Contents

1 Introduction

This report explores how changes in the Air Quality Index (AQI), particularly due to PM2.5 levels, affect consumer behavior in the air purifier market. Air pollution is a significant concern, especially in densely populated cities like Delhi, which has seen hazardous pollution levels recently. As air quality worsens, people are increasingly aware of the health risks posed by fine particulate matter, such as PM2.5, known to harm the respiratory system.

In light of these health risks, demand for air purifiers has risen sharply during periods of unhealthy or hazardous AQI readings. This creates a direct link between environmental conditions and consumer buying habits. Our report aims to connect environmental data analysis with business strategies using advanced time series modeling techniques.

We have two main objectives:

1. Develop accurate forecasting models for PM2.5-based AQI values using historical data. These forecasts are essential for public health planning and can benefit businesses sensitive to environmental changes.

2. Examine how air quality trends influence air purifier sales. By analyzing the relationship between AQI levels and consumer purchasing behavior, we aim to provide insights that can enhance inventory management, marketing strategies, and sales forecasts for companies in the industry.

Our analysis shows the practical value of time series forecasting beyond traditional environmental monitoring. By linking AQI predictions to business outcomes, we illustrate how data-driven approaches can help companies adapt to changing market conditions. Ultimately, this report emphasizes the importance of integrating environmental analytics into business decisions to better serve customers while promoting public health and well-being.

2 Modeling the Delhi AQI Time Series

2.1 Data Preprocessing and AQI Derivation

The AQI dataset originates from historical pollution data, including PM2.5 concentrations recorded daily over a span of more than 10 years. However, for modeling purposes, the analysis focuses on the subset from January 1, 2018, to March 31, 2023.

AQI Explanation

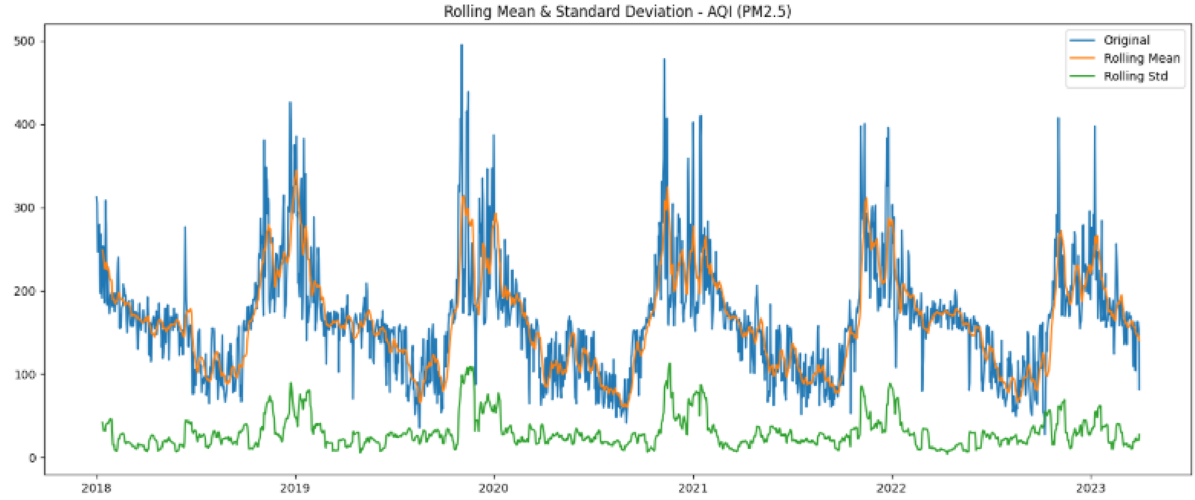
Primary Contributor	Estimated Contribution
Vehicular Emissions	41%
Industrial Activities	18%
Construction and Road Dust	21.5%
Stubble Burning (Seasonal)	30–38%

Table 1: Pollution Sources and Their Estimated Contribution to Air Quality Degradation

Year	Avg. PM2.5 (µg/m^3)	Avg. AQI	Severe AQI Days	Key Events
2018	209	338	24	GRAP implemented
2019	205	345	19	Supreme Court bans petcoke
2020	94	185	6	COVID-19 lockdowns
2021	209	462	15	Post-Diwali AQI peak
2022	100	204	15	CAQM directives
2023	100.9	367	13	Low winter winds

Table 2: Annual summary of average PM2.5 concentration, average AQI, number of severe AQI days, and key air quality events in Delhi from 2018 to 2023.

