"NEW DELHI WEATHER FORECASTING WITH TIME SERIES ANALYSIS"

:Submitted by:

Roll :-MC/UG/S-VI/24 No-1851

Reg. No.:- 2021-0069 of 2021-2022

Introduction:

- New-Delhi, the Capital of India, experiences a diverse range of weather conditions throughout the year including hot summers, monsoon rains, and cold winters specially. The project seeks to leverage advanced weather forecasting techniques and technologies to improve the accuracy of weather prediction and provide valuable information to residents, businesses and policy-markets.

- The weather forecasting system for New-Delhi is a project aimed at providing accurate and timely weather predictions for the city.

- This project will help in forecasting the Avg. temperature of whole year of 2024,on the basis of past 13 historic year data with the help of Time Series Analysis.

Objective of Project :

- The main objective of a weather forecasting project using time series analysis is to accurately predict future weather conditions based on historical data.

- This involves building a model that can capture the patterns, trends, and seasonality in the data to make reliable forecast. Implement advance weather forecasting model, to analyse the collected data and generate accurate weather.

- Here we are using the previous 13 year Historical weather data(imaginary) of New Delhi, India from 2010 to 2023 to predict a the next year 2024 by Long-Term Forecasting.

- We can also use it in Meteorological Dept. real data.

- Give further conclusion of weather behaviour for advance help in agriculture, urban planning, disaster preparation(heat-wave, cyclone etc) to the local bodies or business sectors.

Data Description :

This dataset includes a variety of meteorological variables for each day from 2010 to 2023, such as temperature, humidity, wind speed, atmospheric recorded pressure, and precipitation. Though these are not real value, but an imaginary values :

Date, Year, Month of the recorded weather data
Max Temperature: Maximum temperature (°F),
Avg. Temperature: Average temperature (°F), Min Temperature: Minimum temperature (°C),
Max Dew Point: Maximum dew point (°F),
Avg. Dew Point: Average dew point (°F), Min Dew Point: Minimum dew point (°C),
Max Humidity: Maximum relative humidity (%),
Avg. Humidity: Average relative humidity (%), Min Humidity: Minimum relative humidity (%),
Max Wind Speed: Maximum wind speed (km/h),
Avg. Wind Speed: Average wind speed (km/h), Min Wind Speed: Minimum wind speed (km/h)
Max Pressure: Maximum atmospheric pressure (hPa),
Avg. Pressure: Average atmospheric pressure (hPa),
Min Pressure: Minimum atmospheric pressure (hPa)

Data Sample :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Date	Year	Month	Max Temp	Avg Tempe	Min Tempe	Max Dew I	Avg Dew P	Min Dew P	Max Humi	Avg Humid	Min Humic	Max Wind	Avg Wind S	Min Wind S	Max Press	Avg Pressu	Min Pressure
2	1	2010	1	70	55.7	45	52	47.5	41	100	77.5	43	8	4.9	0	29.2	29.2	29.1
3	2	2010	1	59	50.6	46	54	49.6	46	100	96.7	77	8	4.8	3	29.3	29.3	29.2
4	3	2010	1	57	52.8	48	57	52.1	48	100	97.3	82	12	7.4	3	29.3	29.2	29.2
5	4	2010	1	55	51	48	50	48.4	48	100	91.9	77	15	11	7	29.3	29.2	29.1
6	5	2010	1	63	53.1	46	52	47.9	46	100	85.1	55	14	5.7	0	29.2	29.1	29.1
7	6	2010	1	64	54.4	45	55	49.7	45	100	85.7	56	8	2.5	0	29.2	29.2	29.1
8	7	2010	1	64	53	46	54	50.1	46	100	90.8	56	6	3	0	29.3	29.2	29.2
9	8	2010	1	54	50.2	46	54	50.1	46	100	99.6	93	9	5.2	0	29.3	29.2	29.2
10	9	2010	1	63	51.7	46	54	47.6	46	100	87.7	59	9	5.8	2	29.3	29.2	29.1
11	10	2010	1	50	46.8	45	50	46.6	43	100	98.8	93	7	5.1	3	29.3	29.2	29.2
12	11	2010	1	55	48.4	43	46	44.3	43	100	85.3	67	7	4.1	0	29.3	29.2	29.2
13	12	2010	1	57	49.6	45	48	45.2	43	100	85.2	63	8	4.6	0	29.3	29.2	29.1
14	13	2010	1	61	50.6	45	55	47.7	45	100	91.2	59	8	5.1	0	29.3	29.3	29.1
15	14	2010	1	59	49.9	45	52	47.8	45	100	93.4	72	12	6.7	5	29.4	29.3	29.3
16	15	2010	1	61	51.6	39	52	47.9	39	100	88.7	63	10	5.5	0	29.4	29.4	29.3
17	16	2010	1	68	53.7	39	50	46.7	39	100	80.6	46	12	4.2	0	29.4	29.3	29.3
18	17	2010	1	61	50.7	46	52	49.4	46	100	95.8	72	7	4.6	2	29.4	29.3	29.3
19	18	2010	1	64	53.1	46	50	47.1	45	100	82.6	49	8	3.7	0	29.4	29.3	29.3
20	19	2010	1	66	53	43	50	46.6	43	100	82.6	49	6	2.6	0	29.3	29.2	29.2
21	20	2010	1	66	51.3	45	52	47.8	45	100	90	49	6	2.6	0	29.3	29.3	29.3
22	21	2010	1	64	51.9	45	50	47.5	45	100	87.4	55	6	2.3	0	29.3	29.2	29.2
23	22	2010	1	66	54.3	46	52	49.3	46	100	85.4	56	6	3.5	0	29.3	29.2	29.2
24	23	2010	1	72	57.8	48	55	51.6	48	100	82.7	46	8	3.5	0	29.3	29.2	29.1
25	24	2010	1	73	60.8	48	57	53.7	48	100	80.4	46	10	5.8	0	29.2	29.2	29.1
26	25	2010	1	75	63.9	54	55	53.4	50	100	71.8	41	10	2.7	0	29.3	29.2	29.1

NewDelhiWeatherInsights(2010-20

Methodology :

- **Graphical Plot :** First plot graphically the huge data by variable' **Daily Avg. Temperature'** with Heat Map and boxplot, and visually understand the data.

- **ARIMA (Auto Regressive Integrated Moving Average):**

This model is commonly used for time series forecasting. It combines the concepts of autoregression (AR), Integrated(I), and moving average (MA) to capture different aspects of the time series data.

