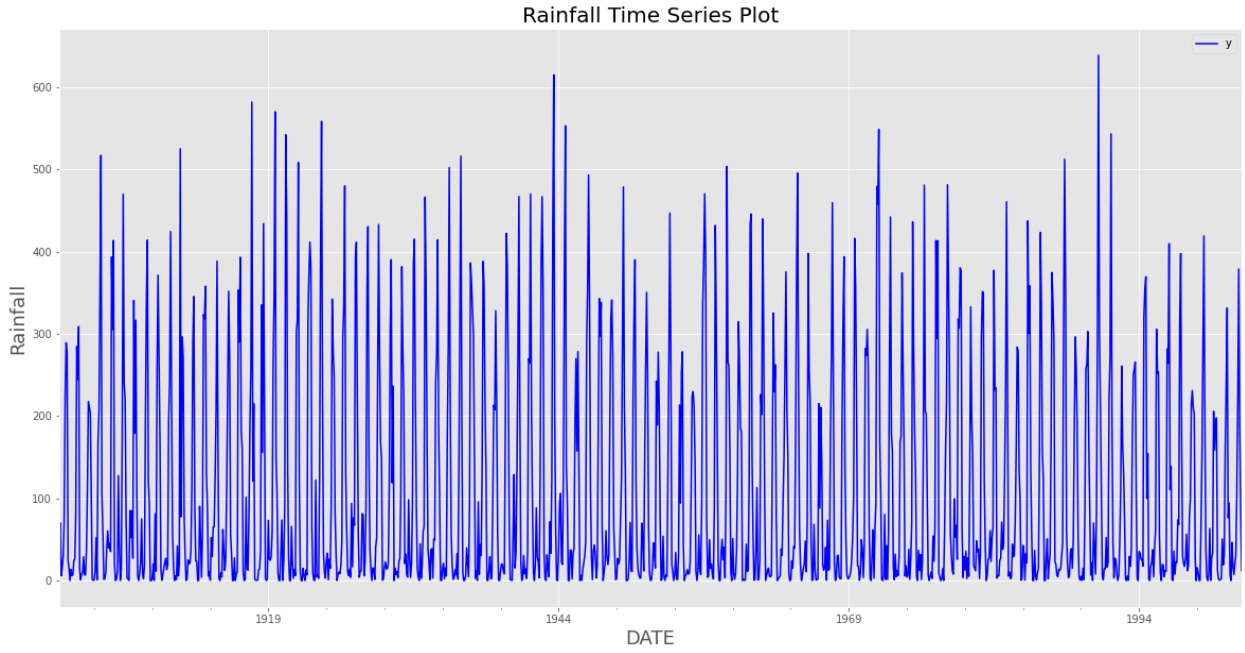
NAME : AMMAAR AHMAD
ROLL NO: 1801CS08
CS575 Project
Date : 2nd December 2021

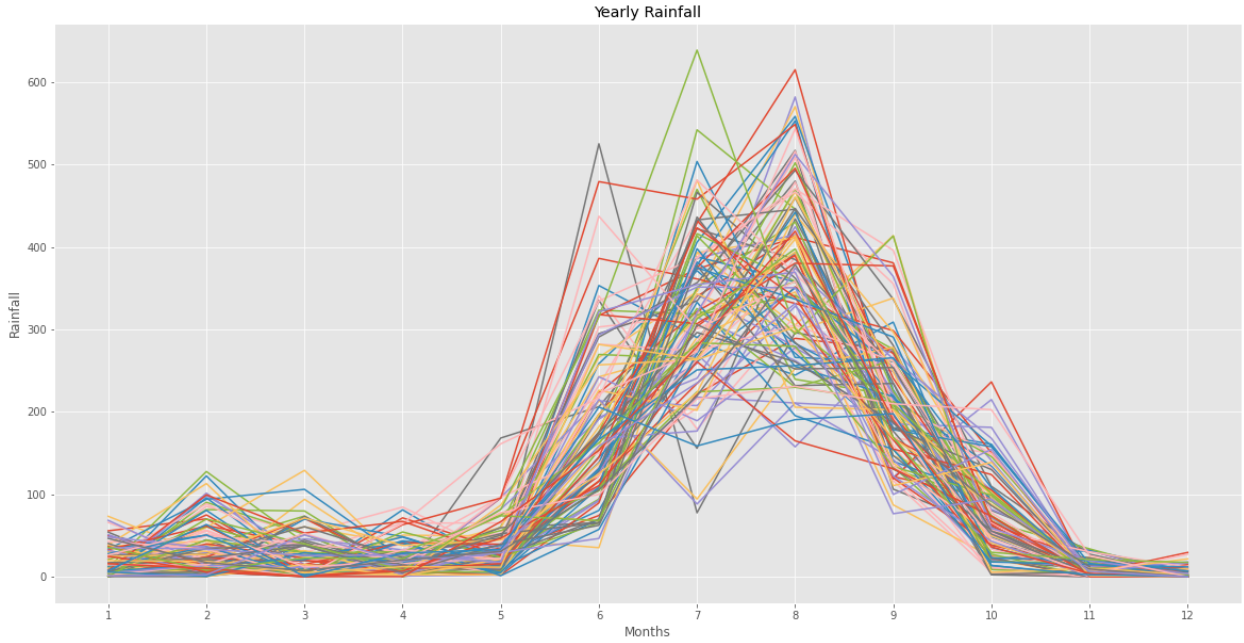
Dataset: Ranchi_Monthly_Rainfall_Data_1901_to_2002
Source:
https://raw.githubusercontent.com/ammaarahmad1999/Time_Series_Dataset/main/rainfall.csv
Project Colab File:
https://colab.research.google.com/drive/1G4foDNIiKCvtg2Nu0GWtaSYqaqB__WhU?usp=sharing