The AQI is computed from PM2.5 using the EPA's breakpoint method, which transforms raw PM2.5 concentrations into AQI scores on a scale from 0 to 500. This transformation is non-linear and piecewise, based on pre-defined concentration intervals and their corresponding AQI values. After calculating the AQI, the dataset undergoes a forward-fill interpolation to handle missing values. Only the daily AQI values are retained, and the index is converted into a daily frequency time series. By the end of this process, the AQI time series is clean, continuous, and ready for analysis.



2.2 Stationarity Check

Before fitting any model, assessing the stationarity of the AQI time series is crucial, as ARMA models require the data to have constant statistical properties over time. The Augmented Dickey-Fuller (ADF) test is applied, which evaluates the null hypothesis that the time series has a unit root, implying non-stationarity.

2.3 ADF Test

Augmented Dickey-Fuller (ADF) Test

The Augmented Dickey-Fuller (ADF) test is a statistical test used to determine whether a given time series is stationary or has a unit root, implying non-stationarity. The ADF test is an augmented version of the Dickey-Fuller test which includes lagged differences of the time series to account for higher-order correlation.

The general form of the ADF regression equation is given by:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta y_{t-i} + \varepsilon_t \tag{1}$$

where:

- y_t is the time series at time t,
- Δ denotes the first difference operator,
- α is a constant,
- βt is the coefficient on a time trend,
- γ is the coefficient used to test for stationarity,
- p is the number of lagged difference terms,
- ε_t is white noise.

Hypotheses:

- **Null Hypothesis (H_0):** The time series has a unit root (i.e., it is non-stationary).
- **Alternative Hypothesis (H_1):** The time series does not have a unit root (i.e., it is stationary).

For performing this test, we use a confidence level of 0.95. This means that if the p-value obtained from the test is less than 0.05, we reject the null hypothesis in favor of the alternative, concluding that the series is stationary. The same method will be used in the future to assess the stationarity of time series.

The result shows a test statistic of approximately -4.35 and a p-value around 0.0004. Since the p-value is far below the threshold of 0.05, the null hypothesis is rejected, and the series is considered stationary. This means the AQI values can be modeled without differencing.

Metric	**Value**
ADF Statistic	-4.347
p-value	0.00037
Critical Value (1%)	-3.434
Critical Value (5%)	-2.863
Critical Value (10%)	-2.568

Table 3: Results of the Augmented Dickey-Fuller test for stationarity on the daily AQI time series.

2.4 ARMA Model Fitting on Raw AQI

The ARMA (AutoRegressive Moving Average) model is a widely used forecasting method in time series analysis that combines two components: the autoregressive (AR) part and the moving average (MA) part. It is typically used to model stationary time series data.

