MATH1318 Time Series Analysis
Final Project Report

Declaration of contributions:

No	Name of Team Member	Contribution to the project
1	Bharat Narayanan Venkatesh	1/3
2	Monish Chezhian	1/3
3	Nachiappan Nachiappan	1/3
	Sum:	1

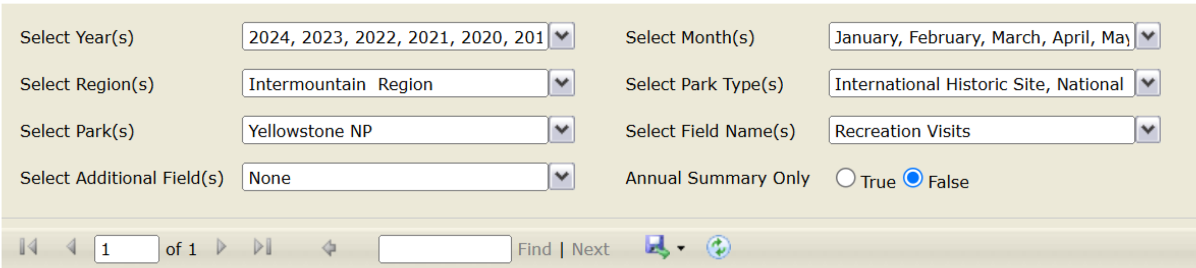
RMIT University
School of Computing Technologies

2025

Table of Contents

1. Introduction

Yellowstone National Park attracts millions of visitors each year, yet accurately forecasting monthly foot traffic remains challenging for park managers tasked with staffing, maintenance and conservation planning. The dataset used in this analysis comprises monthly visitor counts to Yellowstone National Park from January 2000 to December 2024. This data was obtained from the official Integrated Resource Management Applications (IRMA) portal of the U.S. National Park Service, accessible at: [IRMA Stats Portal.](https://irma.nps.gov/Stats/SSRSReports/National%20Reports/Query%20Builder%20for%20Public%20Use%20Statistics%20(1979%20-%20Last%20Calendar%20Year)) {https://irma.nps.gov/Stats/SSRSReports/National%20Reports/Query%20Builder%20for%20Public%20Use%20Statistics%20(1979%20-%20Last%20Calendar%20Year) }

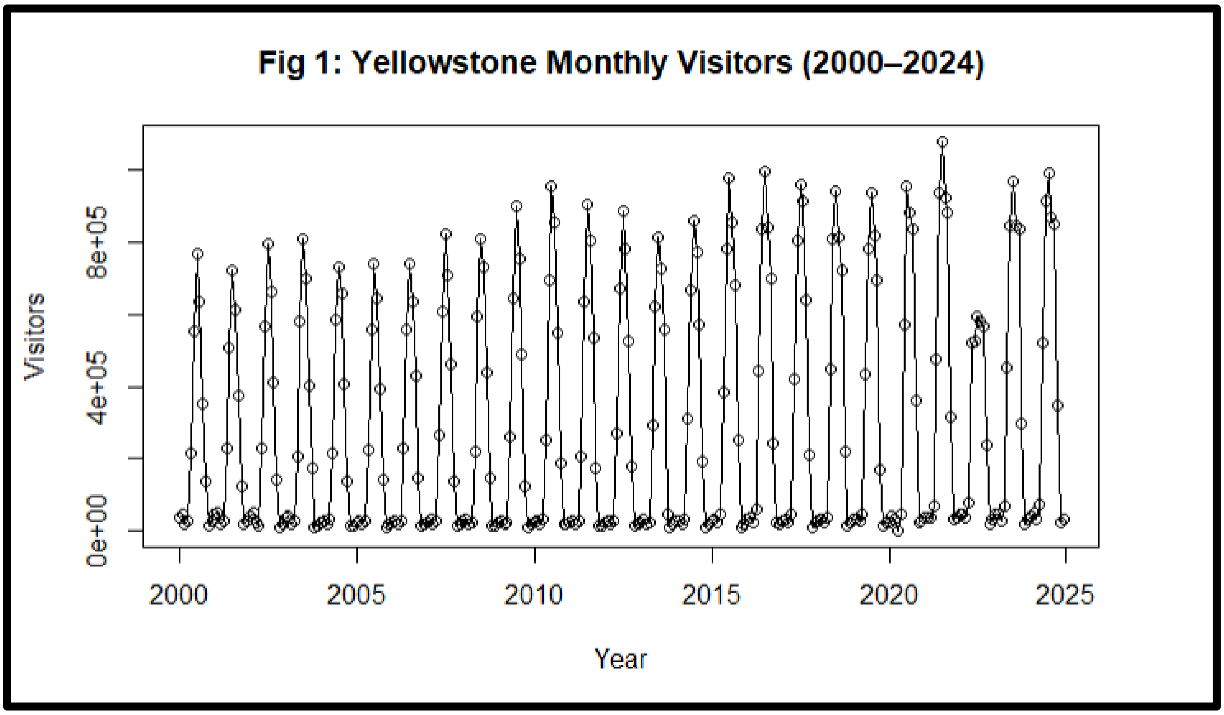


After clicking the above link, select years 2000 to 2024 from "Select Year(s)", "Intermountain Region" from "Select Region(s)", "Yellowstone NP" from "Select Park(s)" and click "View Report".

We considered 3 columns from the data namely Year, Month and Recreation Visits. The attribute "RecreationVisits" was converted to numeric type for our analysis.

2. Descriptive Statistics

The smallest value in our dataset is 0, indicating that at least one observation has no value. 25% of the data values are below 25,573, the Median is 70,516 and the Mean is 295,567. The mean is substantially higher than the median, indicating a right-skewed distribution. The range (max - min = 1,080, 767 - 0 = 1,080,767) is very large, showing significant variability in the data. 75% of the months had fewer than ~582.9k visitors, further highlighting the **wide variability** in monthly counts. At its peak, Yellowstone received over 1 million visitors in a single month, likely during peak summer season.



- **Trend:** Overall upward trajectory in annual peaks from ~700 K (2000) to ~1 M+ (2024). Slight plateau around 2006-2009, then renewed growth

- **Change in Variance:** Early years show moderate month-to-month swings (±100 K). Post-2010 variance widens (peaks ±200 K), indicating more pronounced seasonality

- **Behaviour:** High-visitation months tend to be followed by high months next year. Since we could see seasonality, it is hard to comment on AR/MA.

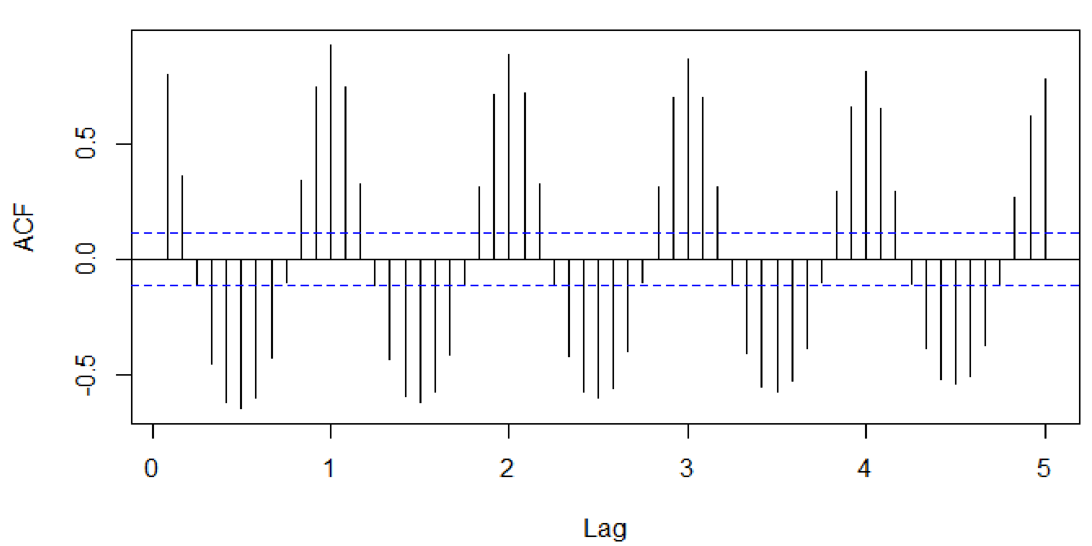
- **Seasonality:** Consistent annual cycle with peaks in July-August, low in January-February. Seasonal amplitude increasing over time

- **Change Points:** Sharp downturn in early 2020 (COVID-19 impact). Rapid recovery beginning mid-2021, marking a new normal of higher visitation

Fig 2: Lag 1 Scatter Plot

There is a general upward trend where months with higher visitor counts tend to follow months that also had high visitor counts. However, the relationship is not strictly linear, and there's considerable dispersion, especially in the mid-range. Some points are densely packed near the bottom left while others form a loose diagonal pattern toward the top right.

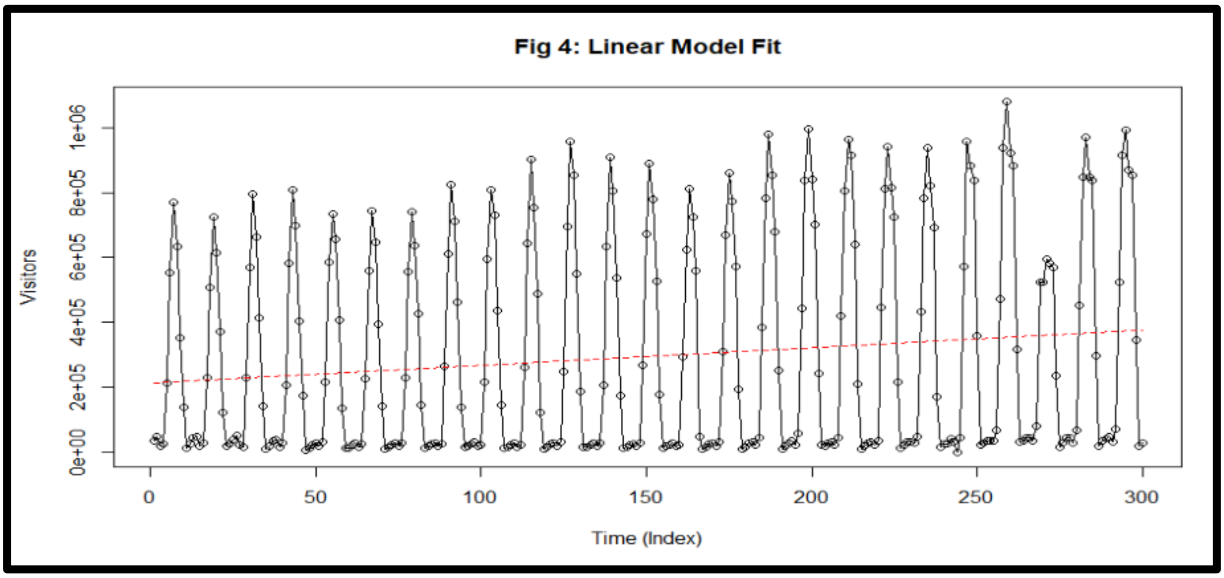
Fig 3: ACF plot



The ACF plot shows a prominent spike at lag 12, which indicates a repeating correlation structure every 12 units. Since our data is based on months, each lag represents one month. A strong autocorrelation at lag 12 suggests that the series has a recurring pattern every 12 months.

3. Trend Models

3.1 Linear Model




```
Call:
lm(formula = yellowstone_ts ~ t)

Residuals:
    Min      1Q  Median      3Q     Max 
-355159 -270139 -174136  323008  726219 

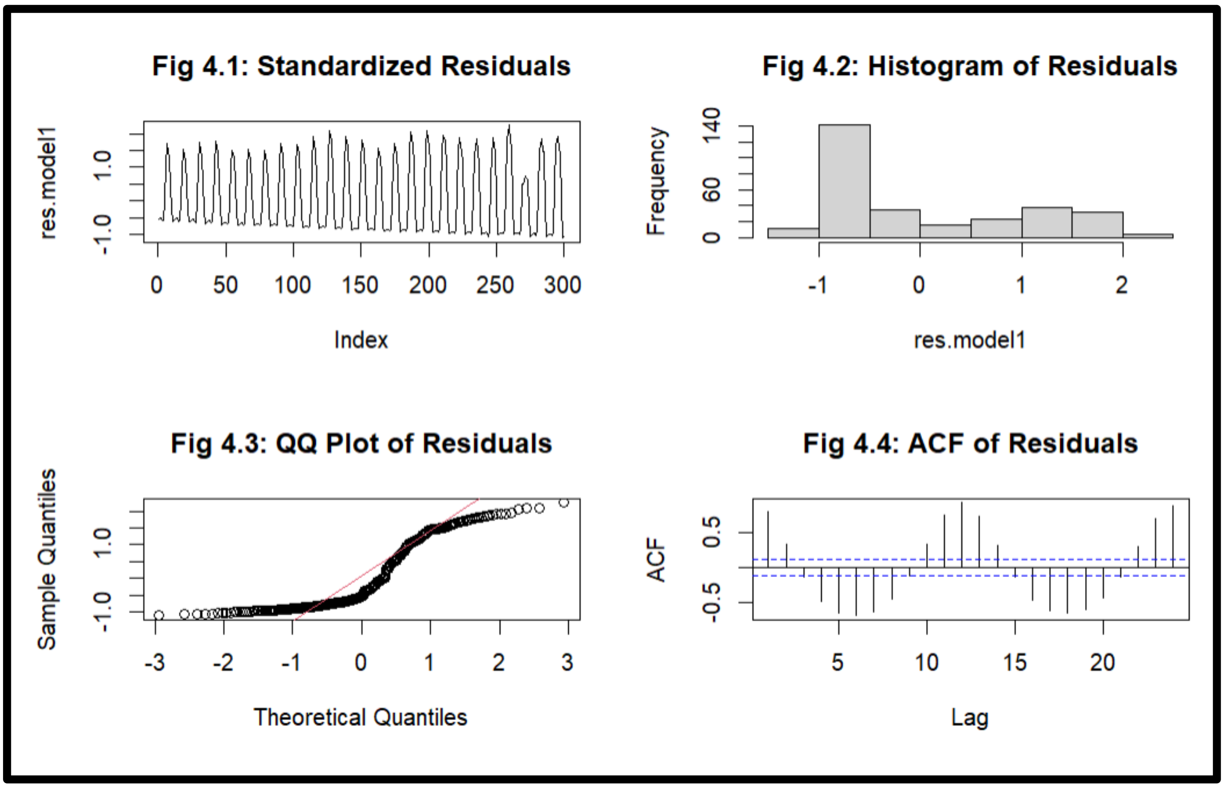
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 213755.2    37691.9   5.671 3.35e-08 ***
t              543.6      217.1   2.504   0.0128 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 325600 on 298 degrees of freedom
Multiple R-squared:  0.02061,	Adjusted R-squared:  0.01732 
F-statistic: 6.271 on 1 and 298 DF,  p-value: 0.01281
```

Table 1 - z test of coefficients

To identify and remove the underlying trend in the Yellowstone visitor time series, we first fitted a simple linear regression model. The linear regression model fitted to the Yellowstone monthly visitor data indicates a statistically significant upward trend over time, with a slope of approximately 543.6. This suggests that, on average, visitor counts increase by about 544 per month.

The intercept is estimated at 213,755.2, representing the expected visitor count at the start. p-value of 0.0128 for the slope and a highly significant F-statistic (p = 0.01281), an R-squared value of just 0.0206 meaning the model accounts for only about 2.06% of the total variation in visitor count. The linear fit, shown in Figure 4, visually confirms that the fitted line poorly tracks the sharp rises and falls in the data. As a whole, we can say that the linear model does not adequately explain the structure of the series.



In residual analysis, the standardized residual plot (Fig 4.1) reveals a clear repetitive structure, suggesting non-randomness. In the Histogram graph (Fig 4.2), the x-axis values range from -2 to 3, indicating there is no symmetric value. From the QQ-plot (Fig 4.3), both the start and end tails are way off. In ACF plot (Fig 4.4), all the lags are above confidence interval line meaning they are significantly autocorrelated which means the model is not able to capture the data.

Hypothesis Test for Shapiro-Wilk Test

H_0: Residuals are normally distributed

H_A: Residuals are not normally distributed

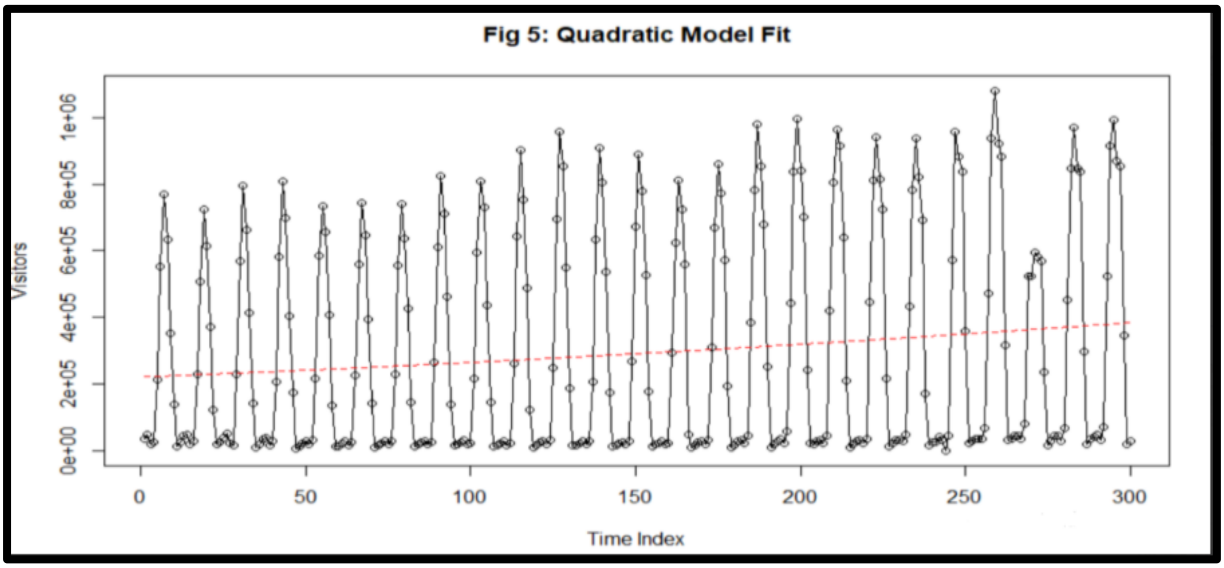
Shapiro-Wilk normality test

data: res.model1

W = 0.8404, p-value < 2.2e-16

The Shapiro-Wilk test result supports this conclusion, with W = 0.8404 and a p-value < 2.2e-16, which strongly rejects the null hypothesis of normality.

3.2 Quadratic Model




```
Call:
lm(formula = yellowstone_ts ~ t + t2)

Residuals:
    Min      1Q  Median      3Q     Max 
-362804 -266592 -179162  326587  723975 

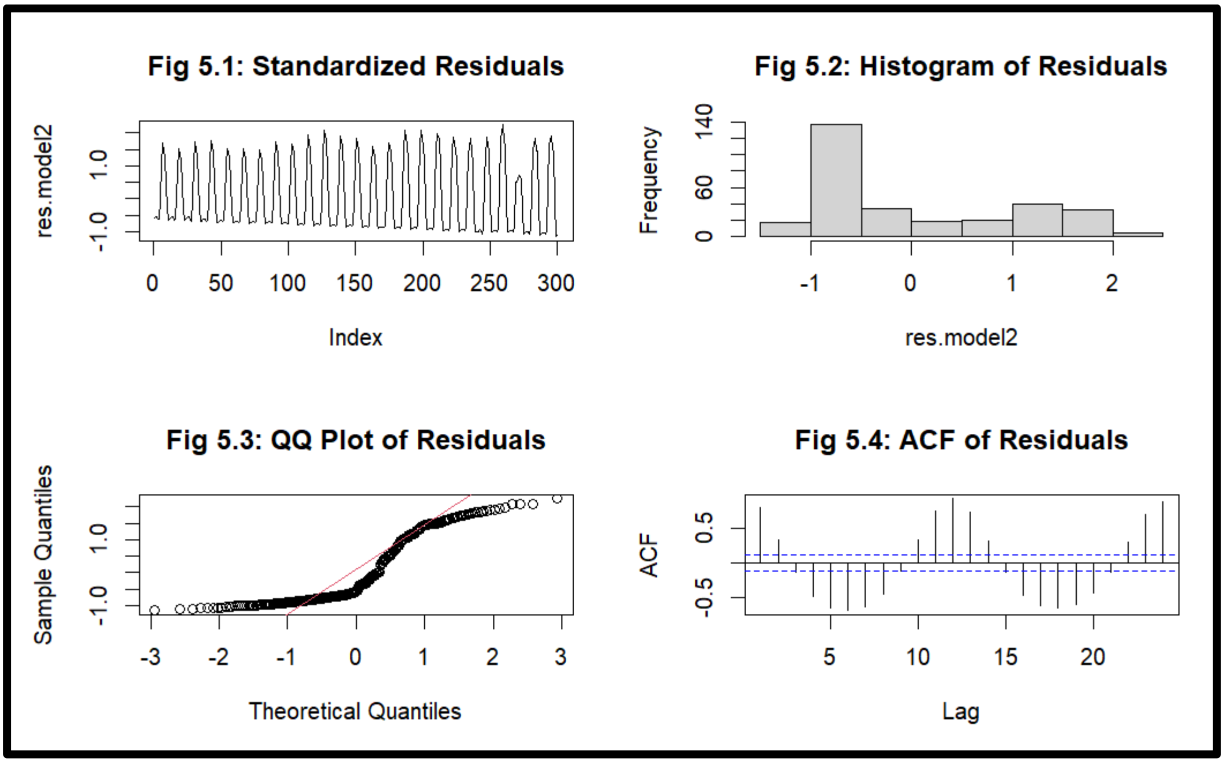
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.217e+05  5.687e+04   3.899  0.00012 ***
t           3.855e+02  8.724e+02   0.442  0.65893    
t2          5.253e-01  2.807e+00   0.187  0.85167    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 326100 on 297 degrees of freedom
Multiple R-squared:  0.02073,   Adjusted R-squared:  0.01413 
F-statistic: 3.143 on 2 and 297 DF,  p-value: 0.04459
```

Table 2 - z test of coefficients

From Fig 5, the fitted line exhibits a slight upward curvature but does not closely follow the observed data points. This indicates that the quadratic model does not effectively capture the patterns in the data, despite being statistically significant as a whole.