Fitting an ARIMA model involves selecting the appropriate parameters (p, d, q) for the autoregressive (AR)(q), differencing (I)(d), and moving average (MA)(p) components of the model. Here's a general outline of how you can fit an ARIMA model:

1. **Visualize the Time Series:** Plot the time series data to understand its patterns, trends, and seasonality.

2. **Find Stationary Series:** ARIMA models require the time series to be stationary, meaning its mean, variance, and autocorrelation structure do not change over time. If the series is not stationary, it may need to difference it until it becomes stationary.

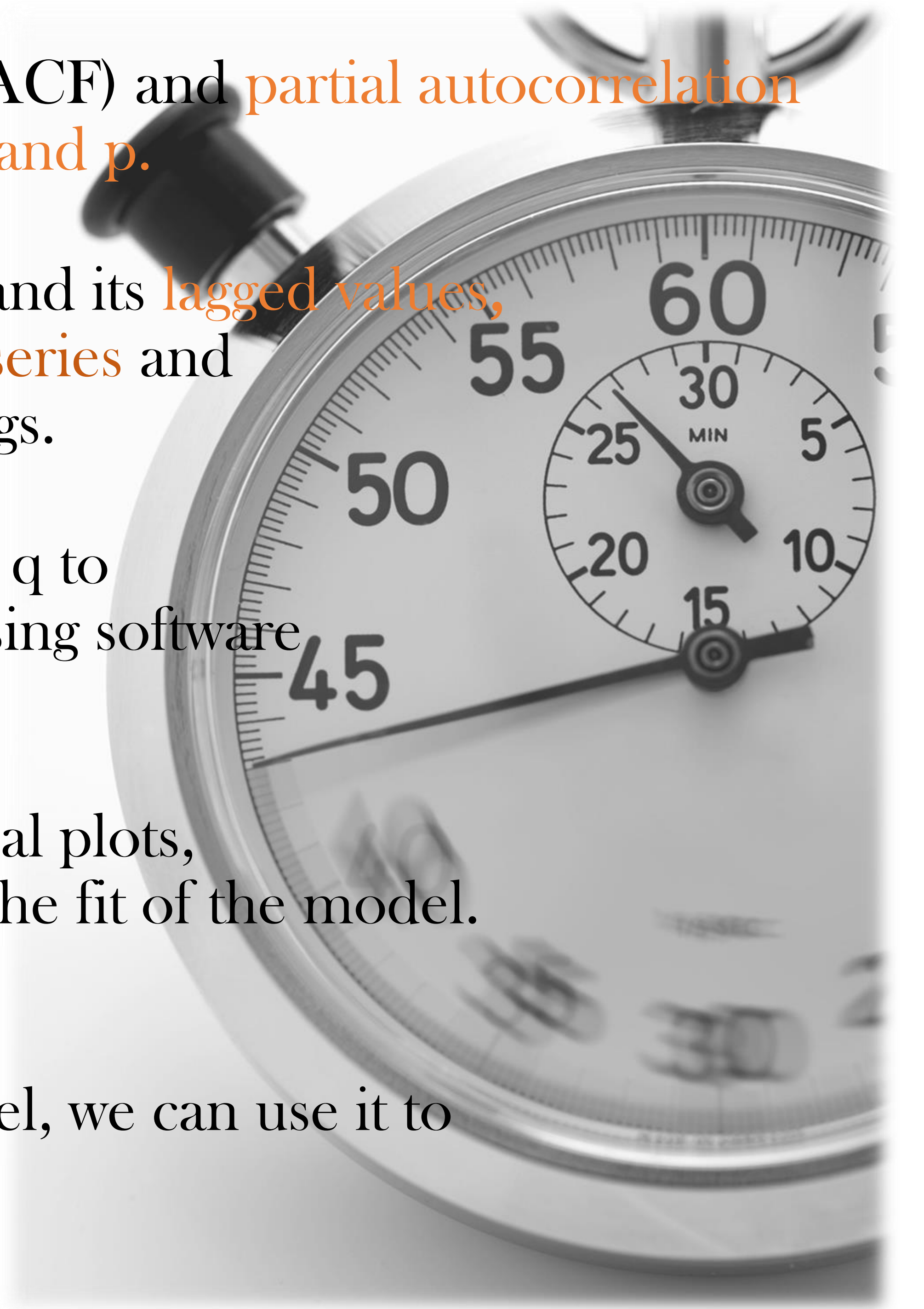
3. Identify Parameters: Use autocorrelation function (ACF) and partial autocorrelation function (PACF) plots to identify potential values of q and p.

The ACF plot shows the correlation between a series and its lagged values, while the PACF plot shows the correlation between a series and its lagged values, excluding the effects of intervening lags.

4. Fit the Model: Use the identified values of p, d, and q to fit the ARIMA model to the data. This can be done using software R's '**forecast**' package.

5. Evaluate the Model: Use diagnostic tools like residual plots, ACF plots of residuals, and statistical tests to evaluate the fit of the model. Adjust the model if necessary.