Dataset Statistics :
Standard Deviation = 136.6736

Rainfall Distribution in Ranchi

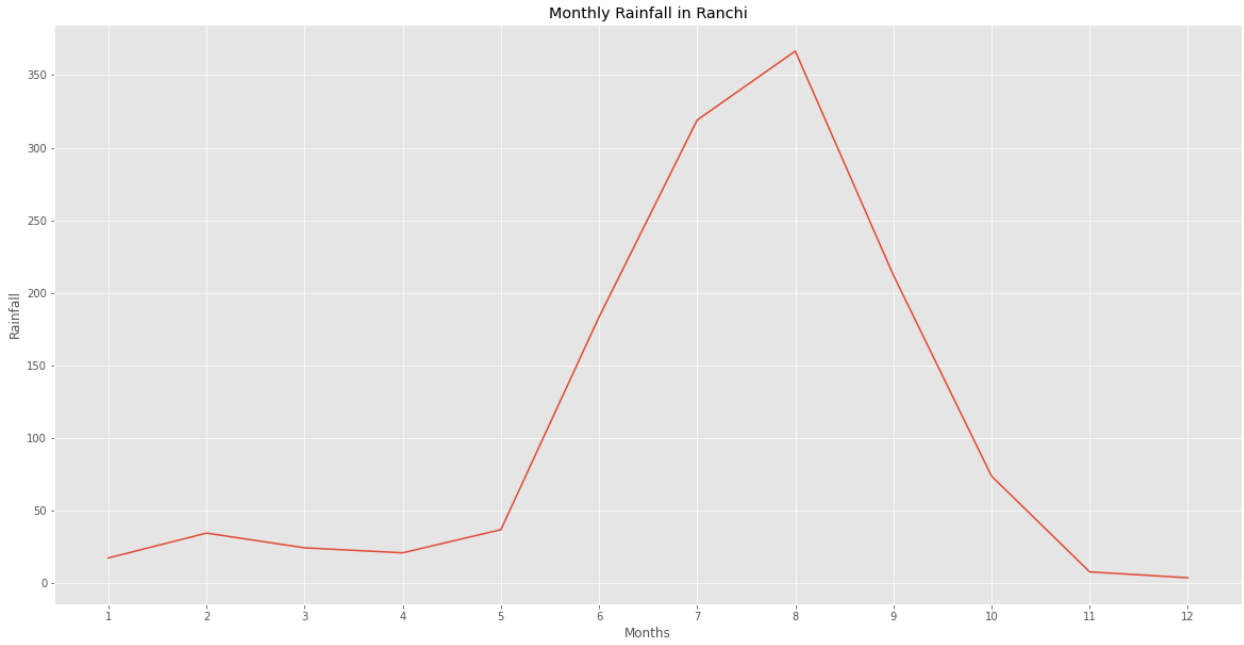


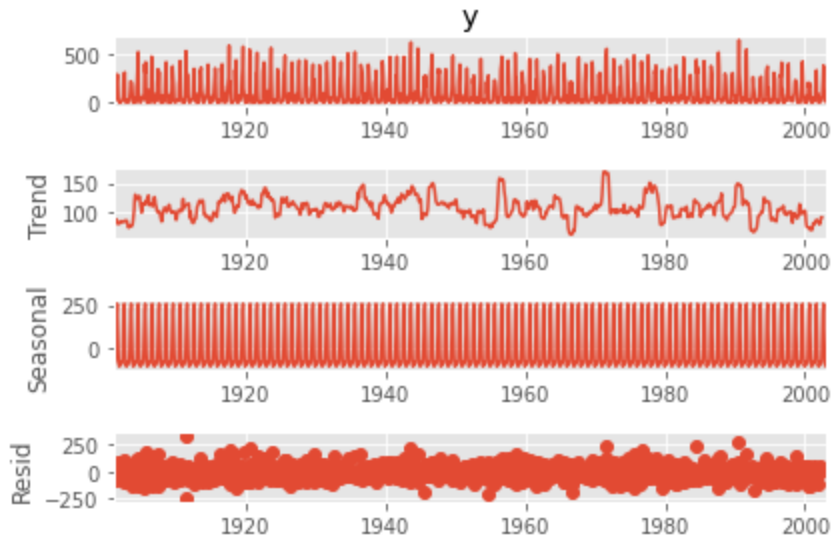
Yearly Rainfall Across Months



Clearly : There is a Seasonality as expected in rainfall distribution

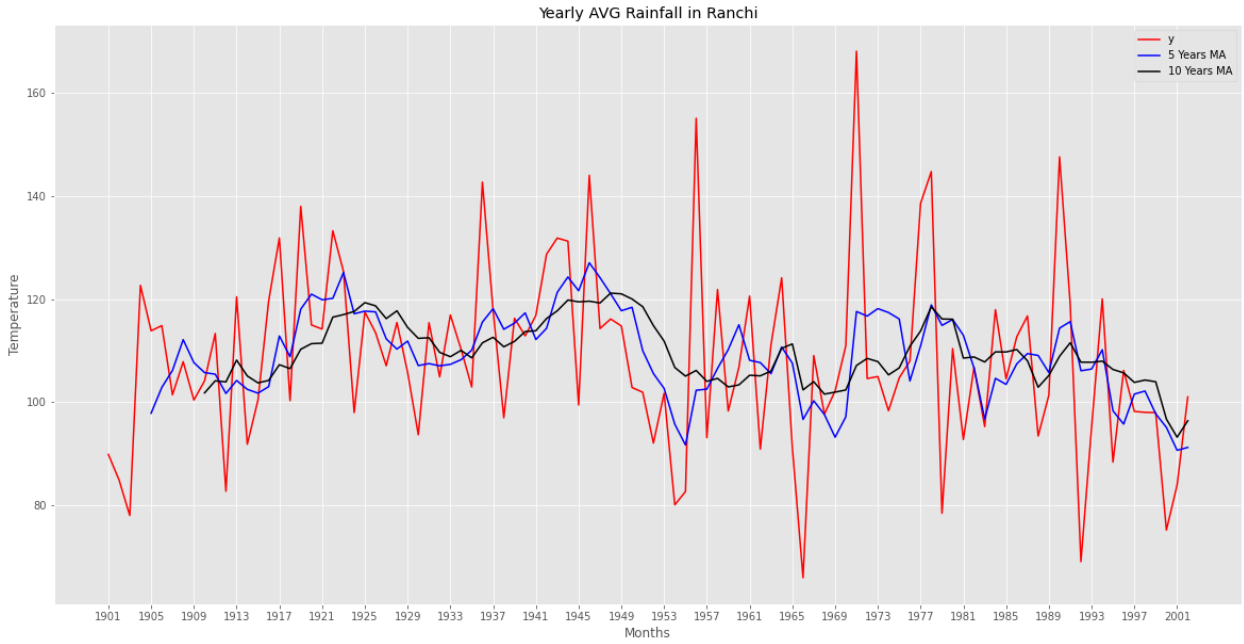
Average Monthly Rainfall in Ranchi





Seasonal Decompose (Original, Trend Component, Seasonal Component and Residuals)

Yearly Average Rainfall in Ranchi (along with 5 years and 10 years moving average)



Baseline Model: We define the baseline as y(t) = y(t-1)
i.e. Predicted Rainfall for this month = Actual Rainfall in the last month.