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{2}$$

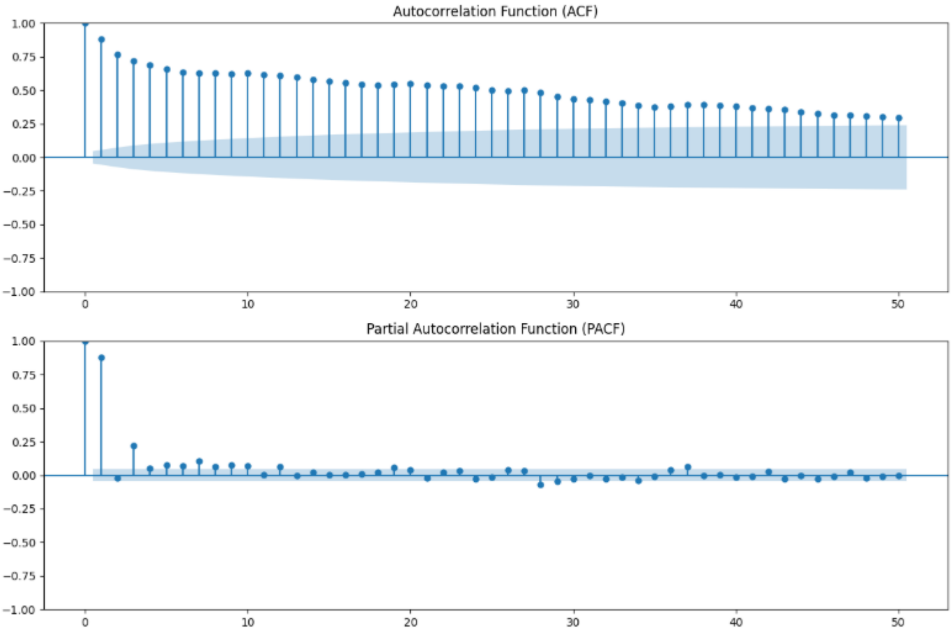
Model Parameters:

- μ: Constant term (mean of the time series).
- ϕ_i: AR coefficients for lagged values of the time series.
- θ_j: MA coefficients for lagged error terms.
- ε_t: White noise (error term) at time t.

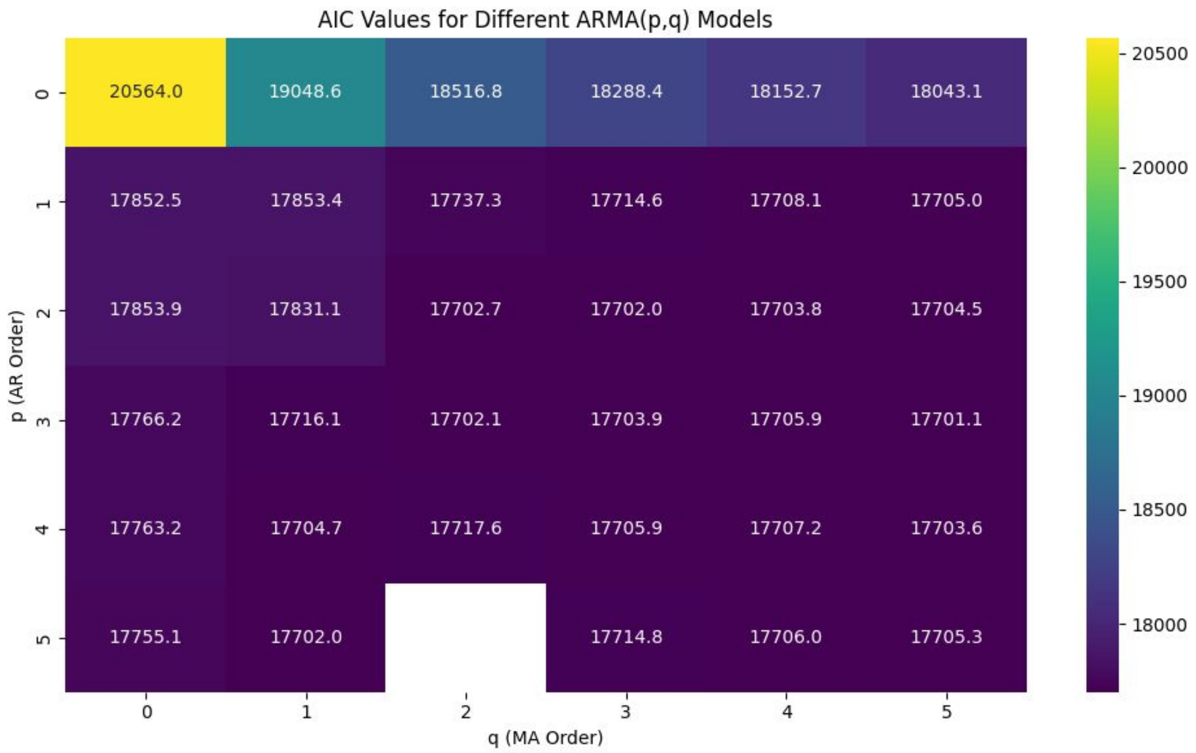
ARMA Model Components:

- **AR (AutoRegressive) part:** The current value of the time series is explained by a linear combination of its past values. The parameter p denotes the number of lag observations included in the model.
- **MA (Moving Average) part:** The current value is also influenced by the past forecast errors, represented by q lags of the forecast errors.

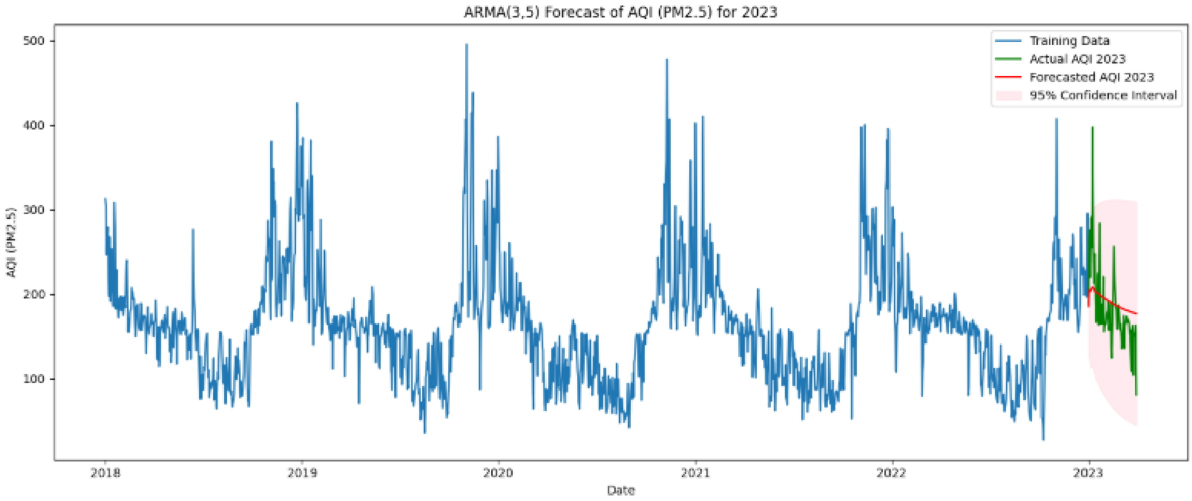
Autocorrelation and partial autocorrelation plots are used to visually inspect the data's memory structure.



However, to select the optimal ARMA(p, q) configuration more systematically, a grid search over all combinations of p and q in the range [0, 5] is performed. Each model is evaluated based on the Akaike Information Criterion (AIC), a measure that balances model fit and complexity.



The best model identified by AIC is ARMA(3, 5), with an AIC score of approximately 17,701. This model is fitted on the training period from 2018 to 2022, and then used to forecast AQI for the entire year of 2023. The model also provides 0.95 confidence intervals for each predicted value, allowing us to assess the uncertainty of the forecast. The root mean square error (RMSE) of the ARMA(3, 5) forecast against the actual AQI values of 2023 is calculated as 40.9. Although predictions generally follow the trend of actual values, error and residual plots reveal that the model struggles to fully account for seasonal effects and long-term trends.



