## Time Series Analysis

---

# Analysing Sunspot Activity through Time Series Analysis

## Problem Statement:

Develop a predictive model to forecast the Monthly Mean Total Sunspot Number based on historical data, considering the cyclic nature of sunspot activity and its correlation with solar magnetic field fluctuations. This model should accurately capture the variations in sunspot numbers over time, accounting for the 11-year solar cycle and the influence of factors such as solar flares and magnetic storms.
The presence of the "date" attribute indicates a sequential ordering of observations, while the "Monthly Mean Total Sunspot Number" attribute represents a continuous variable measured at regular intervals. This alignment with time-based data and numerical values makes the Sunspot dataset ideal for time series analysis, enabling the exploration of temporal patterns, trends, and forecasting of future sunspot activity.

### Background:

Sunspots are temporary phenomena on the Sun's photosphere that appear as spots darker than the surrounding areas. They are directly linked to the Sun's magnetic activity and are indicative of magnetic storms and solar flares. Sunspot activity follows an 11-year cycle known as the solar cycle, characterized by periods of high and low sunspot numbers. Understanding and predicting sunspot activity is crucial for various applications, including space weather forecasting, satellite communication, and understanding the Sun's influence on Earth's climate.

### Objective:

The objective of this project is to develop a predictive model that accurately forecasts the Monthly Mean Total Sunspot Number based on historical data. This model will take into account the cyclic nature of sunspot activity, primarily the 11-year solar cycle, and its correlation with solar magnetic field fluctuations. Additionally, the model will consider the influence of factors such as solar flares and magnetic storms on sunspot activity.

By achieving these objectives, the predictive model will contribute to our understanding of solar dynamics and improve our ability to forecast sunspot activity, thereby enhancing space weather prediction capabilities and supporting various applications reliant on solar activity forecasts.

## Initial Analysis:

### Data Collection:

Historical data on the Monthly Mean Total Sunspot Number can be obtained from sources like the Solar Influences Data Analysis Centre (SIDC) or NASA's Space Weather Prediction Centre (SWPC). This data typically includes the number of sunspots observed each month over a period of several decades.

## Data Preprocessing:

Cleaning the data: Handling missing values, outliers, and any inconsistencies in the dataset.

Exploring the data: Conducting descriptive statistics and visualizations to understand the distribution and patterns of sunspot activity over time.

Time series decomposition: Decompose the time series data into its trend, seasonal, and residual components to understand the underlying patterns and cyclic behavior, such as the 11year solar cycle.

## Model Selection:

Time series forecasting models: Models such as Autoregressive Integrated Moving Average (ARIMA), Exponential Smoothing State Space Model (ETS) can be considered for forecasting sunspot activity.

Model Deployment and Monitoring: Once a suitable forecasting model is developed, it can be deployed to generate monthly forecasts of the Mean Total Sunspot Number. Continuous monitoring of the model's performance is essential, and periodic retraining may be necessary to adapt to changing patterns in sunspot activity.

This initial analysis provides a roadmap for developing a predictive model to forecast the Monthly Mean Total Sunspot Number, considering the cyclic nature of sunspot activity and its correlation with solar magnetic field fluctuations.

# Analysis:

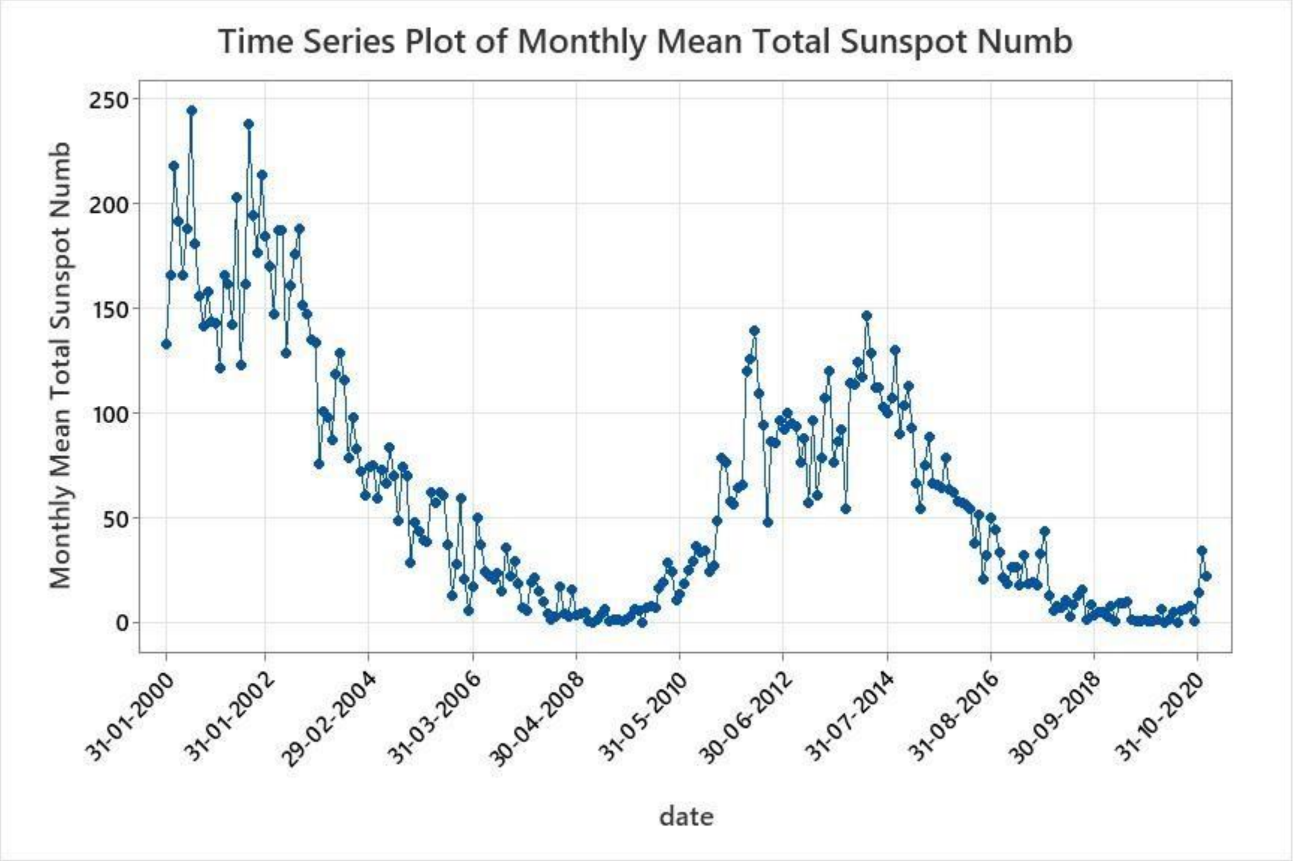
## EDA:

### Descriptive Statistics: Monthly Mean Total Sunspot Number Statistics

| | | | | SE | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Variable** | **N** | **N*** | **Mean** | **Mean** | **StDev** | **Minimum** | **Q1** | **Median** | **Q3** |
| Monthly Mean Total Sunspot Numb | 252 | 0 | 62.4742 | 3.65284 | 57.9870 | 0 | 11.3 | 49.5 | 97.875 |

| **Variable** | **Maximum** |
|---|---|
| Monthly Mean Total Sunspot Numb | 244.3 |

