Time Series Analysis – Predicting Number of Failures

Submitted By: **Vithika Agrawal**

Problem Summary:

The timestamp of failures were given in JSON file. I had the following task in hand:

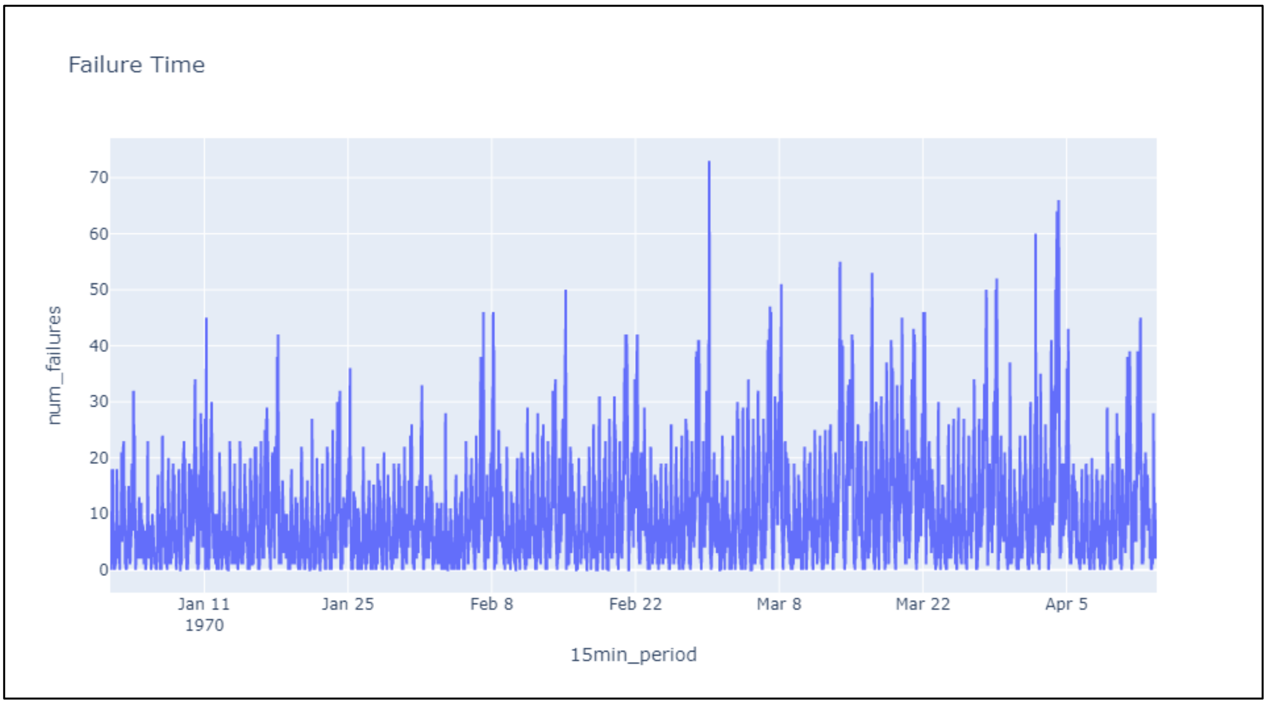
1. Aggregate timestamps on 15 minute time intervals
2. Visualise and find underlying patterns of the number of failures. Discuss the insights on these visualisations
3. Implement a forecasting method at the same 15-minute granularity over the next hour (4 periods ahead) and discuss the method in detail explaining its accuracy

Steps taken for the above tasks to be implemented:

1. Converted the string timestamps into a date time data frame using Pandas.
2. Aggregated the time stamps into 15 minutes time intervals using Grouper function in Pandas and then counted the number of failures in those intervals.

	15min_period	num_failures
0	1970-01-01 20:00:00	2
1	1970-01-01 20:15:00	6
2	1970-01-01 20:30:00	9
3	1970-01-01 20:45:00	7
4	1970-01-01 21:00:00	1

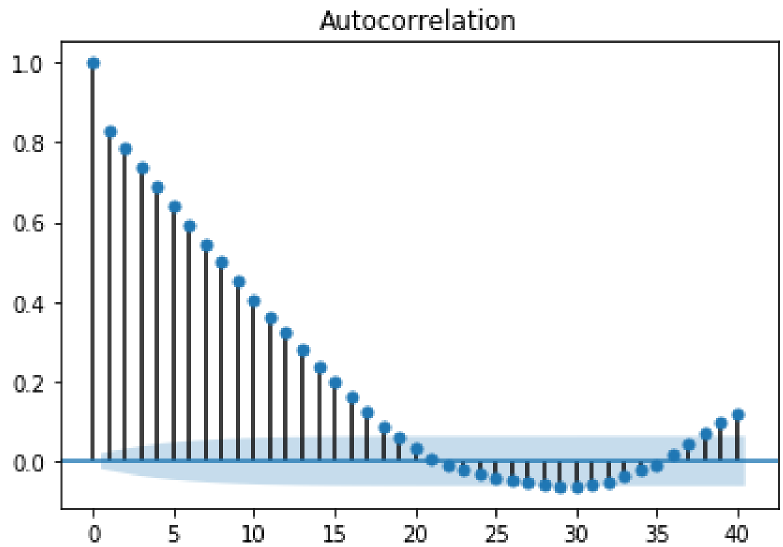
3. Initial Visualization of the entire data:



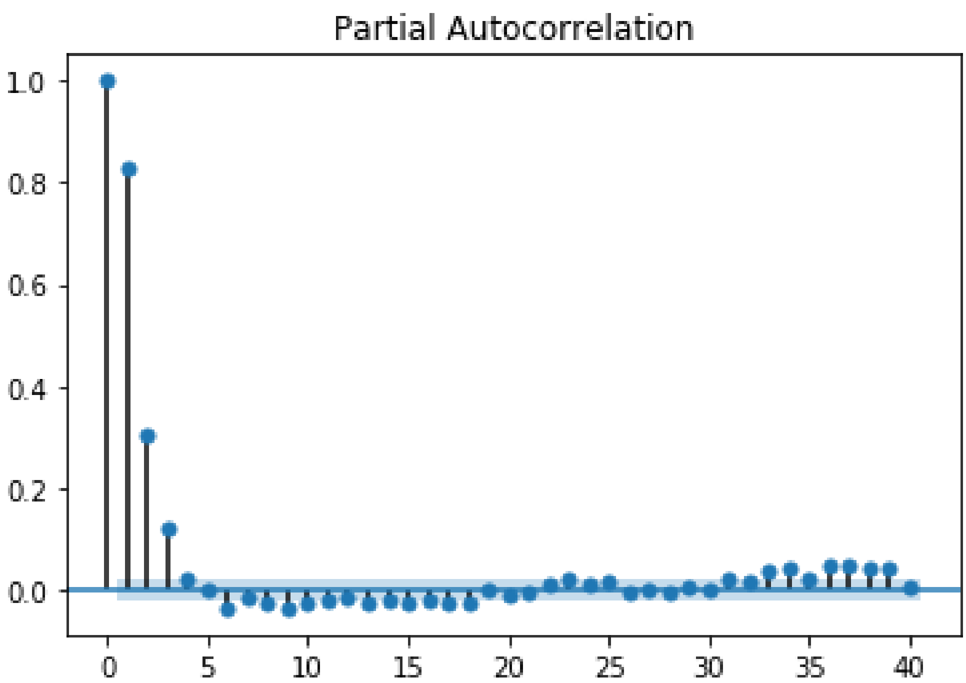
Here, the data seems to be stationary (i.e. the mean of all the intervals seems to be constant). In addition to this, we can't see any trend projection in the data (upward or downward) however, the data seems to show seasonality. To understand better seasonal decomposition of data is required.

4. Observing the seasonal decomposition, I can now confidently say with that the number of failures exhibit stationarity with no upward and downward trend.
5. In order to choose between Moving Average model, Auto Regressive Model, ARIMA and ARMA models, I created an ACF plot and PACF plot, which helps us determine which model to go forward with for forecasting.

   ACF plot shows how the given time series is correlated with itself. We can see negative correlation at lag 21, which is one of the criteria on choosing a moving average model.

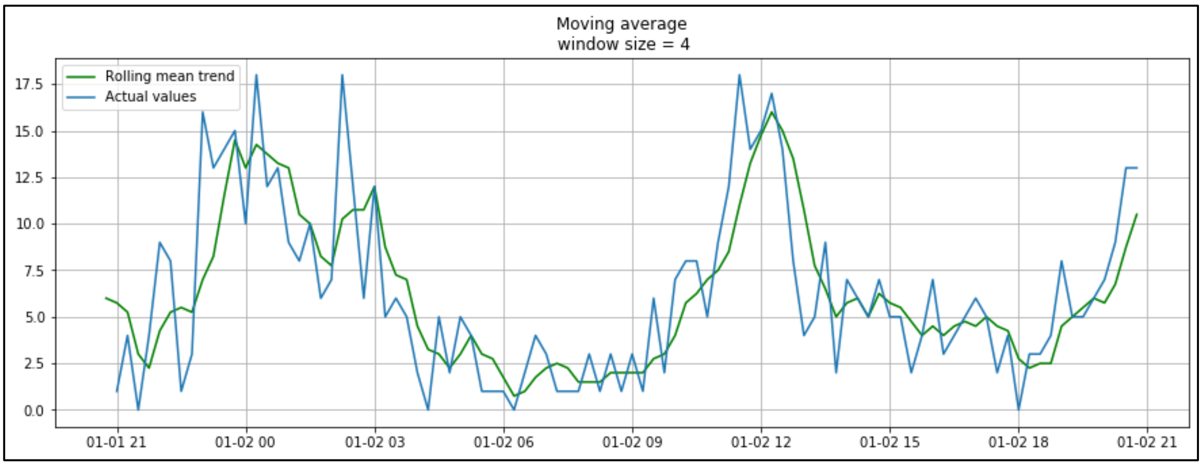
   

   In order to the satisfy the second criteria of choosing a moving average forecasting method, I needed to draw a PACF plot. We can see from the plot below that PACF decreases more gradually. Hence, I decided to move forward with a moving average model.