The quadratic model's Multiple R-squared value is 0.02073, and the Adjusted R-squared is 0.01413, meaning the model explains only around 2% of the total variance in the data. The residual standard error is 326,100, which indicates a substantial average deviation between the observed and predicted values. The F-statistic of 3.143 with a p-value of 0.04459 suggests that the overall model is statistically significant at the 5% level, although the individual predictors are not.



In residual analysis, the standardized residual plot (Fig 5.1) reveals a clear repetitive structure, suggesting non-randomness. In the Histogram graph (Fig 5.2), the x-axis values range from -2 to 3, indicating there is no symmetric value. From the QQ-plot (Fig 5.3), both the start and end tails are way off. In ACF plot (Fig 5.4), all the lags are above confidence interval line meaning they are significantly autocorrelated which means the model is not able to capture the data.

Hypothesis Test for Shapiro-Wilk Test

H_0: Residuals are normally distributed

H_A: Residuals are not normally distributed

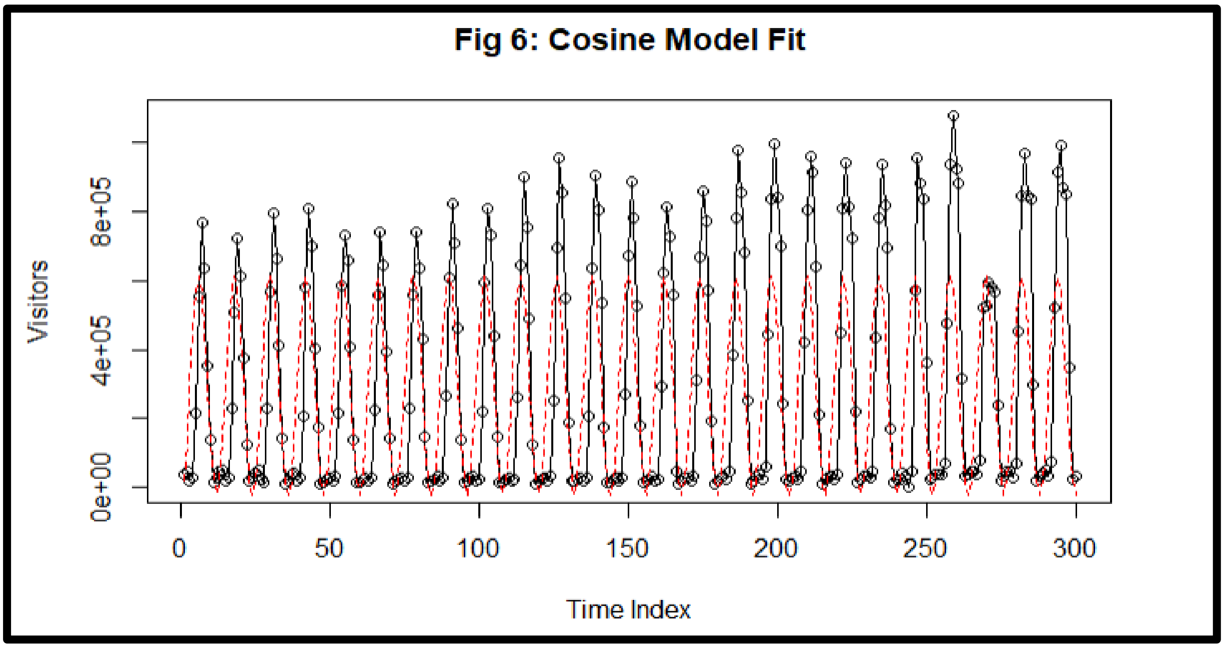
Shapiro-Wilk normality test

data: res.model2

W = 0.8417, p-value < 2.2e-16

The Shapiro-Wilk test result supports this conclusion, with W = 0.8417 and a p-value < 2.2e-16, which strongly rejects the null hypothesis of normality.

3.3 Cosine Model




```
Call:
lm(formula = yellowstone_ts ~ cos_term)

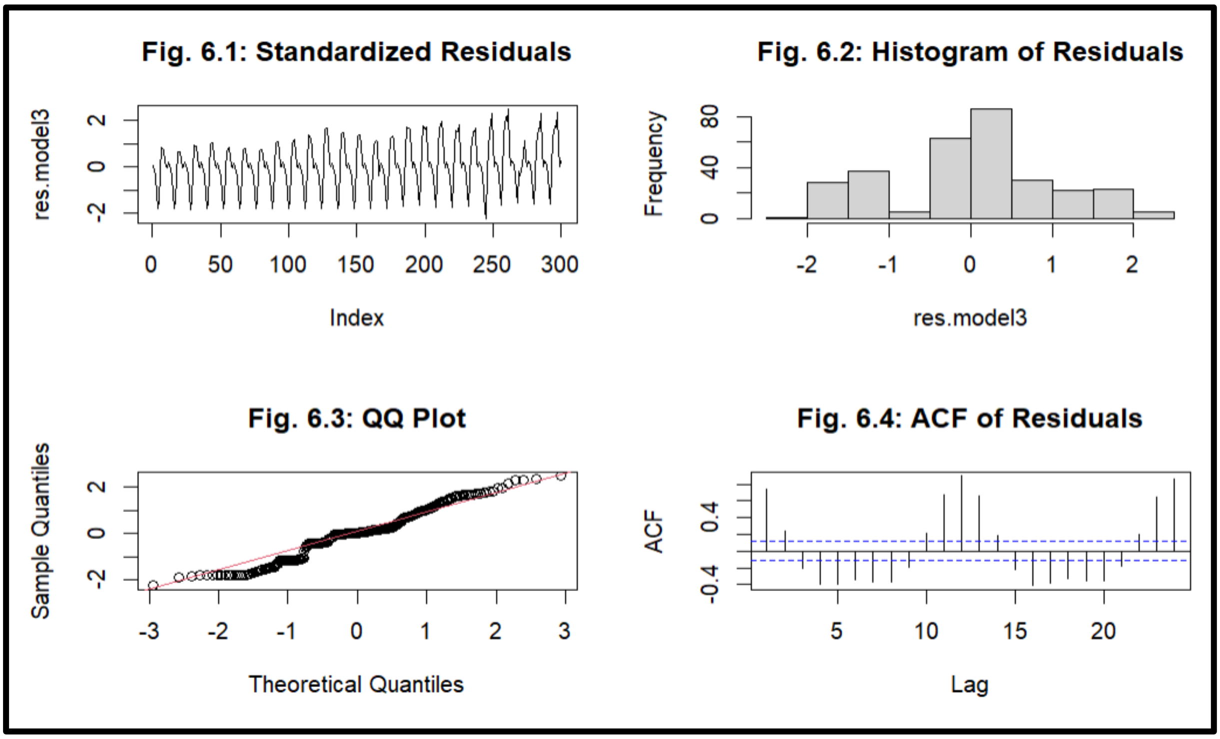
Residuals:
    Min      1Q  Median      3Q     Max 
-526053 -107487    6673  159084  586511 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   295567      13768   21.47   <2e-16 ***
cos_term     -319512      19470  -16.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 238500 on 298 degrees of freedom
Multiple R-squared:  0.4747,	Adjusted R-squared:  0.4729 
F-statistic: 269.3 on 1 and 298 DF,  p-value: < 2.2e-16
```

Table 3 - z test of coefficients

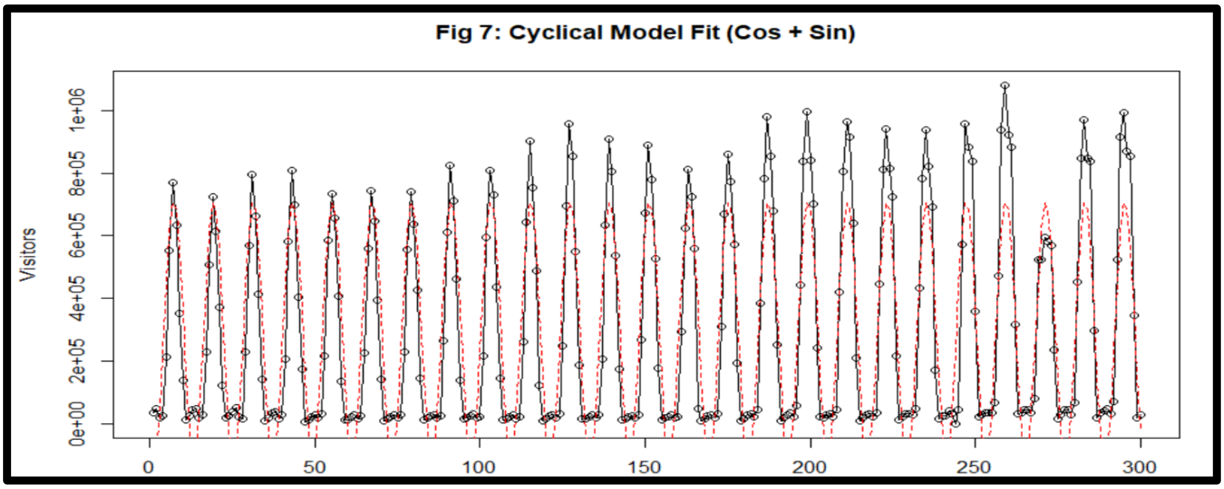
In the corresponding plot titled "Fig 6: Cosine Model Fit", the red dashed line representing the fitted cosine function shows a regular wave-like pattern over time. However, while it follows a repetitive structure, it does not closely align with the magnitude or peaks of the observed values (black circles), indicating that although the cosine model captures periodicity, it does not fully account for the amplitude or complexity of the real fluctuations in visitor numbers. In the residual analysis of cosine model, the intercept is estimated at 295,567 and the coefficient for the cosine term is −319,512, both of which are statistically significant with p-values < 2e-16. The Multiple R-squared value is 0.4747 and the Adjusted R-squared is 0.4729, indicating that the model explains approximately 47.3% of the variation in the Yellowstone visitor data. The F-statistic of 269.3 with p < 2.2e-16 confirms that the model is highly statistically significant. The residual standard error is 238,500 on 298 degrees of freedom, which is lower than that of the linear and quadratic models.



The residual diagnostics for the cosine model indicate some improvement in capturing the structure of the data, though model assumptions are still not fully met. The standardized residual plot (Fig 6.1) shows a repeating cyclical pattern, suggesting that periodic fluctuations remain in the residuals. The histogram of residuals (Fig 6.2) appears approximately centered around zero but is not symmetric, indicating mild skewness. In the Q-Q plot (Fig 6.3), the residuals deviate from the red line, particularly in the lower and upper tails, showing signs of non-normality. This is supported by the Shapiro-Wilk test result, which gives

W = 0.97041 with a p-value of 7.732e-06, leading to the rejection of the null hypothesis of normality. The ACF plot (Fig 6.4) reveals that several autocorrelations are still above the significance threshold, meaning that residuals are not entirely independent. Overall, the cosine model fits better than other two previous models.

3.4 Cyclical Model




```
Call:
lm(formula = yellowstone_ts ~ cos_term + sin_term)

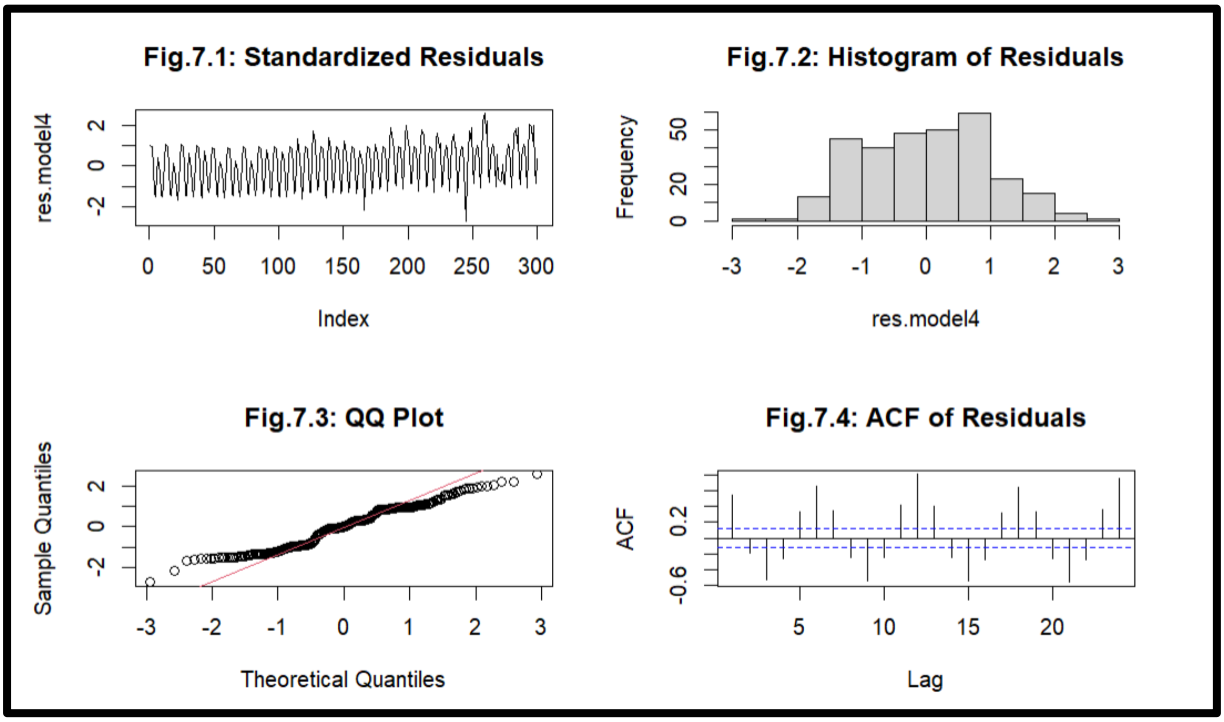
Residuals:
    Min      1Q  Median      3Q     Max 
-394084 -137597    2599  127049  376525 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   295567       8538   34.62   <2e-16 ***
cos_term     -319512      12075  -26.46   <2e-16 ***
sin_term     -263938      12075  -21.86   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 147900 on 297 degrees of freedom
Multiple R-squared:  0.7986,	Adjusted R-squared:  0.7973 
F-statistic: 588.9 on 2 and 297 DF,  p-value: < 2.2e-16
```

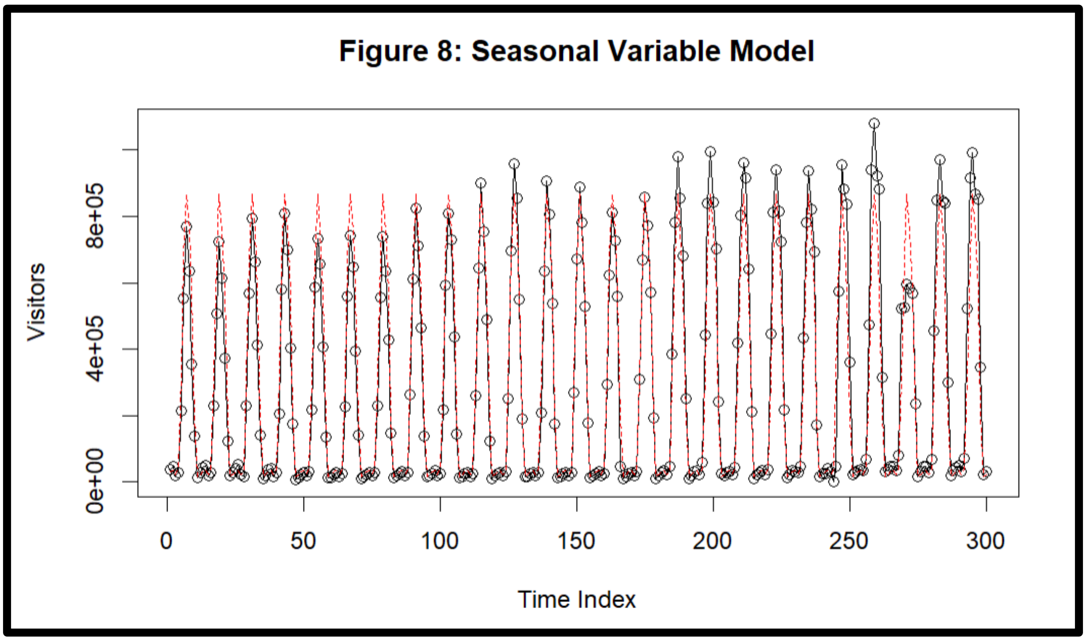
Table 4 - z test of coefficients

Fig 7 represents the overall fit of the cyclical model. The red line traces a smooth wave that rises and falls in a regular twelve-month cycle. This wave closely follows the timing and general amplitude of the observed peaks and troughs. The cyclical model's R-squared is 0.7986 (adjusted $R^2 = 0.7973$), indicating it explains about 79.9% of the variance in visitor counts, and the F-statistic is 588.9 (df = 2, 297, $p < 2.2 \times 10^{-16}$), confirming overall significance.



The residual diagnostics for the cyclical model (cosine + sine) show that, while much of the periodic structure has been captured, the residuals still violate key assumptions. In Fig 7.1, the standardized residuals hover randomly around zero but retain a faint cyclical pattern, indicating some unmodeled seasonality remains. The histogram in Fig 7.2 is roughly bell-shaped yet slightly skewed, suggesting departures from symmetry. In the Q-Q plot (Fig 7.3), most points lie near the reference line but extreme values in the tails deviate, evidencing non-normality. This is confirmed by the Shapiro-Wilk test ($W = 0.96877$, $p = 4.319 \times 10^{-6}$), which rejects normality. Finally, the ACF of the residuals (Fig 7.4) displays a few significant autocorrelations at lags around 6 and 12, showing that residuals are not fully independent. Together, these diagnostics indicate that although the cyclical model captures the main annual cycle, it does not produce residuals that are both normally distributed and uncorrelated.

3.5 Seasonal Model




```
Call:
lm(formula = yellowstone_ts ~ month_factors)

Residuals:
    Min      1Q  Median      3Q     Max
-270567  -21838   -3683   12529  309028

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)        31121      17091   1.821   0.0697 .
month_factors2      5008      24171   0.207   0.8360
month_factors3     -7990      24171  -0.331   0.7412
month_factors4      6386      24171   0.264   0.7918
month_factors5    279999      24171  11.584  < 2e-16 ***
month_factors6    643684      24171  26.631  < 2e-16 ***
month_factors7    836009      24171  34.587  < 2e-16 ***
month_factors8    730105      24171  30.206  < 2e-16 ***
month_factors9    541929      24171  22.421  < 2e-16 ***
month_factors10   162975      24171   6.743  8.5e-11 ***
month_factors11   -15714      24171  -0.650   0.5161
month_factors12    -9032      24171  -0.374   0.7089
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

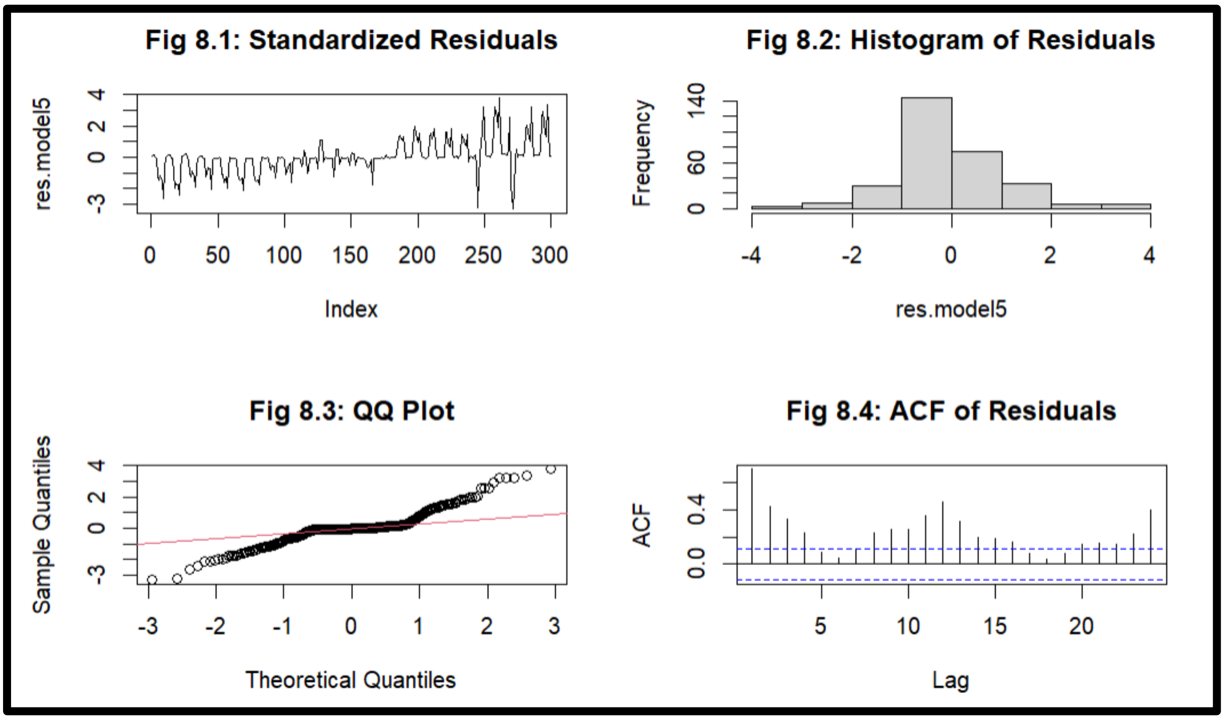
Residual standard error: 85460 on 288 degrees of freedom
Multiple R-squared:  0.9348,    Adjusted R-squared:  0.9323
F-statistic: 375.4 on 11 and 288 DF,  p-value: < 2.2e-16
```

Table 5 - z test of coefficients

In the fit plot (Figure 8), the red dashed line connects the model’s fitted values for each month and virtually overlays the observed data points, accurately tracing both the timing and magnitude of every monthly peak and valley throughout the series.

