1,1)(0,1,0)_{12}$ model. Therefore it is the most accurate model.