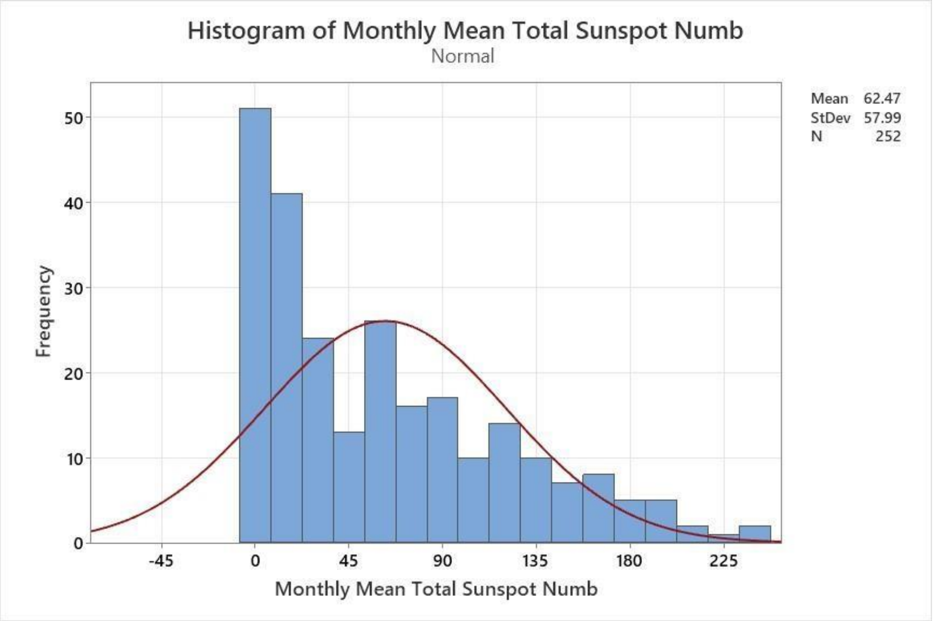
## Time Series Plot of Monthly Mean Total Sunspot Number:



## Interpretation:

Based on the time series plot of the monthly mean total sunspot number, it appears that the sunspot number first increases from the year 2000 to around 2014, and then decreases until 2020. Therefore, it is not accurate to say that the sunspot number is consistently decreasing or increasing throughout this time period. Instead, there are fluctuations in the sunspot number over time and adheres to a cyclical trend, as there is seen an absence of Secular, Seasonal and Irregular Trend. The descriptive statistics also show that the mean sunspot number is 62.4742 with a standard deviation of 57.9870, indicating a significant variability in the data.

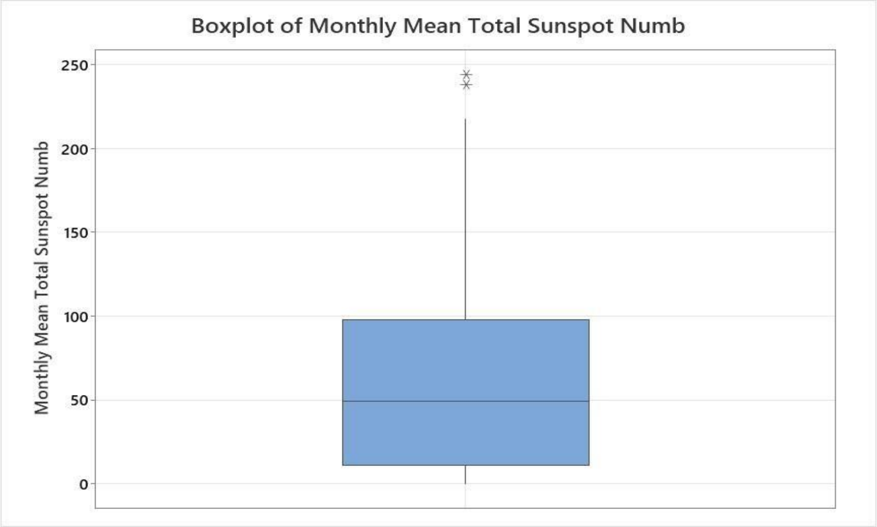
## Histogram of Monthly Mean Total Sunspot Number:



## Interpretation:

The Histogram Plot shows the monthly mean total sunspot number from 2000 to 2020. The sunspot number increases until around 2014, then decreases, with fluctuations throughout the time period. The sunspot number is not consistently increasing or decreasing, but rather varies over time.

## Boxplot of Monthly Mean Total Sunspot Number:



## Interpretation:

The boxplot above shows the monthly mean total sunspot number. The blue line in the graph represents the sunspot number. The boxplot shows the distribution of the data, with the box representing the interquartile range (IQR) and the line inside the box representing the median. The whiskers extend to the minimum and maximum values, excluding any outliers.

# Trend Analysis for Monthly mean

* NOTE * Zero values of Yt exist; MAPE calculated only for non-zero Yt.

## Method

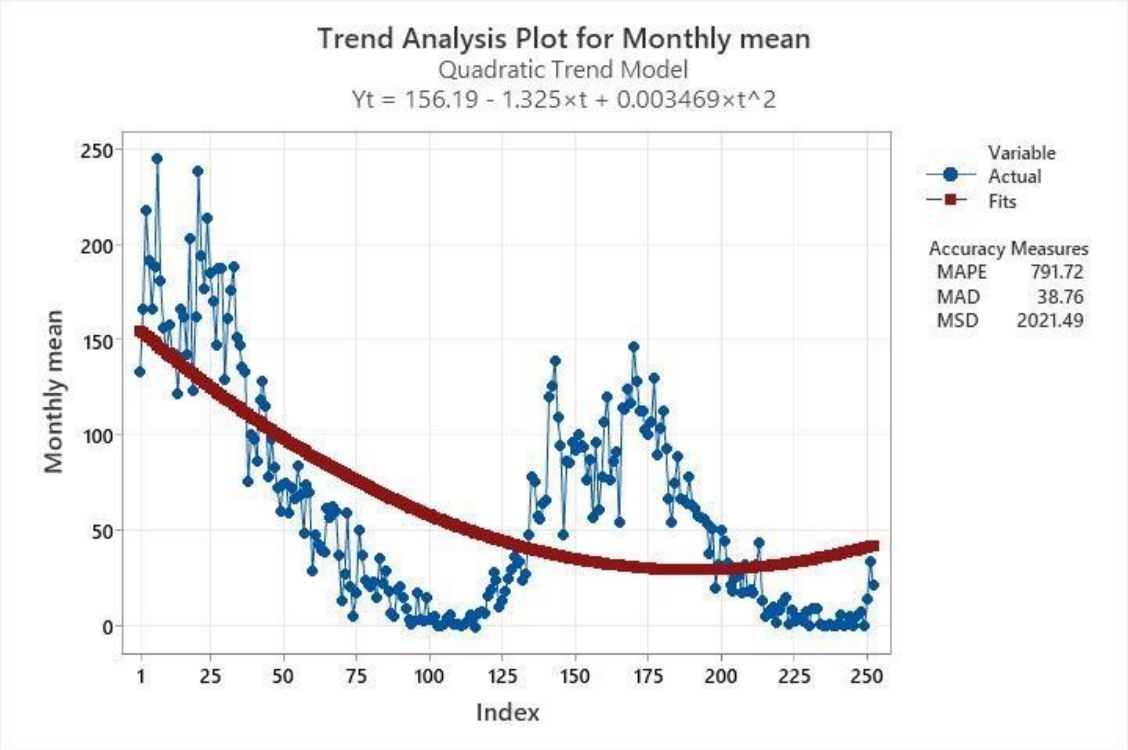
| | |
|---|---|
| Model type | Quadratic Trend Model |
| Data | Monthly mean |
| Length | 252 |
| NMissing | 0 |