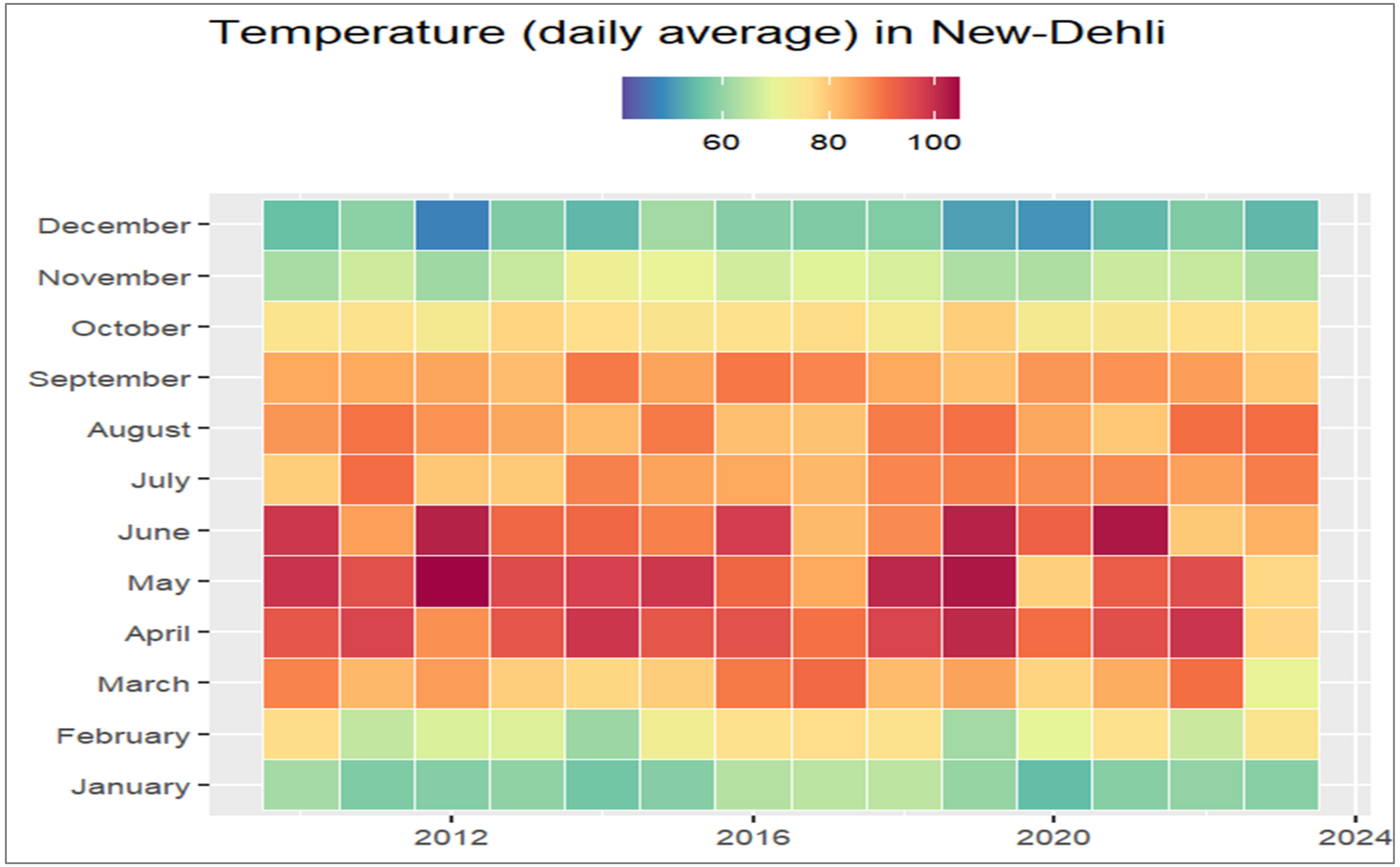
6. Forecast: Once we have a satisfactory data and model, we can use it to forecast future values of the time series.



Graphical representations :

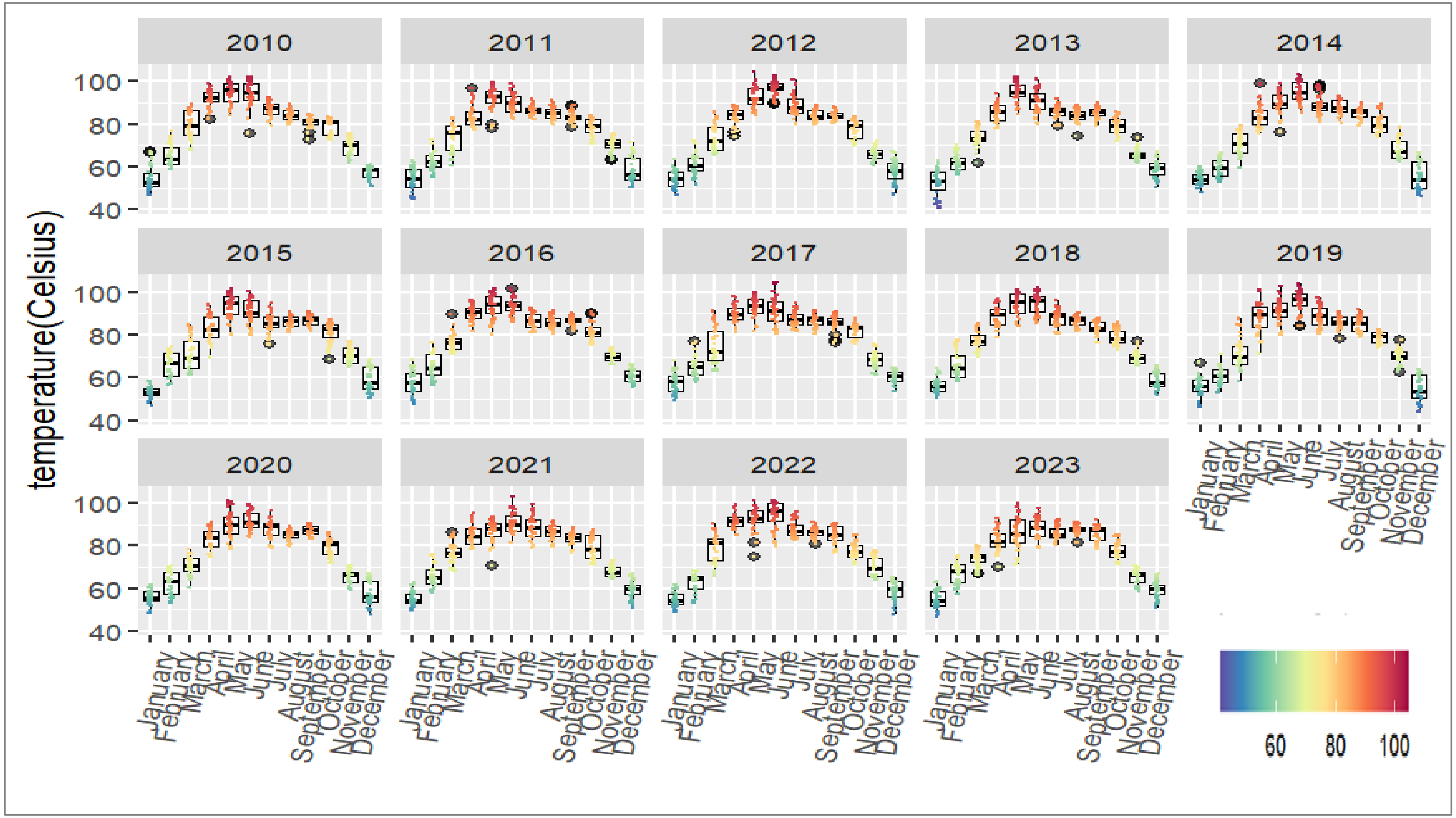
(i) Heat-Map:

The heatmap of daily average temperature visually represents temperature variations across a geographic area over a specific time period(years). It uses color gradients to show different temperature ranges, with warmer temperatures depicted in shades of red (>100)or orange(>80), and cooler temperatures in shades of blue(<60) or green(>60).



