Time Series Analysis on Air Quality Index (AQI) Data

1st Shravan Kakadiya
Information and Communication Technology
DAU
Gandhinagar , India
202201333@daiict.ac.in

2nd Krishil Jayswal
Mathematics and Computing
DAU
Gandhinagar , India
202203040@daiict.ac.in

***Abstract*—In this project, we analyze time series data for Air Quality Index (AQI) over several years using classical statistical models. We examine key properties of the data including mean, variance, autocorrelation, data distribution, trends, seasonality, and stationarity using methods like fitting different distributions, Rolling Mean, Data Transformations, Moving Averages and Differencing and use models like AR, MA, and ARMA, find the best fit and forecast AQI behavior. Our conclusions from this analysis are...**

I. Introduction

In this project, we have Air Quality Index (AQI) (on daily basis) dataset of **Gandhinagar** scraped from the Central Control Room for Air Quality Management website and use time series analysis to understand the hidden patterns in this dataset. Our primary goal is to extract these hidden patterns (trend and seasonality) to make them stationary and do the forecasting on the dataset. Here is a quick snapshot of the dataset.

	Date	AQI
0	2019-05-01 00:00:00+00:00	140
1	2019-05-02 00:00:00+00:00	214
2	2019-05-03 00:00:00+00:00	144
3	2019-05-04 00:00:00+00:00	130
4	2019-05-05 00:00:00+00:00	116

Fig. 1: First few rows of the dataset representing AQI values at a given timestamp.

From the figure, we can see that there are two columns in the dataset **Date** and **AQI**. We collected the AQI data from year **2019** to **2025** on **daily** basis. There are some interpretations of the value of the AQI, the following table in Fig. 2 shows that.

AQI	Remark	Color Code	Possible Health Impacts
0-50	Good		Minimal impact
51-100	Satisfactory		Minor breathing discomfort to sensitive people
101-200	Moderate		Breathing discomfort to the people with lungs, asthma and heart diseases
201-300	Poor		Breathing discomfort to most people on prolonged exposure
301-400	Very Poor		Respiratory illness on prolonged exposure
401-500	Severe		Affects healthy people and seriously impacts those with existing diseases

Fig. 2: Label of AQI at different range of values.

we plotted the data against time and tried to derive some observations.

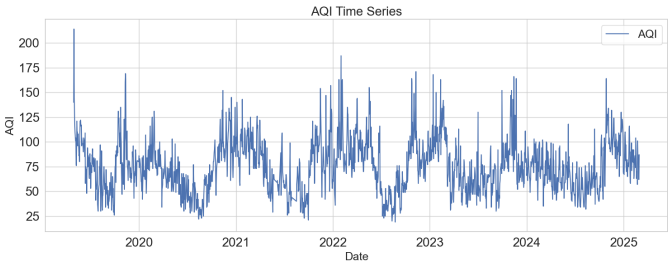

Fig. 3: Plot of AQI over the time.

Observations :-

- Seasonality: Recurring peaks suggest seasonal trends, likely higher AQI in winters.
- Early 2019 Dip: Sharp decline, possibly due to lockdown or external events.
- Post-2021 Fluctuations: Increased variability with more extreme AQI values.

- Late 2024–2025 Drop: Notable decline, possibly due to improved air quality or policy impact.
- AQI Range: Values range from 19 to 214, indicating periods of both good and poor air quality.

Now that we have a brief understanding of the dataset, we can start to play with the data and apply different techniques and transformations.

II. Distributional View of Data

In this section, we will try to determine the underlying distribution of the data and try to estimate it (if possible) with the available distributions like **Gaussian** and **Cauchy**.

we plot the histogram and get an idea of the distribution of the data.