## Fitted Trend Equation

Yt = 156.19 - 1.325×t + 0.003469×t^2

## Accuracy Measures

| | |
|---|---|
| MAPE | 791.72 |
| MAD | 38.76 |
| MSD | 2021.49 |



### Interpretation:

In above trend analysis, the accuracy measures are-

MAPE-791.72 | MAD-38.76| MSD-2021.49

To find the best possible match, we need to consider the minimal values of MAPE, MAD, and MSD. After performing trend analysis on the data, it was found that the quadratic trend model has the lowest value and provided the best fit, displaying a non-linear relationship between the variables with an increasing rate of change over time. The data have a curvature, which indicates that the rate of change varies over time.

## Single Exponential Smoothing for Monthly mean:

* NOTE * Zero values of Yt exist; MAPE calculated only for non-zero Yt. **Method**

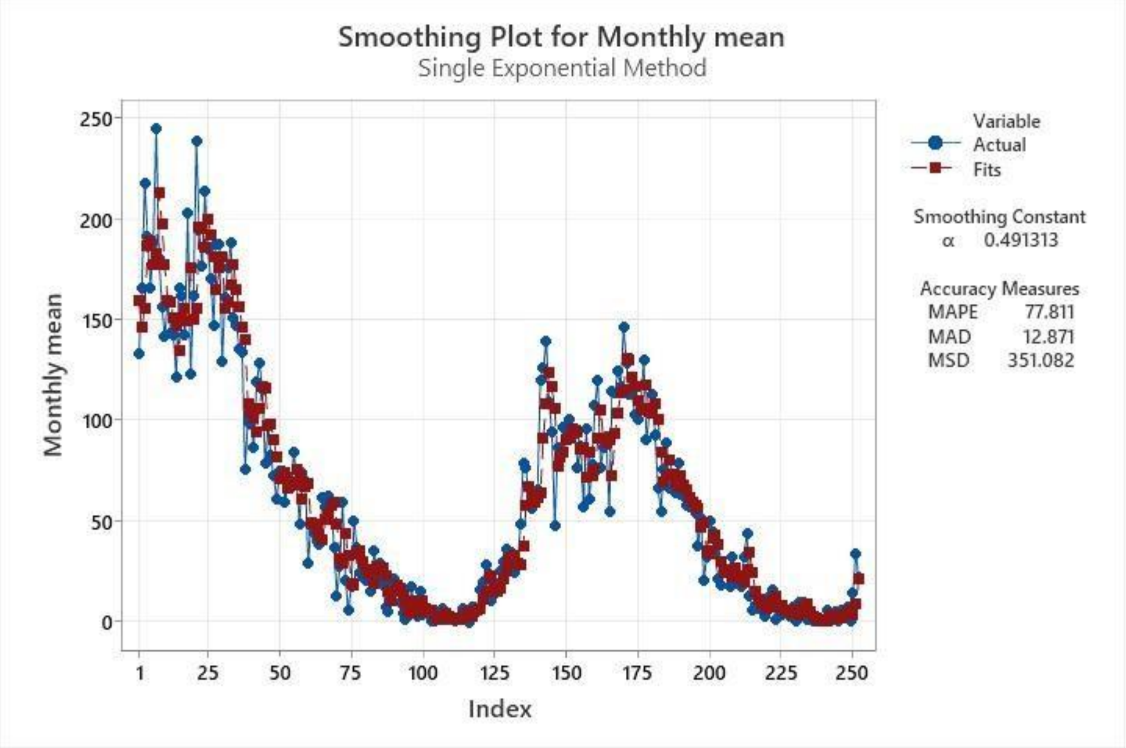
Data Monthly
mean Length 252

**Smoothing Constant** α

0.491313

**Accuracy Measures**

MAPE 77.811
MAD 12.871
MSD 351.082



## Double Exponential Smoothing for Monthly mean:

* NOTE * Zero values of Yt exist; MAPE calculated only for non-zero Yt.

### Method

Data Monthly mean Length
252

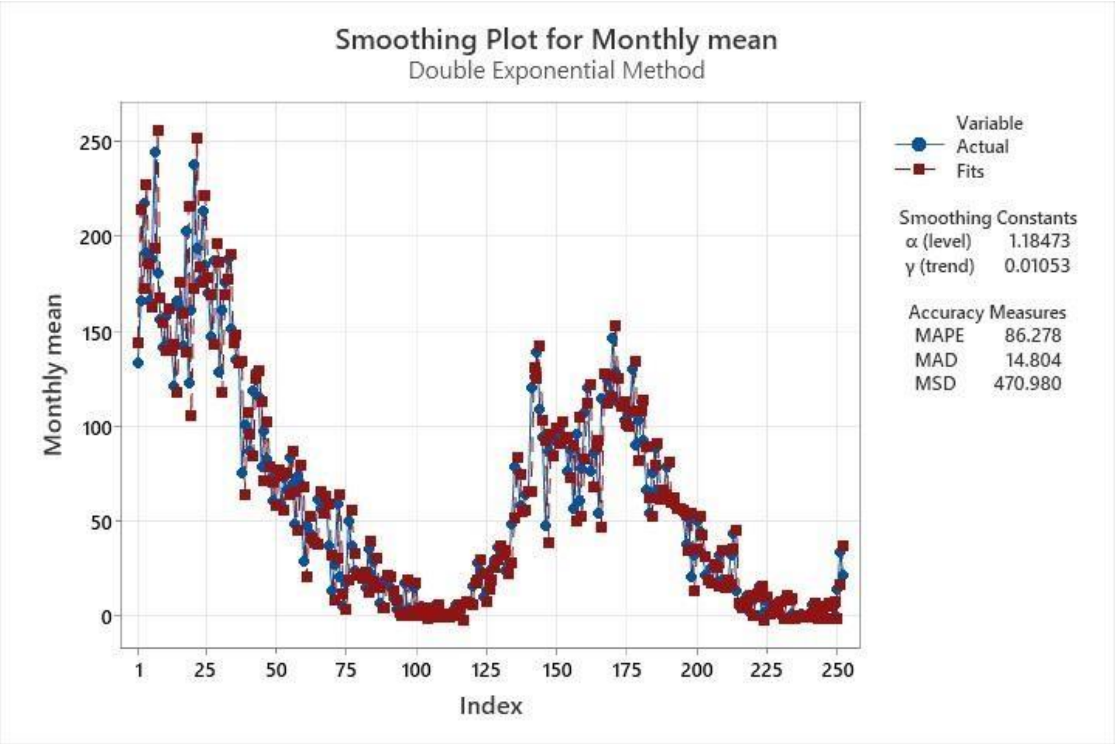
### Smoothing Constants α

1.18473

(level)
γ 0.01053
(trend)

### Accuracy Measures

MAPE 86.278
MAD 14.804
MSD 470.980



## Interpretation:

MAPE measures predicted accuracy as a proportion of actual values. a lower MAPE suggests improved accuracy.

