Time series Analysis & Hypothesis Testing with on Precipitation data for Pune City (1965-2002).

Abstract :

Statistical Analysis on time series precipitation data of the city.

Submitted by :
Manpreet Singh (A21017)

Problem statement

Meteorological datasets for yearly aggregated precipitation values for Pune districts of India from 1965 to 2002. Data is used to analyse the temporal distribution of annual precipitation across the district and predict rainfall patterns. Furthermore, differences in the warm and cold season distributions are presented.

Introduction

The objective of this study is to do Statistical analysis on the time variant data to find various insights and predict the future values based on historical data.

Dataset

Climate parameters included are: Precipitation (millimeters (mm).

Number of Years	37 years
Variable Description	Precipitation (mm)
Period of measurement	Monthly

Tools used

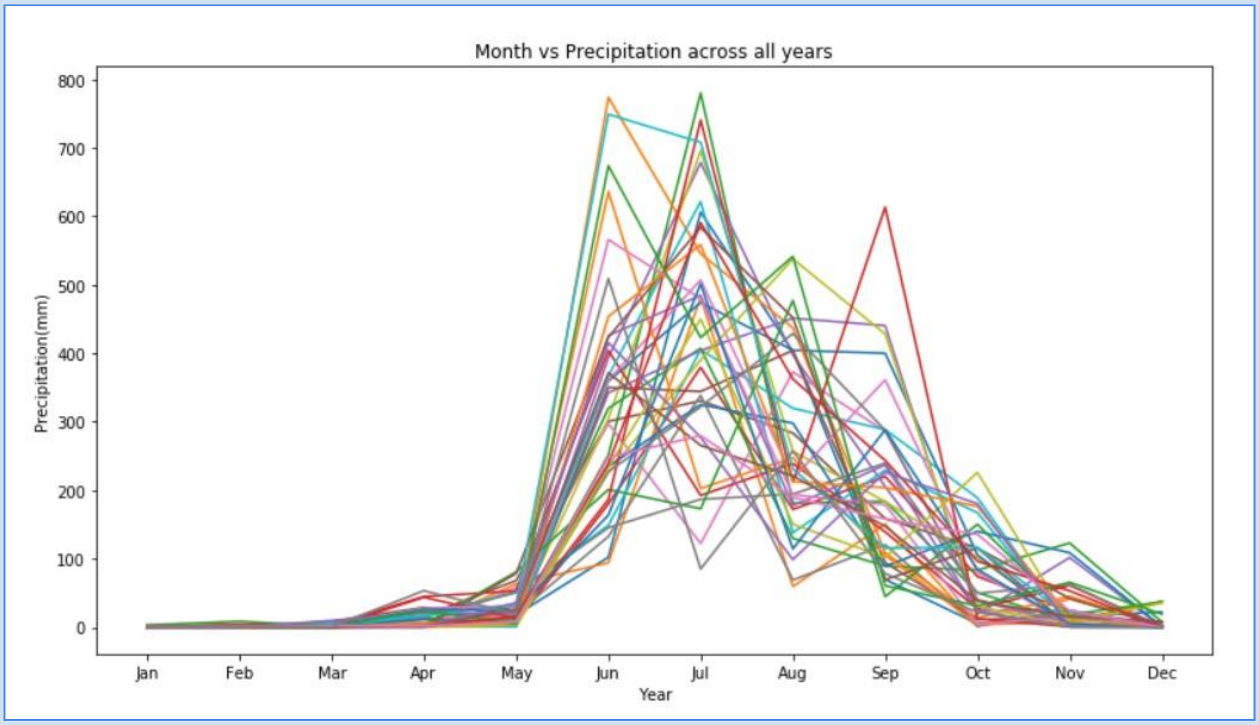
Python, MS Excel & Power BI etc.

Exploratory Data Analysis

Graphs and Insights :