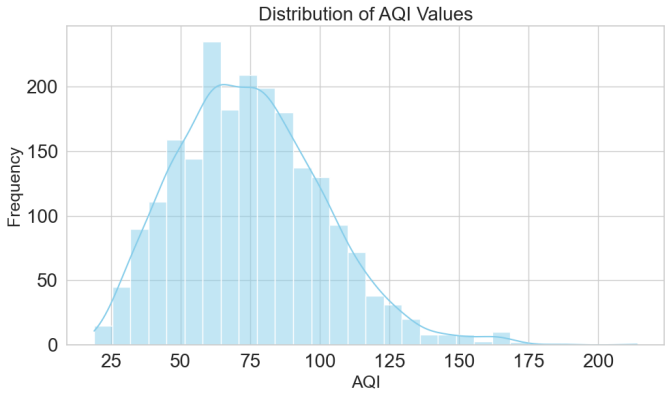

Fig. 4: Histogram plot of the AQI data.

Observation :- From Fig. 4, we can see that our distribution is kind of normal but slightly skewed.

In the next step we try to fit Gaussian and Cauchy Distribution to it.

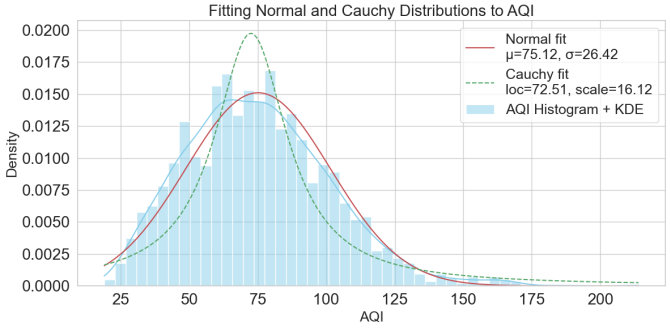

Fig. 5: Gaussian and Cauchy distribution fit to the original distribution of the data.

From Fig. 5, we see that Cauchy Distribution doesn't fit very well with the data. Gaussian fits better but still there are some miscoverages and a slight shift of the curve from the original distribution.

So we have to go with the other route that is weak stationarity.

III. Statistical Properties with Time

From the last section we are unable to determine the underlying distribution of the data so we try to check the presence of weak stationaroty. We plot the rolling mean, variance and Autocorrelation of the AQI data.

we used a rolling window of 30 days because our AQI data is on a daily scale, and a 30-day window helps us capture broader monthly trends while smoothing out short-term fluctuations.

In air quality studies, 30-day windows are often used because monthly patterns (festivals, weather cycles, crop burning) affect AQI. This makes it easier to analyze seasonal variations and check for consistency or changes in the mean and variance over time.

A 7-day window would be too sensitive to weekly changes it might catch short-term noise rather than meaningful trends.

On the other hand, a 60-day window would smooth the data too much and potentially hide important month-wise patterns.

So, 30 days is a balanced choice long enough to reduce noise, but short enough to capture meaningful variation in AQI on a monthly basis.

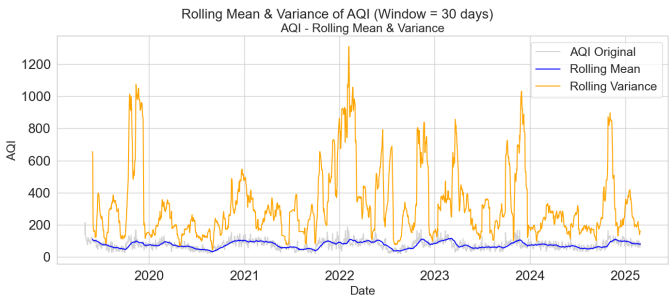

Fig. 6: Plot of rolling Mean and Variance with time of original data.

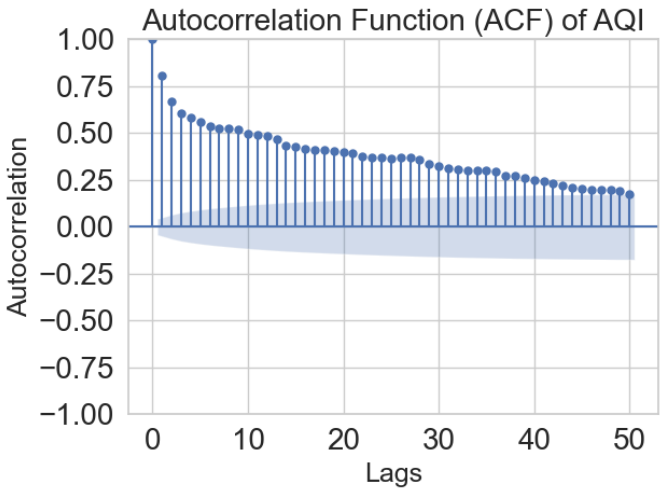

Fig. 7: Plot of Autocorrelation with lags of original data.