MAD calculates the average absolute difference between forecasted and actual data. a lower MAD suggests improved precision

MSD calculates the average squared difference between forecasted and actual data. a lower MSD suggests improved accuracy and precision.

In terms of accuracy and precision, the single exponential technique has the lowest MAPE, MAD, and MSD. The single exponential smoothing method outperforms the double exponential smoothing method. Single exponential smoothing captures the trend in the data by assigning exponentially decreasing weights to older observations.

## Stationarity:

**Augmented Dickey-Fuller Test for Monthly mean:**

### Method

Maximum lag order for terms
15 in the regression model
Criterion for selecting lag order Minimum
AIC

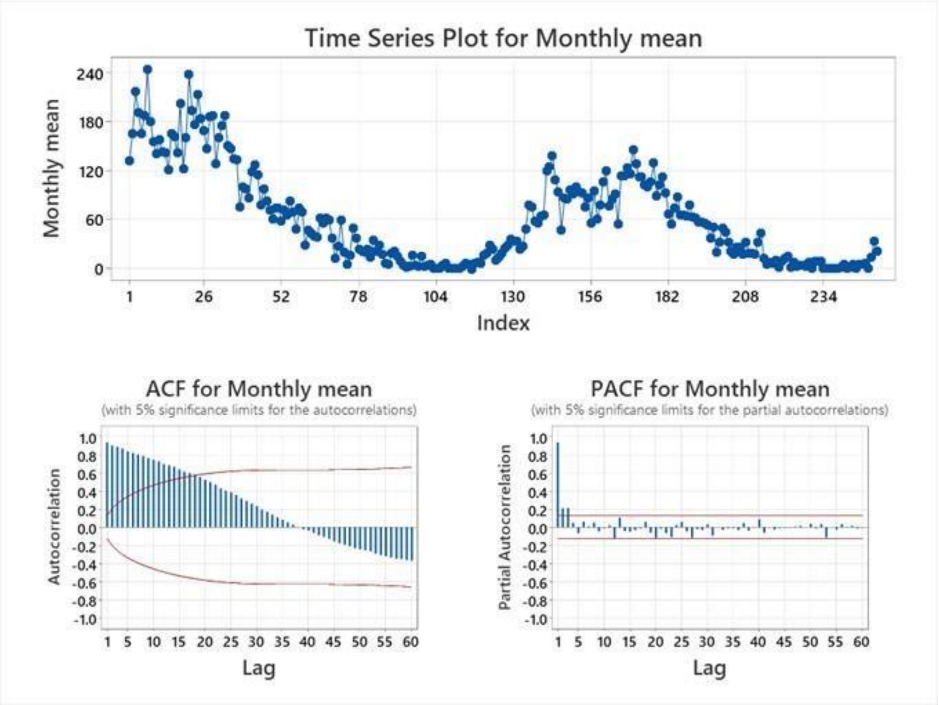
| | |
|---|---|
| Additional terms | Constant |
| Selected lag order | 11 |
| Rows used | 252 |

### Augmented Dickey-Fuller Test

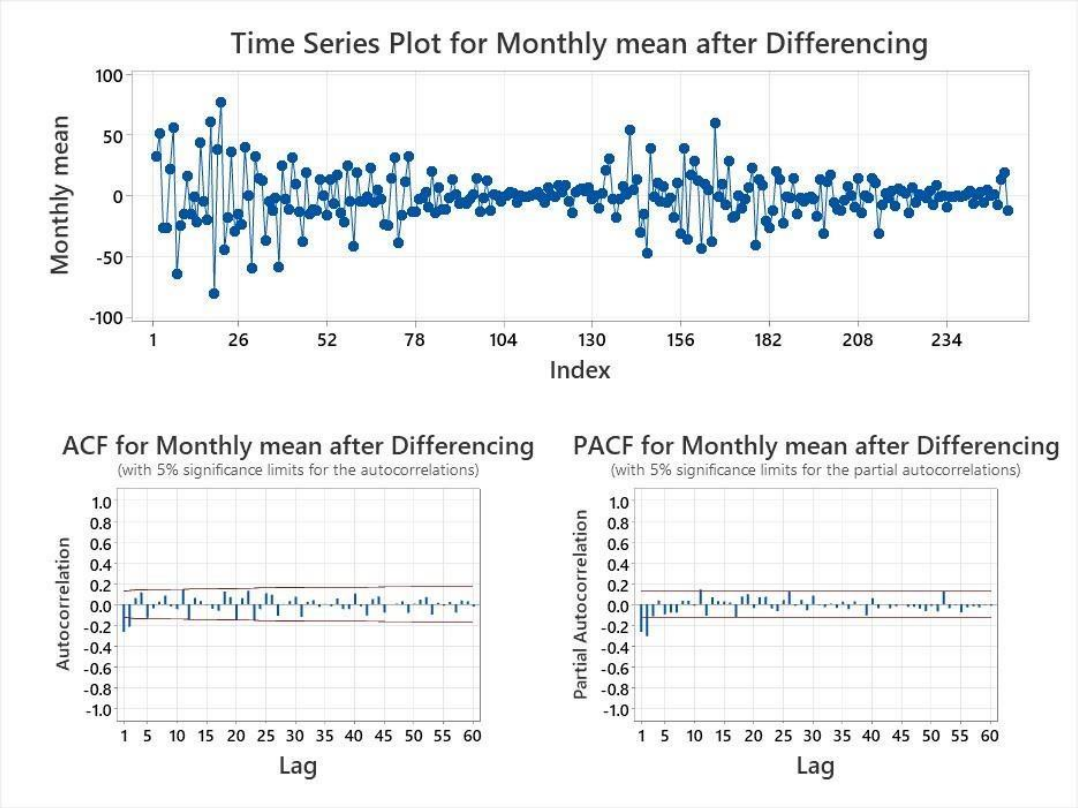
Null hypothesis: Data are nonstationary
Alternative Data are hypothesis:
stationary

**Test PStatistic Value Recommendation**

-
1.89950 0.332 Test statistic > critical value of -
2.87366.
Significance level = 0.05
Fail to reject null hypothesis.
Consider differencing to make
data stationary.



As we can see the p-value of the given test is 0.332 and Test Statistic is -1.89950, hence we accept the null-hypothesis which is that our data is non-stationary. Then we transform our data to stationary to find the ACF and PACF



Now after differencing the data becomes stationary so we get ahead with autocorrelation tests.