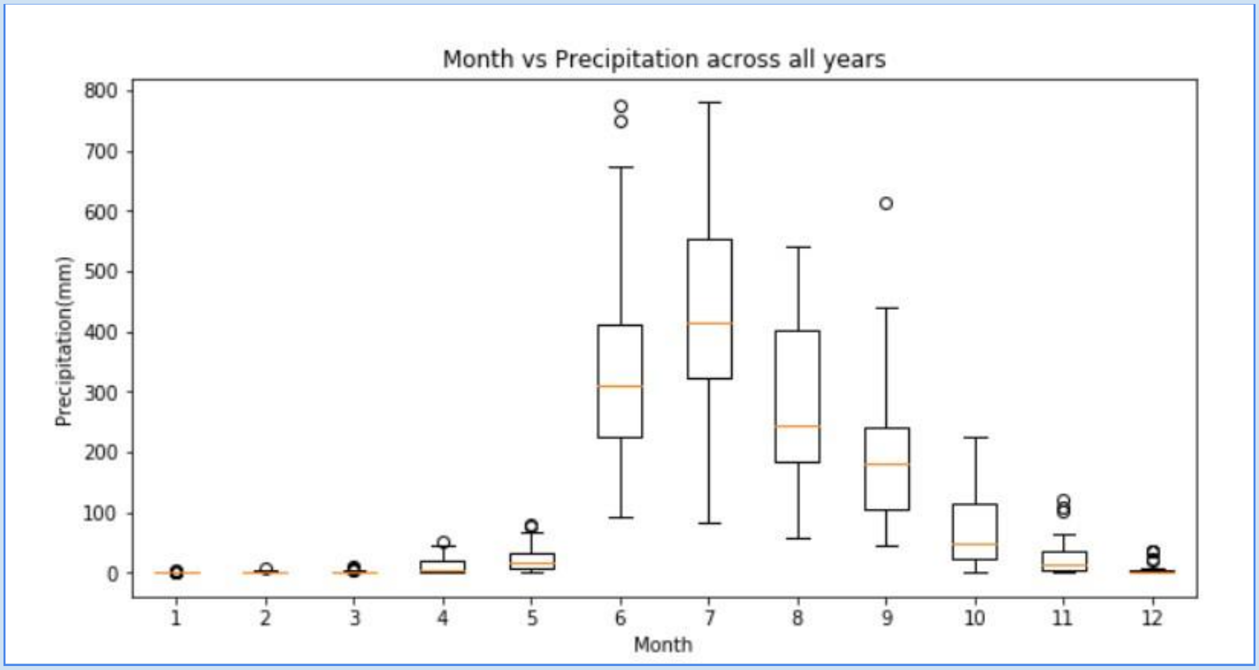
Month vs Precipitation across all years

Line Plot



Line plot for yearly data for all months

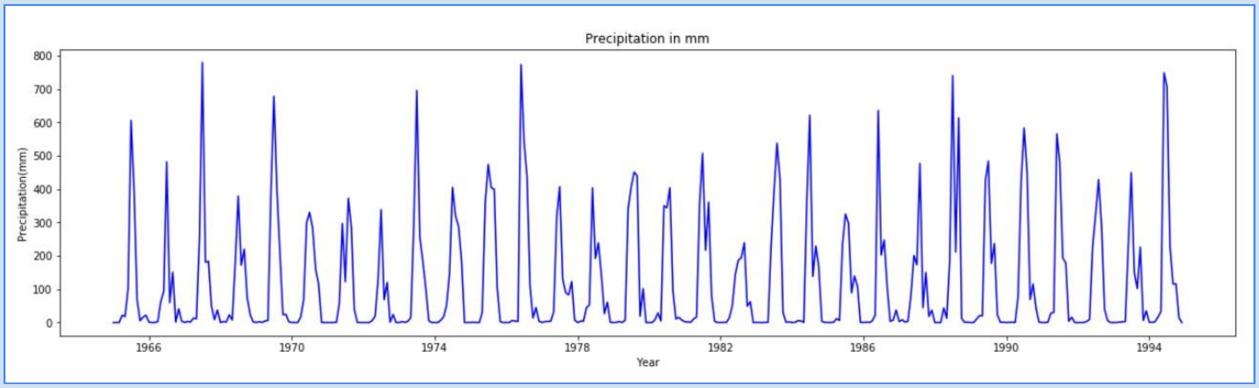
Box Plot



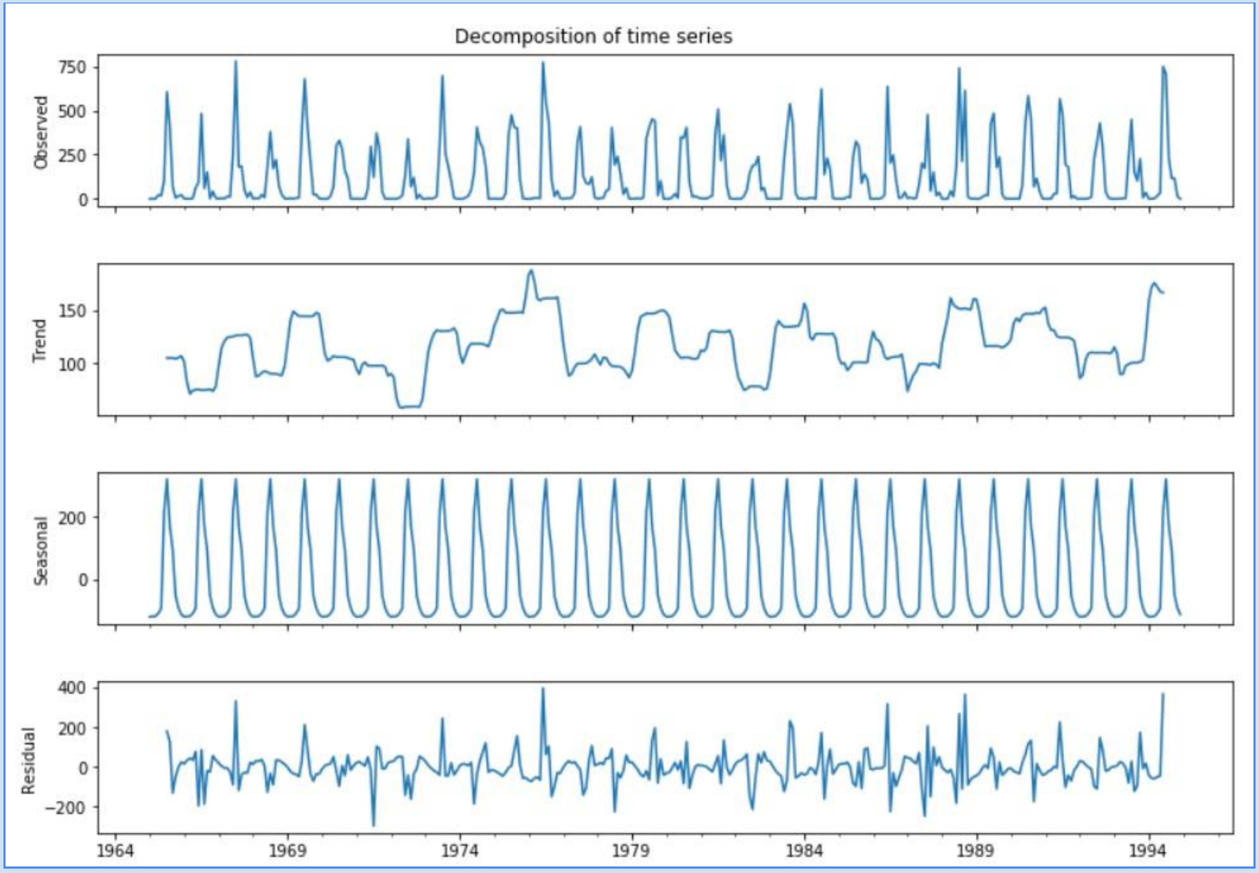
Insights

- The rainfall in the months November, December, January, February, March and April is very less.
- The rainfall in the months June, July and August are high compared to rainfall in other months of the year.
- We can observe the seasonality effect.