2.5 STL Decomposition of AQI Series

To improve forecasting performance, the AQI series is decomposed into three components using STL (Seasonal-Trend decomposition using Loess): trend (T_t), seasonal (S_t), and residual (R_t). It uses locally weighted regression (LOESS) to estimate the trend and seasonal components of the series. The STL (Seasonal-Trend decomposition using Loess) method is an additive decomposition technique that splits a time series y_t into three distinct components:

- T_t: Trend component
- S_t: Seasonal component
- R_t: Remainder (residual or irregular) component

Main Decomposition Formula:

$$y_t = T_t + S_t + R_t \tag{3}$$

Detrending Formula:

$$y_t - T_t = S_t + R_t \tag{4}$$

Deseasonalizing Formula:

$$y_t - S_t = T_t + R_t \tag{5}$$

Residual Calculation:

$$R_t = y_t - T_t - S_t \tag{6}$$

In Seasonal-Trend decomposition using Loess (STL), the **Loess smoothing technique** plays a central role as a non-parametric method used to smooth the components of a time series. Each component is estimated using iterative **Loess smoothing**, which fits local weighted regressions to subsets of the data. The Loess smoother used in STL assigns weights w_{ij} to observations based on their distance from the target point x_i, using a kernel function K and a bandwidth parameter h, as in:

$$\hat{y}_i = \sum_{j=1}^{n} w_{ij} y_j, \quad \text{where} \quad w_{ij} = \frac{K\left(\frac{x_j - x_i}{h}\right)}{\sum_{k=1}^{n} K\left(\frac{x_k - x_i}{h}\right)}$$

This allows STL to flexibly adapt to local variations in the trend and seasonality without assuming a fixed parametric model.

This residual series is then subjected to another ADF test, which confirms that it is stationary (test statistic -10.71, p-value < 0.0001). The residuals are then modeled separately using an AR(2) model, as determined by further AIC-guided selection.