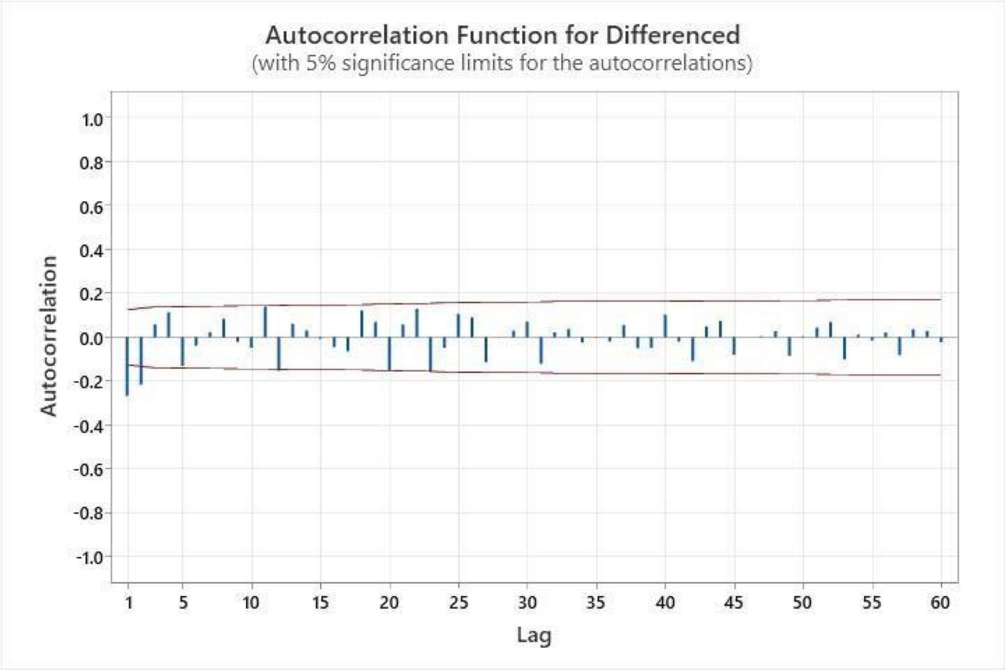
# ACF:

# Autocorrelation Function: Differenced

## Autocorrelations

| Lag | ACF | T | LBQ |
|---|---|---|---|
| 1 | -0.265817 | -4.21 | 17.95 |
| 2 | -0.216223 | -3.21 | 29.87 |
| 3 | 0.059333 | 0.85 | 30.77 |
| 4 | 0.113437 | 1.61 | 34.08 |
| 5 | -0.130070 | -1.83 | 38.45 |
| 6 | -0.038556 | -0.54 | 38.83 |
| 7 | 0.024048 | 0.33 | 38.98 |
| 8 | 0.083506 | 1.16 | 40.81 |
| 9 | -0.021128 | -0.29 | 40.92 |
| 10 | -0.047845 | -0.66 | 41.53 |
| 11 | 0.139671 | 1.92 | 46.69 |
| 12 | -0.152383 | -2.07 | 52.86 |
| 13 | 0.062021 | 0.83 | 53.88 |
| 14 | 0.031284 | 0.42 | 54.15 |
| 15 | -0.006307 | -0.08 | 54.16 |
| 16 | -0.044224 | -0.59 | 54.69 |
| 17 | -0.063325 | -0.84 | 55.77 |
| 18 | 0.122030 | 1.62 | 59.83 |
| 19 | 0.069793 | 0.91 | 61.17 |
| 20 | -0.151276 | -1.98 | 67.46 |
| 21 | 0.057994 | 0.75 | 68.39 |
| 22 | 0.129810 | 1.67 | 73.06 |
| 23 | -0.156979 | -1.99 | 79.92 |
| 24 | -0.048046 | -0.60 | 80.57 |
| 25 | 0.105645 | 1.32 | 83.70 |
| 26 | 0.090235 | 1.12 | 86.00 |
| 27 | -0.112280 | -1.38 | 89.58 |
| 28 | 0.001593 | 0.02 | 89.58 |
| 29 | 0.030226 | 0.37 | 89.84 |
| 30 | 0.071987 | 0.88 | 91.33 |
| 31 | -0.120688 | -1.47 | 95.53 |
| 32 | 0.022101 | 0.27 | 95.67 |

| | | | |
|---|---|---|---|
| 33 | 0.038285 | 0.46 | 96.10 |
| 34 | -0.024024 | -0.29 | 96.27 |
| 35 | 0.003270 | 0.04 | 96.27 |
| 36 | -0.018869 | -0.23 | 96.38 |
| 37 | 0.054786 | 0.66 | 97.27 |
| 38 | -0.048483 | -0.58 | 97.97 |
| 39 | -0.047803 | -0.58 | 98.65 |
| 40 | 0.103719 | 1.25 | 101.89 |
| 41 | -0.018288 | -0.22 | 101.99 |
| 42 | -0.107312 | -1.28 | 105.49 |
| 43 | 0.049334 | 0.59 | 106.23 |
| 44 | 0.074353 | 0.88 | 107.93 |
| 45 | -0.079736 | -0.94 | 109.89 |
| 46 | -0.000493 | -0.01 | 109.89 |
| 47 | 0.004198 | 0.05 | 109.90 |
| 48 | 0.028316 | 0.33 | 110.15 |
| 49 | -0.083310 | -0.98 | 112.33 |
| 50 | 0.005458 | 0.06 | 112.34 |
| 51 | 0.044462 | 0.52 | 112.97 |
| 52 | 0.070109 | 0.82 | 114.53 |
| 53 | -0.099478 | -1.16 | 117.71 |
| 54 | 0.013227 | 0.15 | 117.76 |
| 55 | -0.014934 | -0.17 | 117.84 |
| 56 | 0.022278 | 0.26 | 118.00 |
| 57 | -0.080210 | -0.93 | 120.10 |
| 58 | 0.035942 | 0.42 | 120.53 |
| 59 | 0.028701 | 0.33 | 120.80 |
| 60 | -0.022022 | -0.25 | 120.96 |



## Interpretation:

The Autocorrelation Function (ACF) plot shows the correlation between a time series and its lagged values. In other words, it helps to identify the relationship between each observation in a time series and its past observations at various lag intervals.

- This graphic shows the partial correlation coefficients between the series and its lagged values, while accounting for the values of the time series at all shorter delays.

- The majority of the spikes fall inside the confidence interval (blue horizontal lines), indicating that there are no significant relationships at any lag. This suggests that the first differencing may have eliminated any autocorrelation in the data.