Heat Map of Daily Average Temperature (°F) in New Delhi

Conclusion: Heatmap shows the same seasonality i.e. Temperatures are higher in the summer (although it seems the max. temperature is reached around May/June)of every year from 2010 to 2023.



Box Plot of Daily Average Temperature (°F) of New Delhi

(ii) Box Plot (daily average temperature):

Boxplot is a graphical summary of the distribution of a dataset(daily avg. temp) over time(years).

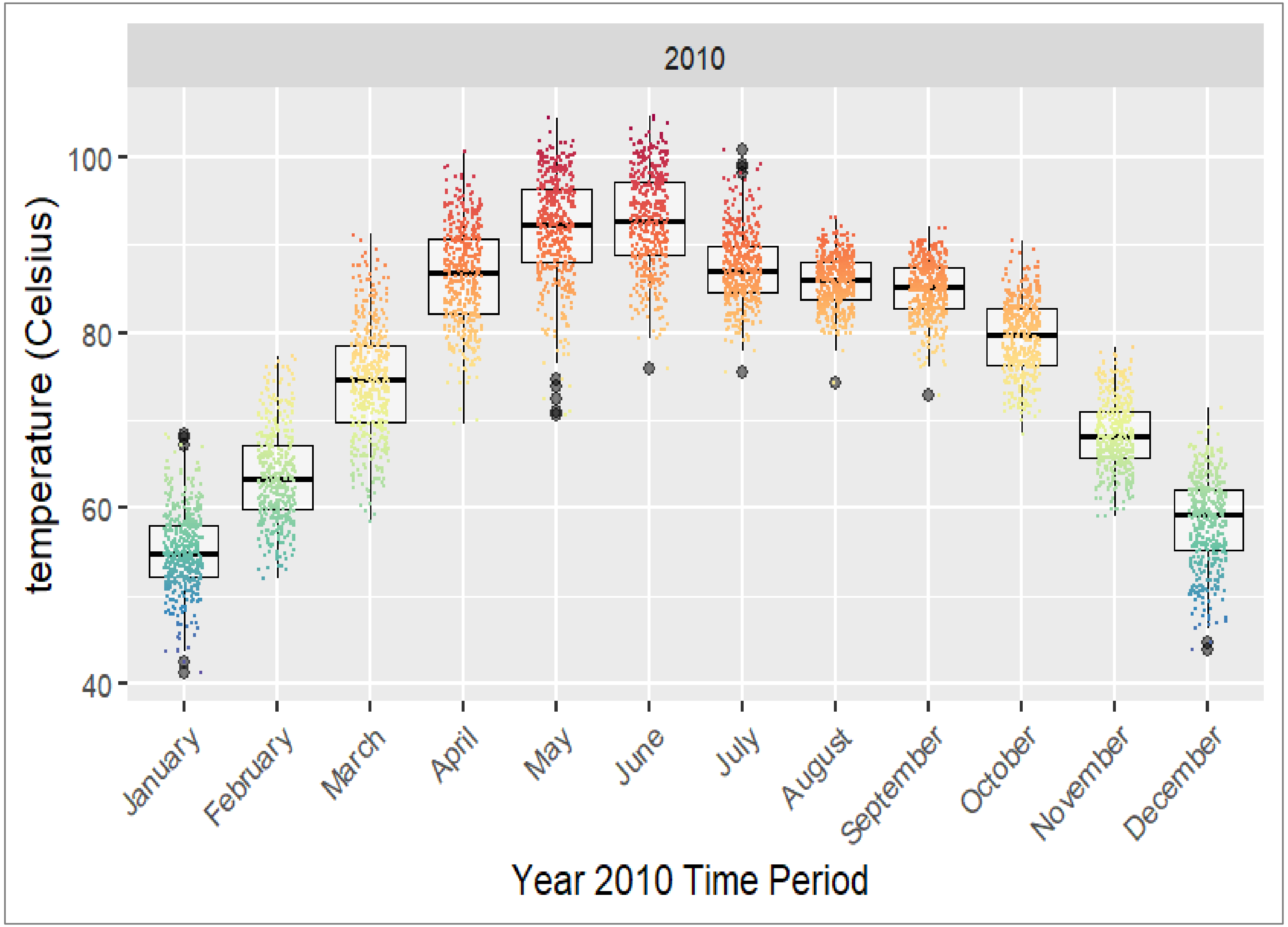
It displays key statistical measures such as the median, quartiles, and potential outliers in a compact format. The plot consists of a box, which represents the interquartile range (IQR) where the middle 50% of the data lies, with a line inside marking the median.

Conclusion: It's too difficult to make any clear conclusion from all the boxplots given clumsy visuals throughout all the years, still can see seasonality in increasing temp. of every year.

(iii) Box Plot(monthly):

Since the previous Box Plot gives very clumsy visuals throughout each the years, So we take single year(2010) boxplot to check monthly changes.

The median is taken over the data within a day similarly, and observe all the monthly changes.