Precipitation in mm



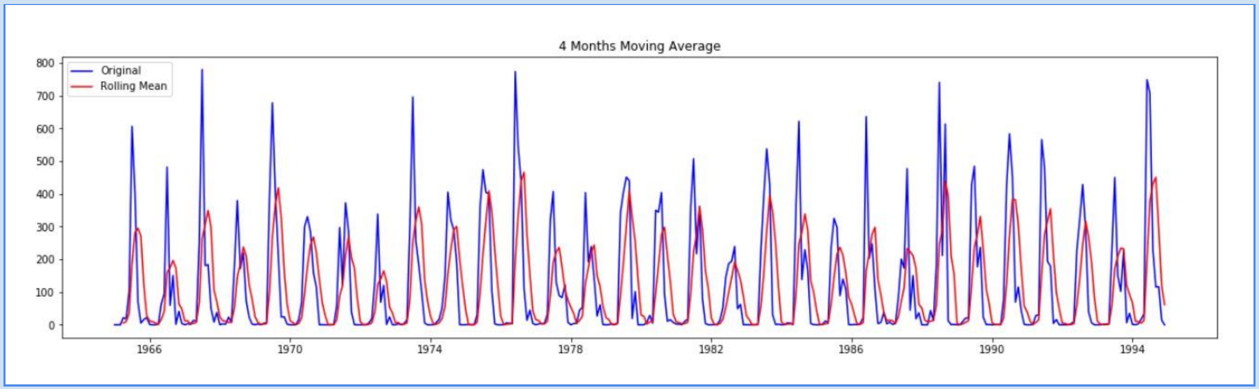
Decomposition of time series

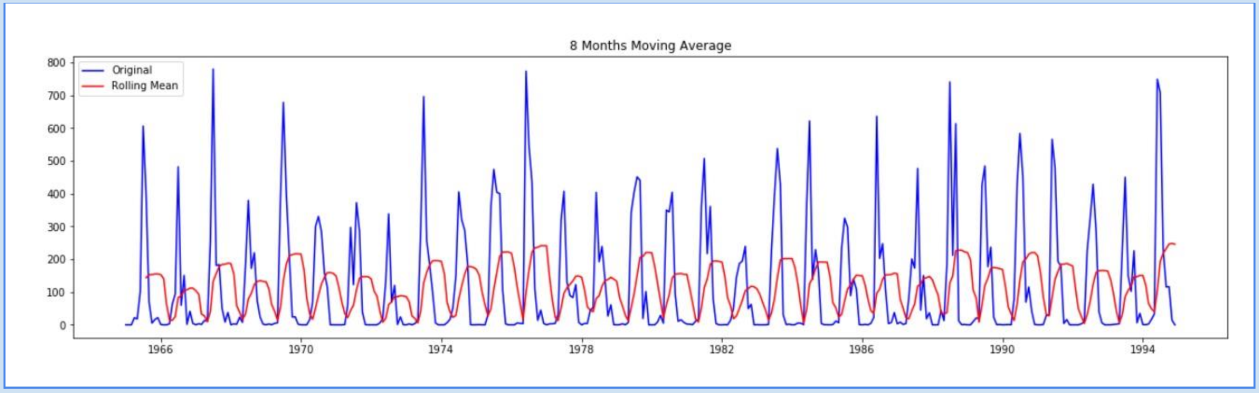


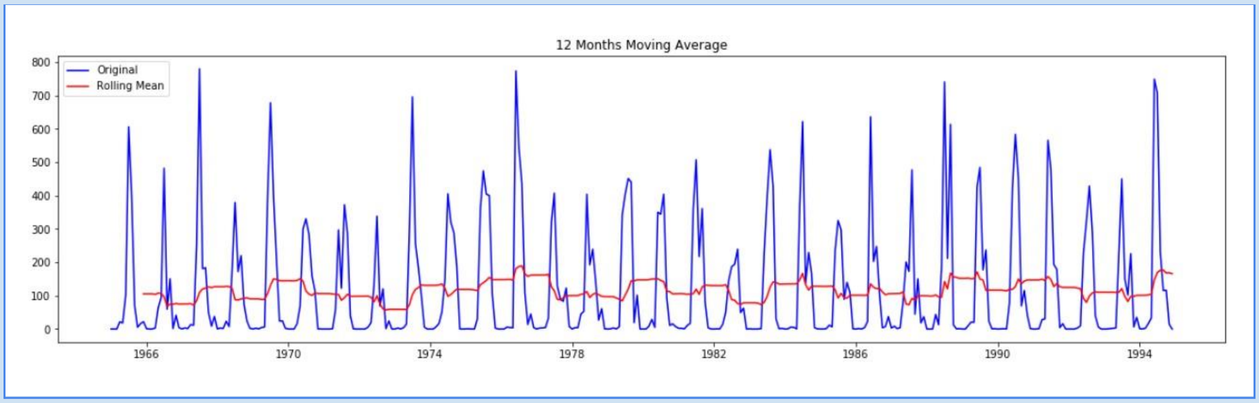
Seasonality window

Moving Averages

4,8 and 12 month moving averages:







Insights from Moving average

- As we could see in the above plots, the 12-month moving average could produce a wrinkle free curve when compared to other moving averages.
- Therefore, s=12.

- This is to find the period of seasonality.

Hypothesis Testing

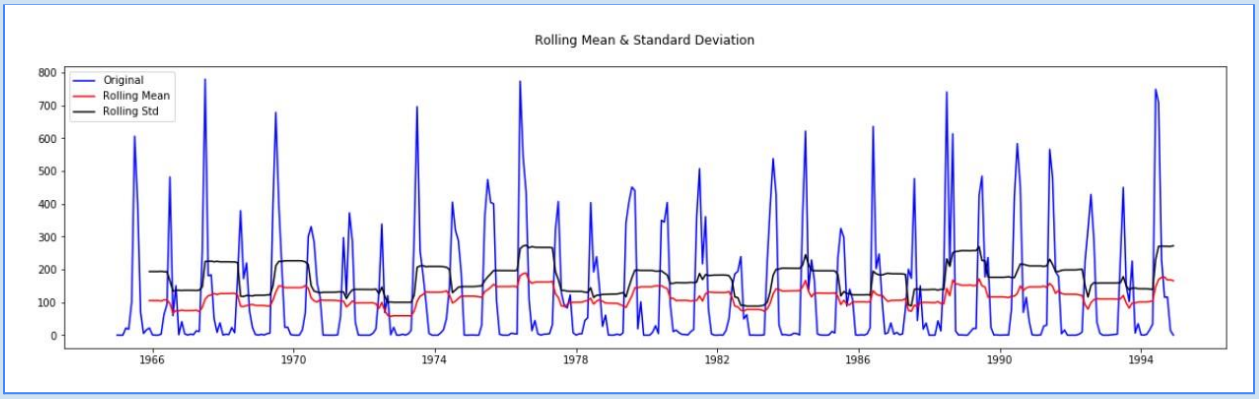
Next step would be to check for Stationarity