# PACF:

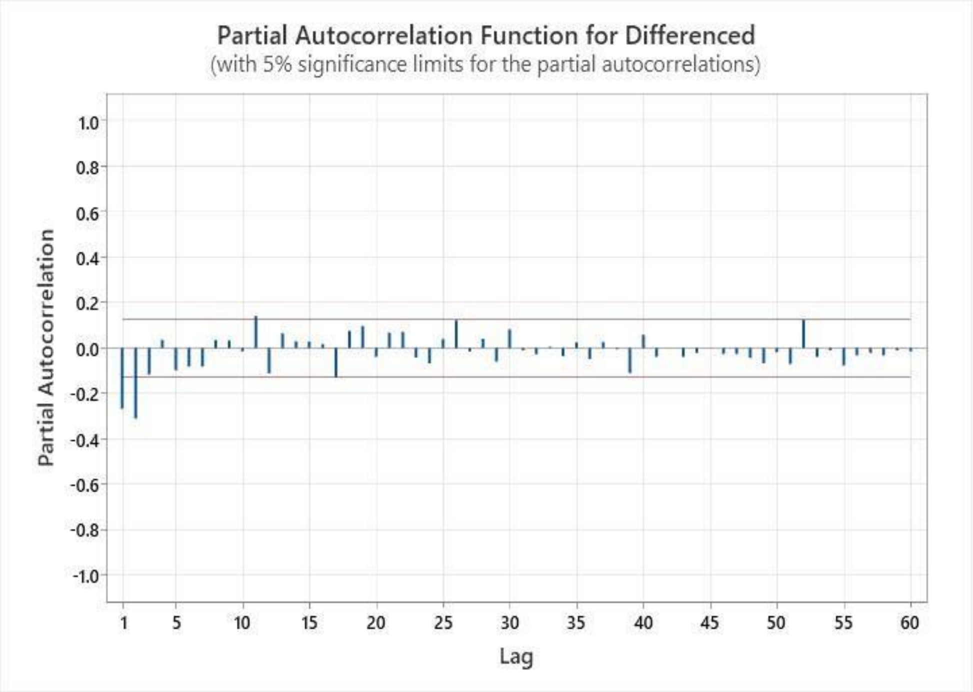
## Partial Autocorrelation Function: Differenced

### Partial Autocorrelations

| Lag | PACF | T |
|---|---|---|
| 1 | -0.265817 | -4.21 |
| 2 | -0.308694 | -4.89 |
| 3 | -0.116489 | -1.85 |
| 4 | 0.035855 | 0.57 |
| 5 | -0.096229 | -1.52 |
| 6 | -0.080282 | -1.27 |

| | | |
|---|---|---|
| 7 | -0.081278 | -1.29 |
| 8 | 0.034431 | 0.55 |
| 9 | 0.032734 | 0.52 |
| 10 | -0.013665 | -0.22 |
| 11 | 0.141726 | 2.25 |
| 12 | -0.110724 | -1.75 |
| 13 | 0.065476 | 1.04 |
| 14 | 0.030218 | 0.48 |
| 15 | 0.028745 | 0.46 |
| 16 | 0.017134 | 0.27 |
| 17 | -0.128880 | -2.04 |
| 18 | 0.075702 | 1.20 |
| 19 | 0.096402 | 1.53 |
| 20 | -0.037345 | -0.59 |
| 21 | 0.067770 | 1.07 |
| 22 | 0.071471 | 1.13 |
| 23 | -0.041415 | -0.66 |
| 24 | -0.066530 | -1.05 |
| 25 | 0.039229 | 0.62 |
| 26 | 0.122784 | 1.95 |
| 27 | -0.013635 | -0.22 |
| 28 | 0.040573 | 0.64 |
| 29 | -0.058171 | -0.92 |
| 30 | 0.083212 | 1.32 |
| 31 | -0.008792 | -0.14 |
| 32 | -0.027156 | -0.43 |
| 33 | 0.005994 | 0.09 |
| 34 | -0.035069 | -0.56 |
| 35 | 0.024601 | 0.39 |
| 36 | -0.047425 | -0.75 |
| 37 | 0.027009 | 0.43 |
| 38 | -0.005360 | -0.08 |

| | | |
|---|---|---|
| 39 | -0.108786 | -1.72 |
| 40 | 0.058080 | 0.92 |
| 41 | -0.037595 | -0.60 |
| 42 | 0.001101 | 0.02 |
| 43 | -0.038837 | -0.62 |
| 44 | -0.020145 | -0.32 |
| 45 | -0.002247 | -0.04 |
| 46 | -0.025316 | -0.40 |
| 47 | -0.026130 | -0.41 |
| 48 | -0.042584 | -0.67 |
| 49 | -0.067105 | -1.06 |
| 50 | -0.016754 | -0.27 |
| 51 | -0.070601 | -1.12 |
| 52 | 0.124409 | 1.97 |
| 53 | -0.037884 | -0.60 |
| 54 | -0.009135 | -0.14 |
| 55 | -0.076546 | -1.21 |
| 56 | -0.031283 | -0.50 |
| 57 | -0.020690 | -0.33 |
| 58 | -0.031145 | -0.49 |
| 59 | -0.009052 | -0.14 |



## Interpretation:

The Partial Autocorrelation Function (PACF) plot shows the partial correlation between a time series and its lagged values, accounting for the influence of intermediate lags. In essence, the PACF plot helps to identify the direct relationship between each observation in a time series and its past observations, while controlling for the effects of intervening lags.

Similar to the PACF plot, most of the autocorrelation coefficients are within the 5% significance limits, suggesting that the first differencing has been effective in making the series stationary. second-order differencing, or twofold differencing, is a technique used to stabilize the variance of a time series by removing trends and seasonality, and because the dataset that we have chosen has no seasonality, we do not apply second-order smoothing to it.

# Forecast with Best ARIMA Model for Monthly mean

ARIMA stands for Autoregressive Integrated Moving Averages. Basically, after differencing the data, we can predict the further observation which will be more significant rather than doing direct forecasting on the non-stationary data. For that we need to determine three parameters; p, d and q.

Where, p = order of the auto-regressive part, which indicates how many past values are used in the model

d = degree of differencing, which indicates how many times the data are differenced to make them stationary
q = order of the moving average part, which indicates how many past errors are used in the model

## Method

Criterion for Minimum best
model AICc Rows used
252
Rows unused 0

## Model Selection

| Model (d = 1) | LogLikelihood | AICc | AIC | BIC |
|---|---|---|---|---|
| p = 0, q =2* | -1087.69 | 2183.55 | 2183.39 | 2197.49 |
| p = 2, q = 1 | -1087.29 | 2184.82 | 2184.58 | 2202.20 |
| p = 1, q = 2 | -1087.31 | 2184.87 | 2184.63 | 2202.25 |
| p = 2, q = 2 | -1086.58 | 2185.51 | 2185.16 | 2206.32 |
| p = 2, q = 0 | -1088.69 | 2185.54 | 2185.37 | 2199.48 |
| p = 1, q = 1 | -1089.14 | 2186.44 | 2186.28 | 2200.38 |
| p = 0, q = 1 | -1091.93 | 2189.95 | 2189.85 | 2200.43 |
| p = 1, q = 0 | -1102.40 | 2210.90 | 2210.80 | 2221.37 |
| p = 0, q = 0 | -1111.67 | 2227.39 | 2227.34 | 2234.39 |

The ARIMA (0,1,2) model, with differencing parameter d=1, autoregressive order p=0, and moving average order q=2, has been shown to be the best fit for forecasting the monthly mean total sunspots number.