From Fig. 6 and 7, we can clearly state that the **statistical properties varies with time** and hence our time series data is **non-stationary**. Now to make this stationary we will apply some techniques in the next section to make it stationary.

IV. Detrending and Deseasoning for Stationarity

First we will plot the additive and multiplicative decomposition with period size of **365 days** as observed from Fig. 3.

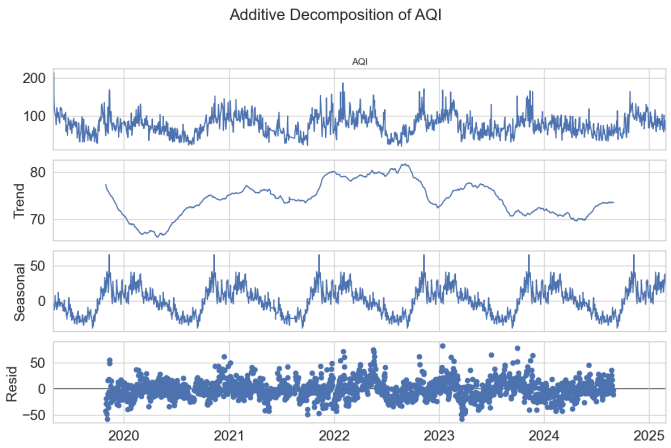

Fig. 8: Additive Decomposition of the AQI data.

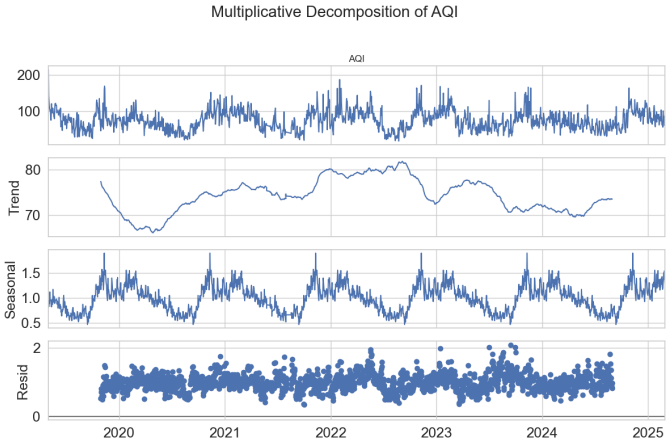

Fig. 9: Multiplicative Decomposition of the AQI data.

The reason of plotting this both plots is to just see the nature of trend(linear/non-linear) and sesonality(period) of the series if any. From both decompositions, we can observe that we have a **non-linear** trend and a seasonality of period **365 days.**

From section 2, we have observed that our data follows a skewed distribution.

Also from section 3, we observed that the variance varies very rapidly with time. So this both observation implies that we have to stabilize the variance first.

To stabilize the variance we use the **log** transformation. After the transformation, the rolling mean and variance will become as follows:

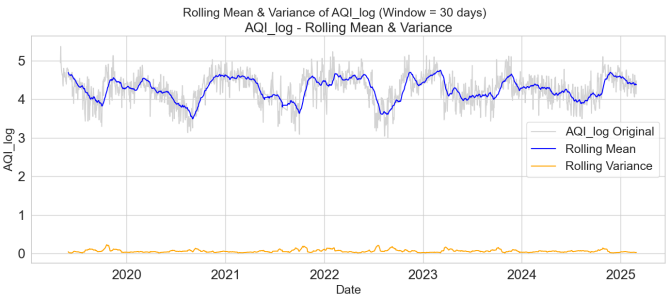

Fig. 10: Plot of Mean and Varicance with time after log transformation.