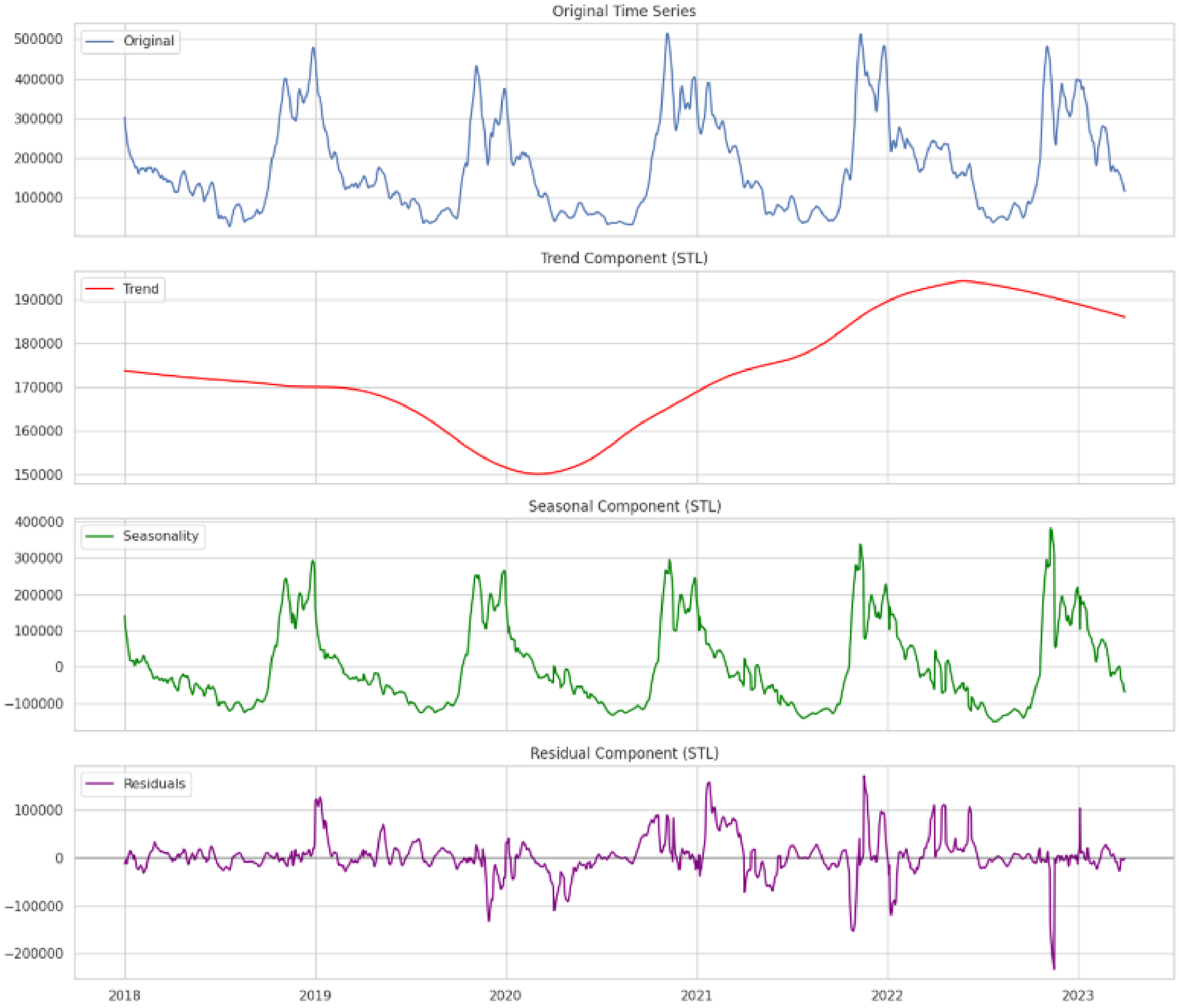


Figure 1: Time series decomposition

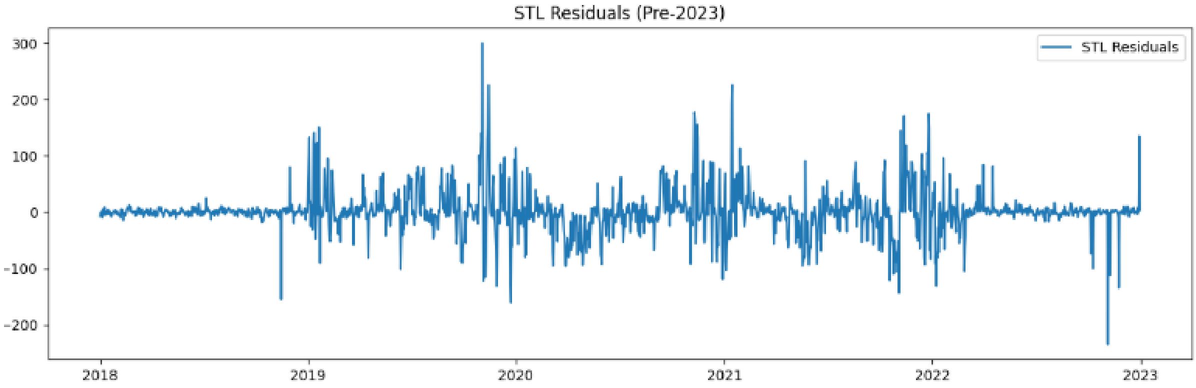


Figure 2: STL residuals (Pre-2023)

2.6 Forecast Reconstruction and Performance

The final forecast is reconstructed by summing the forecasts of the individual components:

$$\hat{X}_t = \hat{T}_t + \hat{S}_t + \hat{R}_t$$

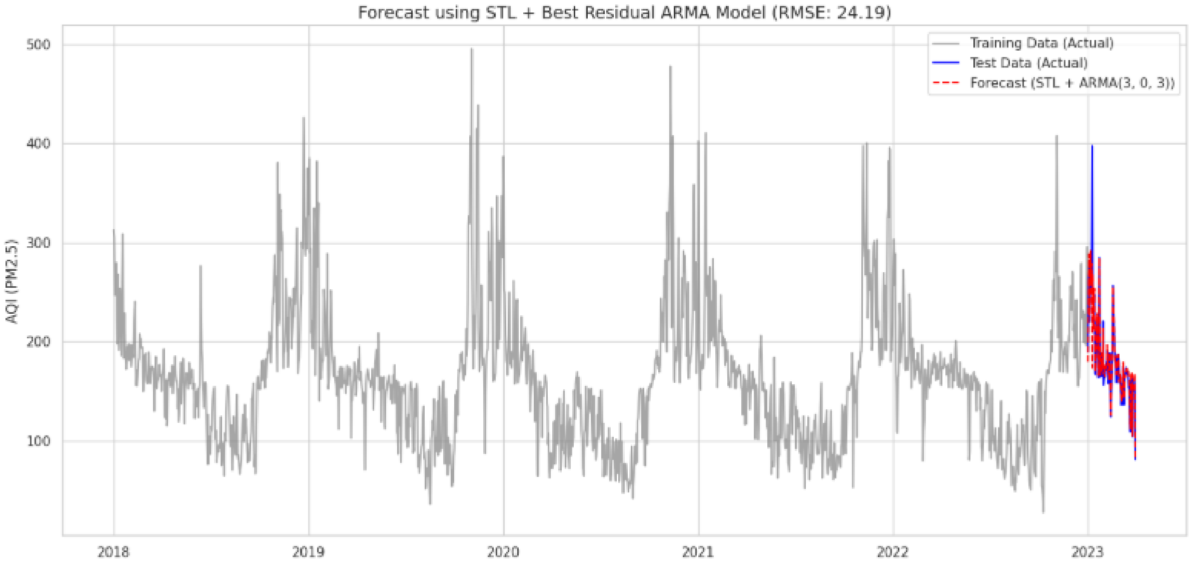
Where:

- $\hat{T}_t$: Trend component forecast, typically extrapolated from the last known trend values.
- $\hat{S}_t$: Seasonal component forecast, often obtained by repeating the last full seasonal cycle.
- $\hat{R}_t$: Residual component forecast, obtained from the ARMA model fitted to the residuals.

Predicted residuals from ARMA(3,3)
Evaluation:

- New RMSE (2023): 24.19
- Improvement: 41% reduction from initial ARMA-only forecast

This shows that STL + ARMA on residuals captures seasonality and irregular fluctuations better than direct ARMA modeling.



3 Modeling the Revenue Time Series

3.1 Data Cleaning and Structure

The revenue dataset contains daily total revenue figures from January 1, 2018, to March 31, 2023. Initial preprocessing involves replacing missing entries with zero values , then interpolating based on time and applying a centered rolling median with a window size of 5 to smooth out noise. This ensures continuity and reduces the effect of irregular spikes or drops in the data. Once cleaned, the time series is set to a daily frequency, with forward-filling used to handle any residual gaps. At this point, the revenue data is ready for decomposition and modeling.

3.2 Stationarity Assessment

The revenue series undergoes the Augmented Dickey-Fuller test, returning a test statistic of -4.37 and a p-value of 0.0003. As with the AQI data, this indicates the series is stationary and can be modeled using ARMA without differencing.

3.3 STL Decomposition of Revenue Series

As done with AQI, the revenue series is decomposed using STL into its trend, seasonal, and residual components. This decomposition reveals cyclical patterns in revenue—possibly due to monthly, quarterly, or holiday-related effects—and highlights the presence of long-term growth or decline in economic activity.

The residual component, isolated after removing trend and seasonality, is subjected to modeling using ARMA.