The seasonal dummy regression model's monthly visitor counts have a separate coefficient for each month of the year (with January as the baseline). From the output, the intercept (January level) is 31,121 (SE = 17,091, $p = 0.0697$). Among the monthly effects, May through October have highly significant positive coefficients: May = 279,999, June = 643,684, July = 836,009, August = 730,105, September = 541,929, and October = 162,975 (all $p < 2 \times 10^{-16}$ except October's $p = 8.5 \times 10^{-11}$).

The coefficients for February, March, April, November, and December are small and not significant ($p > 0.05$). The model's Multiple R-squared is 0.9348 (adjusted $R^2 = 0.9323$), indicating it explains about 93.5% of the variation in monthly visitors. The residual standard error is 85,460 on 288 degrees of freedom, and the F-statistic is 375.4 (df = 11, 288, $p < 2 \times 10^{-16}$), confirming overall significance.



In Fig 8.1, the standardized residuals fluctuate around zero but exhibits periods of increased volatility particularly in the later observations rather than remaining randomly scattered. The histogram (Fig 8.2) is roughly centered but displays a flattened top and longer tails, indicating a distribution that is not perfectly bell-shaped. The Q-Q plot (Fig 8.3) further shows that residuals diverge from the theoretical line at both extremes, especially in the upper tail, signalling heavy-tailed behaviour. The ACF of the residuals (Fig 8.4) shows several significant spikes at lags corresponding to the seasonal period, implying residual autocorrelation. Finally, the Shapiro-Wilk test yields $W = 0.90417$ with $p = 7.184 \times 10^{-13}$, strongly rejecting the null hypothesis of normality.

3.6 Forecast based on Best Trend Model

3.6.1 Seasonal vs Cyclical Model:

Seasonal model aligns precisely with sharp spikes whereas cyclical model often underestimates or misses these spikes. It is more like underfitting the pattern and Cyclical model is not able to capture abrupt seasonal spikes or drops. The seasonal model adapts to changing peak heights across the time index. The cyclical model assumes a fixed amplitude and frequency, which makes it less adaptive to real-world seasonal variation.

From the summary of Cyclical model:

Residual standard error is 147900 on 297 degrees of freedom Multiple R-squared is 0.7986, Adjusted R-squared is 0.7973 F-statistic is 588.9 on 2 and 297 DF, p-value is < 2.2e-16

Shapiro-Wilk normality test

data: res.model4

W = 0.96877, p-value = 4.319e-06

From the summary of Seasonal Model:

Residual standard error is 85460 on 288 degrees of freedom Multiple R-squared is 0.9348, Adjusted R-squared is 0.9323 F-statistic is 375.4 on 11 and 288 DF, p-value is < 2.2e-16

Shapiro-Wilk normality test

data: res.model5
W = 0.90417, p-value = 7.184e-13

The seasonal model explains 93.2% of the variability in the data, compared to just 79.7% for the cyclical model. The seasonal model produces more accurate predictions, with residuals (errors) being significantly smaller nearly 42% lower than the cyclical model. Based on a significantly higher R-squared, lower residual error, and better fit to the observed seasonal structure, the Seasonal Dummy Variable Model is clearly superior to the cyclical model even though the residuals are slightly less normal. It captures the underlying trend and fluctuations much more effectively for forecasting purposes.

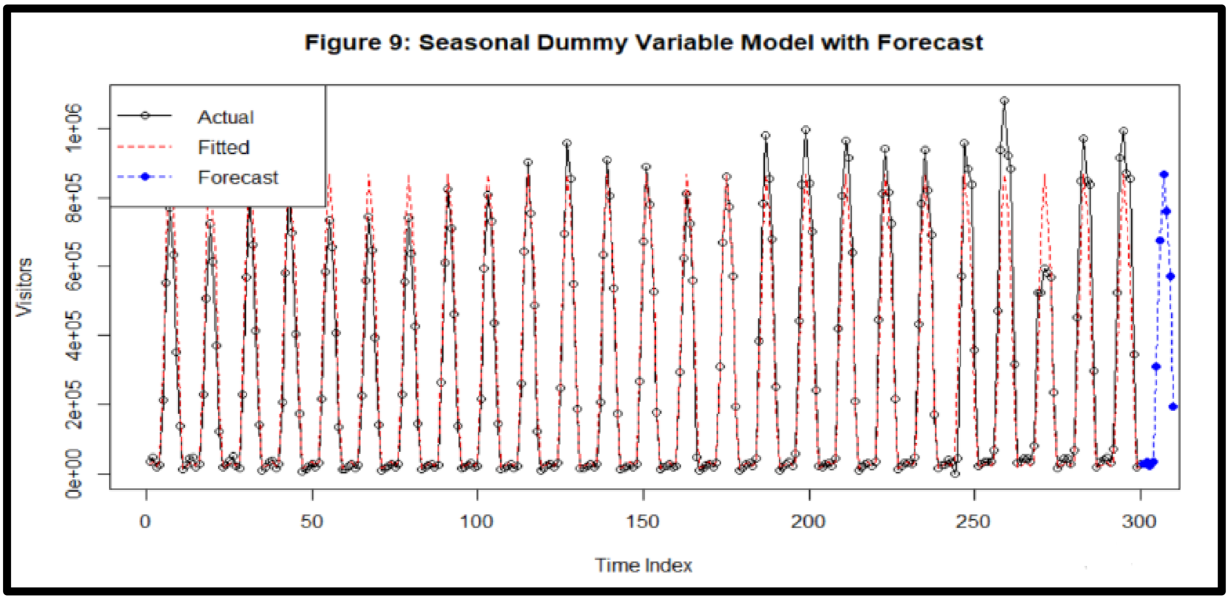
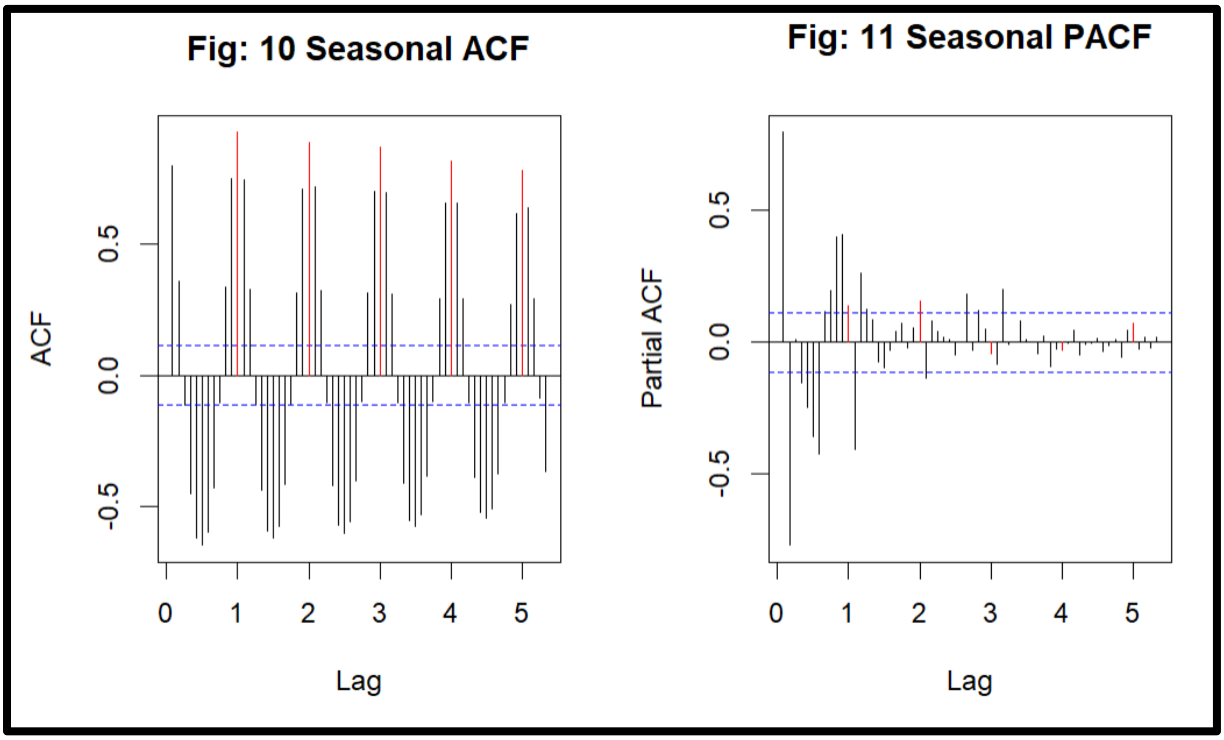


Fig 9.1: Forecast Table for Seasonal Model

	Time_Index <int>	Month <dbl>	Forecast_Visitors <dbl>
1	301	1	31121
2	302	2	36128
3	303	3	23130
4	304	4	37507
5	305	5	311120
6	306	6	674805
7	307	7	867129
8	308	8	761226
9	309	9	573050
10	310	10	194096

Fig 9, shows the overall forecast fit of Seasonal Model and the 10-month forecasted values of the Seasonal model are shown in Fig 9.1. The size and timing of the summer peak are consistent with the long run mean for those months. The model predicts an extremely quiet winter and a sharp jump beginning in May, culminating in a July peak of ~ 0.87 million visits. Overall, the model has appreciably predicted the forecast of the Yellowstone National Park.

4. SARIMA Model



The sample-ACF plot(Fig 10) displays very large positive spikes at lags 12, 24, 36, 48, 60 (highlighted in red) and equally large negative spikes at lags half-way between those multiples. The pattern repeats every 12 lags and decays only slightly, which confirms the presence of strong annual cycle in our monthly data. The sample-PACF plot(Fig 11) shows a single prominent spike at lag 1 and a smaller, but a significant spike at lag 12 and after that, the partial autocorrelations drop quickly. Together, the ACF and PACF confirms that the series has a period of 12 months.

Augmented Dickey-Fuller Test

data: yellowstone_ts

Dickey-Fuller = -23.638, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary

Phillips-Perron Unit Root Test

data: yellowstone_ts

Dickey-Fuller Z(alpha) = -86.394, Truncation lag parameter = 5, p-value = 0.01
alternative hypothesis: stationary

KPSS Test for Level Stationarity

data: yellowstone_ts

KPSS Level = 0.28985, Truncation lag parameter = 5, p-value = 0.1

Shapiro-Wilk normality test

data: yellowstone_ts

W = 0.79434, p-value < 2.2e-16

Stationarity tests are used to determine whether a time-series fluctuates around a constant mean and variance which is a prerequisite for most ARIMA-type models. The adf test (Dickey-Fuller = -23.64, $p < 0.01$) and the pp test (Z = -86.39, $p < 0.01$) both reject the null hypothesis of a unit root, while the kpss test (statistic = 0.29, $p > 0.10$) fails to reject the null stationarity. Thus, these three tests indicate that the series can be treated as stationary in level without additional non-seasonal differencing.

Normality, however, is not supported as the Shapiro-Wilk test yields W = 0.794 and $p < 2.2 \times 10\text{-}16$, strongly rejecting the hypothesis that the monthly visitor counts follow a Gaussian distribution. Thus, while the series is stationary for SARIMA purposes, its residual distribution is markedly non-normal, which may motivate variance-stabilising transformations or robust modelling techniques in later steps.

4.1 Box-Cox Test Results

Shapiro-Wilk normality test

data: ts_transformed

W = 0.84762, p-value < 2.2e-16

Augmented Dickey-Fuller Test

data: ts_transformed

Dickey-Fuller = -13.314, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary

Phillips-Perron Unit Root Test

data: diff_yellowstone

Dickey-Fuller Z(alpha) = -93.07, Truncation lag parameter = 5, p-value = 0.01
alternative hypothesis: stationary

KPSS Test for Level Stationarity

data: diff_yellowstone

KPSS Level = 0.0098883, Truncation lag parameter = 5, p-value = 0.1

Since we did not achieve normality and in order to stabilise variance, we went for Box-Cox transformation. We got the optimal lambda value 0.36 from Box-Cox. The unit-root tests still gave consistent evidence of stationarity with ADF (-13.31, $p < 0.01$) and PP (-92.40, $p < 0.01$) reject a unit root, while KPSS fails to reject level-stationarity. In terms of normality, the Shapiro-Wilk test on the transformed series yields $W = 0.848$ with $p < 2.2 \times 10^{-16}$, so the null hypothesis of normality is still strongly rejected. Box-Cox has shifted the distribution towards normality with W being improved from ~0.79 to ~0.85 but not enough to achieve normality.

4.2 Residual Diagnostics:

Fig: 12 Time series plot of the residuals

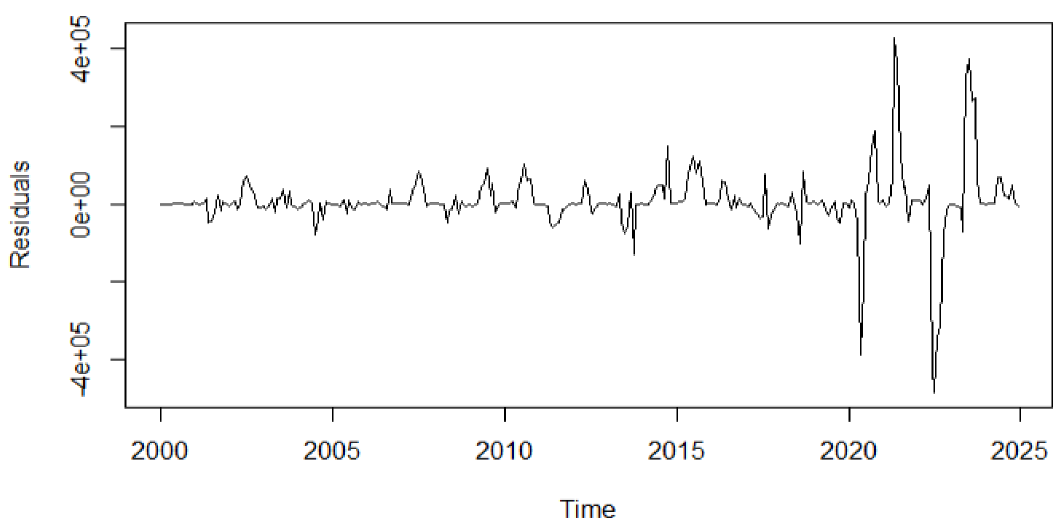


Fig: 13 ACF of m1 residuals

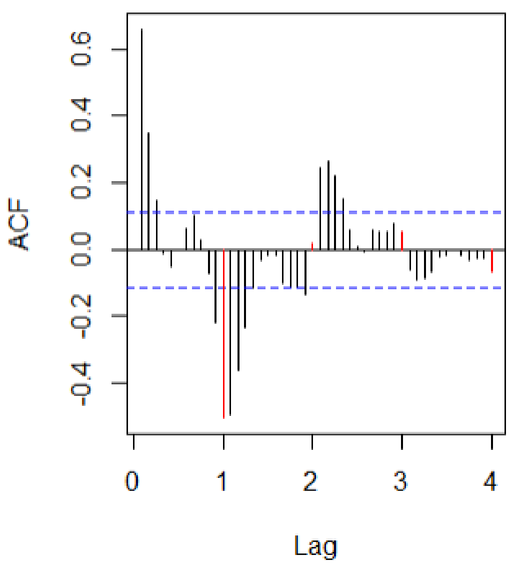
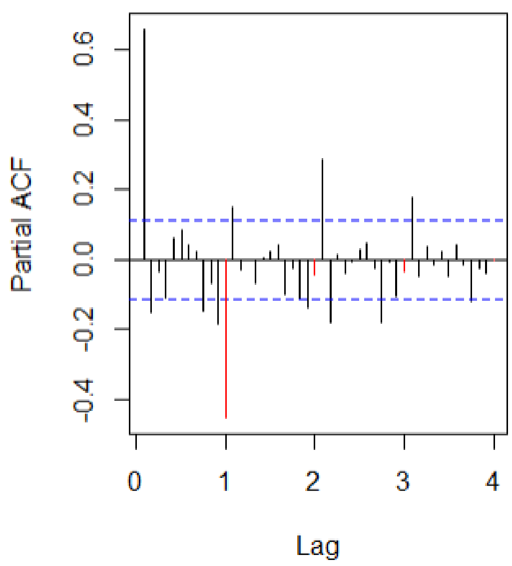


Fig: 14 PACF of m1 residuals



We fitted the plain model with only the first seasonal difference with order D = 1 to get rid of the seasonal trend effect by inspecting the autocorrelation structure of the residuals. By doing so, we could see 1 significant seasonal lag in both ACF and PACF (Fig 13 and 14).

Fig: 15 Time series plot of m2 residual

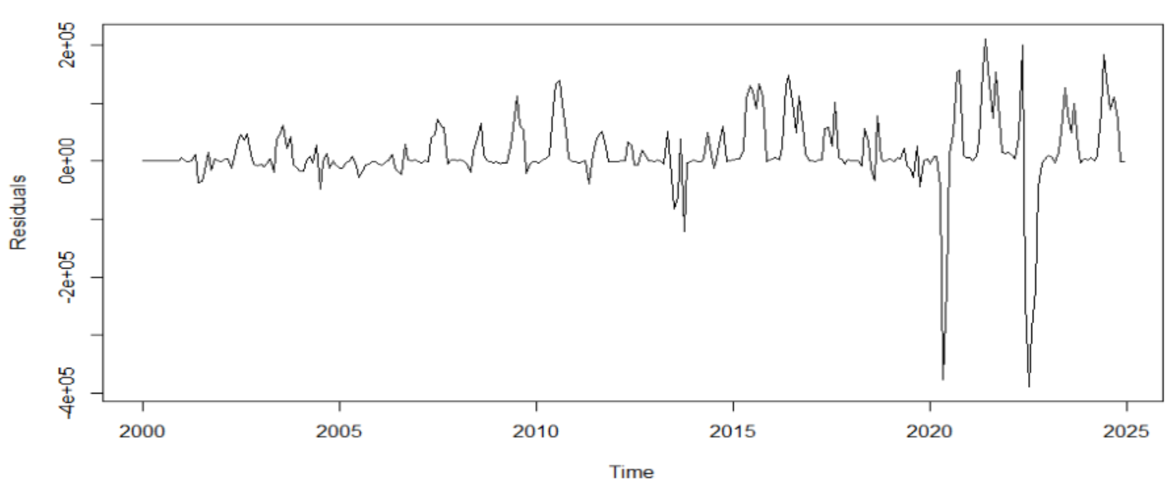


Fig: 16 ACF of m2 residuals

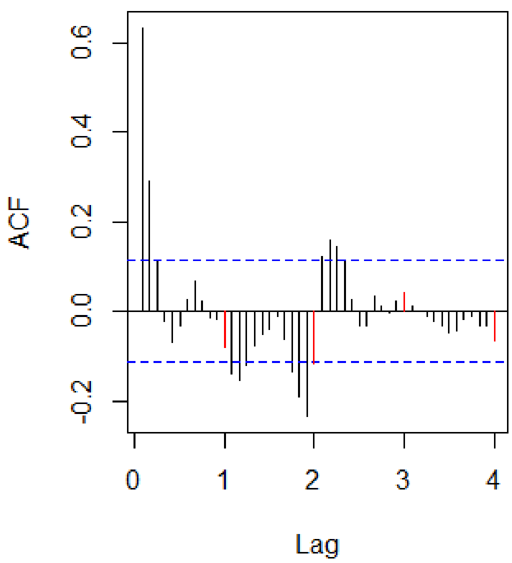
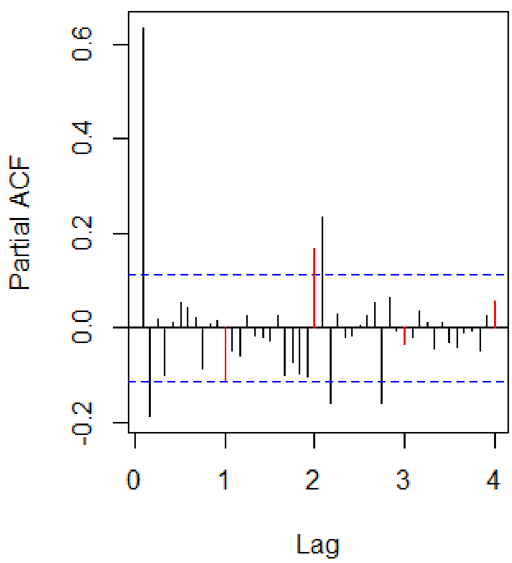


Fig: 17 PACF of m2 residuals



Since we got 1 significant seasonal lag in both ACF and PACF, we incremented P, Q (Seasonal orders) values from 0 to 1 with D already being 1. This results in the removal of seasonal lags from both the plots which is evident in Fig 16 and 17. Now that the seasonal lag from both the plots were taken care of, we moved on to correct the normal ARIMA orders. In m2's ACF and PACF plot (Fig 16 and 17), we could see that the 1st lag in both the plots was very high. Also, ACF plot shows a slowly decaying pattern.