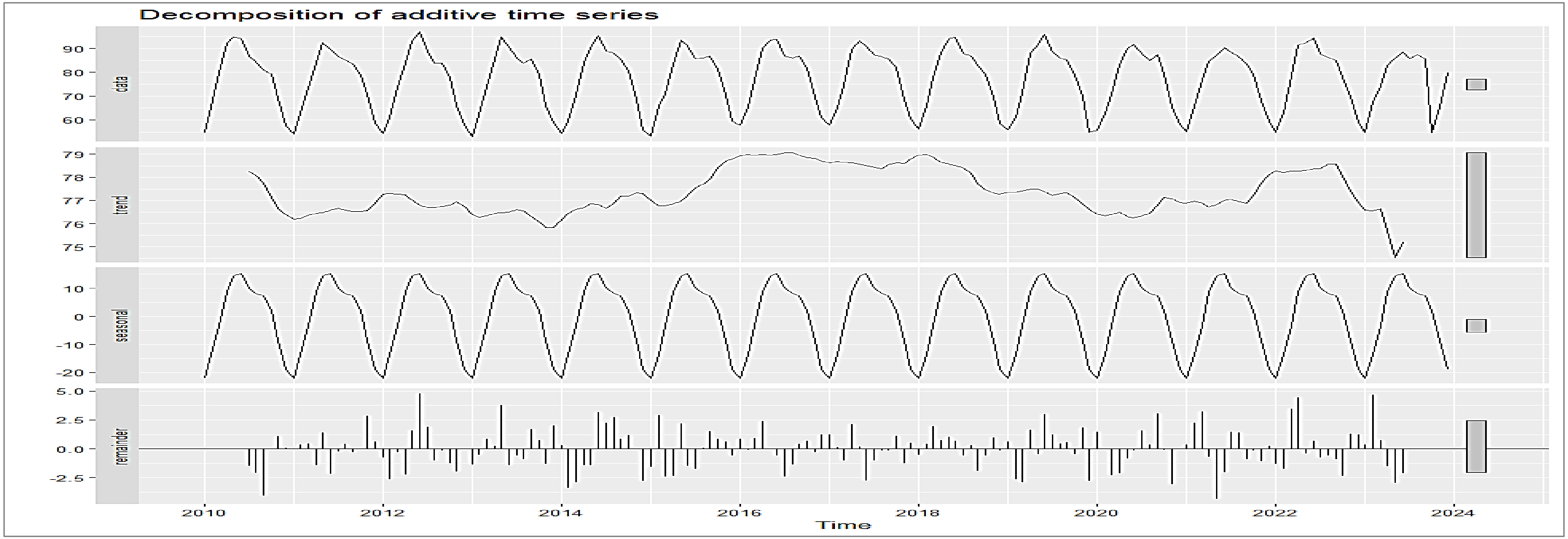
Box Plot of Avg. Temperature(($^{\circ}$F) of New Delhi for Year 2010

Conclusion: The Above Box-Plots for the single year(2010) shows more clearly the daily avg. temperatures by each month from January,2010 to December,2010 by the color scale, and Boxplot shows the same seasonality i.e. Temperatures are higher in the summer (although it seems the max. temperature is reached around May/June).

Data Analysis Output & Interpretations of output:

(i) **Time series decomposition :** is a statistical method used to break down a time series dataset into its underlying components: trend, seasonal variation, and residual (or noise).

Remarks: A daily timeseries would have been a bit of an overkill for a time series so a monthly one can capture the fluctuations (seasonality) and shows if there is a trend vs. Year (increase, stability, decrease)



Decomposition of Additive Time Series Model of Monthly Avg. Temperature (°F)

(ii)Stationarity check with Augmented Dickey- Fuller Test :