The AICc values of the model (0,1,2) is lowest compare to the other models. **Therefore, ARIMA(0,1,2) is best fitted model for the data.**

** Best model with minimum AICc. Output for the best model follows.*

## Final Estimates of Parameters

| | | SE | T- | P- |
|---|---|---|---|---|
| Type | Coef | Coef | Value | Value |
| MA1 | 0.3945 | 0.0620 | 6.36 | 0.000 |
| MA 2 | 0.1949 | 0.0623 | 3.13 | 0.002 |
| Constant | -0.596 | 0.479 | -1.24 | 0.215 |

*Differencing: 1 Regular*

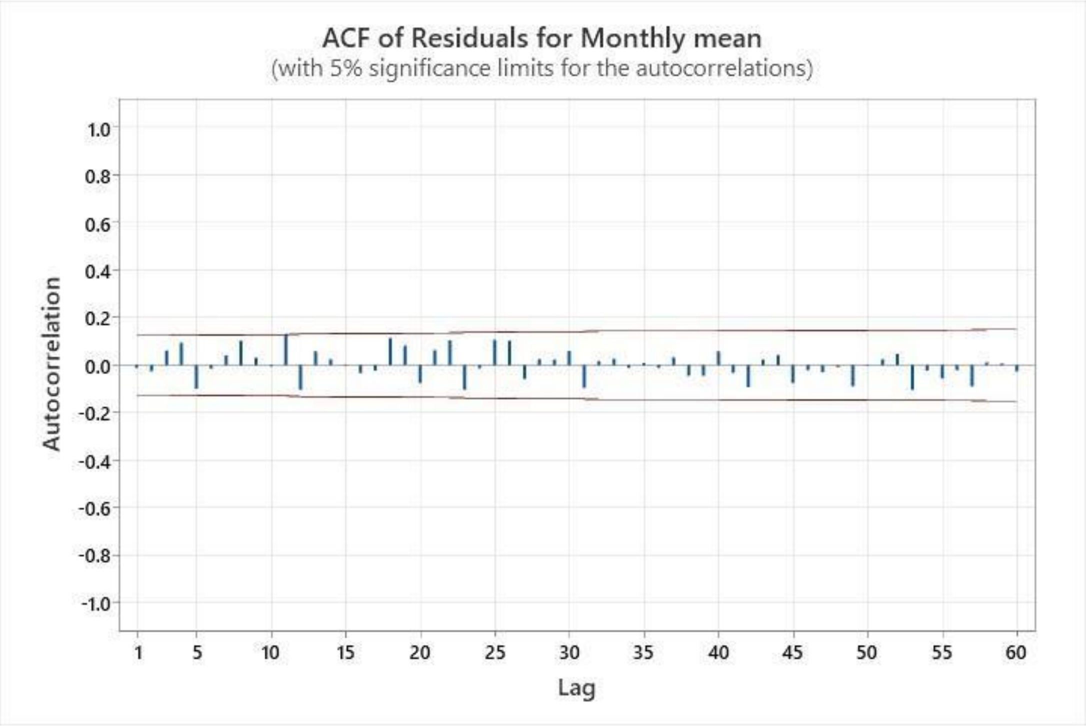
## Model Summary

| DF | SS | MS | MSD | AICc | AIC | BIC |
|---|---|---|---|---|---|---|
| 248 | 84343.4 | 340.094 | 336.029 | 2183.55 | 2183.39 | 2197.49 |

*MS = variance of the white noise series*

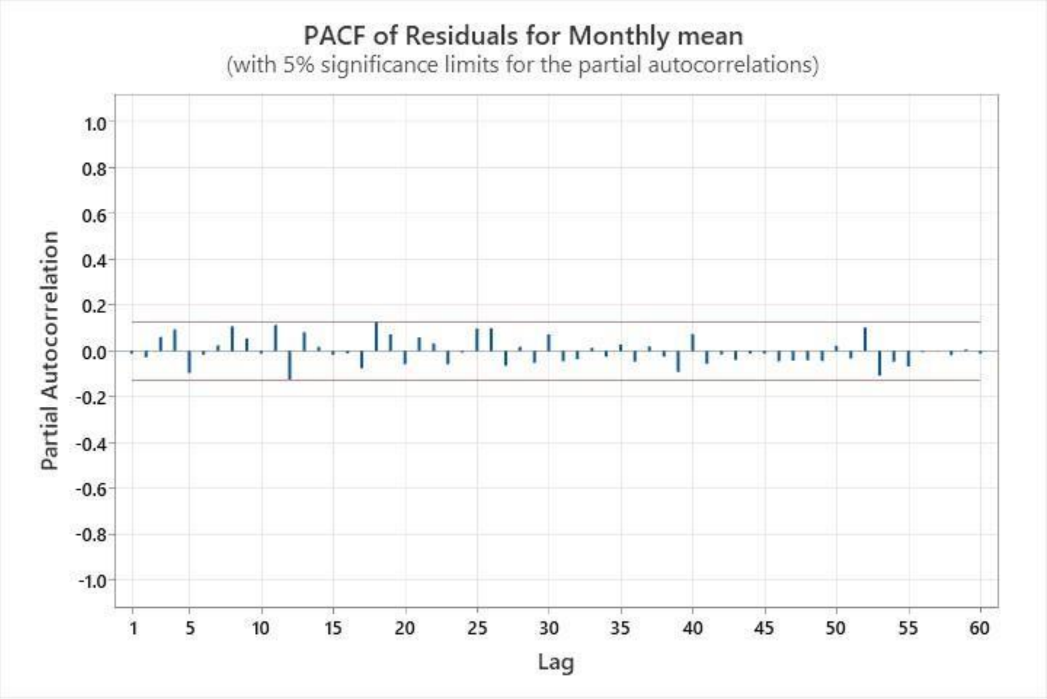
## Modified Box-Pierce (Ljung-Box) Chi-Square Statistic

|  | 12 | 24 | 36 | 48 | Lag |
|---|---|---|---|---|---|
| Chi-Square | 16.85 | 32.55 | 44.04 | 52.34 |  |
| DF | 9 | 21 | 33 | 45 |  |
| P-Value | 0.051 | 0.051 | 0.095 | 0.210 |  |