Fig: 18 Time series plot of m3 residual

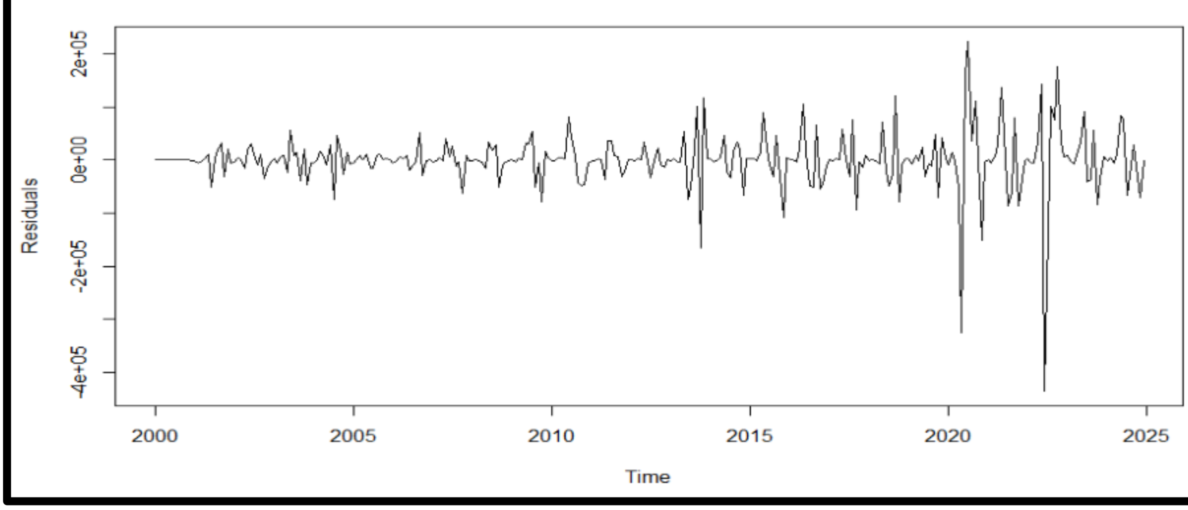


Fig: 19 ACF of m3 residuals

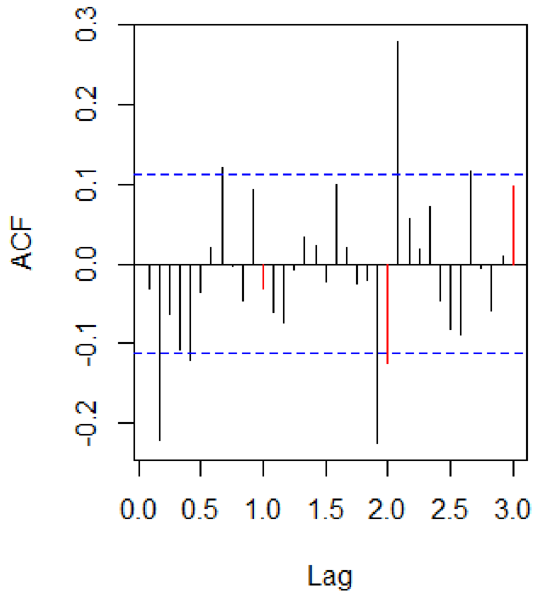
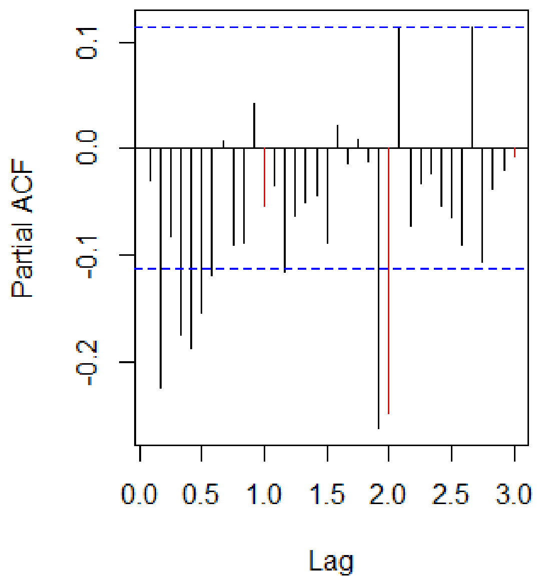
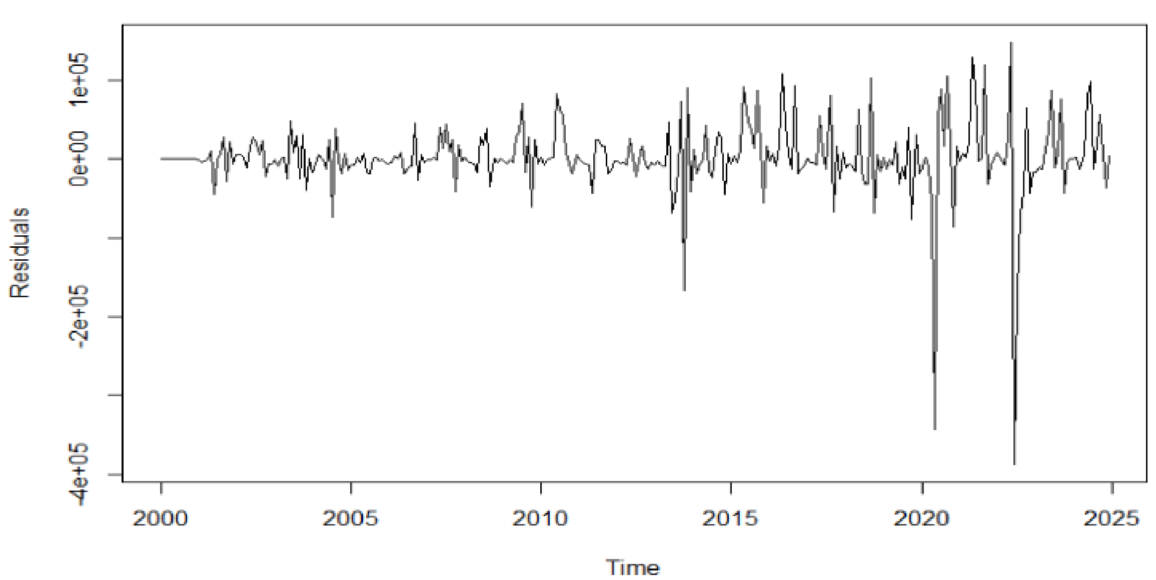


Fig: 20 PACF of m3 residuals



To rectify the above-mentioned issues, we performed 1st order differencing (d=1) and the corresponding results are shown in Figures 18,19 and 20. To clear out all the normal order lags from m3(Fig 19 and 20), we counted the significant lags which are way past the confidence interval line. In PACF, we have 5 significant lags and in ACF plot, we have 2 significant lags.

Fig: 21 Time series plot of m4 residual



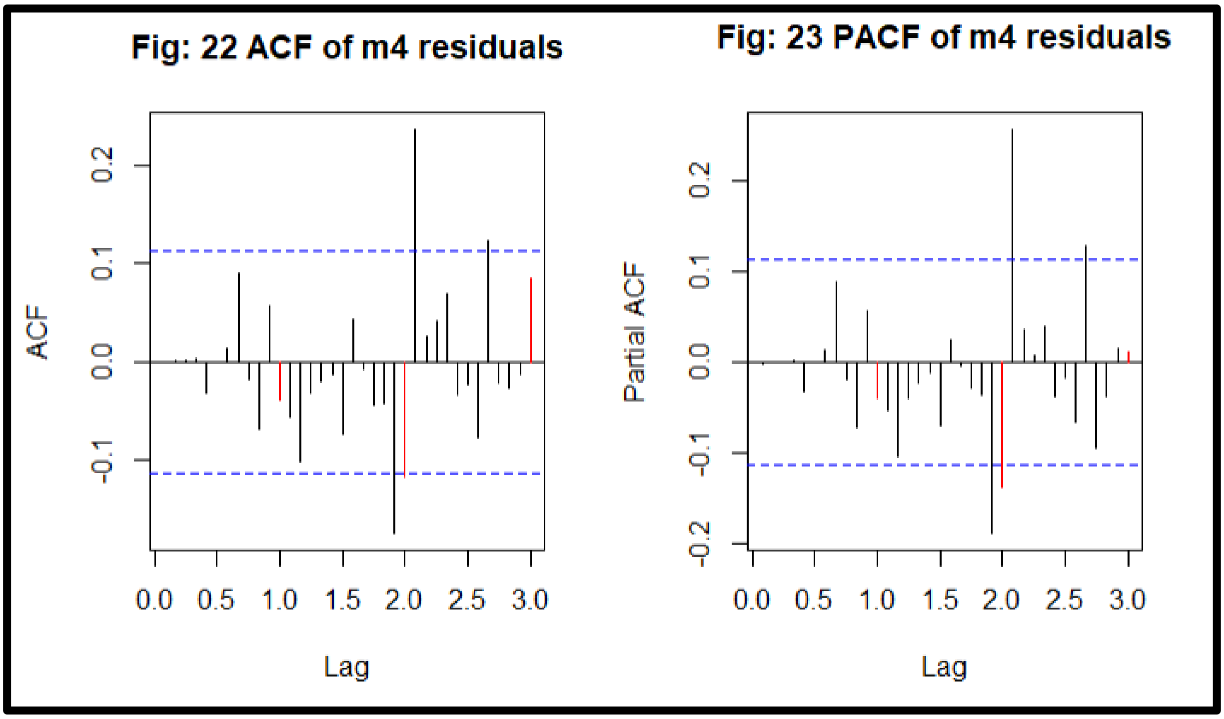


Fig: 22 ACF of m4 residuals

Fig: 23 PACF of m4 residuals

We fitted the last residual model with p=5 and q=2. By following this allocation, from Figures 22 and 23, we could clearly see that all the significant lags were removed.

The model that we got from this ACF and PACF plot is

SARIMA (5,1,2) x (1,1,1)$_{12}$

4.3 EACF

Fig: 24 - EACF plot

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

Reading the EACF from top-left corner, the first contiguous block in which every entry is a 0 down a row and across a column occurs at rows 1-2 and columns 2-3. Specifically, the cells for (p,q) = (1,2), (1,3), (2,2) and (2,3) form a 2 × 2 rectangle of uninterrupted 0's, whereas adjacent cells contain x's.

Hence, the chosen models from EACF are:

SARIMA (1,1,2) x (1,1,1)12
SARIMA (2,1,2) x (1,1,1)12
SARIMA (1,1,3) x (1,1,1)12
SARIMA (2,1,3) x (1,1,1)12

4.4 BIC Table

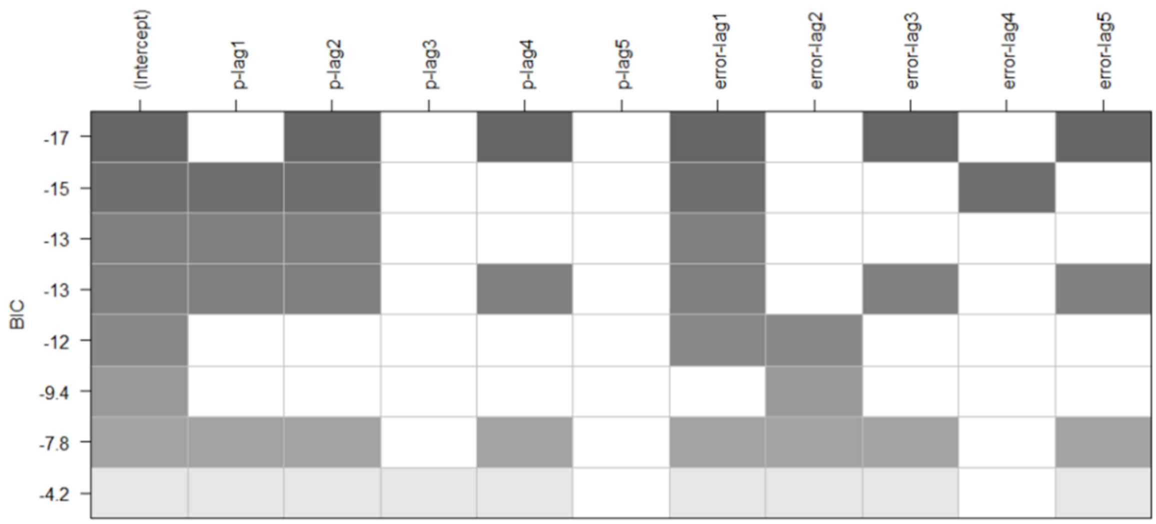


Fig:25 BIC Table

Each cell in the Fig 25 BIC table represents the BIC value of a model with a specific autoregressive (AR) and moving average (MA) order combination. Darker shades indicate lower BIC values, which represent better-fitting, more parsimonious models. The table reveals that models involving low-order combinations tend to perform best. By following this approach, the possible models are:

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

In total, the following models were considered:

SARIMA (5,1,2) x (1,1,1)$_{12}$
SARIMA (1,1,2) x (1,1,1)$_{12}$
SARIMA (2,1,2) x (1,1,1)$_{12}$
SARIMA (1,1,3) x (1,1,1)$_{12}$
SARIMA (2,1,3) x (1,1,1)$_{12}$
SARIMA (1,1,1) x (1,1,1)$_{12}$

SARIMA (2,1,1) x (1,1,1)12

4.5 Parameter Estimation

4.5.1 SARIMA (5,1,2) x (1,1,1)12 - ML

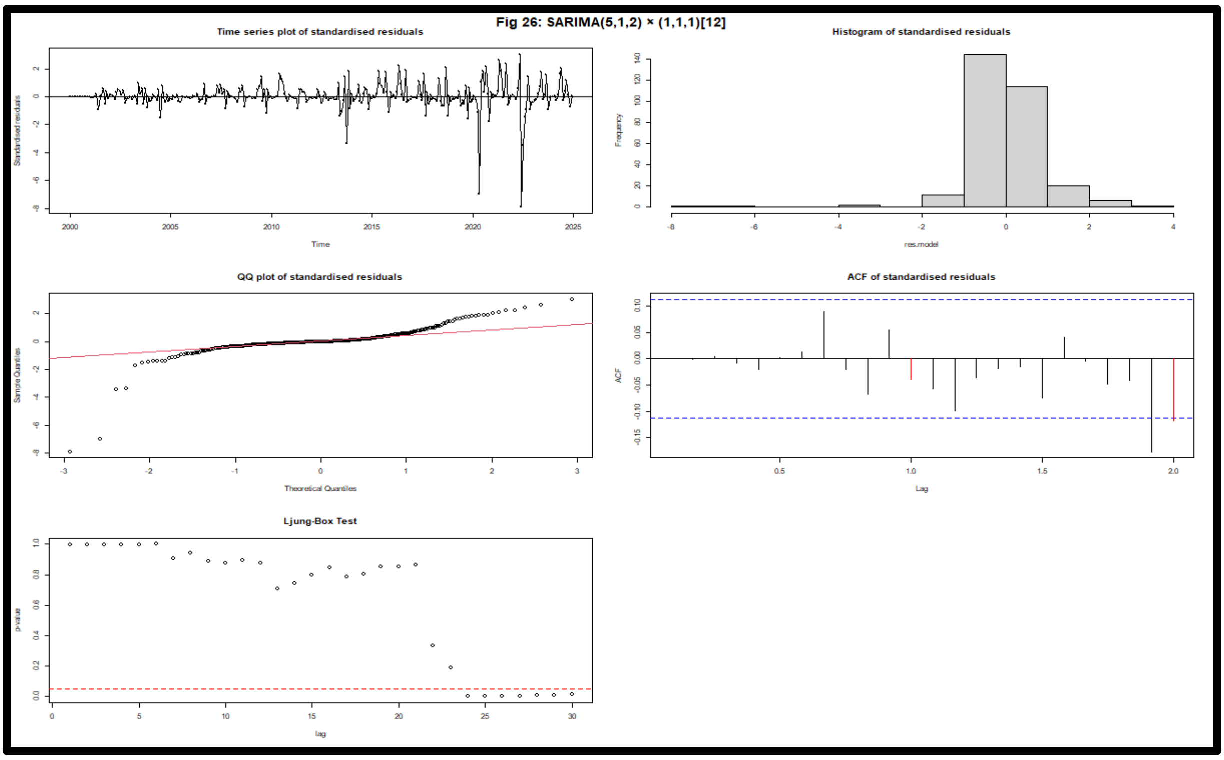
```
z test of coefficients:

      Estimate Std. Error  z value Pr(>|z|)    
ar1   0.155989   1.089031   0.1432   0.8861    
ar2   0.232699   0.826329   0.2816   0.7782    
ar3  -0.039700   0.253433  -0.1566   0.8755    
ar4  -0.029410   0.130575  -0.2252   0.8218    
ar5  -0.077559   0.135332  -0.5731   0.5666    
ma1  -0.399143   1.087838  -0.3669   0.7137    
ma2  -0.600813   1.087717  -0.5524   0.5807    
sar1 -0.081279   0.075215  -1.0806   0.2799    
sma1 -0.638482   0.053228 -11.9952   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        Shapiro-Wilk normality test

data:  res.model
W = 0.71506, p-value < 2.2e-16
```

Table 6.1 - z test of coefficients



The fitted model contains five non-seasonal AR terms (ar1-ar5), two non-seasonal MA terms (ma1-ma2), one seasonal AR term (sar1) and one seasonal MA term (sma1) after one regular and one seasonal difference. ar1-5 and ma1-2 all are not statistically significant, because their p-values are much higher than

0.05. This implies that the model might be over-parameterised. Seasonal MA(1) is highly significant.

Residuals fluctuate around zero with isolated large negative spikes (2020) but otherwise a moderate variance is there. Histogram shows a slightly left skewed distribution. In QQ plot, both the tails are way off from the red line indicating the presence of outliers. There are no autocorrelations in ACF plot, as all the lags are within the 95% confidence interval. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

4.5.2 SARIMA (5,1,2) x $(1,1,1)_{12}$ - CSS

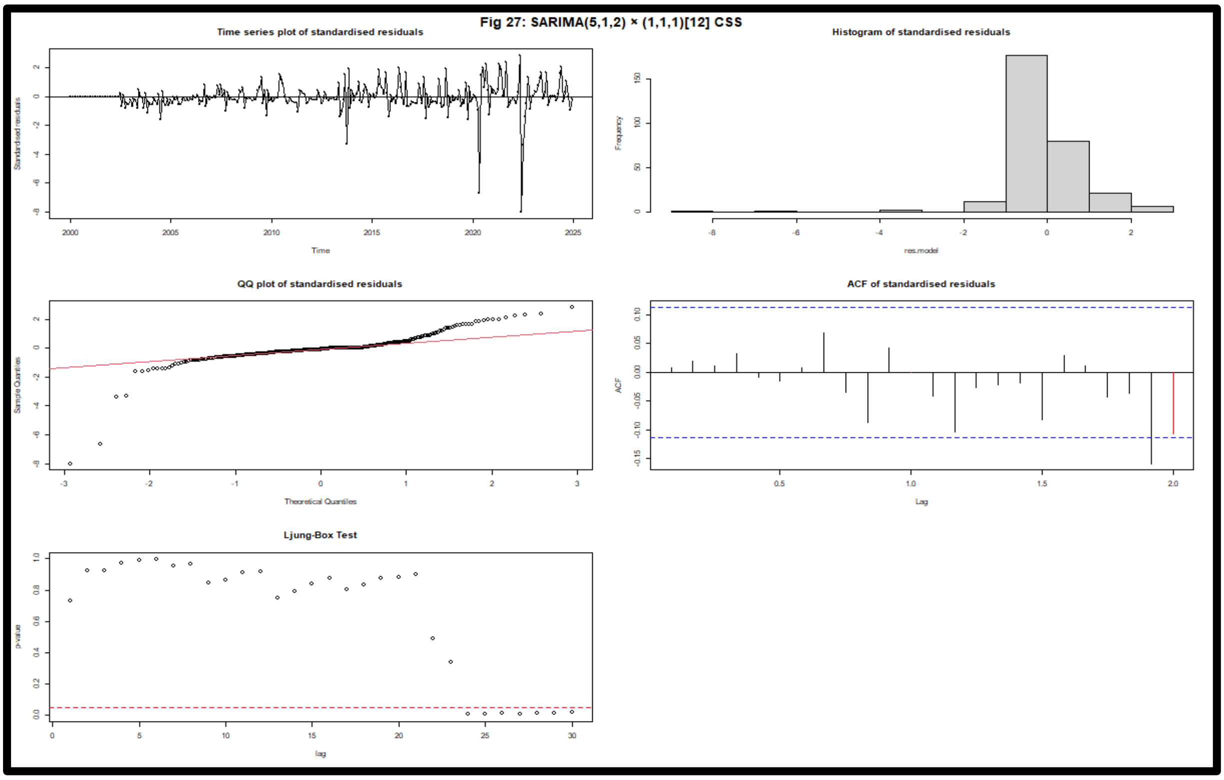
```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)    
ar1  -0.170003   0.118871  -1.4301 0.1526754    
ar2   0.372225   0.105917   3.5143 0.0004409 ***
ar3  -0.127545   0.068201  -1.8701 0.0614647 .  
ar4  -0.048258   0.063339  -0.7619 0.4461204    
ar5  -0.132068   0.061180  -2.1587 0.0308741 *  
ma1  -0.020467   0.111063  -0.1843 0.8537935    
ma2  -0.793573   0.115707  -6.8585 6.959e-12 ***
sar1 -0.146199   0.073290  -1.9948 0.0460635 *  
sma1 -0.631595   0.053547 -11.7951 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


        Shapiro-Wilk normality test

data:  res.model
W = 0.74003, p-value < 2.2e-16
```

Table 6.2 - z test of coefficients



ar2, ar5, ma2, sar1 and sma1 are all statistically significant, because their p-values are less than 0.05. Residuals fluctuate around zero with a few large negative spikes (2020-2021), but no visible trend or seasonality. Histogram shows a slightly left skewed distribution. In QQ plot, both the tails are way off from the red line indicating the presence of outliers. There are no autocorrelations in ACF plot, as all the lags are within the 95% confidence interval. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