- The Augmented Dickey-Fuller (ADF) test is a statistical test used to determine if a time series dataset is stationary or non-stationary.
- Stationarity refers to a series having constant statistical properties over time, including a stable mean and variance.
- The "augmented" version of the test includes additional lagged terms in the regression equation to account for possible autocorrelation and improve the accuracy of the test.
- The ADF test is commonly used in econometrics and time series analysis to assess the stationarity of data before applying statistical models like ARIMA (Auto-Regressive Integrated Moving Average) for forecasting.

```
> adf_results

	Augmented Dickey-Fuller Test

data:  myts
Dickey-Fuller = -13.461, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

- Unit Root: The test examines whether a unit root is present in the data. A unit root indicates non-stationarity, implying that the series exhibits trends or cycles that can lead to unreliable statistical conclusions.
- Null Hypothesis: The null hypothesis of the ADF test is that the series has a unit root, meaning it is non-stationary.
- Test Statistic: The ADF test calculates a test statistic that compares the amount of autocorrelation in the series to that expected under the null hypothesis of a unit root.

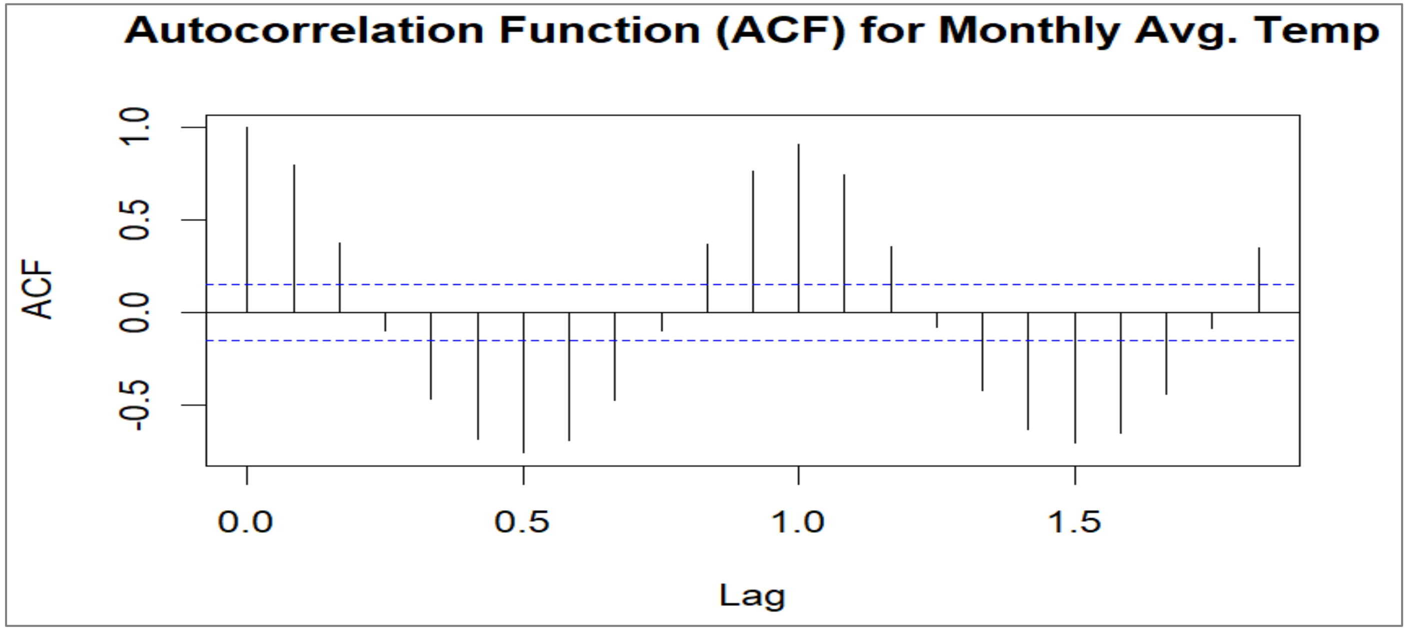
The significance of the test statistic is compared against critical values from statistical tables to determine if the null hypothesis should be rejected.

Conclusion: Hence by the ADF test in R, the data is stationary since p-value=0.01<0.05,
we get the Time series data (Monthly Avg. Temperature over Different years) is Stationary.
So, we can use ARIMA model.

(iii) ACF & PACK check :

❑ ACF Check:

The ACF measures the correlation between a series and its lagged values. It helps identify the order of the Moving Average (MA) component (q) in an ARIMA model. Significant autocorrelations at higher lags suggest a non-zero MA order.

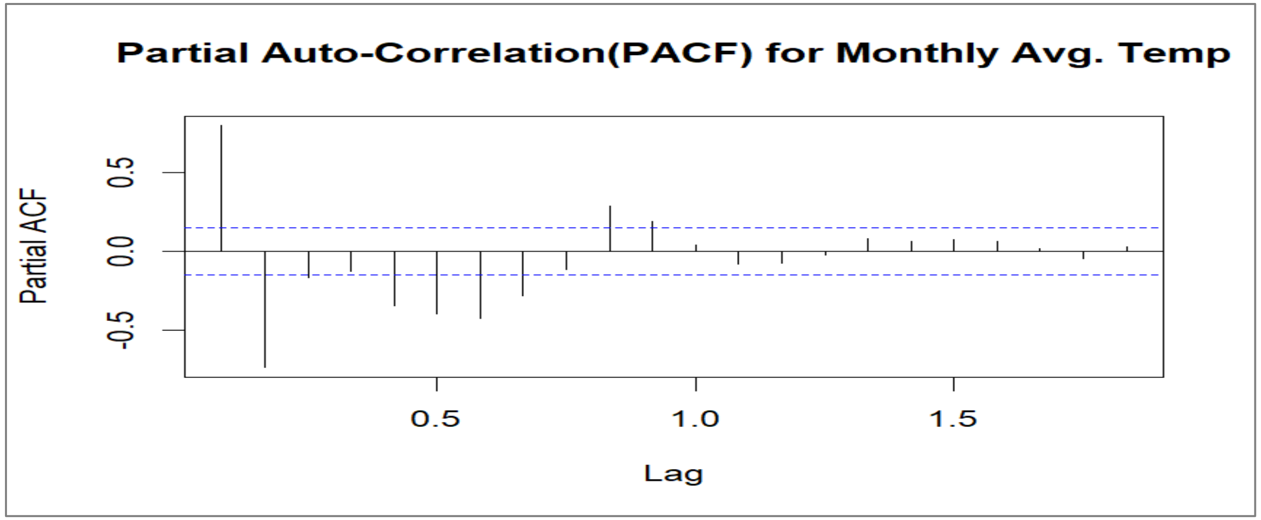


Conclusion:

Here the significant spike is at lag 1 in ACF plot, this suggests MA (1) model i.e. q=1.

❑ PACF check:

The PACF measures the correlation between a series and its lagged values, removing the effect of intermediate lags. It helps identify the order of the Auto-Regressive (AR) component (p) in an ARIMA model. Significant partial autocorrelations at higher lags suggest a non-zero AR order.

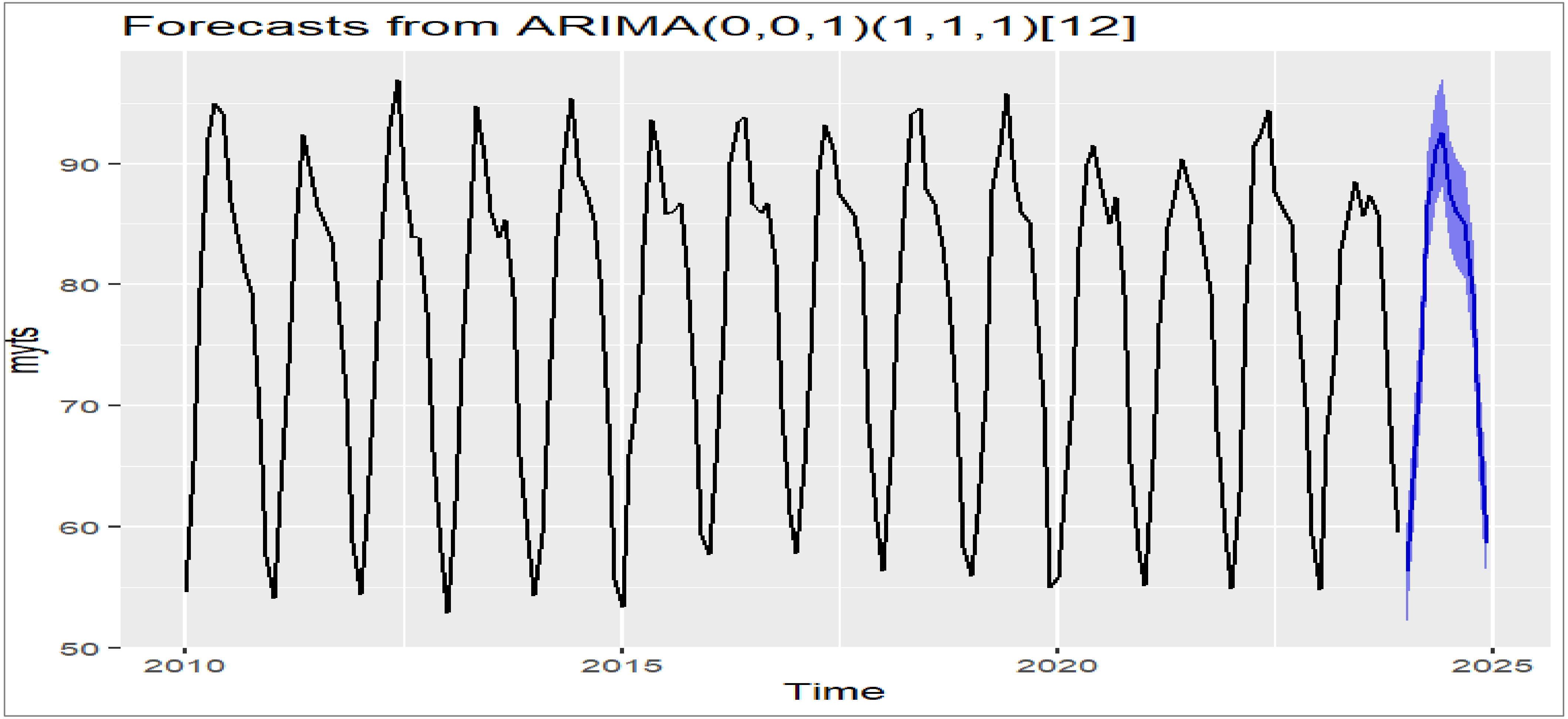


Conclusion:

Here the significant spike is at lag 1 in PACF plot, this suggests AR (1) model i.e. p=1.

(iv) ARIMA Model Fit to forecast :

ARIMA (1,1,1) Fitting: By the above procedure mentioned in 'Methodology' we fit the ARIMA (1,1,1) model to the time series data and get the result below with a forecasted Blue Line shaded area for year 2024.Which shows a possible pattern curve, what the temperature of the year Jan,2024 to Dec,2024 would possibly occur.



Auto-Arima Fitting and Forecast 2024 Graphical Data over 12 months

(v) Summary(Forecast):