Here we can see that the variance is stabilized, but the mean is still varying with time. For that part we will again decompose the series and observe the trend.

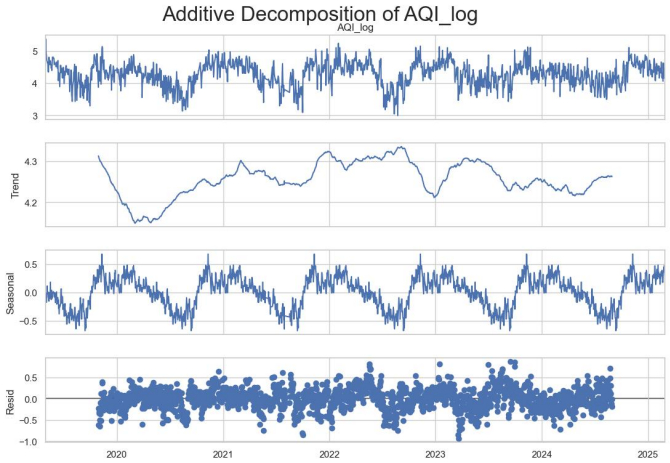

Fig. 11: Additive Decomposition after Log Transformation.

From the decomposition, we can observe that the scale of fluctuations of trend has decreased because of log transformation and also we can observe a local linear trend in Fig. 11.

So if we assume a locally linear trend(original locally exponential trend) then, we can use the method of differencing to remove that trend.

So to stabilize the mean, we can use the **first differencing**. After the transformation, the statistical properties becomes as follows:

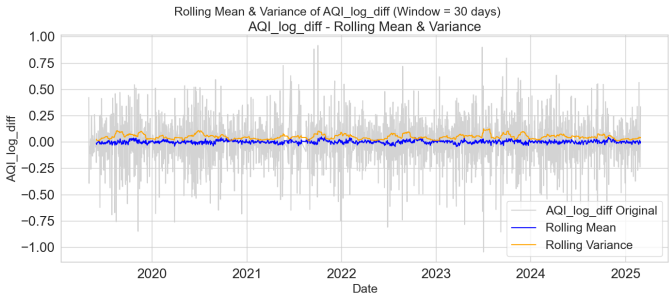

Fig. 12: Plot of Mean and Variance with time after first differencing on the log transformed data.

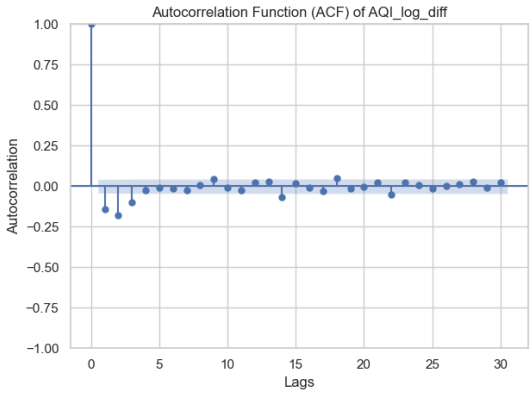

Fig. 13: Plot of Autocorrelation after first differencing on log transformed data.

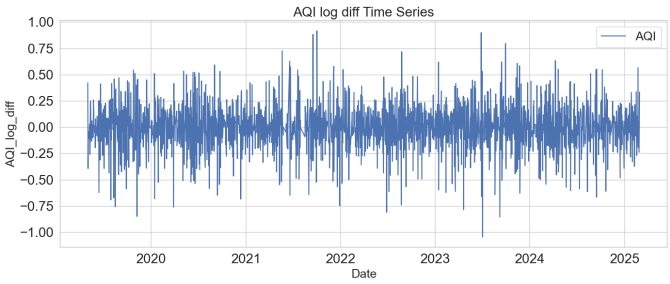

Fig. 14: Plot of log differenced AQI over time.

From Fig. 12, we can see that our rolling mean and variance have stabilized and doesn't vary much with time. So we can treat them as constant.

From Fig. 13, we can see that our autocorrelation also depends only on lag.