4.5.3 SARIMA (5,1,2) x $(1,1,1)_{12}$ - CSSML

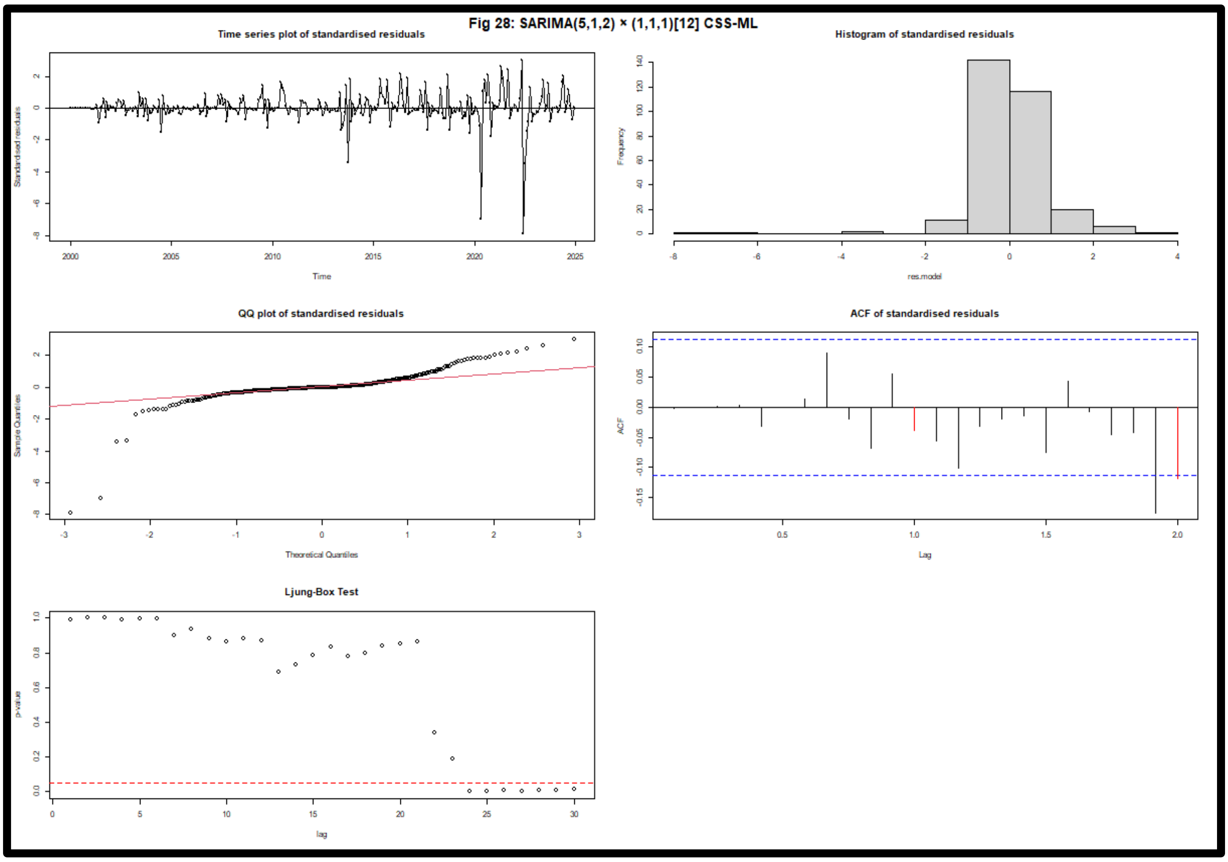
```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)    
ar1  -0.128709        NaN      NaN       NaN    
ar2   0.448183        NaN      NaN       NaN    
ar3  -0.100674   0.031576  -3.1883  0.001431 ** 
ar4  -0.011699   0.059963  -0.1951  0.845315    
ar5  -0.092399   0.059437  -1.5546  0.120046    
ma1  -0.113881        NaN      NaN       NaN    
ma2  -0.886087        NaN      NaN       NaN    
sar1 -0.082486   0.075217  -1.0966  0.272800    
sma1 -0.637160   0.053126 -11.9935 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


        Shapiro-Wilk normality test

data:  res.model
W = 0.71571, p-value < 2.2e-16
```

Table 6.3- z test of coefficients



Only ar3 and sma1 are statistically significant, because their p-values are less than 0.05. The highly significant seasonal MA(1) term captures the annual shock pattern. The MA parameters cannot be trusted because their variances are not estimable, and the other AR and seasonal AR terms are not statistically significant which suggests over-parameterisation. Residuals oscillate around zero with a few large negative spikes (2020-2021) that could relate to pandemic related shocks, but no visible trend or seasonality. Histogram shows a slightly heavy-tailed distribution. In QQ plot, the points follow the line in the centre but diverge in the tails indicating the presence of outliers. There are no autocorrelations in ACF plot, as all the lags are within the 95% confidence interval. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

4.5.4 SARIMA (1,1,2) x $(1,1,1)_{12}$ - ML

```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ar1    0.456795   0.081126   5.6307 1.795e-08 ***
ma1   -0.693383   0.088484  -7.8362 4.642e-15 ***
ma2   -0.306616   0.087768  -3.4935 0.0004768 ***
sar1  -0.086022   0.074734  -1.1510 0.2497150
sma1  -0.635408   0.052125 -12.1900 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        Shapiro-Wilk normality test

data:  res.model
W = 0.71825, p-value < 2.2e-16
```

Table 6.4 - z test of coefficients

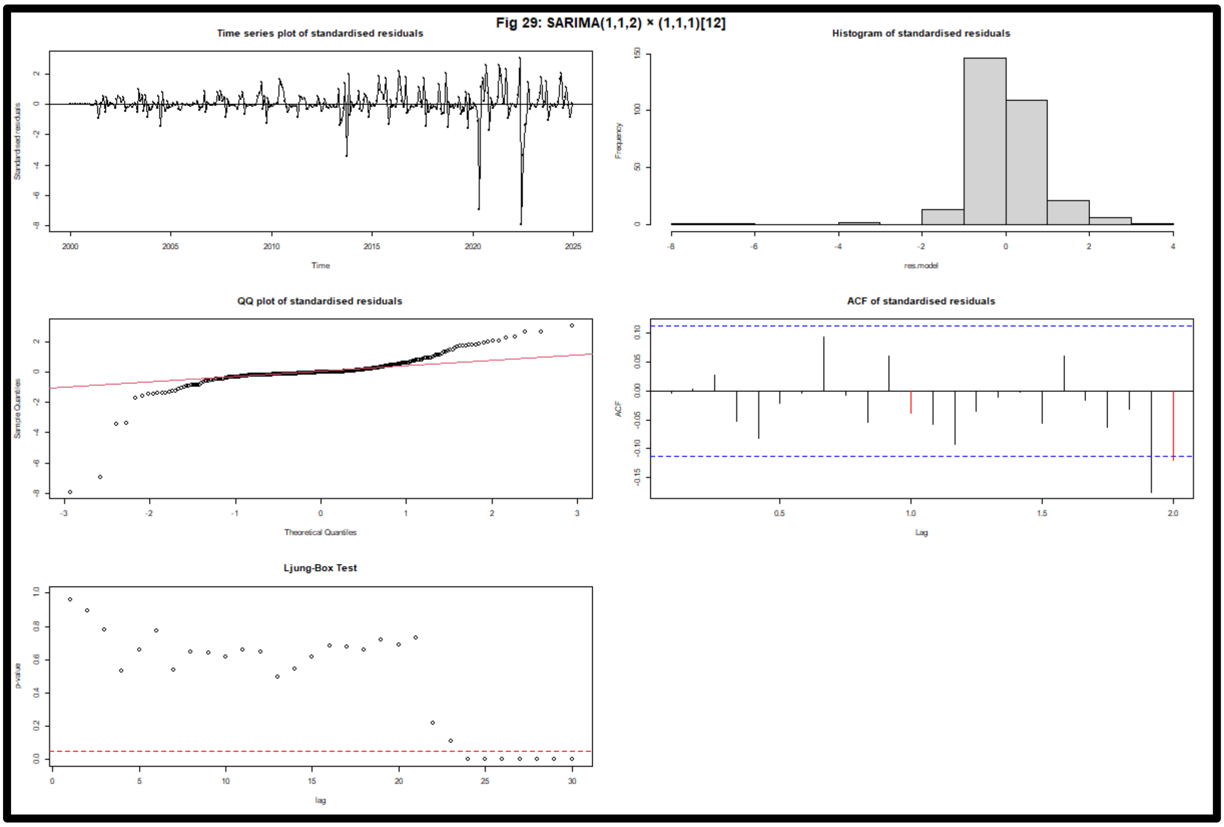


Fig 29: SARIMA(1,1,2) × (1,1,1)[12]

ar1, ma1, ma2 and sma1 are all highly significant as their p-values are less than 0.05. The highly significant seasonal MA(1) term captures the annual shock pattern. Residuals centered on zero with a few large negative spikes (2020-2021) that could relate to pandemic related shocks, but no visible trend or seasonality. Histogram shows a slightly heavy-tailed distribution. In QQ plot, good fit along line could be observed in the centre but diverge in the tails indicating the presence of outliers. There are no autocorrelations in ACF plot, as all the lags are within the 95% confidence interval. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

4.5.5 SARIMA (2,1,2) x (1,1,1)12 -ML

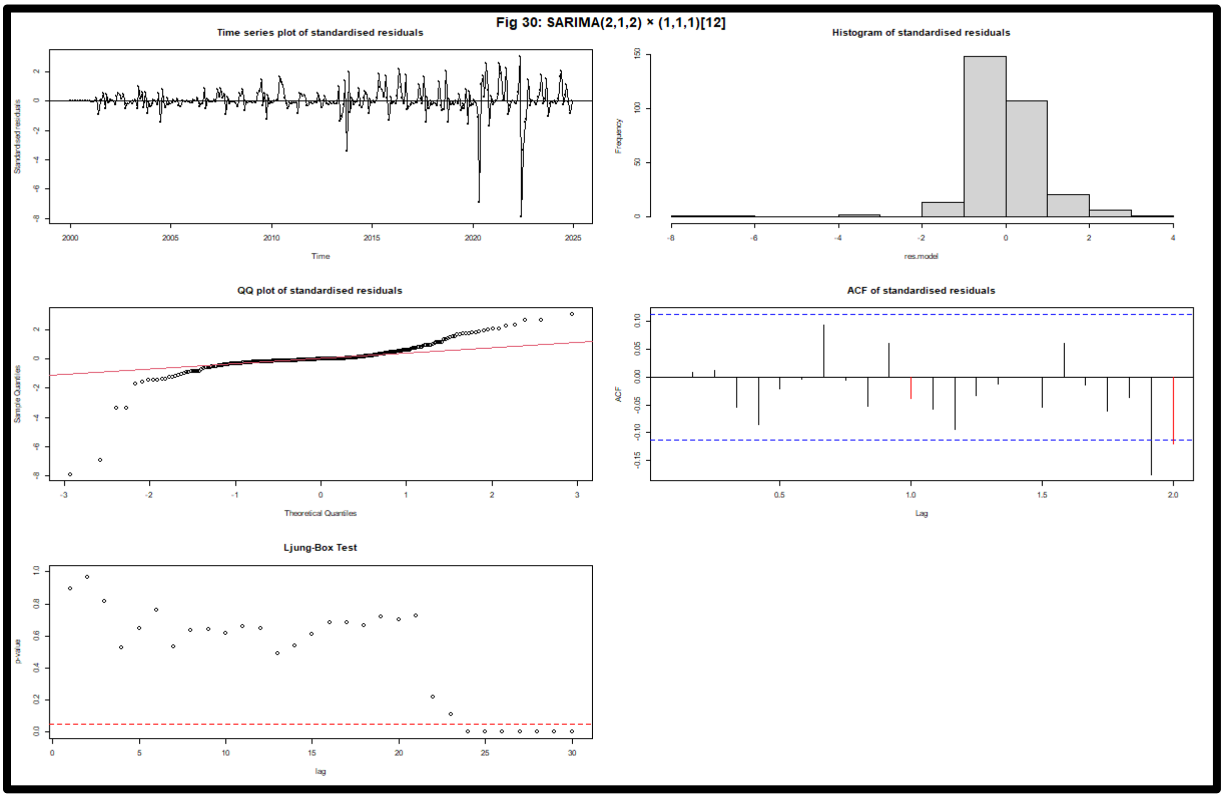
```
z test of coefficients:

      Estimate Std. Error  z value Pr(>|z|)
ar1   0.386502   0.299086   1.2923  0.19626
ar2   0.049589   0.207975   0.2384  0.81154
ma1  -0.624546   0.290865  -2.1472  0.03178 *
ma2  -0.375450   0.290661  -1.2917  0.19646
sar1 -0.086165   0.074737  -1.1529  0.24895
sma1 -0.635211   0.052173 -12.1751  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


        Shapiro-Wilk normality test

data:  res.model
W = 0.71942, p-value < 2.2e-16
```

Table 6.5 - z test of coefficients



ma1 is significant and sma1 is highly significant as their p-values are below 0.05. All AR parameters and the second MA term add no measurable explanatory power. This indicates the model is over-parameterised. Residuals oscillate around zero and two large negative spikes in 2020-21 could be observed possibly related to pandemic outliers, not serial correlation. Histogram shows a slightly heavy-tailed distribution. In QQ plot, good fit along line could be observed in the centre but diverge in the tails indicating the presence of outliers. No significant autocorrelations could be observed in ACF plot, as all the lags lie inside 95% limits. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

4.5.6 SARIMA (2,1,2) x (1,1,1)12 -CSS

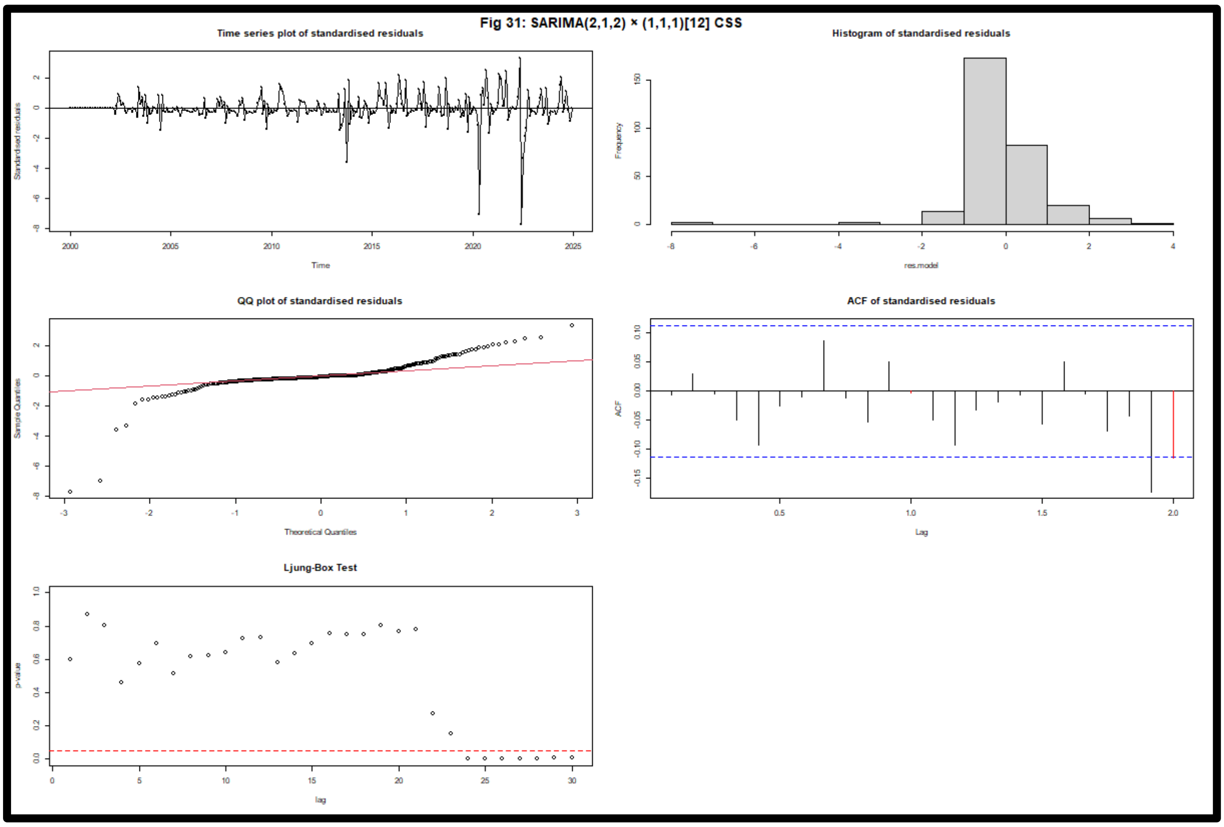
```
z test of coefficients:

      Estimate Std. Error  z value Pr(>|z|)
ar1   0.308156        NaN      NaN      NaN
ar2   0.079372        NaN      NaN      NaN
ma1  -0.566645        NaN      NaN      NaN
ma2  -0.467677        NaN      NaN      NaN
sar1 -0.128471   0.056187  -2.2865  0.02223 *
sma1 -0.635051   0.048723 -13.0339  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


        Shapiro-Wilk normality test

data:  res.model
W = 0.72487, p-value < 2.2e-16
```

Table 6.6 - z test of coefficients



For the model fitted with CSS, the information matrix could not be inverted for the non-seasonal parameters, so their standard errors appear as NA. Seasonal AR(1) is significant. Seasonal MA(1) is highly significant, dominating the annual dynamics. The inability to estimate non-seasonal standard errors suggests the model is over-parameterised or numerically unstable.

Residuals oscillate around zero and two large negative spikes in 2020-21 could be observed possibly related to pandemic outliers. Histogram shows a slightly heavy-tailed distribution. In QQ plot, points follow the red line in the centre but diverge in the tails indicating the presence of outliers. No significant autocorrelations could be observed in ACF plot, as all the lags lie inside 95% limits. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

4.5.7 SARIMA (2,1,2) x $(1,1,1)_{12}$ – CSS-ML

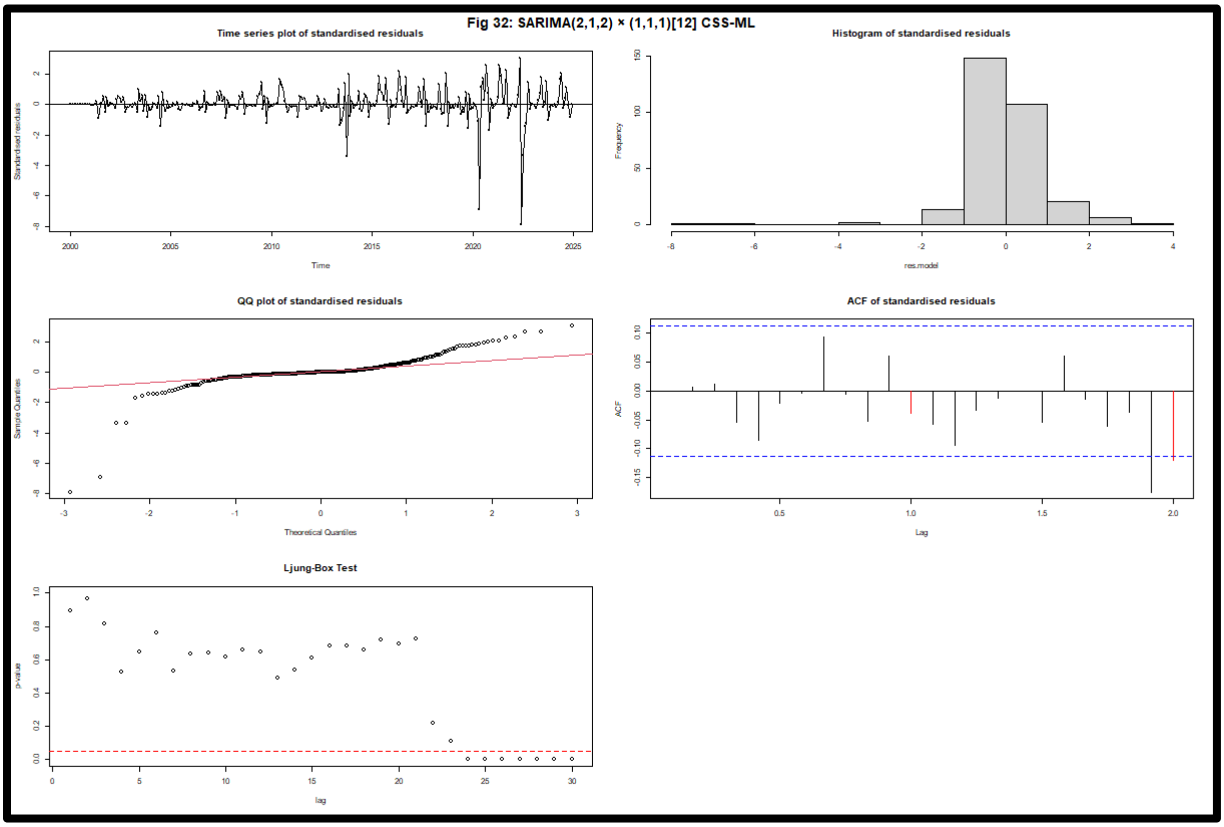
```
z test of coefficients:

      Estimate Std. Error  z value Pr(>|z|)
ar1   0.387836   0.300436   1.2909  0.19673
ar2   0.048731   0.208784   0.2334  0.81545
ma1  -0.625926   0.292366  -2.1409  0.03228 *
ma2  -0.374049   0.292153  -1.2803  0.20043
sar1 -0.085968   0.074740  -1.1502  0.25005
sma1 -0.635371   0.052147 -12.1842  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


        Shapiro-Wilk normality test

data:  res.model
W = 0.71939, p-value < 2.2e-16
```

Table 6.7 - z test of coefficients



MA(1) is significant and Seasonal MA(1) is highly significant as their p-values are below 0.05. All AR terms, the MA(2) term, and the seasonal AR(1) are not statistically significant suggesting the model is over-parameterised. Residuals oscillate around zero and two large negative spikes in 2020-21 could be observed possibly related to pandemic outliers. Histogram shows a slightly heavy-tailed distribution. In QQ plot, points follow the red line in the centre but diverge in the tails indicating the presence of outliers. No significant autocorrelations could be observed in ACF plot, as all the lags lie inside 95% limits. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

4.5.8 SARIMA (1,1,3) x (1,1,1)12-ML

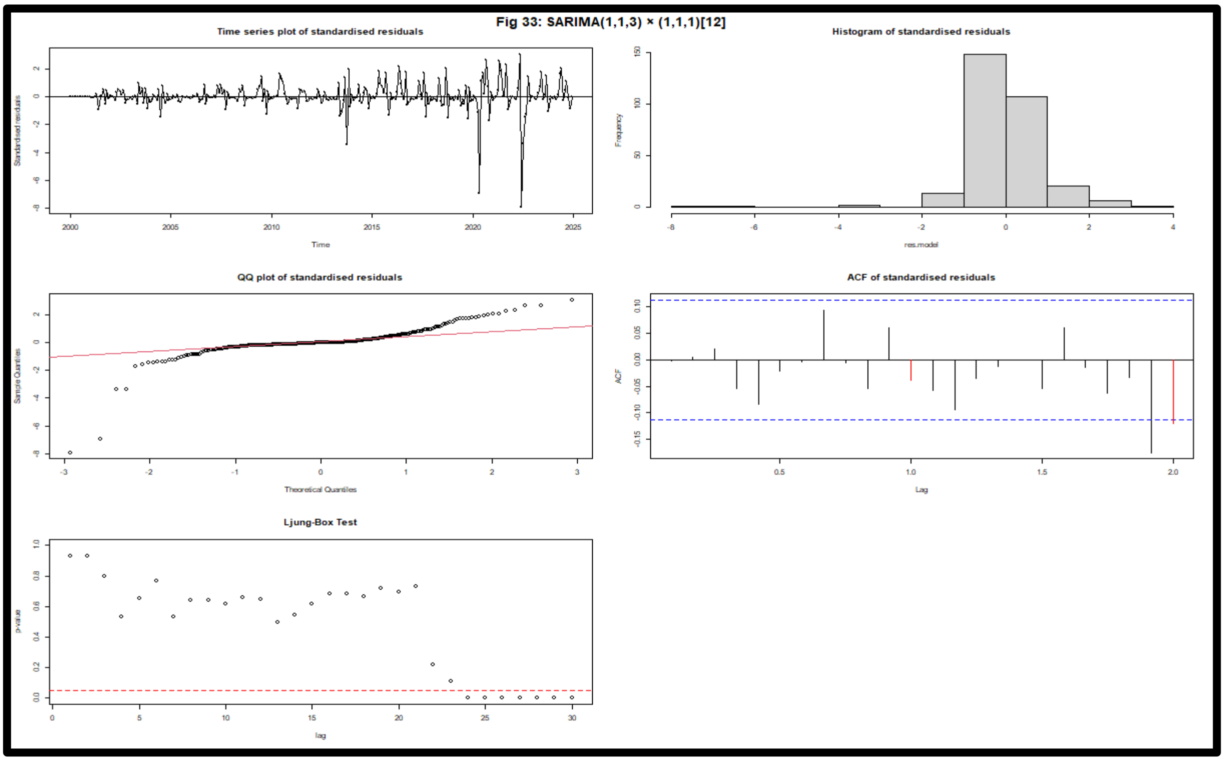
```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)    
ar1   0.473454   0.140966   3.3586 0.0007833 ***
ma1  -0.710384   0.148619  -4.7799 1.754e-06 ***
ma2  -0.304153   0.088122  -3.4515 0.0005575 ***
ma3   0.014562   0.104080   0.1399 0.8887282    
sar1 -0.085957   0.074726  -1.1503 0.2500214    
sma1 -0.635691   0.052070 -12.2084 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        Shapiro-Wilk normality test

data:  res.model
W = 0.71882, p-value < 2.2e-16
```

Table 6.8 - z test of coefficients



ar1, ma1, ma2 and sma1 are all highly significant as their p-values are less than 0.05. Residuals oscillate around zero and two large negative spikes in 2020-21 could be observed possibly related to pandemic outliers. Histogram shows a slightly heavy-tailed distribution. In QQ plot, points follow the red line in the centre but diverge in the tails indicating the presence of outliers. No significant autocorrelations could be observed in ACF plot, as all the lags lie inside 95% limits. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag. A simpler SARIMA (1,1,2) × $(0,1,1)_{12}$ had achieved the same residual behaviour.