Root Mean Square Error for Baseline Model(RMSE) = 80.4078

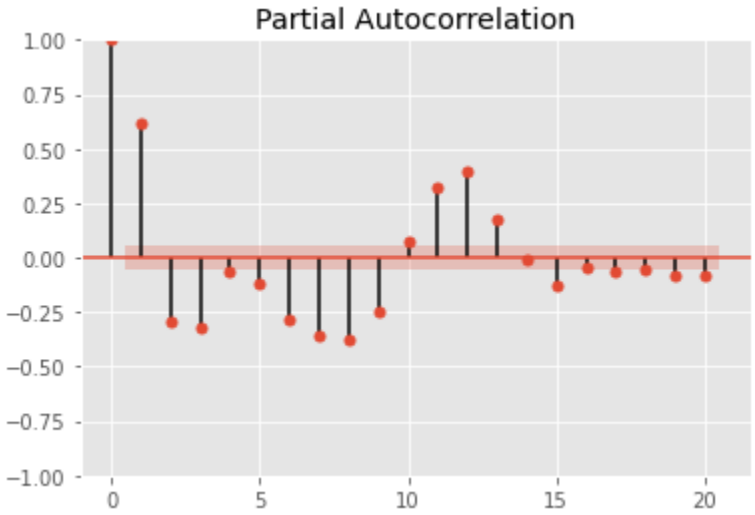
Dickey Fuller Test on Rainfall each month shows the data is stationary in the long run. => Mean, Variance of large dataset is constant (AIC was used as parameter to minimize)

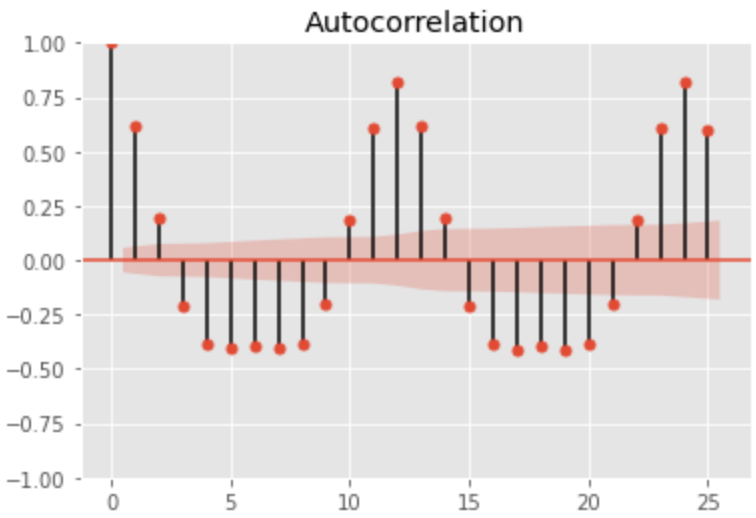
Our Baseline Model Mean Square Error = 88.4078

```
[30]  1 stationary_test = adfuller(df1['y'].values, autolag='AIC')
      2 print(f'P value = {stationary_test[1]}')
```

```
P value = 3.7110646679365726e-07
```

Since P Value < 0.05 => Reject the Null Hypothesis that it has a unit root. => Data is stationary



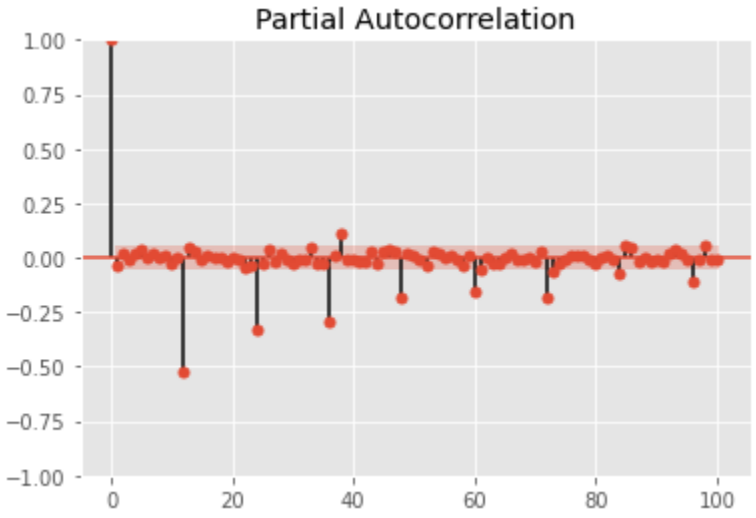


PACF and ACF plots of y. Cyclic character of ACF shows that there is a seasonality in the dataset. From the Plot m = 12

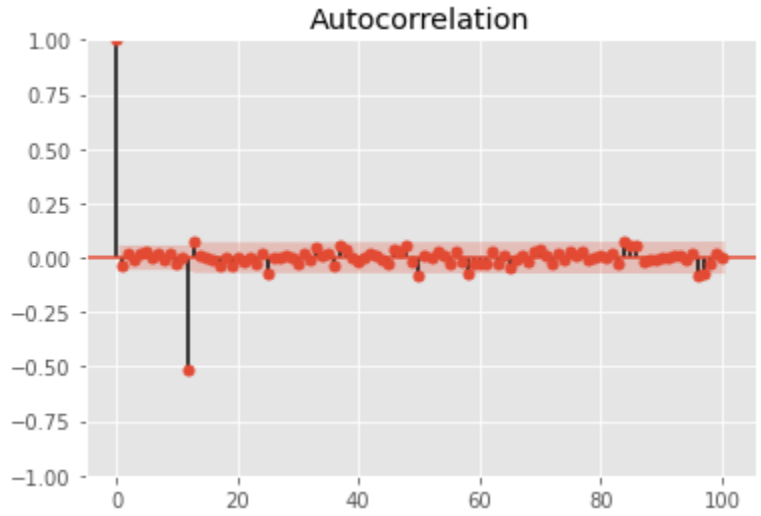
```
[34]  1 stationary_test = adfuller(df1['y'].diff(12).dropna(), autolag='AIC')
      2 print(f'P value = {stationary_test[1]}')

P value = 5.2811242599321435e-24
```