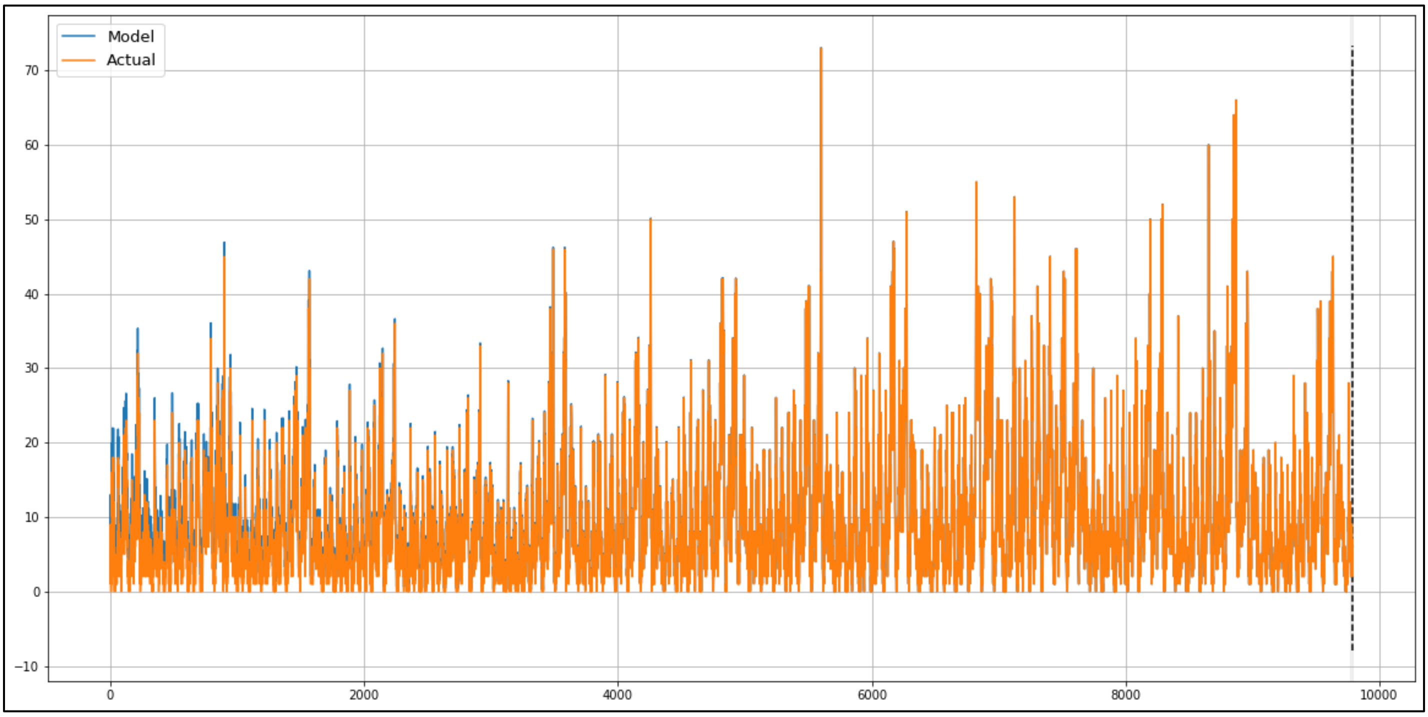
   

6. Hence, I used moving average method for forecasting the next hour. Before actually forecasting the next hour failure times in 15-minute granularity, I wanted to evaluate if the prediction from this model aligns with the true values or not. For that matter, first I saw the moving average with n=4 and plotted its trend.

7. From the graph below, I could conclude that Moving Average is a fairly good method for forecasting but it is not a smooth forecast. In addition to this, I had to predict the values of next hour for 4 periods at 15 minutes granularity. I was not able to achieve that using Simple Moving Average model.



8. For that matter, instead of using a Simple Moving averages, I opted for going for going towards a weighted average method.
9. Initially I decided to go forward with Single Exponential Smoothening method. However, as already observed, the data has a presence of seasonality. Therefore, the best method that I could address smoothening as well as seasonality to get an accurate forecast was Holts Winter Forecasting model. I considered this model for prediction because this model takes care of level, trend (though it is not required for our data) as well as the seasonality. This model is considered to be one of the best, when the time series data is very repetitive in terms of seasonality.
10. On implementing this model, I could successfully see that the model had fit the data perfectly.



11. Therefore, I forecasted the next four periods using this model and got the following result:

15min_period	num_failures	next_hr_prediction
1970-04-13 18:00:00	5	4
1970-04-13 18:15:00	2	1
1970-04-13 18:30:00	7	6
1970-04-13 18:45:00	6	5

At 7:00PM, the number of failures would be 4 times, at 7:15PM it would be 1 time, at 7:30PM it will be 6 times and at 7:45PM it will happen 5 times.