```
Forecast method: ARIMA(0,0,1)(1,1,1)[12]
Model Information:
Series: myts
ARIMA(0,0,1)(1,1,1)[12]
Coefficients:
         ma1      sar1      sma1
      0.4433   -0.0435   -0.8567
s.e.  0.0714    0.1095    0.1102

sigma^2 = 4.274:  log likelihood = -341.55
AIC=691.09   AICc=691.36   BIC=703.29

Error measures:
                      ME     RMSE      MAE        MPE
Training set -0.02607379 1.972834 1.557079 -0.06046625
    MAPE      MASE
 2.066159 0.7406791
         ACF1
Training set 0.03837686
```

Forecasts:

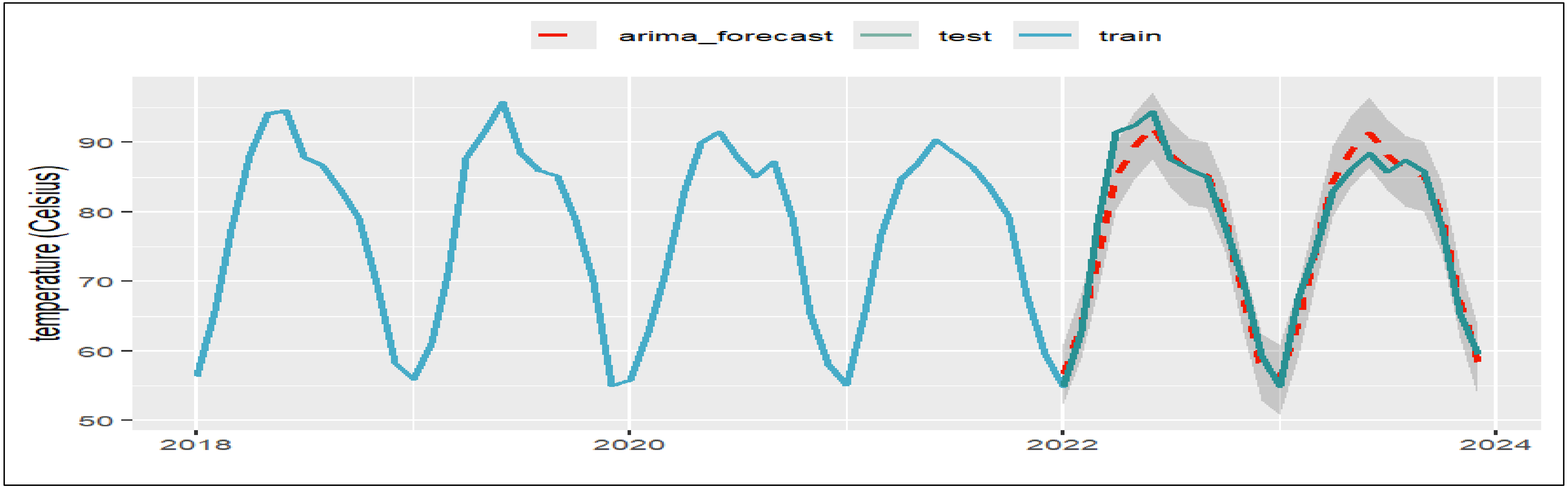
	Point Forecast	Lo 95	Hi 95
Jan 2024	56.27784	52.21825	60.33744
Feb 2024	63.97855	59.53840	68.41869
Mar 2024	74.65307	70.21293	79.09322
Apr 2024	86.54915	82.10900	90.98929
May 2024	91.19727	86.75712	95.63742
Jun 2024	92.54463	88.10448	96.98477
Jul 2024	87.43887	82.99873	91.87902
Aug 2024	86.03982	81.59968	90.47997
Sep 2024	85.03061	80.59047	89.47076
Oct 2024	79.28753	74.84738	83.72767
Nov 2024	68.24838	63.80823	72.68853
Dec 2024	58.60252	54.16273	63.04231

(vi) Test the Accuracy of forecasting of Model for Year Interval 2023-24 :

Because we have data over 13 years, we can split them into a train/test sample :

- Train sample : Data of Year 2010 to 2021
- Test sample : Year 2022-2023
- Run ARIMA prediction over train sample
- compare with (true) test data

Conclusion: From the above test we see Auto-Arima(red) tends to underestimate the true data (green). Hence, we can say our ARIMA model almost accurate forecast for Temperature of year 2022 and 2023. Therefore, our previous ARIMA (1,1,1) model also predicted most possible and accurate temperatures for the whole year 2024.



ARIMA Forecasting of Year 2023 by Train and Testing Sample

(vii) Correlation with other features :

Though there are some other remaining variables like Wind Speed, Wind pressure, Dew Points and Humidity. If we can find similarly the forecasted values for these weather variables for Year 2024 with the help of ARIMA model.

Then we can correlate these variables with each other like (Temperature, Humidity), (Temperature, Wind Pressure), (Wind Speed, Wind Pressure),

(Dew Points, Humidity) etc. Thus, we can get total forecasted possible Weather data for the whole Year 2024.

- Avg. Temp & Avg. Humidity
- Avg. temp & Avg. Wind Speed
- Avg. temp & Avg. Dew points
- Avg. temp & Avg. Wind Pressure
- Wind Speed & Wind Pressure

Summary :

- In ‘Data Visualisation’ Both the Heatmap and Boxplot shows the same seasonality i.e. Temperatures are higher in the summer (although it seems the max. temperature is reached around May/June). We can’t say only by eye observation a potential increase of this behaviour vs. Year; 13 years data is not enough to notice a global increase ‘warming’ in New Delhi.
- After ‘time series decomposition’ we get the seasonal and trend of data, there we check the data stationarity by ADF test and got it Stationary.