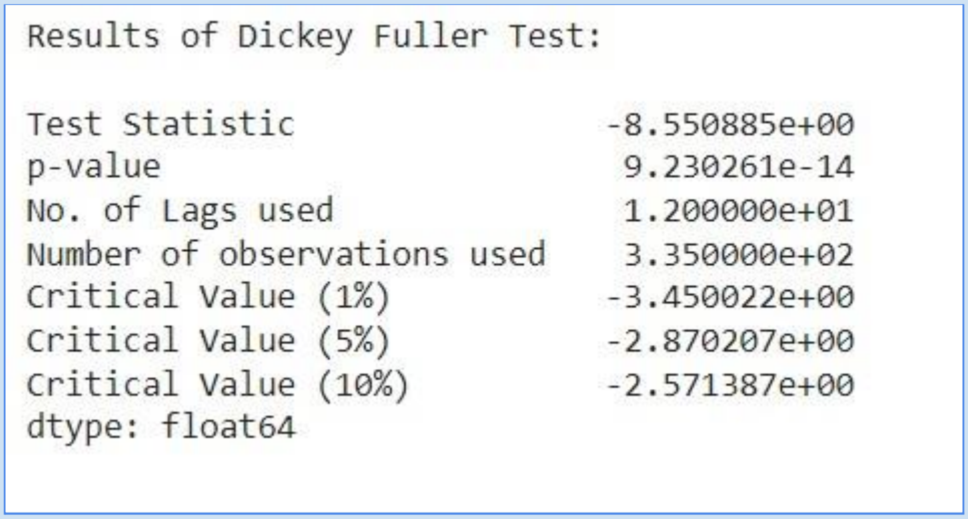
Dickey Fuller test

This is one of the statistical tests for checking stationarity. Critical Values are for different confidence levels. If the ‘Test Statistic’ is less than the ‘Critical Value’, we can reject the null hypothesis and say that the series is stationary.

Null hypothesis : Time series is non-stationary.

Alternate Hypothesis : Time series is stationary.





```
Results of Dickey Fuller Test:

Test Statistic                  -8.550885e+00
p-value                          9.230261e-14
No. of Lags used                 1.200000e+01
Number of observations used      3.350000e+02
Critical Value (1%)             -3.450022e+00
Critical Value (5%)             -2.870207e+00
Critical Value (10%)            -2.571387e+00
dtype: float64
```

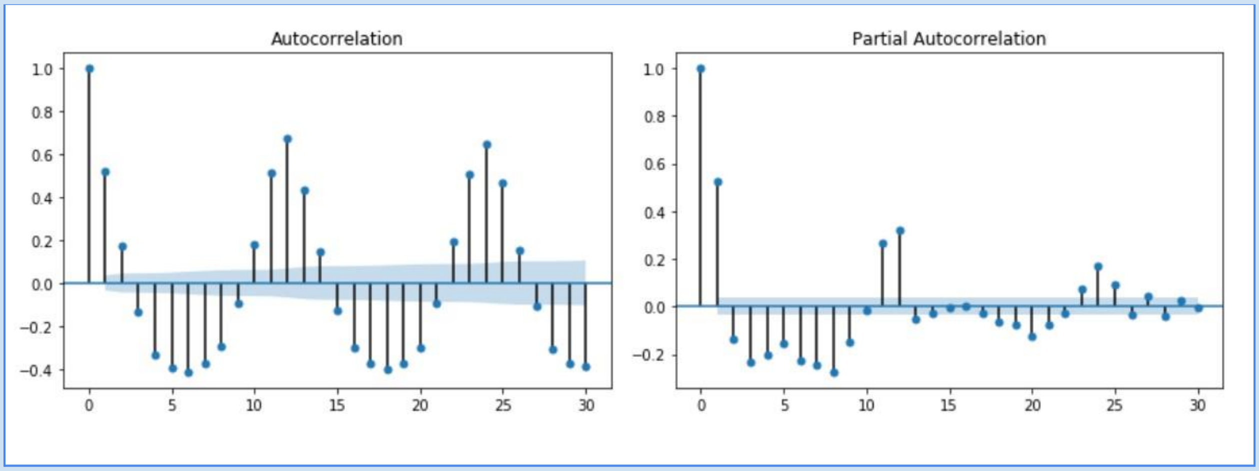
Analysis :

- As we could see, the p-value is very less. Also, "Test statistic" is less compared to "Critical Value".
- Therefore, Null hypothesis is rejected, which means, Time series is stationary.

Conclusion : As time series is stationary, differencing is not required.