4.5.9 SARIMA (1,1,3) x $(1,1,1)_{12}$ -CSS

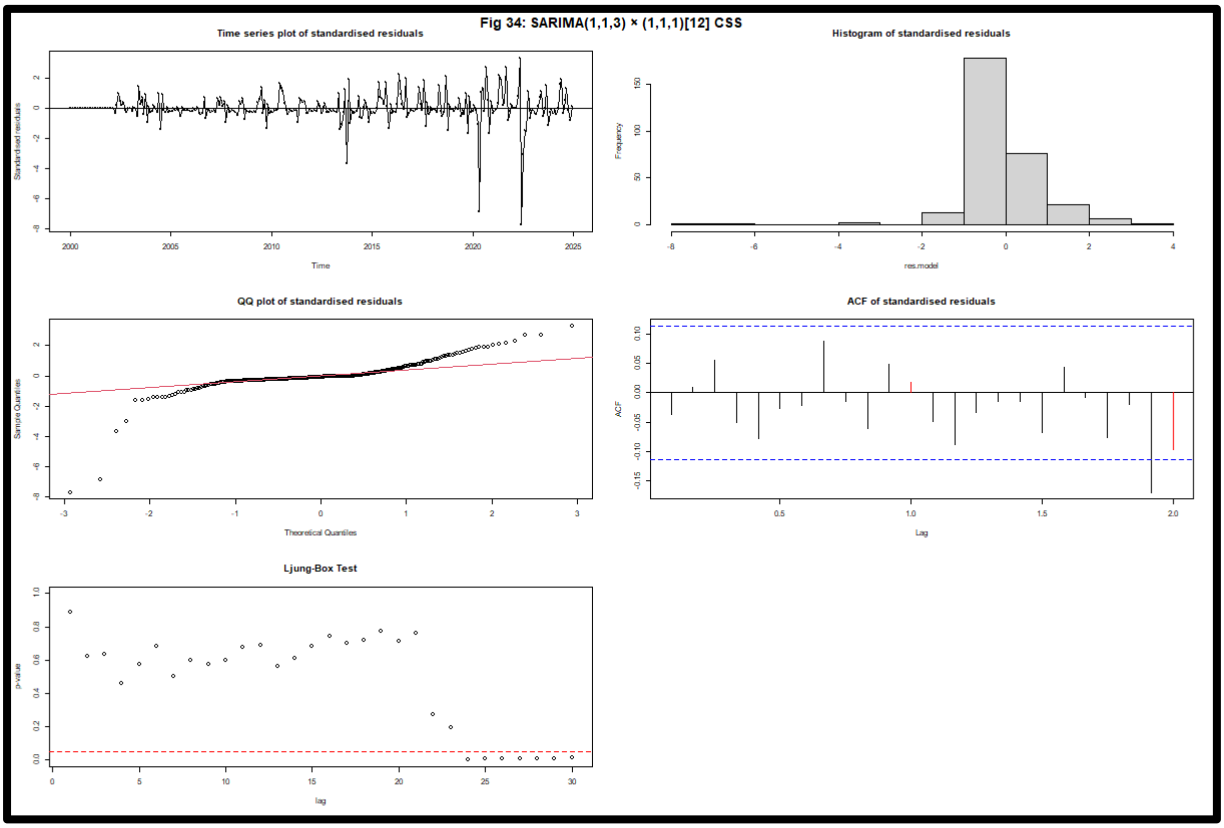
```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)    
ar1   0.330118   0.018372  17.9690 < 2.2e-16 ***
ma1  -0.563907   0.021410 -26.3386 < 2.2e-16 ***
ma2  -0.387853   0.066604  -5.8233  5.77e-09 ***
ma3  -0.088129   0.028527  -3.0894  0.002006 ** 
sar1 -0.122727   0.060307  -2.0350  0.041847 *  
sma1 -0.662068   0.042985 -15.4023 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


        Shapiro-Wilk normality test

data:  res.model
W = 0.72807, p-value < 2.2e-16
```

Table 6.9 - z test of coefficients



All six coefficients are statistically significant at the 5% level, so every component contributes measurably to explain the series. Compared with the ML version, CSS delivers smaller standard errors, making even **ma3** and **sar1** significant. Residuals oscillate around zero and two large negative spikes in 2020-21 could be observed possibly related to pandemic outliers. Histogram shows a slightly heavy-tailed distribution. In QQ plot, points follow the red line in the centre but diverge in the tails indicating the presence of outliers. No significant autocorrelations could be observed in ACF plot, as all the lags lie inside 95% limits. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

4.5.10 SARIMA (2,1,3) x $(1,1,1)_{12}$ -ML

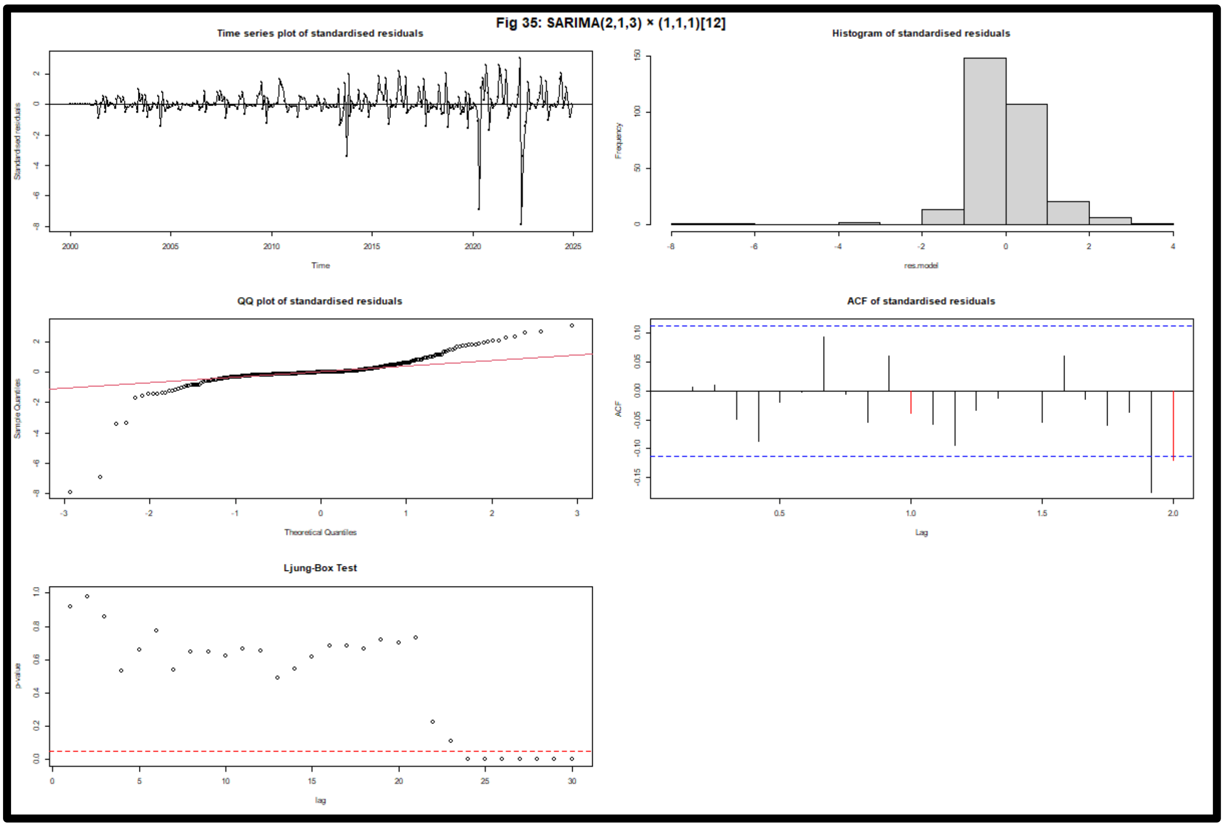
```
z test of coefficients:

      Estimate Std. Error  z value Pr(>|z|)
ar1   0.047382   1.027207   0.0461   0.9632
ar2   0.206361   0.455559   0.4530   0.6506
ma1  -0.286089   1.028208  -0.2782   0.7808
ma2  -0.610758   0.690332  -0.8847   0.3763
ma3  -0.103116   0.347191  -0.2970   0.7665
sar1 -0.085891   0.074791  -1.1484   0.2508
sma1 -0.634657   0.052290 -12.1373   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


        Shapiro-Wilk normality test

data:  res.model
W = 0.7193, p-value < 2.2e-16
```

Table 6.10 - z test of coefficients



All AR and MA terms have p-value more than 0.05, so they are not statistically significant. Model is clearly over-parameterized. Seasonal MA(1) is highly significant. We could see 2 large dips in 2020 - 2021 in standardized residual plot, which could be related to the pandemic. Histogram has range from -8 to 4. This could be due to the presence of outliers and the graph is left skewed. In QQ plot, points follow the red line in the centre but diverge in the tails indicating the presence of outliers. No significant autocorrelations could be observed in ACF plot, as all the lags lie inside 95% limits. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

4.5.11 SARIMA (2,1,3) x $(1,1,1)_{12}$ -CSS

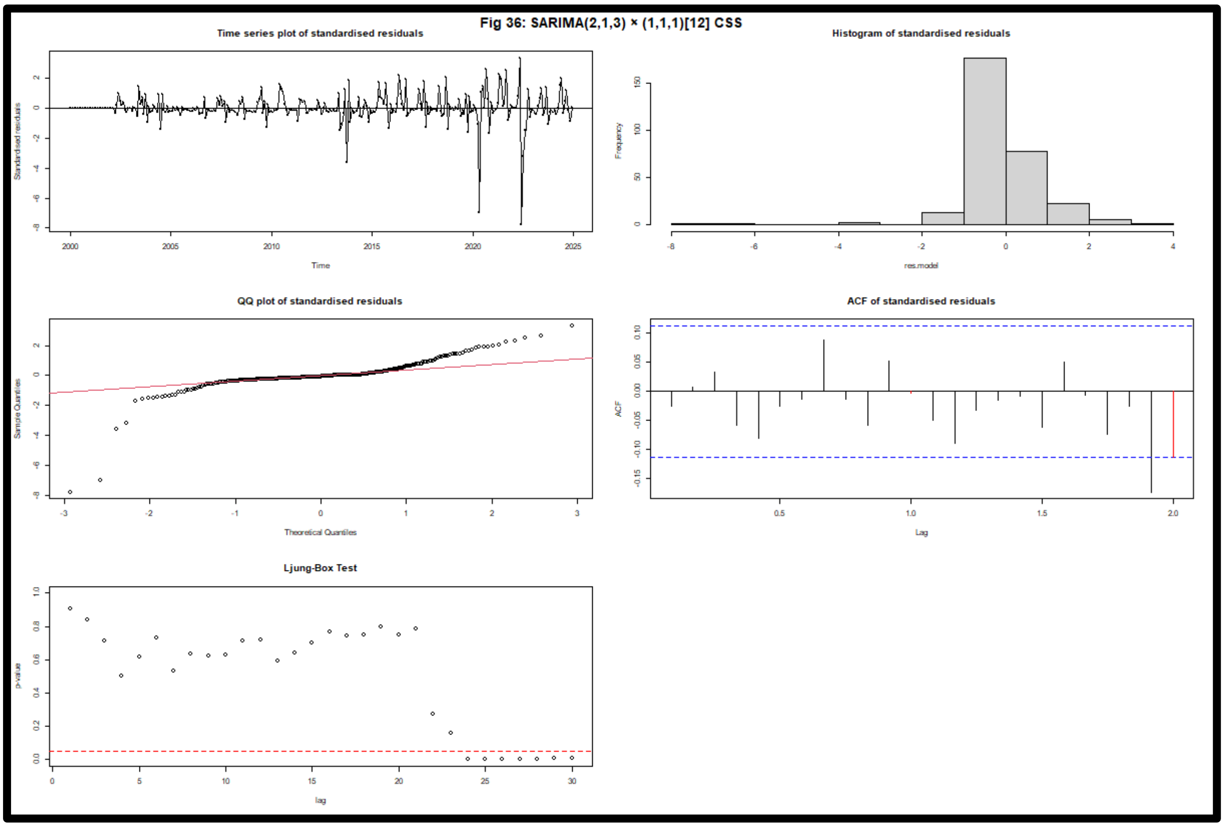
```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ar1   0.011315   0.042388   0.2669 0.7895101
ar2   0.177626   0.047709   3.7231 0.0001968 ***
ma1  -0.254466        NaN      NaN       NaN
ma2  -0.628053   0.060741 -10.3399 < 2.2e-16 ***
ma3  -0.166163   0.030331  -5.4784 4.293e-08 ***
sar1 -0.122670   0.067528  -1.8166 0.0692813 .
sma1 -0.638520   0.050234 -12.7110 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


        Shapiro-Wilk normality test

data:  res.model
W = 0.72096, p-value < 2.2e-16
```

Table 6.11 - z test of coefficients



ar2, ma2 and ma3 are all highly significant with p-value lesser than 0.05, remaining all other parameters are not significant. ma1 returns NA standard error, meaning it could not reliably estimate. Residuals fluctuate around 0, with 2 large spikes in 2020-2021. Histogram has range from -8 to 4. This could be due to the presence of outliers and the graph is left skewed. In QQ plot, most of the points follow the red line in the centre but diverge in the tails indicating the presence of outliers. No significant autocorrelations could be observed in ACF plot, as all the lags lie inside 95% limits. The Ljung-Box test shows that all the points are well above the Confidence Interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

4.5.12 SARIMA (2,1,3) x $(1,1,1)_{12}$ -CSSML

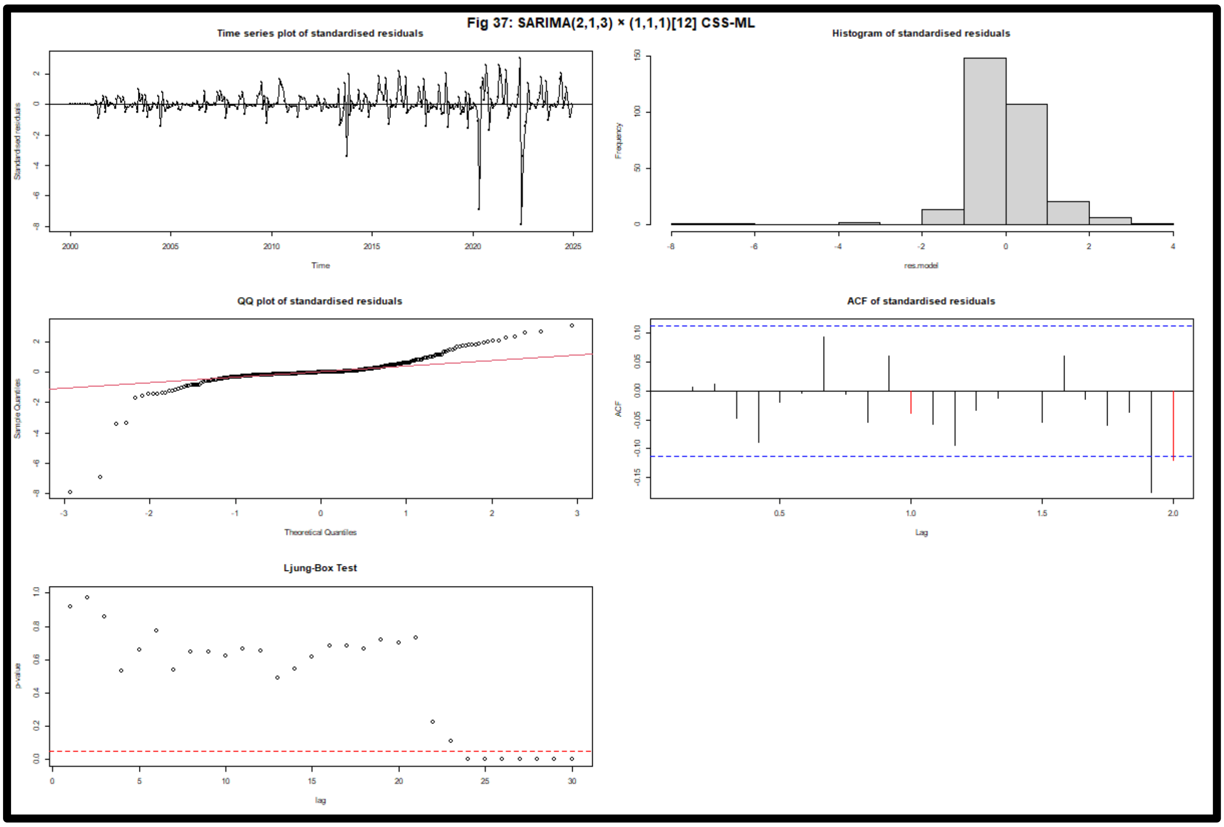
```
z test of coefficients:

      Estimate Std. Error  z value Pr(>|z|)
ar1   0.0077071  1.1410547   0.0068   0.9946
ar2   0.2233512  0.4969552   0.4494   0.6531
ma1  -0.2463650  1.1430326  -0.2155   0.8293
ma2  -0.6371262  0.7581529  -0.8404   0.4007
ma3  -0.1164931  0.3930457  -0.2964   0.7669
sar1 -0.0860562  0.0747866  -1.1507   0.2499
sma1 -0.6346171  0.0522870 -12.1372   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        Shapiro-Wilk normality test

data:  res.model
W = 0.71929, p-value < 2.2e-16
```

Table 6.12 - z test of coefficients



All AR and MA parameters have p-value more than 0.05, so they are not statistically significant. Model is clearly over-parameterized. We could see 2 large dips in 2020 - 2021 in standardized residual plot, which could be due to the pandemic. Histogram has range from -8 to 4. This could be due to the presence of outliers and the graph is left skewed. In QQ plot, points follow the red line in the centre but diverge in the tails indicating the presence of outliers. No significant autocorrelations could be observed in ACF plot, as all the lags lie inside 95% limits. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