- After fitting ARIMA (1,1,1) model we forecast the monthly average temperature of the Year 2024 and got the forecasted values.
- To check the ARIMA model accuracy we train it with data of year 2010 to 2021 and then test forecast of year 2022 to 2023. There we see Auto-Arima(red) tends to underestimate the true data (green). Hence, we can say our ARIMA model almost accurate forecast for Temperature of year 2022 and 2023.
- Therefore, our previous ARIMA (1,1,1) model also predicted most possible and accurate temperatures for the whole year 2024.

If we also use the forecast model after stationarity checking of the other remainder variable like Humidity, Wind Speed, Wind Pressure and Dew points, we also got some forecasted values for the year 2024. And we could find some correlation between these variables. Thus we can get the compact possible Forecasted Weather Data for New Delhi for Year 2024.

Scope:

- **Seasonal and Annual Variations:** Time series analysis helps in understanding seasonal variations in weather patterns, such as temperature fluctuations across different seasons or annual cycles in rainfall.
- **Extreme Event Prediction:** Identifying and predicting extreme weather events, such as hurricanes, heatwaves, or heavy rainfall, is crucial for disaster preparedness and risk management. Time series models can detect patterns indicative of such events based on historical data.
- **Short-Term Forecasting:** Time series analysis allows for the prediction of weather conditions in the near future, typically from hours to a few days ahead. Techniques like ARIMA models or exponential smoothing can capture short-term trends and patterns in weather data.
- **Data-Driven Decision Making:** Weather forecasts derived from time series analysis support decision-making in various sectors, including agriculture, transportation, energy, and emergency response. Accurate predictions enable better planning and resource allocation.

References:

- Data Source:
- https://www.kaggle.com/datasets/ashx010/weather-patterns-and-trends
- Website:

'Kaggle' ,'R-pubs', 'Git-hub', 'R-tutorial', 'ChatGPT' etc.

- Books:

1.'Fundamentals of Applied Statistics'-S.C. Gupta, V.K. Kapoor

2.'An Insights into Statistics'- Sarkhel, Dutta

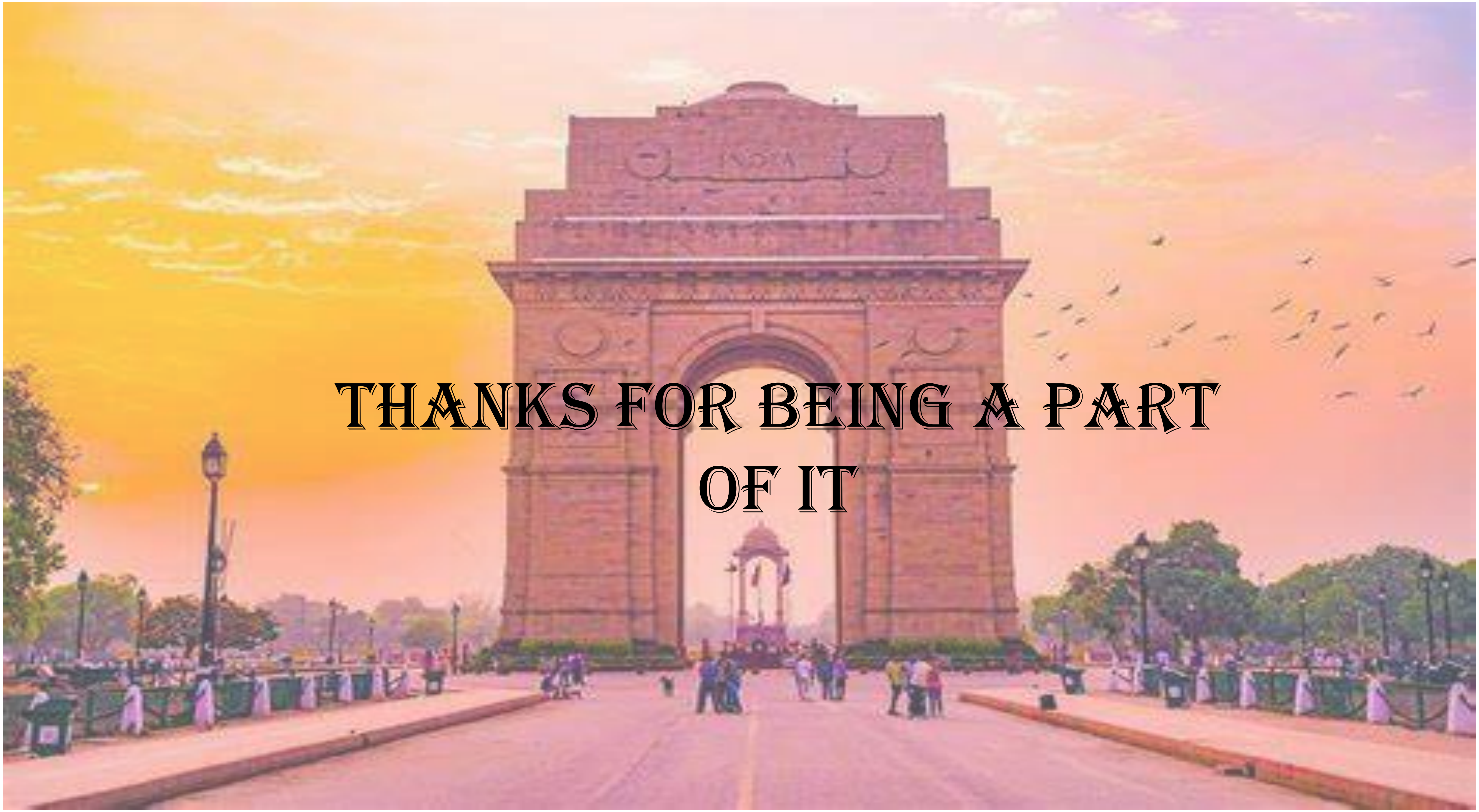
- Image:

'Google'

Acknowledgement :

- I would like to express my gratitude to my teacher Rittik Hazra Sir, who gave me the golden opportunity to do this wonderful project of Statistics on **"TIME SERIES ANALYSIS"**, and helped me in completing my project.

- I would also like to extend my gratitude to the HOD Ma'am, and Professor Bodhoditya Barma Sir, and Miss Swikriti Kundu Ma'am, for providing me with all the facility that was required during this project.


THANKS FOR BEING A PART OF IT