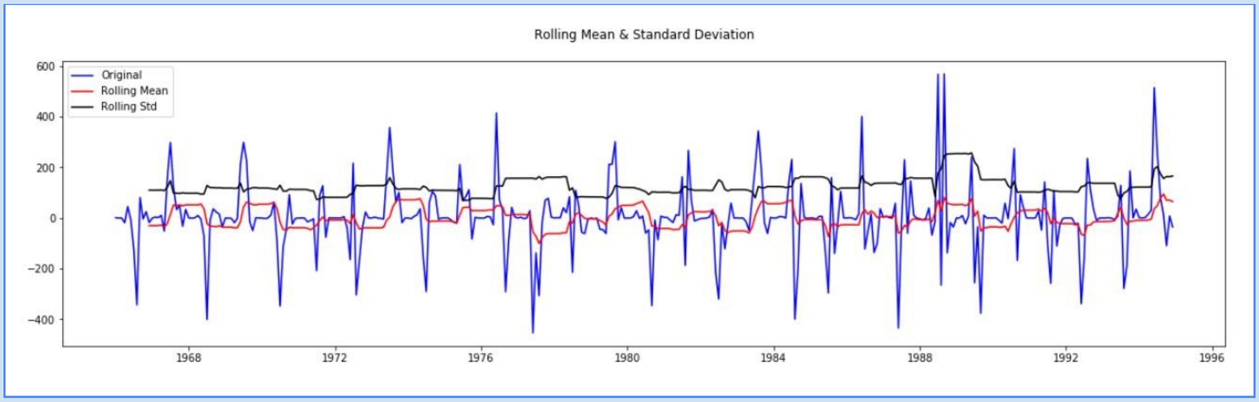
Next Steps : we have to find p and q values by plotting ACF and PACF plots.

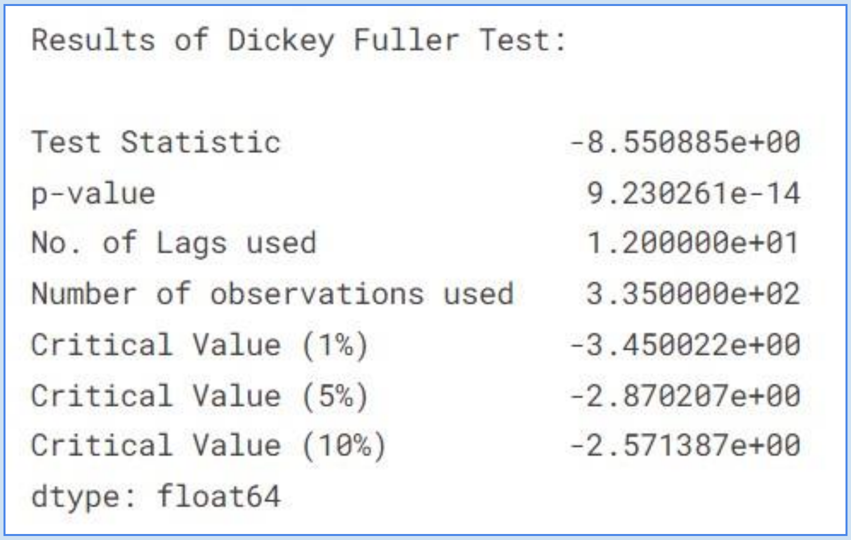
ACF & PACF Plot :



Inference : we could see, there is a seasonality effect.

Remedial : Applying differencing D= 1 and again performing ADF test





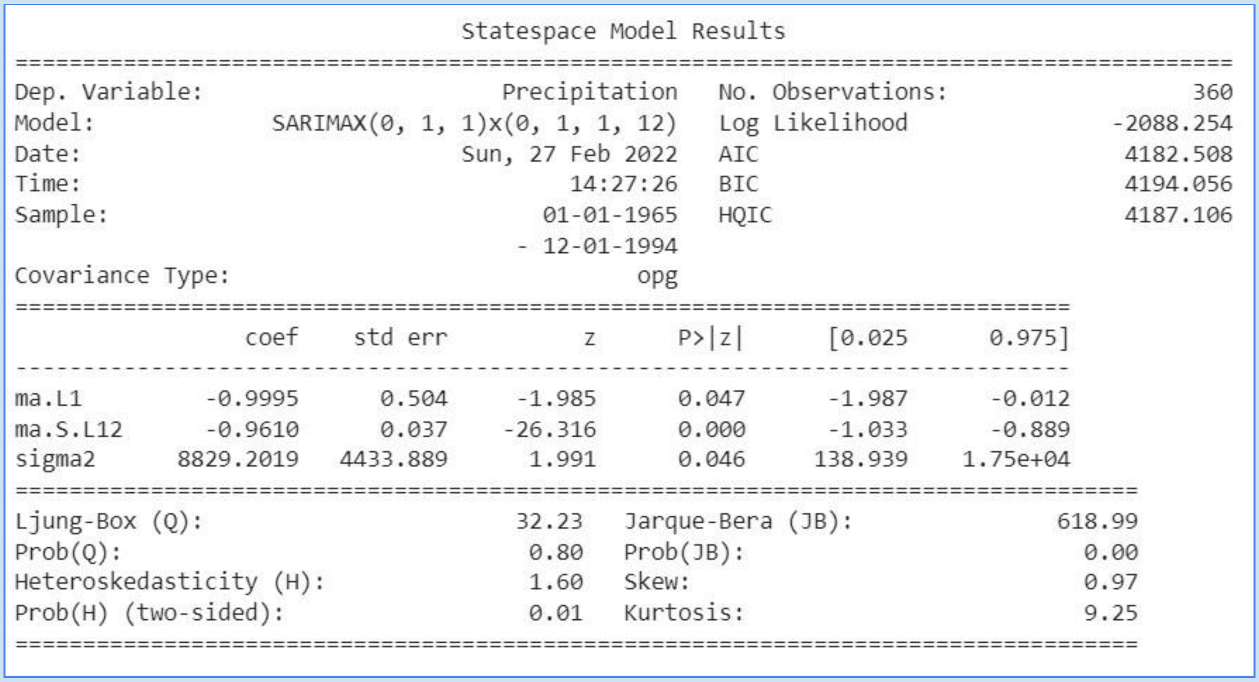
```
Results of Dickey Fuller Test:

Test Statistic                -8.550885e+00
p-value                        9.230261e-14
No. of Lags used               1.200000e+01
Number of observations used    3.350000e+02
Critical Value (1%)           -3.450022e+00
Critical Value (5%)           -2.870207e+00
Critical Value (10%)          -2.571387e+00
dtype: float64
```