So from this observations, we can conclude that our modified series have become **stationary**.

V. Autocorrelation Analysis, Model Selection and Forecasting

In this section, we try to figure out the most suitable model we can use for the obtained stationary time series.

From the autocorrelation plot in the last section, we can see that the autocorrelation is significant upto **3 lags** and then drops off. Also there is some uncertainity at lag 3 whether to consider it or not.

So our series is possibly a **MA(2)** or **MA(3)** process.

To verify this, we fitted both **MA(2)** and **MA(3)** model and used the **AIC** criteria to choose the best

model so that we dont have redundant parameters in our model. Here are the **AIC** scores of both the models.

```
MA(2) AIC: -766.1279150729997
MA(3) AIC: -813.5223440844743
```

Fig. 15: AIC Score of MA(2) and MA(3) model.

We can see that score of **MA(3)** model is much lower than **MA(2)** model. So we choose **MA(3)** model as best fit.

Again to verify this we done the analysis on the residuals and plotted the **Autocorrelation**, **Q-Q** and **Histogram** plot of the residuals.

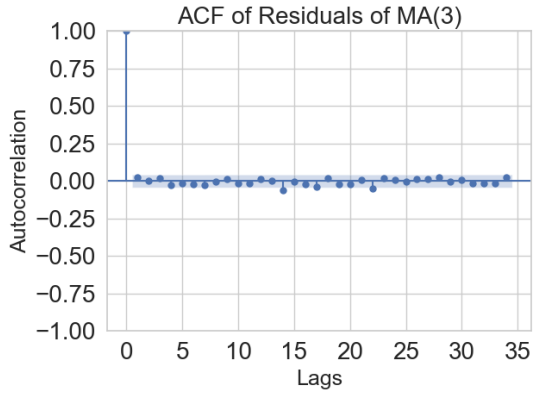

Fig. 16: Autocorrelation plot of the Residuals.

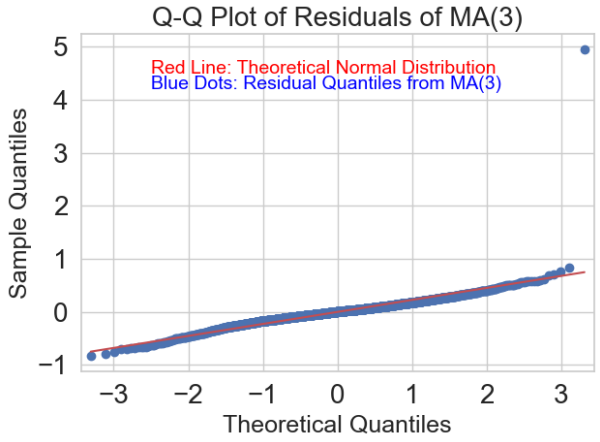

Fig. 17: Q-Q plot of the Residuals.

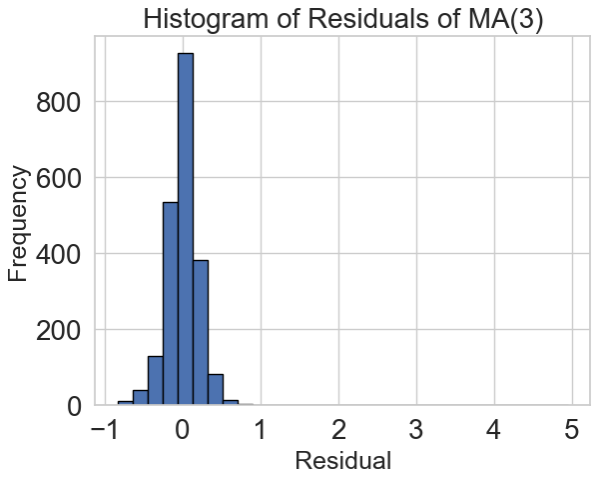

Fig. 18: Histogram plot of the Residuals.

From the plots, the residuals statisfies all the required properties (IID white noise with normal distribution). And we forecasted the AQI value of three days from **27 Feb 2025**. Here are the comparison of predicted vs actual values (actual values taken from the website).