4.5.13 SARIMA (1,1,1) x $(1,1,1)_{12}$ -ML

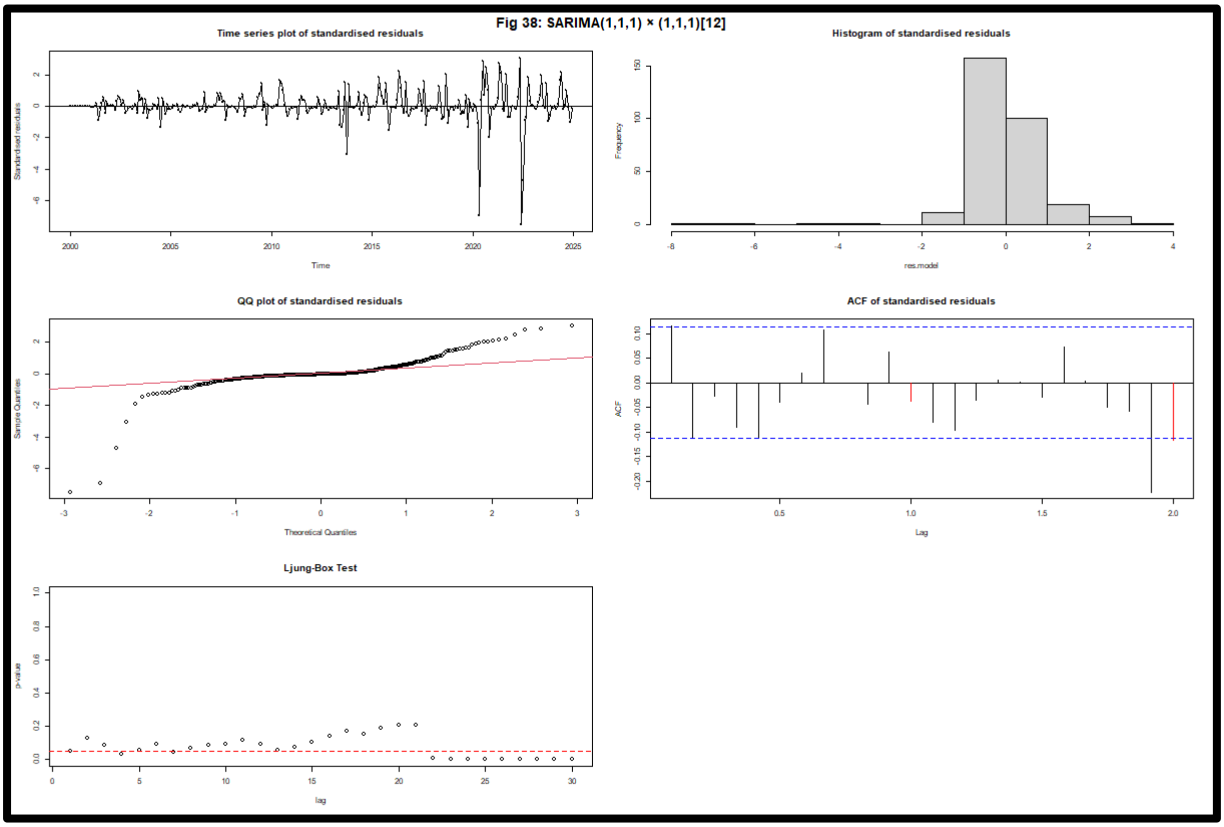
```
z test of coefficients:

      Estimate  Std. Error  z value  Pr(>|z|)
ar1   0.636399   0.046068  13.8144   <2e-16 ***
ma1  -0.999985   0.011089 -90.1744   <2e-16 ***
sar1 -0.070214   0.076142  -0.9221   0.3565
sma1 -0.629679   0.054088 -11.6417   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


        Shapiro-Wilk normality test

data:  res.model
W = 0.70488, p-value < 2.2e-16
```

Table 6.13 - z test of coefficients



ar1, ma1 and sma1 are highly significant with p-value less than 0.05. This could be one of the contenders for best model. We could see 2 large dips in 2020-2021 in standardized residual plot which could be related to the pandemic. Histogram has range from -8 to 4. This could be due to the presence of outliers and the graph is left skewed. In QQ plot, points follow the red line in the centre but diverge in the tails indicating the presence of outliers. No significant autocorrelations could be observed in ACF plot, as all the lags lie inside 95%

limits. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

4.5.14 SARIMA (2,1,1) x (1,1,1)12 -ML

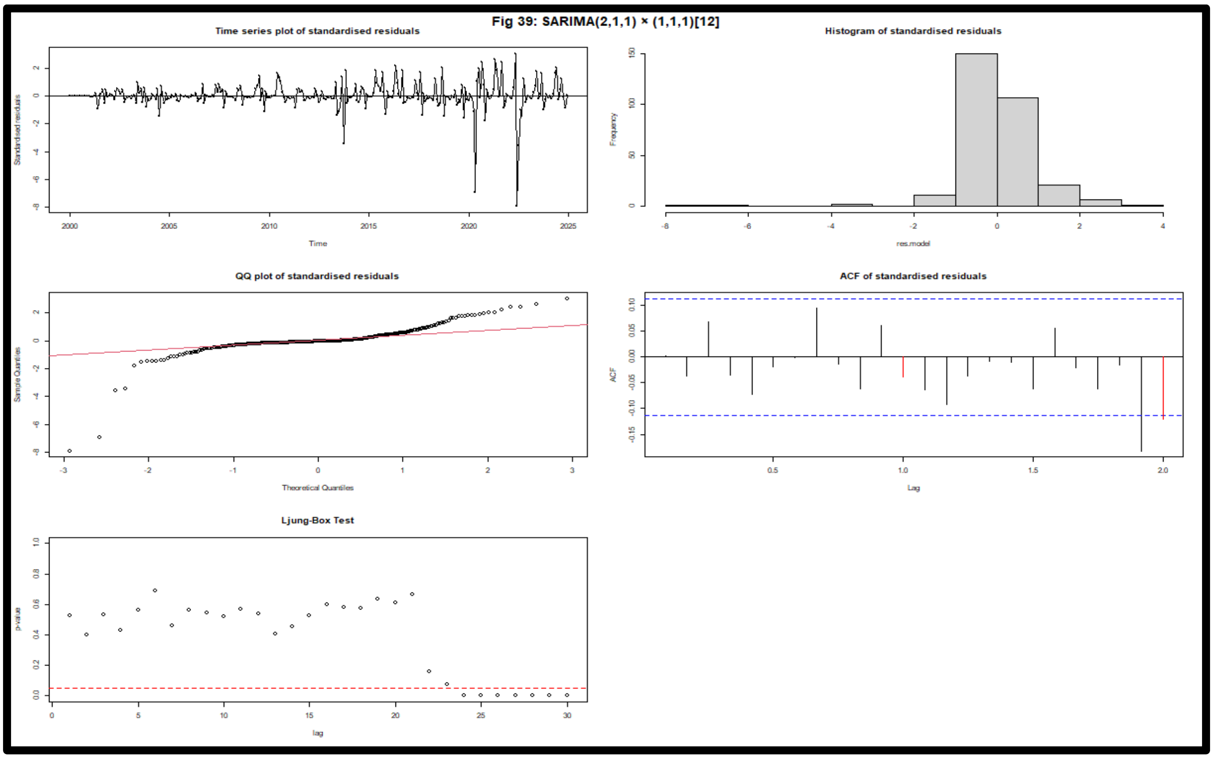
```
z test of coefficients:

      Estimate  Std. Error  z value   Pr(>|z|)    
ar1   0.751961    0.057902  12.9869  < 2.2e-16 ***
ar2  -0.185987    0.058051  -3.2039   0.001356 ** 
ma1  -0.999999    0.011999 -83.3396  < 2.2e-16 ***
sar1 -0.082608    0.074993  -1.1015   0.270665    
sma1 -0.633381    0.052499 -12.0647  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


        Shapiro-Wilk normality test

data:  res.model
W = 0.71202, p-value < 2.2e-16
```

Table 6.14 - z test of coefficients



ar1, ar2, ma1 and sma1 are all significant. This model also could be one of the contenders for best model. We could see 2 large dips in 2020-2021 in standardized residual plot, which could be related to the pandemic. Histogram has range from -8 to 4. This could be due to the presence of outliers and the graph is left skewed. In QQ plot, points follow the red line in the centre but diverge in the tails indicating the presence of outliers. No significant autocorrelations could be observed in ACF plot, as all the lags lie inside 95% limits. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

4.6 Model Selection

4.6.1 AIC, BIC and Accuracy Metrics Table

A tibble: 7 x 9

model <chr>	df <dbl>	AIC <dbl>	BIC <dbl>	ME <dbl>	RMSE <dbl>	MAE <dbl>	MASE <dbl>	ACF1 <dbl>
SARIMA(1,1,2)x(1,1,1)[12]	6	7034.635	7056.592	1480.787	47519.86	25163.70	0.6988666	-0.0032412887
SARIMA(2,1,1)x(1,1,1)[12]	6	7035.329	7057.286	1530.897	47581.78	24998.69	0.6942837	0.0024429872
SARIMA(2,1,2)x(1,1,1)[12]	7	7036.584	7062.201	1460.301	47518.19	25204.11	0.6999889	-0.0007790956
SARIMA(1,1,3)x(1,1,1)[12]	7	7036.612	7062.229	1471.707	47518.78	25183.82	0.6994252	-0.0023911584
SARIMA(2,1,3)x(1,1,1)[12]	8	7038.526	7067.801	1459.957	47515.30	25193.69	0.6996994	-0.0004935366
SARIMA(5,1,2)x(1,1,1)[12]	10	7040.134	7076.729	1627.634	47291.41	24917.03	0.6920157	0.0001157807
SARIMA(1,1,1)x(1,1,1)[12]	5	7043.389	7061.686	1282.794	48487.49	25040.54	0.6954460	0.1158759377

7 rows

Table 7 - AIC, BIC and Accuracy Metrics Table for best model selection

We are going to check the goodness of fit criteria. We will check AIC, BIC, ME, MAE, MASE and RMSE values of all the models. We sorted the SARIMA models based on AIC values in ascending order. The model with least AIC and BIC values are considered as potential candidate models.

SARIMA$(1,1,2)\times(1,1,1)_{12}$ has the lowest AIC and the lowest BIC values. SARIMA$(2,1,1)\times(1,1,1)_{12}$ is almost close to SARIMA$(1,1,2)\times(1,1,1)_{12}$, but the parameter estimate of SARIMA$(1,1,2)\times(1,1,1)_{12}$ was better than SARIMA$(2,1,1)\times(1,1,1)_{12}$. We selected SARIMA$(1,1,2)\times(1,1,1)_{12}$ model from as it has the lowest value.

From Table 7, the SARIMA(5,1,2)×(1,1,1)$_{12}$ model seems to be a top contender in terms of accuracy metrics among all other models . But to select a best model, it needs to be parsimonious as well as needs to be more accurate with all metrics. If we consider SARIMA(5,1,2)×(1,1,1)$_{12}$ as best model, it might have the best accuracy, but it won't follow parsimony due to higher ARIMA orders (p,q). Moreover, in terms of parameter estimation, not many coefficients were significant in all the 3 methods (CSS, ML and CSSML). So, we look ahead to find a better model.

As we saw earlier, SARIMA(1,1,2)×(1,1,1)$_{12}$ has the lowest AIC and BIC values. It's ME, MAE, MASE and RMSE values seems to be very close with SARIMA(5,1,2)×(1,1,1)$_{12}$. ME value of SARIMA(1,1,2)×(1,1,1)$_{12}$ is even lesser than that of the SARIMA(5,1,2)×(1,1,1)$_{12}$. MAE, MASE and RMSE values of SARIMA(1,1,2)×(1,1,1)$_{12}$ are very close to that of the SARIMA(5,1,2)×(1,1,1)$_{12}$. We then checked SARIMA(1,1,2)×(1,1,1)$_{12}$ parameter coefficients too.

```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ar1   0.456795   0.081126   5.6307 1.795e-08 ***
ma1  -0.693383   0.088484  -7.8362 4.642e-15 ***
ma2  -0.306616   0.087768  -3.4935 0.0004768 ***
sar1 -0.086022   0.074734  -1.1510 0.2497150
sma1 -0.635408   0.052125 -12.1900 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        Shapiro-Wilk normality test

data:  res.model
W = 0.71825, p-value < 2.2e-16
```

Table 8 - z test of coefficients for SARIMA(1,1,2)×(1,1,1)$_{12}$ ML model

From Table 8, we could see that ar1, ma1 and ma2 have p-value less than 0.05 and they are statistically significant which was not the case for SARIMA(5,1,2)×(1,1,1)$_{12}$ model. From the residual analysis of SARIMA(1,1,2)×(1,1,1)$_{12}$ model, significant seasonal MA (1) term captures the annual shock pattern. Residuals are centered on zero with few large negative spikes (2020-2021) that could relate to pandemic related shocks, but no visible trend or seasonality. Histogram shows a slightly heavy-tailed distribution. In QQ Plot, good fit along line could be observed in the centre but diverge in the tails indicating the presence of outliers. There are no autocorrelations in ACF plot, as all the lags are within the 95% confidence interval. The Ljung-Box test shows that all the points are well above the confidence interval line, we have some points which are slightly below the interval, but they are all coming way past 20th lag.

Also, there are only six estimated coefficients (df = 6) versus 8-10 in the larger candidates. They are easier to interpret and there is also a lesser risk of over-fitting.

We then checked the neighbours of this model to see if we can get 1 or 2 more significant models.

SARIMA $(1,1,3) \times (1,1,1)_{12}$
SARIMA $(2,1,2) \times (1,1,1)_{12}$

4.7 Overfitting Estimation

4.7.1 SARIMA $(1,1,3) \times (1,1,1)_{12}$

We already have this model in our Parameter Coefficient test. This model was not significant in ML method, but all the parameters were significant in CSS test. Then we checked model's AIC, BIC and accuracy metrics. In both tables, SARIMA$(1,1,3)\times(1,1,1)_{12}$ model has higher AIC and BIC value than SARIMA$(1,1,2)\times(1,1,1)_{12}$ model. So, we did not consider this model for further analysis.

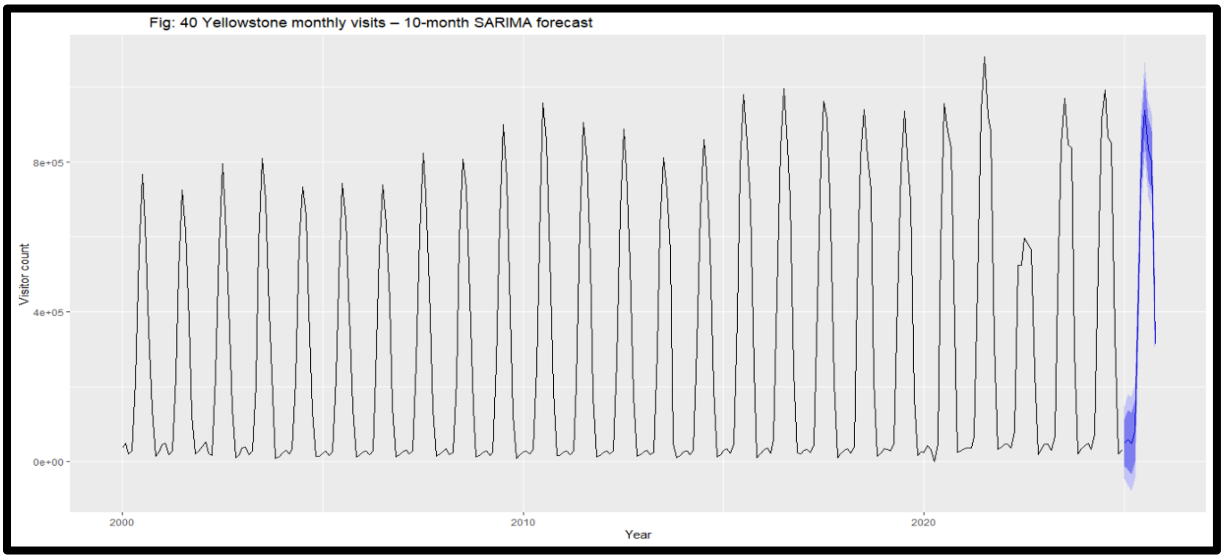
4.7.2 SARIMA $(2,1,2) \times (1,1,1)_{12}$

We also have this model in our Parameter Coefficient test. In all the 3 tests (ML, CSS, CSSML), this model was not significant. Also, we checked model's AIC, BIC and accuracy metrics. In both tables, SARIMA$(2,1,2)\times(1,1,1)_{12}$ model has higher AIC and BIC value than SARIMA$(1,1,2)\times(1,1,1)_{12}$ model. Hence, we did not go with this model too.

From all the analysis, we conclude that **SARIMA$(1,1,2)\text{x}(1,1,1)_{12}$** is the **best model** among all other models as it has the lowest AIC and BIC value, significant coefficients in parameter estimation and has better accuracy metrics.

4.8 Forecasting for SARIMA Model

We will proceed with forecasting for this model. Fig. 40 shows the 10-month forecasting of Yellowstone monthly visits with SARIMA$(1,1,2)\text{x}(1,1,1)_{12}$ model.



	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
Jan 2025	50661.84	-12278.18	113601.9	-45596.61	146920.3
Feb 2025	58345.72	-20992.69	137684.1	-62991.91	179683.4
Mar 2025	47911.01	-34526.88	130348.9	-78166.87	173988.9
Apr 2025	80900.36	-2209.80	164010.5	-46205.66	208006.4
May 2025	472344.90	389076.88	555612.9	344997.45	599692.4
Jun 2025	813600.50	730291.23	896909.8	686189.96	941011.0
Jul 2025	938724.07	855402.38	1022045.8	811294.54	1066153.6
Aug 2025	834540.11	751214.09	917866.1	707103.95	961976.3
Sep 2025	800295.69	716967.97	883623.4	672856.93	927734.5
Oct 2025	313361.68	230033.23	396690.1	185921.82	440801.5

Table 9 - Forecast values

We were able to forecast the seasonality trend which is evident from the Fig 40. The black line is the historical monthly visitor series (1999 - Oct 2024). The solid blue curve is the forecast generated by the fitted SARIMA(1,1,2)×(1,1,1)$_{12}$ model for the next ten months (Jan-Oct 2025). The two semi-transparent blue bands are prediction intervals. From Table 9 we infer :

- Point = model's best single-value forecast.
- Lo/Hi 80 % = bounds that should contain the true value 4 times out of 5 if the model were perfect.
- Lo/Hi 95 % = wider bounds with 95 % nominal coverage.

During the Winter months from (Jan-Mar), we could expect very low counts, mirroring the traditional off-season due to the prevalence of snow. Spring month (Apr) has a jump to nearly 81k which indicates that the park can expect to see some good crowd. Peak season is during the summer months (May-Aug) where all the schools will be closed, and the colleges will have their semester breaks too. This is also indicated by a sharp rise and a plateau in Jun to Aug, with July predicted as the busiest month (~939k). This follows the historical pattern in which July typically draws the largest crowds. Early autumn (Sep) is still busy, only slightly below August. Late autumn (Oct) is the month where the traffic drops quickly (~313k) as they prepare to close for winter.

5. Conclusion

Seasonal dummy model from Trend Model captured the deterministic trend as well as Seasonality pattern and showed that monthly dummies add the most explanatory power among our trend alternatives. SARIMA $(1,1,2) \times (1,1,1)_{12}$ gave the best stochastic fit (lowest AIC/BIC, better residuals) and therefore the most reliable multi-step forecasts.

In short, two models serve different but complementary purposes. The seasonal-dummy regression fitted on the original visitor counts, isolates the deterministic structure of the series as it quantifies the average month-to-month lift (the "seasonal factors") and any underlying growth trend, giving us an easily interpretable picture of how Yellowstone visitation behaves over the calendar year. Once that systematic pattern is removed by the required regular and seasonal differencing, the SARIMA$(1,1,2)\times(1,1,1)_{12}$ model takes over, capturing the remaining short-term autocorrelation in the seasonally adjusted data, its lowest AIC/BIC and better residual diagnostics makes it the most reliable engine for generating multi-step forecasts. Viewed together, the dummy-regression explains why the level data move the way they do, while the SARIMA explains how the residual fluctuations propagate forward, giving us both structural insight and accurate prediction.