- If there is seasonality, it will be better if we try all combinations of different parameters and choose the best set of parameters that gives less AIC score. Parameters - p, d, q
- p=0, d= 1, q=1
- 12 month moving average could produce wrinkle free curve s=12

Time series analysis /prediction

Using the best values of p ,d ,q we build our model.

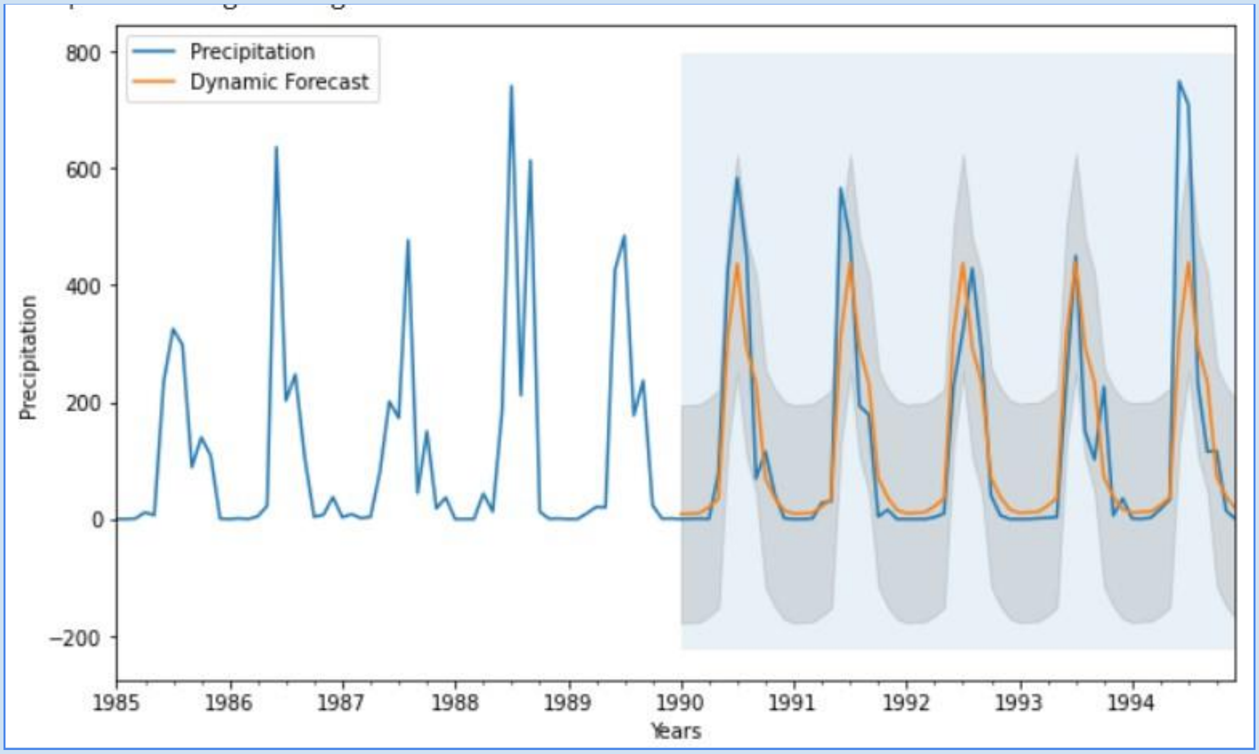
Mode Interpretation



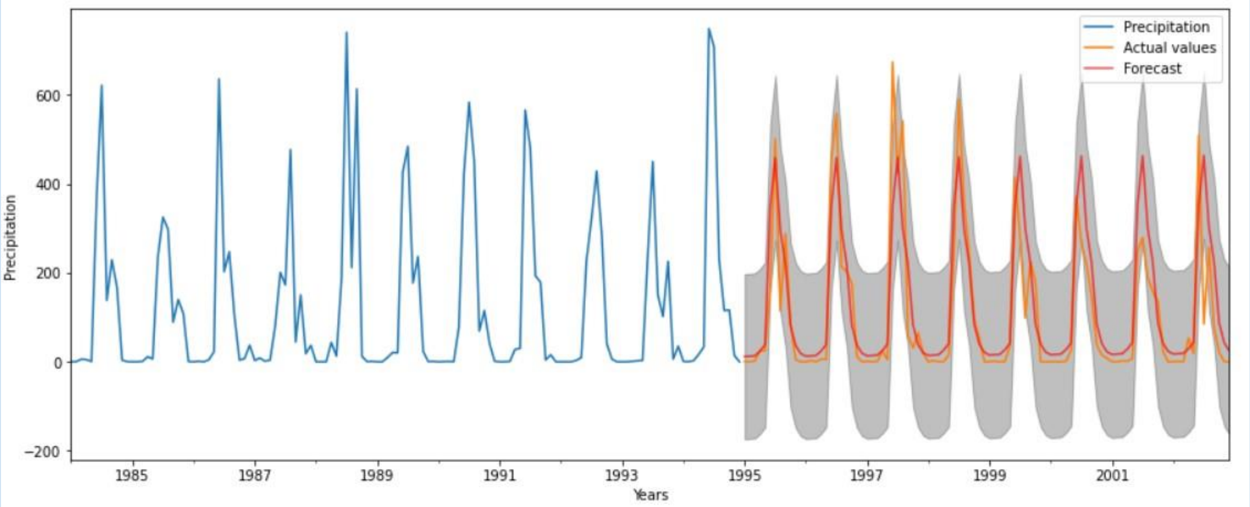
```
                                 Statespace Model Results
==========================================================================================
Dep. Variable:                    Precipitation   No. Observations:                    360
Model:             SARIMAX(0, 1, 1)x(0, 1, 1, 12)   Log Likelihood               -2088.254
Date:                          Sun, 27 Feb 2022   AIC                           4182.508
Time:                                  14:27:26   BIC                           4194.056
Sample:                              01-01-1965   HQIC                          4187.106
                                   - 12-01-1994
Covariance Type:                            opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.9995      0.504     -1.985      0.047      -1.987      -0.012
ma.S.L12      -0.9610      0.037    -26.316      0.000      -1.033      -0.889