On Applying Dickey Fuller Test on Seasonal Differenced Data i.e. [Y(t) - Y(t-12)] shows that it's again stationary
PACF Plot



ACF Plot



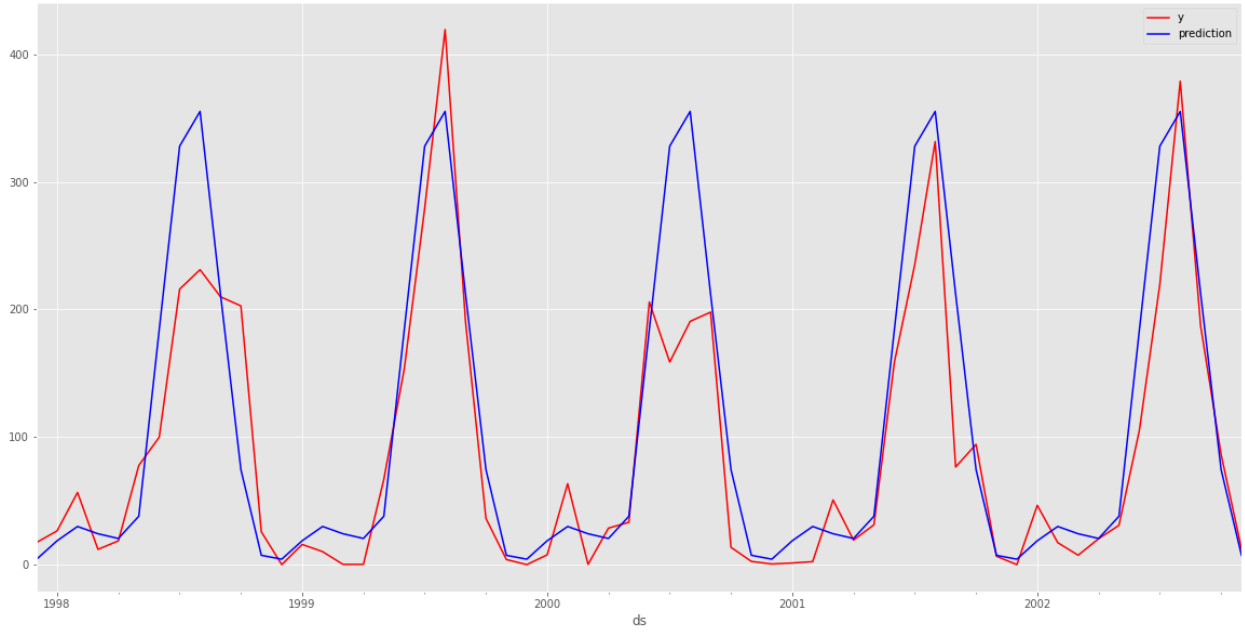
ACF Plot shows there is a strong negative lag at 12 indicating SMA(12) model and MA(0) model

PACF Plot shows a gradual decrease in dependency on 12th, 24th, 36th lags indicating SMA(12) model with no AR component.

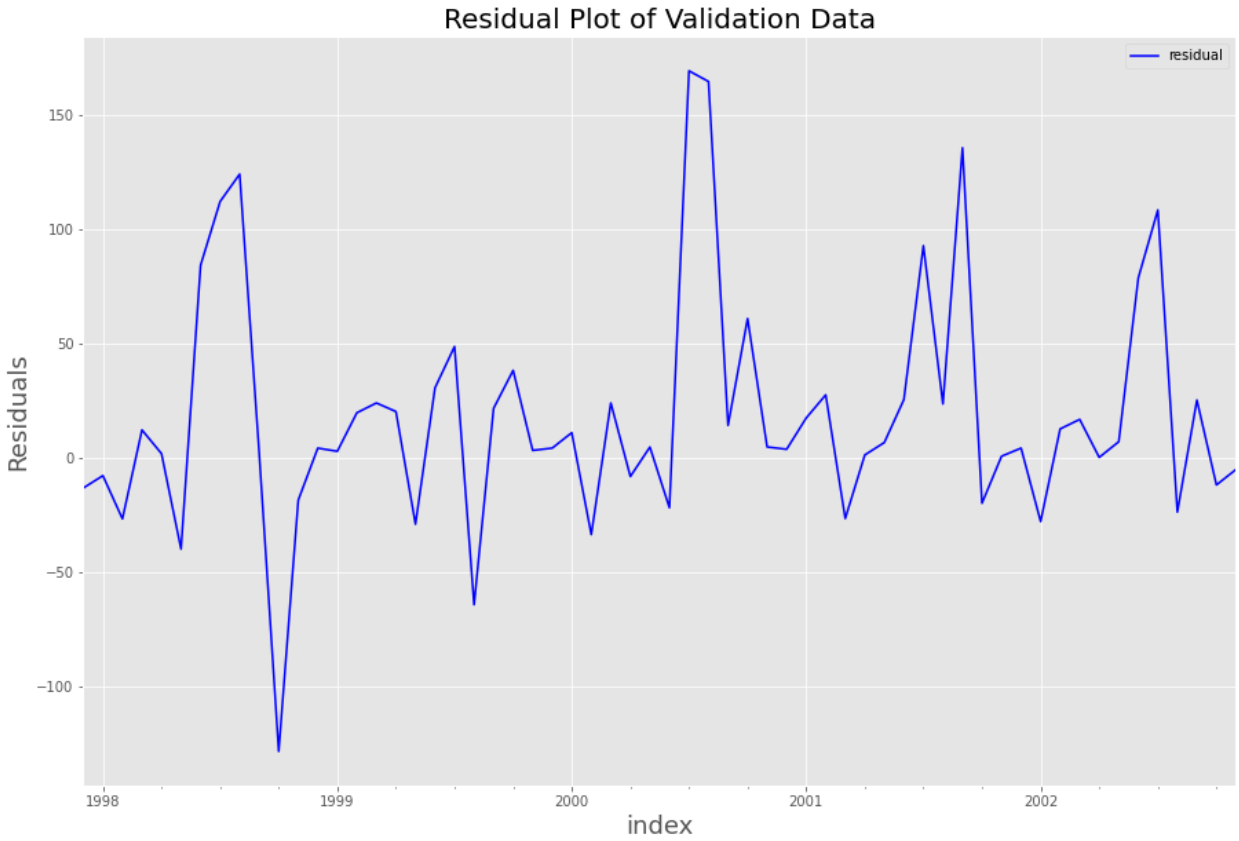
SARIMA Model (0,0,0) (0,1,1,12)