6. References

1. National Park Service. (n.d.). *Public use statistics: Monthly visitation report builder* (IRMA report ID nps. Stats.976). Integrated Resource Management Applications portal. Retrieved June 15, 2025, from https://irma.nps.gov/Stats/SSRSReports/National%20Reports/Query%20Builder%20for%20Public%20Use%20Statistics%20(1979-Last%20Calendar%20Year)
2. Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). OTexts. https://otexts.com/fpp3/

3. Hyndman, R. J., et al. (2024). forecast (Version 8.23) [R package]. Comprehensive R Archive Network. https://cran.r-project.org/package=forecast
4. R Core Team. (2024). R: A language and environment for statistical computing (Version 4.4) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/
5. Shumway, R. H., & Stoffer, D. S. (2017). Time series analysis and its applications: With R examples (4th ed.). Springer.
6. Wickham, H., et al. (2024). ggplot2: Create elegant data visualisations using the grammar of graphics (Version 3.5) [R package]. Comprehensive R Archive Network. https://cran.r-project.org/package=ggplot2
7. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). John Wiley & Sons.
8. https://rmit.instructure.com/courses/140832/pages/week-2-after-class?module_item_id=7091840
9. https://rmit.instructure.com/courses/140832/files/44500708?module_item_id=7223237
10. https://rmit.instructure.com/courses/140832/pages/week-3-after-class?module_item_id=7092159
11. https://rmit.instructure.com/courses/140832/files/44536989?module_item_id=7225020
12. https://rmit.instructure.com/courses/140832/pages/week-4-after-class?module_item_id=7092165
13. https://rmit.instructure.com/courses/140832/files/43692306?module_item_id=7077746
14. https://rmit.instructure.com/courses/140832/pages/week-5-after-class?module_item_id=7092178
15. https://rmit.instructure.com/courses/140832/files/44854127?module_item_id=7250312
16. https://rmit.instructure.com/courses/140832/pages/week-6-after-class?module_item_id=7092201
17. https://rmit.instructure.com/courses/140832/files/45246826?module_item_id=7298823

7. Appendix

```
---
title: "TimeSeriesProject"
output: html_document
date: "2025-05-25"
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r}
# Load necessary libraries
suppressPackageStartupMessages({
  suppressWarnings({
    library(readr)
    library(dplyr)
    library(lubridate)
    library(FitAR)
    library(ggplot2)
    library(forecast)
    library(tibble)
```
```

```
    library(TSA)
    library(fUnitRoots)
    library(FitAR)
    library(lmtest)
    library(tseries)
  })
})
```

```{r}
# Read the CSV file
df <- read_csv("Yellowstone.csv")
```

```{r}
# Cleanning RecreationVisits column
df <- df %>%
  mutate(RecreationVisits = as.numeric(gsub(",", "", RecreationVisits)),
         Date = make_date(Year, Month, 1)) %>%
  arrange(Date)
```

```{r}
```

````
# Creating time series object from January 2000 to December 2024

yellowstone_ts <- ts(df$RecreationVisits, start = c(2000, 1), frequency = 12)

```



```{r}

# Plotting time series

plot(yellowstone_ts, ylab = "Visitors", xlab = "Year", main = "Fig 1: Yellowstone Monthly Visitors (2000–2024)", type = "o")

```



```{r}

# Structure

str(df)



# Summary statistics for RecreationVisits

summary(df$RecreationVisits)



# Missing value check

sum(is.na(df$RecreationVisits))

which(is.na(df$RecreationVisits))

```



```{r}
````

````
# Creating lagged version of the Yellowstone time series
y <- yellowstone_ts
x <- zlag(y)

index <- 2:length(x)

# Computing lag-1 correlation
correlation_lag1 <- cor(y[index], x[index])
print(paste("Lag 1 Correlation:", round(correlation_lag1, 4)))

# Scatter plot of original vs lagged values
plot(y[index], x[index],
    main = "Fig 2: Lag 1 Scatter Plot",
    xlab = "Lag 1 (Previous Month's Visitors)",
    ylab = "Current Month's Visitors")
```


```{r}
acf(yellowstone_ts, lag.max = 60, main = "Fig 3: ACF plot")
```



```{r}
````

```
# Time index
t <- 1:length(yellowstone_ts)

# 1. Linear Model
model1 <- lm(yellowstone_ts ~ t)
summary(model1)

# Fitted values
fitted.model1 <- fitted(model1)

# Plotting
plot(t, yellowstone_ts, type = 'o', ylab = 'Visitors',
    main = 'Fig 4: Linear Model Fit', xlab = 'Time (Index)')
lines(t, fitted.model1, col = 'red', lty = 2)

```

```
```{r}
# Residual Analysis
res.model1 <- rstudent(model1)
```
```

```
par(mfrow = c(2, 2))
plot(res.model1, type = 'l', main = "Fig 4.1: Standardized Residuals")
hist(res.model1, main = "Fig 4.2: Histogram of Residuals")
qqnorm(res.model1, main = "Fig 4.3: QQ Plot of Residuals")
qqline(res.model1, col = 2)
acf(res.model1, main = "Fig 4.4: ACF of Residuals")
par(mfrow = c(1, 1))

# Normality test
shapiro.test(res.model1)
```
```


```{r}
# Time index
t <- 1:length(yellowstone_ts)
t2 <- t^2

# 2. Quadratic Model
model2 <- lm(yellowstone_ts ~ t + t2)
summary(model2)

# Fitted values
fitted.model2 <- fitted(model2)
```
```

```
# Plot
plot(t, yellowstone_ts, type = 'o', ylab = 'Visitors', xlab = 'Time Index',
     main = 'Fig 5: Quadratic Model Fit')
lines(t, fitted.model2, col = 'red', lty = 2)

```

```
```{r}
# Residual Analysis for Quadratic Model
res.model2 <- rstudent(model2)

par(mfrow = c(2, 2))
plot(res.model2, type = 'l', main = "Fig 5.1: Standardized Residuals")
hist(res.model2, main = "Fig 5.2: Histogram of Residuals")
qqnorm(res.model2, main = "Fig 5.3: QQ Plot of Residuals")
qqline(res.model2, col = 2)
acf(res.model2, main = "Fig 5.4: ACF of Residuals")
par(mfrow = c(1, 1))

# Normality test
shapiro.test(res.model2)
```
```

````
```

```{r}
# Cosine Model
t <- 1:length(yellowstone_ts)
cos_term <- cos(2 * pi * t / 12)
model3 <- lm(yellowstone_ts ~ cos_term)
summary(model3)

# Fitted values and plot
fitted.model3 <- fitted(model3)
plot(t, yellowstone_ts, type = 'o', ylab = 'Visitors', xlab = 'Time Index',
    main = 'Fig 6: Cosine Model Fit')
lines(t, fitted.model3, col = 'red', lty = 2)

```

```{r}
res.model3 <- rstudent(model3)

par(mfrow = c(2, 2))
plot(res.model3, type = 'l', main = "Fig.6.1: Standardized Residuals")
hist(res.model3, main = "Fig.6.2: Histogram of Residuals")
````

```
qqnorm(res.model3, main = "Fig.6.3: QQ Plot")
qqline(res.model3, col = 2)
acf(res.model3, main = "Fig.6.4: ACF of Residuals")
par(mfrow = c(1, 1))

shapiro.test(res.model3)

```

```{r}
# Cyclical Model (Cosine + Sine)
sin_term <- sin(2 * pi * t / 12)
model4 <- lm(yellowstone_ts ~ cos_term + sin_term)
summary(model4)

# Fitted values and plot
fitted.model4 <- fitted(model4)
plot(t, yellowstone_ts, type = 'o', ylab = 'Visitors', xlab = 'Time Index',
    main = 'Fig 7: Cyclical Model Fit (Cos + Sin)')
lines(t, fitted.model4, col = 'red', lty = 2)
```

```{r}
```

```
res.model4 <- rstudent(model4)


par(mfrow = c(2, 2))
plot(res.model4, type = 'l', main = "Fig.7.1: Standardized Residuals")
hist(res.model4, main = "Fig.7.2: Histogram of Residuals")
qqnorm(res.model4, main = "Fig.7.3: QQ Plot")
qqline(res.model4, col = 2)
acf(res.model4, main = "Fig.7.4: ACF of Residuals")
par(mfrow = c(1, 1))


shapiro.test(res.model4)
```
```



```{r}
# Creating month factor variable
months <- cycle(yellowstone_ts)
month_factors <- factor(months)


# Seasonal Dummy Model
model5 <- lm(yellowstone_ts ~ month_factors)
summary(model5)
```
```

```
# Fitted values and plot
fitted.model5 <- fitted(model5)
plot(t, yellowstone_ts, type = 'o', ylab = 'Visitors', xlab = 'Time Index',
    main = 'Fig: 8 Seasonal Dummy Variable Model')
lines(t, fitted.model5, col = 'red', lty = 2)

```
```

```{r}
res.model5 <- rstudent(model5)

par(mfrow = c(2, 2))
plot(res.model5, type = 'l', main = "Fig: 8.1 Standardized Residuals")
hist(res.model5, main = "Fig: 8.2 Histogram of Residuals")
qqnorm(res.model5, main = "Fig: 8.3 QQ Plot")
qqline(res.model5, col = 2)
acf(res.model5, main = "Fig: 8.4 ACF of Residuals")
par(mfrow = c(1, 1))

shapiro.test(res.model5)

```
```

```{r}
freq <- frequency(yellowstone_ts)
n_future <- 10
t_future <- (length(yellowstone_ts) + 1):(length(yellowstone_ts) + n_future)
future_months <- ((cycle(yellowstone_ts)[length(yellowstone_ts)] + 1:n_future - 1) %% freq) + 1
future_month_factors <- factor(future_months, levels = 1:freq)
new_data <- data.frame(month_factors = future_month_factors)
forecast_values <- predict(model5, newdata = new_data)



forecast_tbl <- data.frame(
  Time_Index       = t_future,
  Month            = future_months,
  Forecast_Visitors = round(forecast_values, 0)
)


print(forecast_tbl, row.names = FALSE)


plot(t, yellowstone_ts, type = 'o', ylab = 'Visitors', xlab = 'Time Index',
     main = 'Fig 9: Seasonal Dummy Variable Model with Forecast')
lines(t, fitted.model5, col = 'red', lty = 2)
points(t_future, forecast_values, col = 'blue', pch = 19)
```

````
lines(t_future, forecast_values, col = 'blue', lty = 2)

legend("topleft",

       legend = c("Actual", "Fitted", "Forecast"),

       col = c("black", "red", "blue"), lty = c(1, 2, 2), pch = c(1, NA, 19))
```

```{r}
residual.analysis <- function(model, std = TRUE,start = 2, class = c("ARIMA","GARCH","ARMA-GARCH", "fGARCH")[1]){

  library(TSA)

  library(FitAR)

  if (class == "ARIMA"){

    if (std == TRUE){

      res.model = rstandard(model)

    }else{

      res.model = residuals(model)

    }

  }else if (class == "GARCH"){

    res.model = model$residuals[start:model$n.used]

  }else if (class == "ARMA-GARCH"){

    res.model = model@fit$residuals
````

```
  }else if (class == "fGARCH"){
    res.model = model@residuals
  }else {
    stop("The argument 'class' must be either 'ARIMA' or 'GARCH' ")
  }
  par(mfrow=c(3,2))
  plot(res.model,type='o',ylab='Standardised residuals', main="Time series plot of standardised residuals")
  abline(h=0)
  hist(res.model,main="Histogram of standardised residuals")
  qqnorm(res.model,main="QQ plot of standardised residuals")
  qqline(res.model, col = 2)
  seasonal_acf(res.model,main="ACF of standardised residuals")
  print(shapiro.test(res.model))
  k=0
  LBQPlot(res.model, lag.max = 30, StartLag = k + 1, k = 0, SquaredQ = FALSE)
  par(mfrow=c(1,1))
}
```

````
```{r}
helper <- function(class = c("acf", "pacf"), ...) {
````
````

```
# Capture additional arguments
params <- match.call(expand.dots = TRUE)
params <- as.list(params)[-1]

# Calculate ACF/PACF values
if (class == "acf") {
  acf_values <- do.call(acf, c(params, list(plot = FALSE)))
} else if (class == "pacf") {
  acf_values <- do.call(pacf, c(params, list(plot = FALSE)))
}

# Extract values and lags
acf_data <- data.frame(
  Lag = as.numeric(acf_values$lag),
  ACF = as.numeric(acf_values$acf)
)

# Identify seasonal lags to be highlighted
seasonal_lags <- acf_data$Lag %% 1 == 0

# Plot ACF/PACF values
if (class == "acf") {
```

```
    do.call(acf, c(params, list(plot = TRUE)))
  } else if (class == "pacf") {
    do.call(pacf, c(params, list(plot = TRUE)))
  }

  # Add colored segments for seasonal lags
  for (i in which(seasonal_lags)) {
    segments(x0 = acf_data$Lag[i], y0 = 0, x1 = acf_data$Lag[i], y1 = acf_data$ACF[i], col = "red")
  }
}

```

```{r}
# seasonal_acf
seasonal_acf <- function(...) {
  helper(class = "acf", ...)
}

# seasonal_pacf
seasonal_pacf <- function(...) {
  helper(class = "pacf", ...)
```

````
}
```

```{r}
par(mfrow=c(1,2))
seasonal_acf(yellowstone_ts,  lag.max = 64,main="Fig: 10 Seasonal ACF")
seasonal_pacf(yellowstone_ts,  lag.max = 64,main="Fig: 11 Seasonal PACF")
par(mfrow=c(1,1))
```

```{r}
# Test for stationarity
adf.test(yellowstone_ts)
pp.test(yellowstone_ts)
kpss.test(yellowstone_ts)
shapiro.test(yellowstone_ts)
```

```{r}
yellowstone_ts_adj <- yellowstone_ts
yellowstone_ts_adj[yellowstone_ts_adj <= 0] <- 1
````

````
lambda_seq <- seq(0, 2, 0.01)  # or any custom range


# Performing Box-Cox likelihood profile search
BC <- BoxCox.ar(yellowstone_ts_adj, lambda = lambda_seq)


# Extracting optimal lambda
lambda_opt <- BC$lambda[which.max(BC$loglike)]
print(paste("Optimal lambda in range -1 to 1.5:", round(lambda_opt, 3)))
lambda_opt
# Applying Box-Cox transformation with lambda = 0.36
ts_transformed <- BoxCox(yellowstone_ts_adj, lambda = 0.36)

```


```{r}
shapiro.test(ts_transformed)
adf.test(ts_transformed)
pp.test(ts_transformed)
kpss.test(ts_transformed)
```


```{r}
````

```
m1 <- Arima(yellowstone_ts, order = c(0, 0, 0), seasonal = list(order = c(0, 1, 0), period = 12))

res1 <- residuals(m1)

par(mfrow=c(1,1))

plot(res1,xlab='Time',ylab='Residuals',main="Fig: 12 Time series plot of the residuals")

par(mfrow=c(1,2))

seasonal_acf(res1, lag.max = 48, main = "Fig: 13 ACF of m1 residuals")

seasonal_pacf(res1, lag.max = 48, main = "Fig: 14 PACF of m1 residuals")

` ` `



` ` `{r}

m2 <- Arima(yellowstone_ts, order = c(0, 0, 0), seasonal = list(order = c(1, 1, 1), period = 12))

res2 <- residuals(m2)

par(mfrow=c(1,1))

plot(res2,xlab='Time',ylab='Residuals',main="Fig: 15 Time series plot of m2 residual")

par(mfrow=c(1,2))

seasonal_acf(res2, lag.max = 48, main = "Fig: 16 ACF of m2 residuals")

seasonal_pacf(res2, lag.max = 48, main = "Fig: 17 PACF of m2 residuals")

` ` `
```

```{r}
m3 = Arima(yellowstone_ts,order=c(0,1,0),seasonal=list(order=c(1,1,1), period=12))

res3 = residuals(m3);

par(mfrow=c(1,1))

plot(res3,xlab='Time',ylab='Residuals',main="Fig: 18 Time series plot of m3 residual")

par(mfrow=c(1,2))

seasonal_acf(res3, lag.max = 36, main = "Fig: 19 ACF of m3 residuals")

seasonal_pacf(res3, lag.max = 36, main = "Fig: 20 PACF of m3 residuals")
```

```{r}
m4 = Arima(yellowstone_ts,order=c(5,1,2),seasonal=list(order=c(1,1,1), period=12))

res4 = residuals(m4);

par(mfrow=c(1,1))

plot(res4,xlab='Time',ylab='Residuals',main="Fig: 21 Time series plot of m4 residual")

par(mfrow=c(1,2))

seasonal_acf(res4, lag.max = 36, main = "Fig: 22 ACF of m4 residuals")

seasonal_pacf(res4, lag.max = 36, main = "Fig: 23 PACF of m4 residuals")
```

```{r}
eacf(res3)
```

```{r}
par(mfrow=c(1,1))
bic_table = armasubsets(y=res3,nar=5,nma=5,y.name='p',ar.method='ols')
plot(bic_table)
mtext("Fig 25: BIC Table", side = 1, line = 1, cex = 1)
```

```{r}
# SARIMA(5,1,2)x(1,1,1)_12
m3_512.bhemanML = Arima(yellowstone_ts,order=c(5,1,2),seasonal=list(order=c(1,1,1), period=12),method = "ML")
coeftest(m3_512.bhemanML)
residual.analysis(model = m3_512.bhemanML)
```

```{r}
```

```
m3_512.bhemanCSS =
Arima(yellowstone_ts,order=c(5,1,2),seasonal=list(order=c(1,1,1),
period=12),method = "CSS")

coeftest(m3_512.bhemanCSS)

residual.analysis(model = m3_512.bhemanCSS)

```

````
```{r}

m3_512.bhemanCSSML =
Arima(yellowstone_ts,order=c(5,1,2),seasonal=list(order=c(1,1,1),
period=12),method = "CSS-ML")

coeftest(m3_512.bhemanCSSML)

residual.analysis(model = m3_512.bhemanCSSML)

```



```{r}

# SARIMA(1,1,2)x(1,1,1)_12

m3_112.bhemanML =
Arima(yellowstone_ts,order=c(1,1,2),seasonal=list(order=c(1,1,1),
period=12),method = "ML")

coeftest(m3_112.bhemanML)

residual.analysis(model = m3_112.bhemanML)

```



```{r}
````
````

```
# SARIMA(2,1,2)x(1,1,1)_12

m3_212.bhemanML =
Arima(yellowstone_ts,order=c(2,1,2),seasonal=list(order=c(1,1,1),
period=12),method = "ML")

coeftest(m3_212.bhemanML)

residual.analysis(model = m3_212.bhemanML)

```

```{r}

m3_212.bhemanCSS =
Arima(yellowstone_ts,order=c(2,1,2),seasonal=list(order=c(1,1,1),
period=12),method = "CSS")

coeftest(m3_212.bhemanCSS)

residual.analysis(model = m3_212.bhemanCSS)

```

```{r}

m3_212.bhemanCSSML =
Arima(yellowstone_ts,order=c(2,1,2),seasonal=list(order=c(1,1,1),
period=12),method = "CSS-ML")

coeftest(m3_212.bhemanCSSML)

residual.analysis(model = m3_212.bhemanCSSML)

```

```{r}
```

```
# SARIMA(1,1,3)x(1,1,1)_12

m3_113.bhemanML =
Arima(yellowstone_ts,order=c(1,1,3),seasonal=list(order=c(1,1,1),
period=12),method = "ML")

coeftest(m3_113.bhemanML)

residual.analysis(model = m3_113.bhemanML)

```



```{r}

m3_113.bhemanCSS =
Arima(yellowstone_ts,order=c(1,1,3),seasonal=list(order=c(1,1,1),
period=12),method = "CSS")

coeftest(m3_113.bhemanCSS)

residual.analysis(model = m3_113.bhemanCSS)

```



```{r}

# SARIMA(2,1,3)x(1,1,1)_12

m3_213.bhemanML =
Arima(yellowstone_ts,order=c(2,1,3),seasonal=list(order=c(1,1,1),
period=12),method = "ML")

coeftest(m3_213.bhemanML)

residual.analysis(model = m3_213.bhemanML)

```
```
```

```{r}

m3_213.bhemanCSS = Arima(yellowstone_ts,order=c(2,1,3),seasonal=list(order=c(1,1,1), period=12),method = "CSS")

coeftest(m3_213.bhemanCSS)

residual.analysis(model = m3_213.bhemanCSS)

```

```{r}

m3_213.bhemanCSSML = Arima(yellowstone_ts,order=c(2,1,3),seasonal=list(order=c(1,1,1), period=12),method = "CSS-ML")

coeftest(m3_213.bhemanCSSML)

residual.analysis(model = m3_213.bhemanCSSML)

```

```{r}

# SARIMA(1,1,1)x(1,1,1)_12

m3_111.bhemanML = Arima(yellowstone_ts,order=c(1,1,1),seasonal=list(order=c(1,1,1), period=12),method = "ML")

coeftest(m3_111.bhemanML)

residual.analysis(model = m3_111.bhemanML)

```

```{r}

# SARIMA(2,1,1)x(1,1,1)_12

m3_211.bhemanML =
Arima(yellowstone_ts,order=c(2,1,1),seasonal=list(order=c(1,1,1),
period=12),method = "ML")

coeftest(m3_211.bhemanML)

residual.analysis(model = m3_211.bhemanML)

```

```{r}

fits <- list(

  "SARIMA(1,1,2)×(1,1,1)[12]" = m3_112.bhemanML,

  "SARIMA(2,1,2)×(1,1,1)[12]" = m3_212.bhemanML,

  "SARIMA(1,1,3)×(1,1,1)[12]" = m3_113.bhemanML,

  "SARIMA(2,1,3)×(1,1,1)[12]" = m3_213.bhemanML,

  "SARIMA(1,1,1)×(1,1,1)[12]" = m3_111.bhemanML,

  "SARIMA(2,1,1)×(1,1,1)[12]" = m3_211.bhemanML,

  "SARIMA(5,1,2)×(1,1,1)[12]" = m3_512.bhemanML

)

# Helper function to compute "AIC", "BIC", "ME", "RMSE", "MAE",
"MASE", "ACF1" and sort the models by AIC

sarima_metrics <- function(model_list,

                        sort_by = c("AIC", "BIC",
```

```
                        "ME", "RMSE", "MAE", "MASE", "ACF1"))
{
 sort_by <- match.arg(sort_by)


 bind_rows(lapply(names(model_list), function(nm) {
  fit <- model_list[[nm]]


  acc <- accuracy(forecast(fit, h = 1))[1, ]


  tibble(model = nm,
      df    = attr(logLik(fit), "df"),
      AIC   = AIC(fit),
      BIC   = BIC(fit),
      ME    = acc["ME"],
      RMSE  = acc["RMSE"],
      MAE   = acc["MAE"],
      MASE  = acc["MASE"],
      ACF1  = acc["ACF1"])
 })) |>
 arrange(.data[[sort_by]])
}
```

```
# Generate the comparison table

metrics_table <- sarima_metrics(fits, sort_by = "AIC")

print(metrics_table, digits = 3)


```
```

```{r}

# SARIMA(1,1,3)x(1,1,1)_12

m3_113.bhemanML = Arima(yellowstone_ts,order=c(1,1,3),seasonal=list(order=c(1,1,1), period=12),method = "ML")

coeftest(m3_113.bhemanML)

residual.analysis(model = m3_113.bhemanML)

```

```{r}

m4_113.bhemanCSS = Arima(yellowstone_ts,order=c(1,1,3),seasonal=list(order=c(1,1,1), period=12),method = "CSS")

coeftest(m4_113.bhemanCSS)

residual.analysis(model = m4_113.bhemanCSS)

```
````

````
```{r}

# SARIMA(2,1,2)x(1,1,1)_12

m3_212.bhemanML = Arima(yellowstone_ts,order=c(2,1,2),seasonal=list(order=c(1,1,1), period=12),method = "ML")

coeftest(m3_212.bhemanML)

residual.analysis(model = m3_212.bhemanML)

```



```{r}

m4_212.bhemanCSS = Arima(yellowstone_ts,order=c(2,1,2),seasonal=list(order=c(1,1,1), period=12),method = "CSS")

coeftest(m4_212.bhemanCSS)

residual.analysis(model = m4_212.bhemanCSS)

```



```{r}

m4_212.bhemanCSSML = Arima(yellowstone_ts,order=c(2,1,2),seasonal=list(order=c(1,1,1), period=12),method = "CSS-ML")

coeftest(m4_212.bhemanCSSML)

residual.analysis(model = m4_212.bhemanCSSML)

```
````

```{r}
# Fitting the model
fit_112 <- Arima(
  yellowstone_ts,
  order    = c(1, 1, 2),
  seasonal = list(order = c(1, 1, 1), period = 12)
)

# 10-month forecast
fc_10 <- forecast(fit_112, h = 10)

# Plotting the forecast
autoplot(fc_10) +
  ggtitle("                Fig: 40 Yellowstone monthly visits – 10-month SARIMA forecast") +
  ylab("Visitor count") +
  xlab("Year")

# Printing the numeric forecast table
print(fc_10)
```

```
```{r}



```
```