sigma2      8829.2019   4433.889      1.991      0.046     138.939    1.75e+04
===================================================================================
Ljung-Box (Q):                       32.23   Jarque-Bera (JB):                618.99
Prob(Q):                              0.80   Prob(JB):                          0.00
Heteroskedasticity (H):               1.60   Skew:                              0.97
Prob(H) (two-sided):                  0.01   Kurtosis:                          9.25
===================================================================================
```

By selecting the best parameters for p,d,q the model is created.

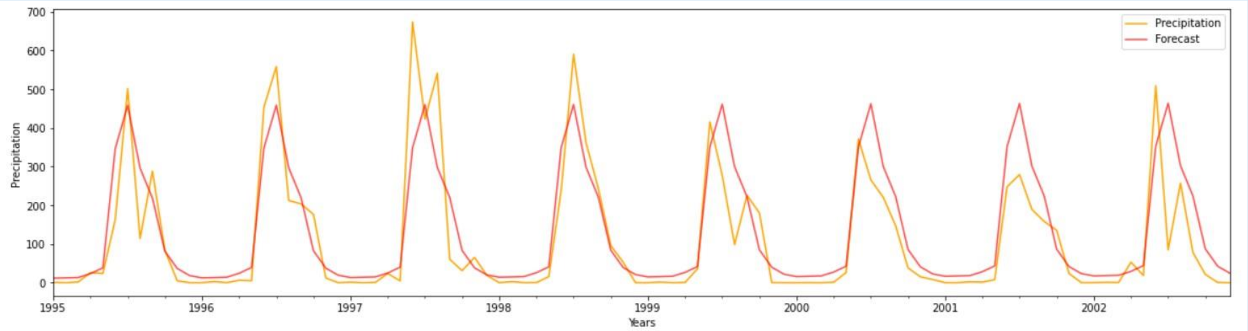
Information criteria, including AIC BIC helps us in deciding the best model built and significance of each variable in the model holds good.



Inference : We have made a forecast for next 7 years and compared it with our test values.



Inference : Plot the forecast along with confidence band.



Inference : Zoomed view of our prediction.

Conclusion and Insights

We have analyzed the change of data over time and designed a forecasting model to identify trends and predict future values of precipitation based on past values. We can take measures to make changes in Crop characters, identify ground water contribution, land and soil characteristics of that region to boost agriculture in those regions and help farmers.