Root Mean Square Error (RMSE) = 54.6410

Predicted vs Actual Values on Validation/Test Set Plot



Residual Plot



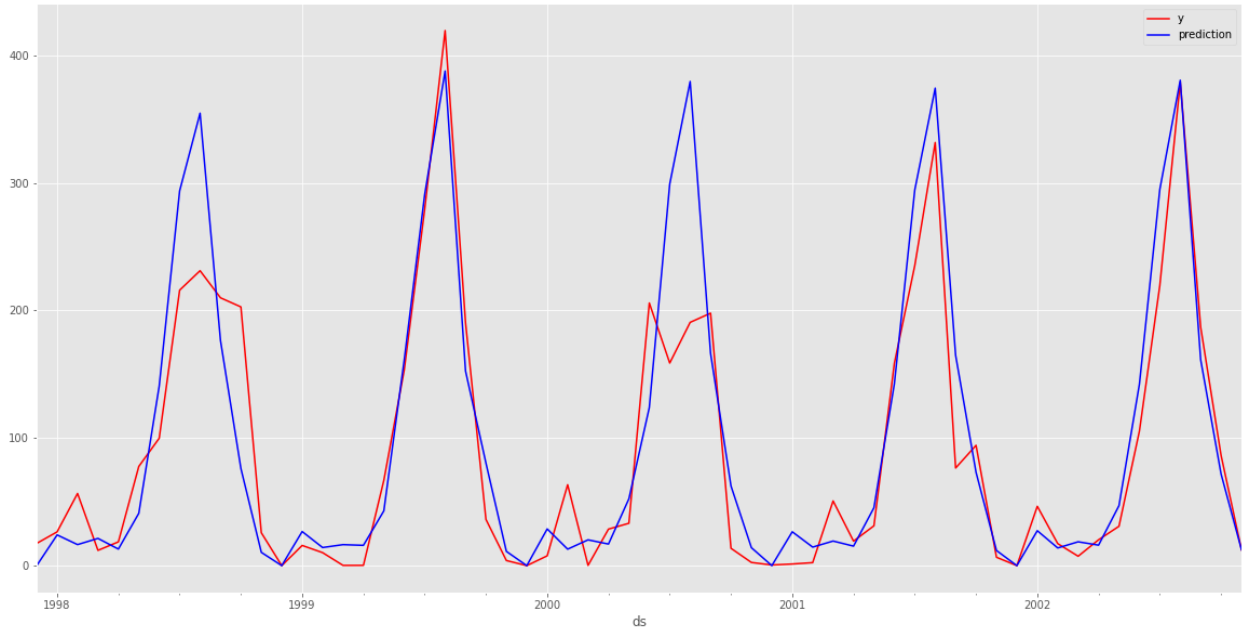
ADFuller Test on Residuals - P Value = 1e-7. Stationary
T-test on Residuals : Null Hypothesis : Mean = 0
P Value = 0.008799 < 0.05 => Reject the null hypothesis
Conclusion : SARIMA (0,0,0) (0,1,1,12) is not a good fit

Using auto_arima to find best SARIMA model

SARIMA (0,0,0) (2,1,0,12) best fit on AIC score

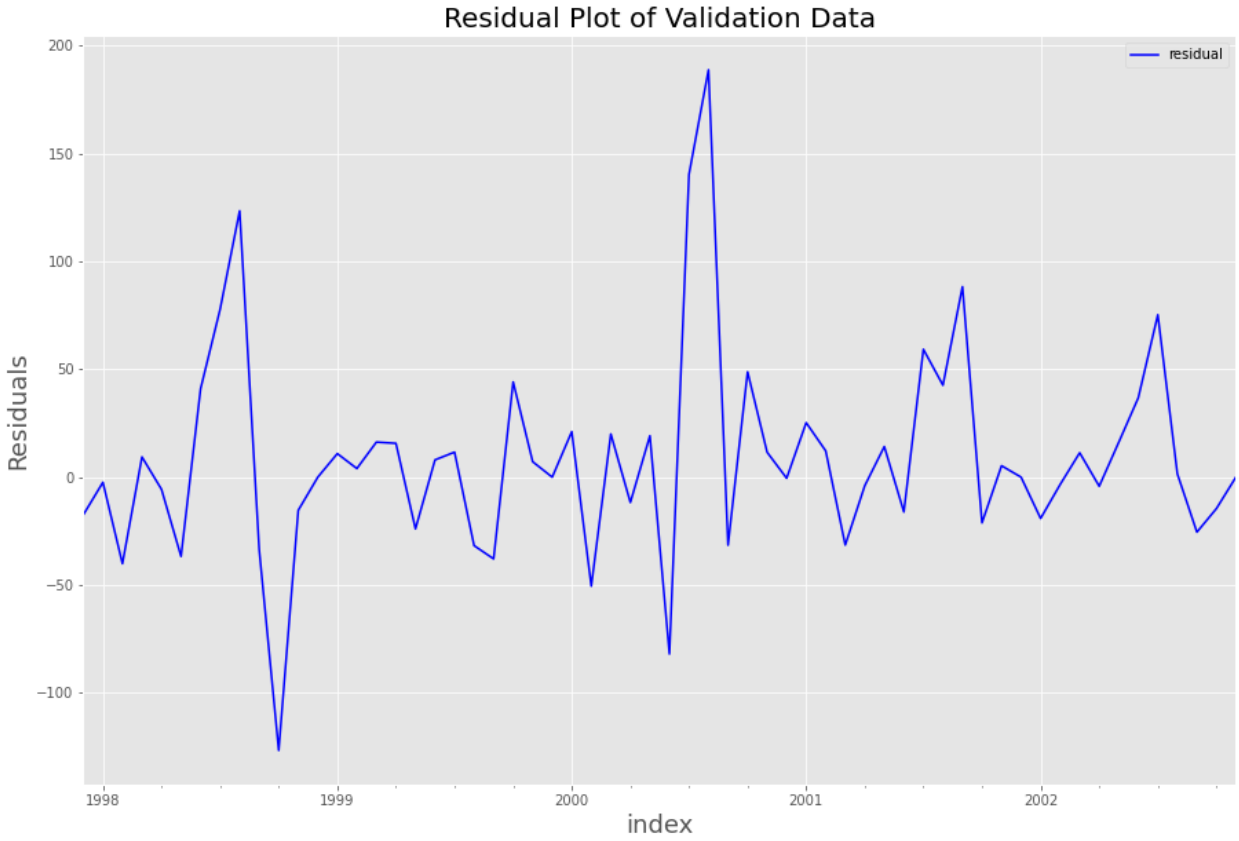
Root Mean Square Error (RMSE) = 48.91455 (Best)
RMSE is 1/3rd of Standard Deviation hence the good fit