	Forecasted AQI	Actual AQI
2025-03-01 00:00:00+00:00	85	86
2025-03-02 00:00:00+00:00	80	77
2025-03-03 00:00:00+00:00	77	65

Fig. 19: 3 days prediction of the AQI value from 27 feb 2025.

VI. Causality and Invertibility

To evaluate the invertibility of the MA(3) process, we examine the roots of its characteristic polynomial:

$$1 + \theta_1 L + \theta_2 L^2 + \theta_3 L^3$$

where L is the lag operator and the estimated MA coefficients are:

$$\theta_1 = -0.2845, \quad \theta_2 = -0.2761, \quad \theta_3 = -0.1532$$

We compute the roots of the polynomial:

$$1 - 0.2845L - 0.2761L^2 - 0.1532L^3 = 0$$

The roots are:

$$-1.50 \pm 1.79i, \quad 1.20$$

Taking the modulus of the roots:

$$|z_1| = |z_2| = 2.33, \quad |z_3| = 1.20$$

Since modulus of every root is greater than 1, the MA(3) process is invertible. This implies the process can be expressed as an infinite-order AR process (AR(∞)).

MA (Moving Average) processes are inherently causal by construction, meaning current values are already expressed in terms of past and current white noise. Therefore, causality is not a concern for MA processes and does not require separate verification.

Thus the stationary series we obtained after transformations is **causal** and **invertible**.

VII. Conclusions

In this project, we conducted a detailed time series analysis of daily Air Quality Index (AQI) data from 2019 to 2025. We began by exploring the distributional characteristics of the raw data and observed a skewed distribution with non-stationary behavior. We also attempted to fit both the Normal and Cauchy distributions to our AQI data but they does not perfectly capture the shape of the AQI distribution. So we are unable to determine the underlying distribution.

From the time series plot and decomposition, it is evident that the AQI series exhibits a clear seasonal pattern with a periodicity of 365 days, along with a nonlinear trend. The trend appears to be locally exponential, as applying a logarithmic transformation followed by differencing (log-diff) renders the series stationary. This reinforces the conclusion that the original series is non-stationary in nature. Furthermore, the stationarity achieved after the log-differencing transformation suggests that the underlying process is multiplicative in form: **Xt=Ut*St*Et**

To address the non-stationarity, we applied logarithmic transformation to stabilize variance, followed by differencing to remove trend components. This resulted in a stationary series suitable for modeling. And as seen earlier that we are unable to determine the underlying distribution of AQI series so stationarity achieved is the weak stationarity. For strong stationarity, distribution of data should be known.

Through autocorrelation analysis, we identified a suitable MA(3) model using the AIC criterion. The model was validated through residual diagnostics, showing the residuals behaved like white noise. The model was then used to successfully forecast AQI for the next three days, and predictions were found to be close to actual values.

We also verified that the MA(3) process is invertible, meaning it can be equivalently represented as an AR(∞) process. Since MA processes are inherently causal, the resulting model is both causal and invertible.

VIII. Future Work

This analysis focused on the AQI as a univariate time series. However, AQI is derived from the maximum of several pollutant concentrations (e.g., PM2.5, PM10, NO_2, O_3, CO, SO_2). A deeper investigation could involve analyzing these underlying pollutant time series individually to understand their behavior, seasonality, and relative contributions to the overall AQI. Studying their cross correlations can provide more insights into pollutant interactions and sources.

Additionally, future work could involve applying Seasonal ARIMA (SARIMA) or SARIMAX models to better capture the seasonal components and incorporate exogenous variables such as temperature, humidity, wind speed, traffic levels, or policy interventions. These models can enhance forecasting performance and help in interpreting the effects of external factors on air quality.

Multivariate models like Vector AutoRegressive (VAR) or VAR with exogenous variables (VARX) could also be considered when dealing with multiple pollutant series and external drivers. This would allow for a more comprehensive and dynamic modeling of the air quality system.

IX. References

1) Government Website Link

2) AQI Dataset Link of Gandhinagar Sector-10

3) Jupyter Notebook Link