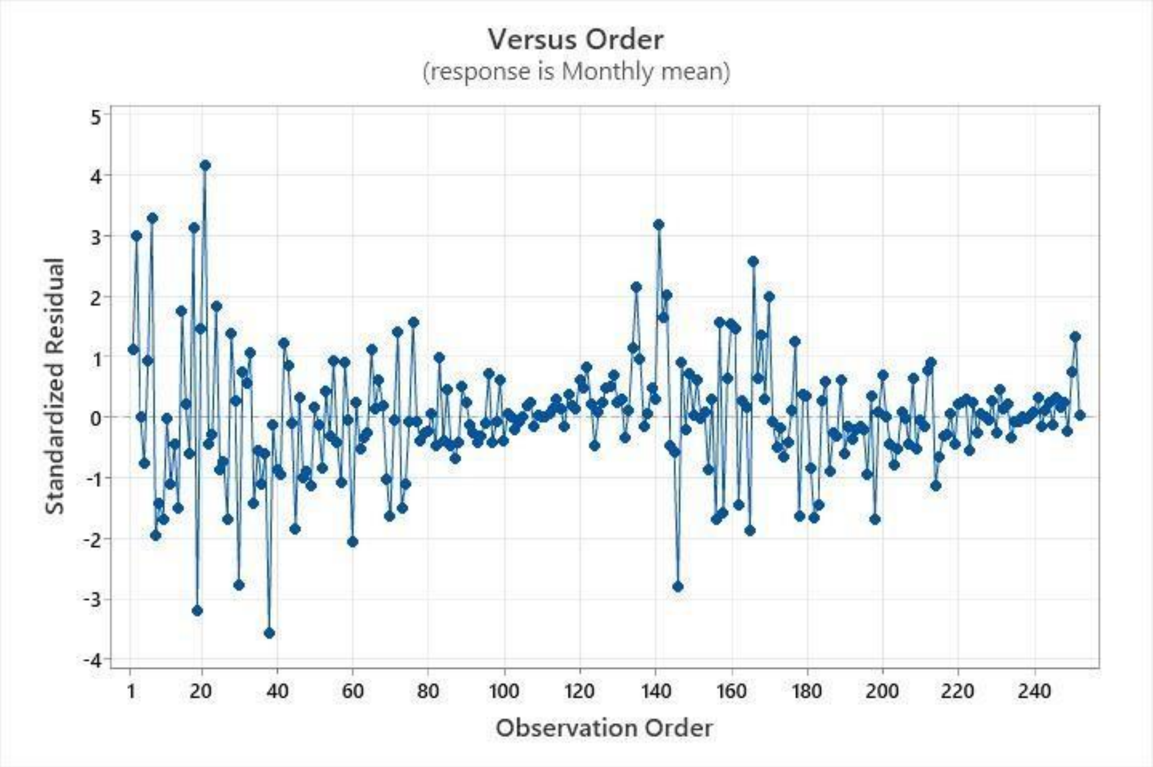
## Interpretation:

The ACF is showing no significant autocorrelation at any lag, i.e all autocorrelations are falling within the confidence interval, indicating randomness. Hence it is a White Noise Process.



## Interpretation:

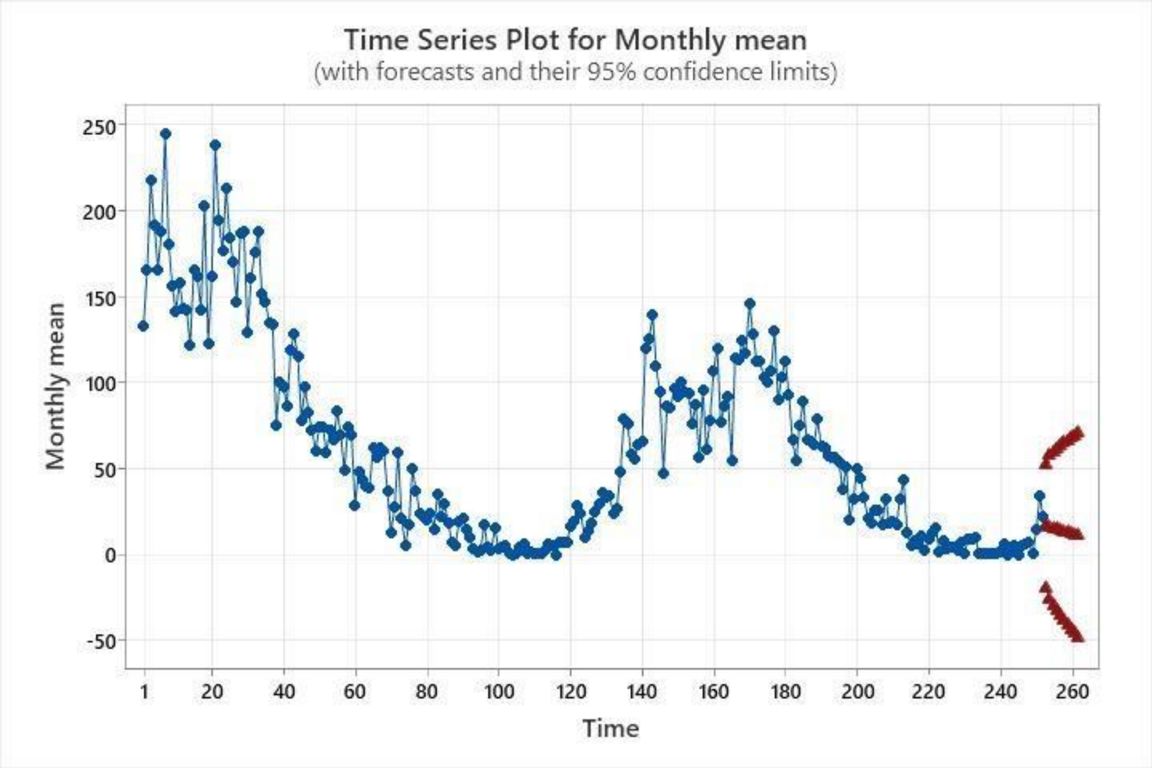
Similar to the ACF, PACF is showing no significant autocorrelation at any lag. Hence it is a white noise process



# Forecasting future values on the best fitted values:

## Forecasts from Time Period 252

| | | 95% Limits | | Time | SE |
|---|---|---|---|---|---|
| **Period** | **Forecast** | **Forecast** | **Lower** | **Upper** | **Actual** |
| 253 | 16.0378 | 18.4416 | -20.1151 | 52.1907 | |
| 254 | 15.2774 | 21.5589 | -26.9866 | 57.5413 | |
| 255 | 14.6813 | 22.8503 | -30.1143 | 59.4770 | |
| 256 | 14.0853 | 24.0726 | -33.1064 | 61.2770 | |
| 257 | 13.4893 | 25.2357 | -35.9826 | 62.9611 | |
| 258 | 12.8932 | 26.3475 | -38.7583 | 64.5447 | |
| 259 | 12.2972 | 27.4143 | -41.4456 | 66.0399 | |
| 260 | 11.7011 | 28.4410 | -44.0545 | 67.4568 | |
| 261 | 11.1051 | 29.4320 | -46.5933 | 68.8034 | |
| 262 | 10.5091 | 30.3907 | -49.0687 | 70.0868 | |



## Interpretation:

This plot provides a visual representation of the trends and variability in the monthly mean data, as well as the uncertainty associated with forecasting future values.

## Conclusion:

Software Use- We have used Minitab as it is a user-friendly statistical software. It is very easy to analyse the time series data using Minitab.

**In conclusion,** single exponential smoothing offers valuable insights into sunspot activity by capturing trends, providing forecasts, and estimating smoothing levels. However, its effectiveness may be limited by its inability to handle complex seasonal patterns. We use the Quadratic trend due to the fact that it covers most of the data points when compared with other types of trends.