Predicted vs Actual Values on Validation/Test Set Plot



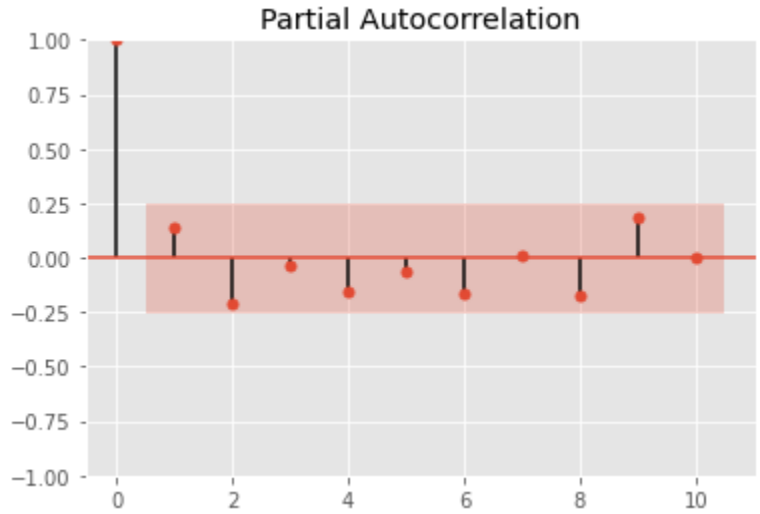
Residual Plot below shows a random walk model.
T-test on Residuals : Null Hypothesis : Mean = 0
P Value = 0.171619 >= 0.05 => Fail to reject the null hypothesis hence there is high probability of Mean = 0

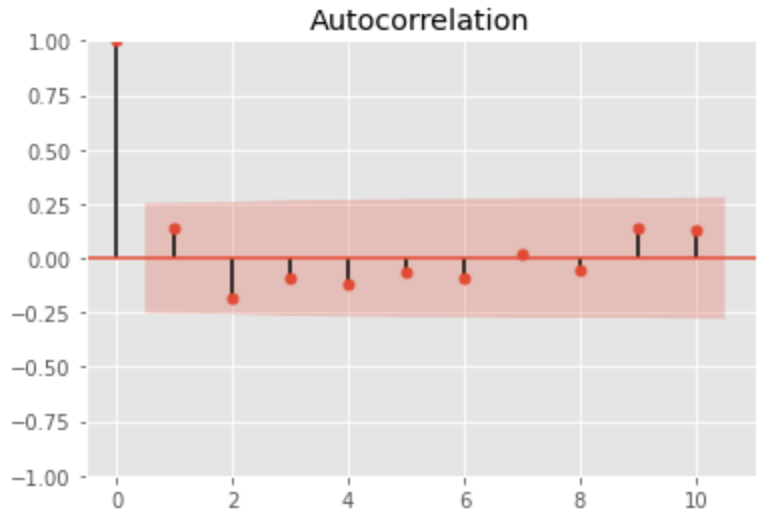
Residual Plot

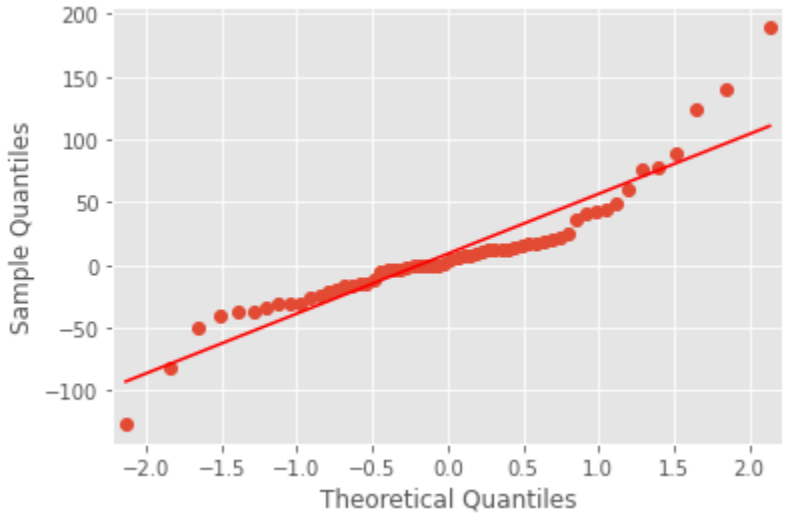


ADFuller Test on Residuals P Value = 6.3779e-09 < 0.05 => Reject the Null Hypothesis => Stationary

PACF, ACF and QQPlot plots of Residual







There is no lag correlation according to ACF and PACF plot and QQPlot shows nearly normal behaviour of Residuals

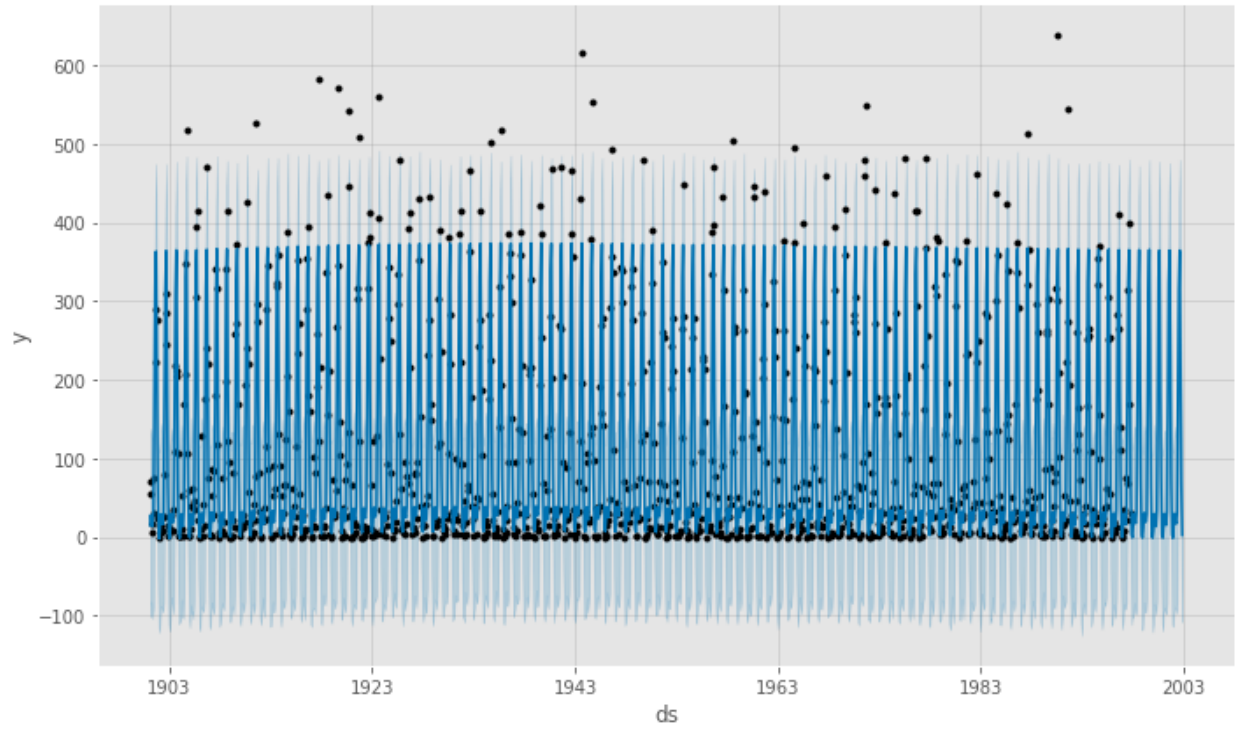
Conclusion : SARIMA (0,0,0) (2,0,1,12) is the good fit on this dataset

We also tried SARIMA (1,0,0) (2,0,1,12) and SARIMA (1,0,0) (1,0,1,12) => but RMSE for these models were high and Residuals didn;t have mean = 0 hence not a good fit

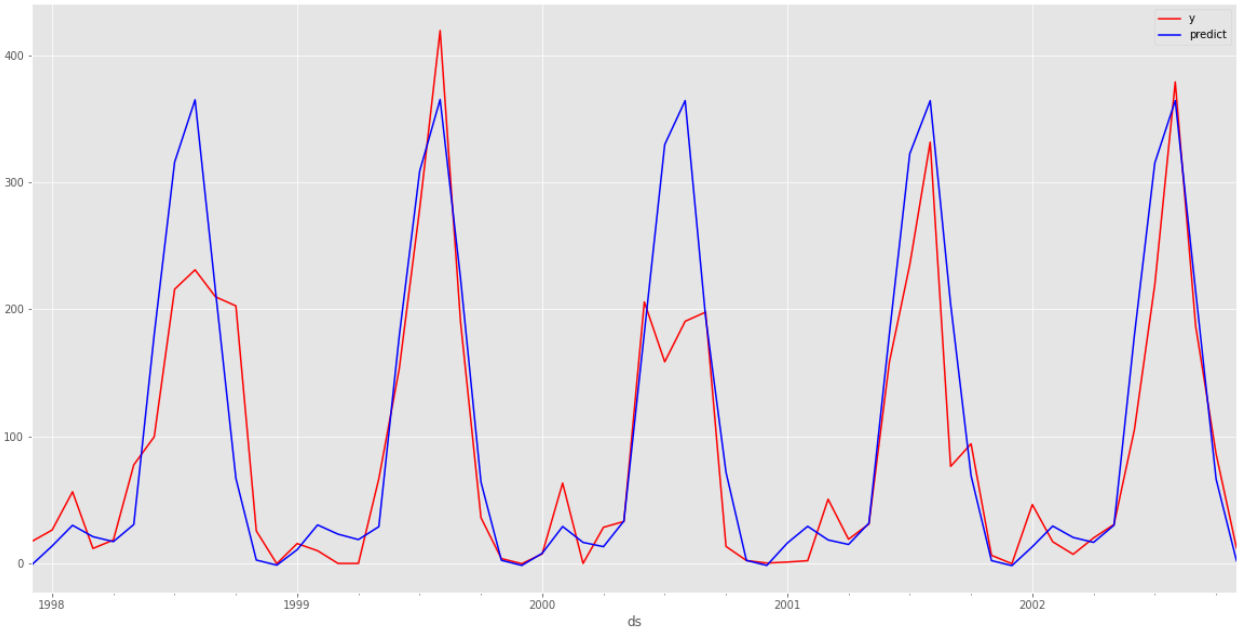
Prophet Model

Root Mean Square Error (RMSE) = 54.26

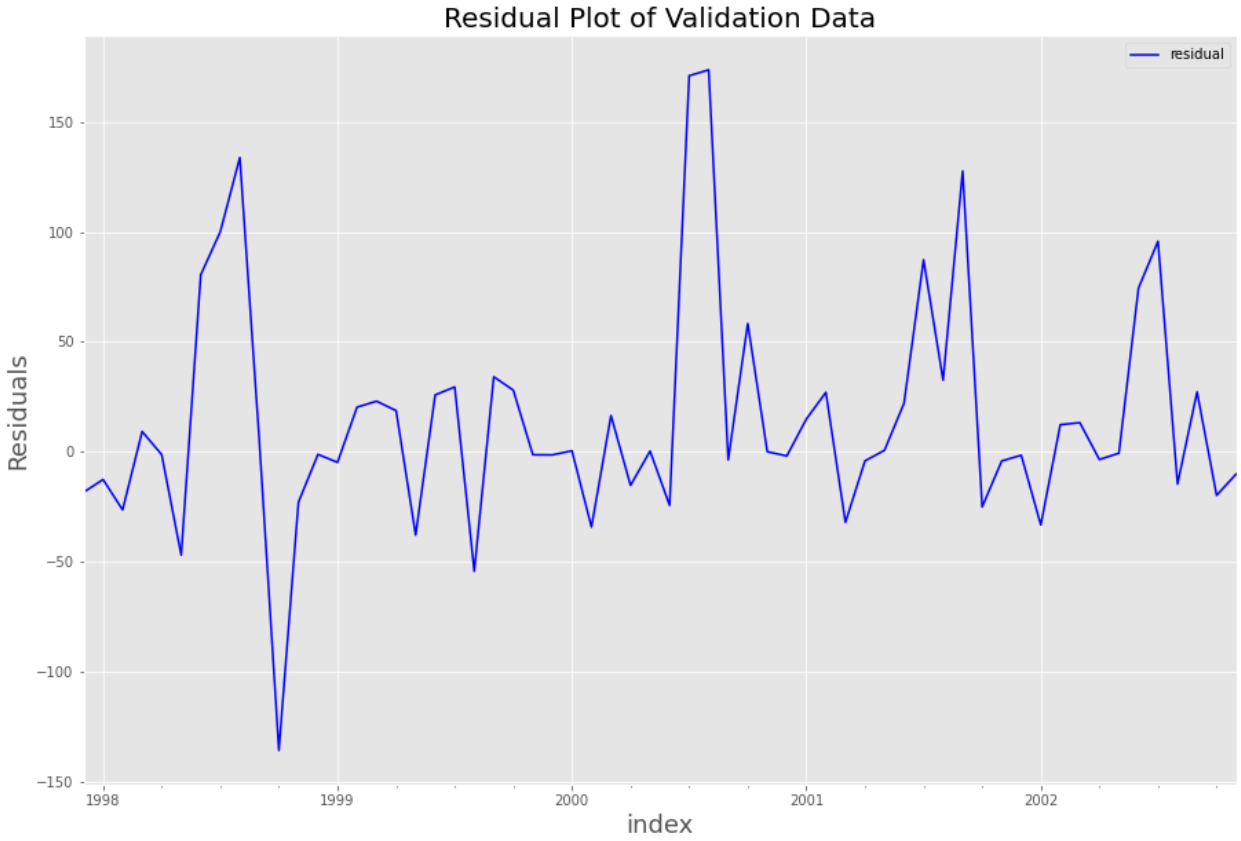
Prophet Model Forecast Plot



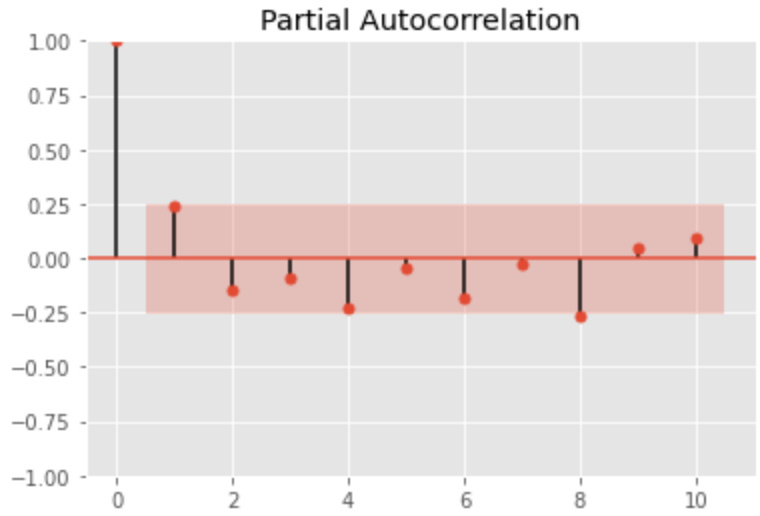
Predicted vs Actual Values on Validation/Test Set Plot

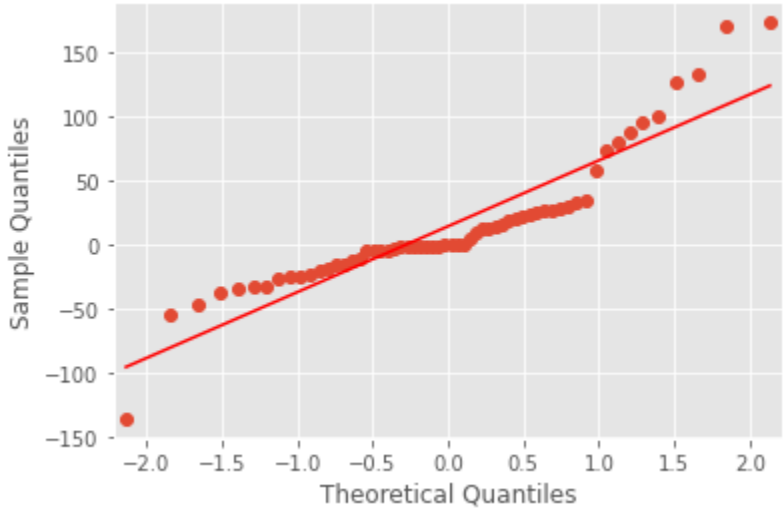


Residual Plot



ADFuller Test on Residuals P Value = 1e-7 => Stationary
T-test on Residuals : Null Hypothesis : Mean = 0
P Value = 0.037 < 0.05 (Reject the Null Hypothesis) => Mean is not 0 => Not a Random Walk Model





Also PACF and QQ Plot shows dependency as well as non-Normal behaviour => Prophet is not a very good fit for the model as residuals aren't independent. Hence we reject this model

We also tried various LSTM models with different lookback = {5,15,25} but all performed poorly. RMSE for these models was worse than Baseline Model hence these were rejected.

Conclusion:

Best Fit Model -> SARIMA (0,0,0) (2,1,0,12)
Observation : Rainfall in a Month doesn't depend on previous month statistics. It only has seasonal components

Motivation:

Weather Forecasting is one of the important Time Series applications. So I was interested to learn how to fit models for weather forecasting. Also there is a trend of Climate Change and Global Warming, I wanted to check it myself using data statistics. But unfortunately the Dataset I used didn't provide any useful information as Average Rainfall in the start of the 20th century didn't differ from late decades.

Challenges:

Finding the best fit model is not easy as there are a lot of random variables and in reality there are multiple factors which affect the rainfall. Also in the Validation/Test set rainfall was less compared to the training set. This natural phenomenon causes the error of 48 in the predictions which is on the high side. But it's still 1/3rd of Standard Deviation of the Rainfall indicating a decent prediction. After trying various models, I concluded that this is the best model to forecast the rainfall in Ranchi with the given dataset.

Thank You