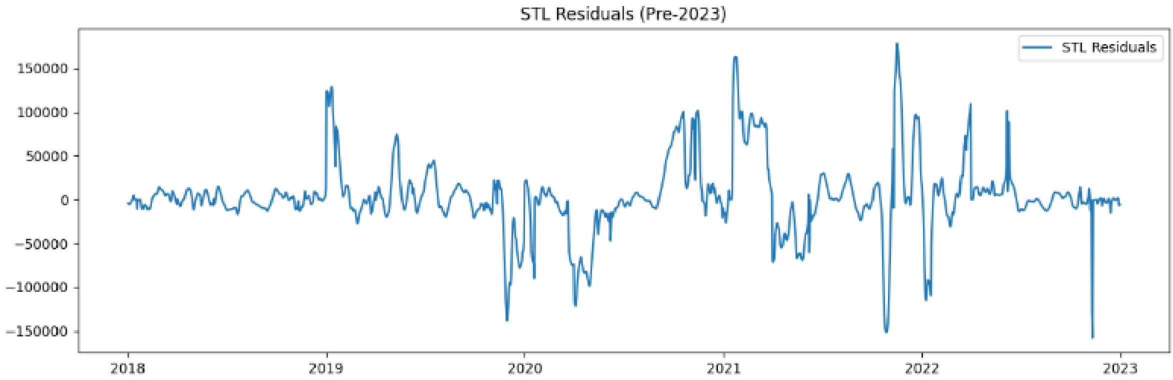


Figure 3: Revenue residuals

First plot the ACF and PACF during the training phase.

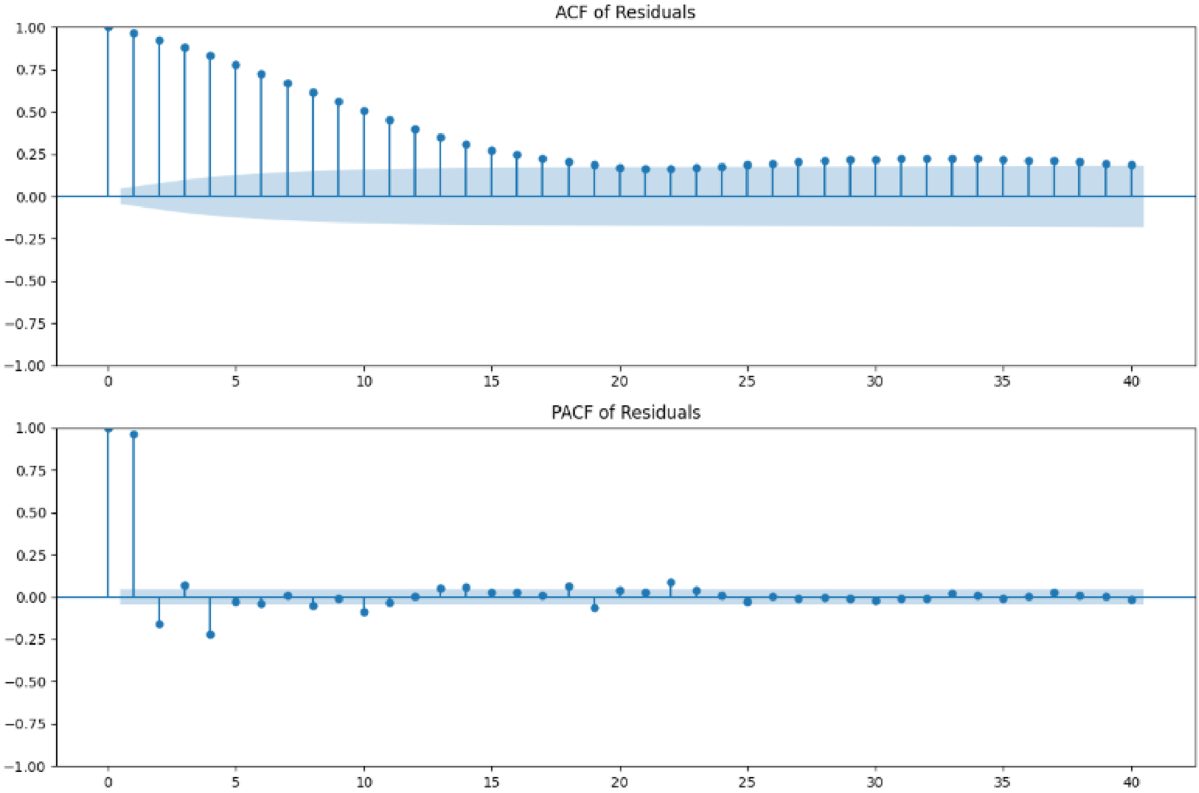
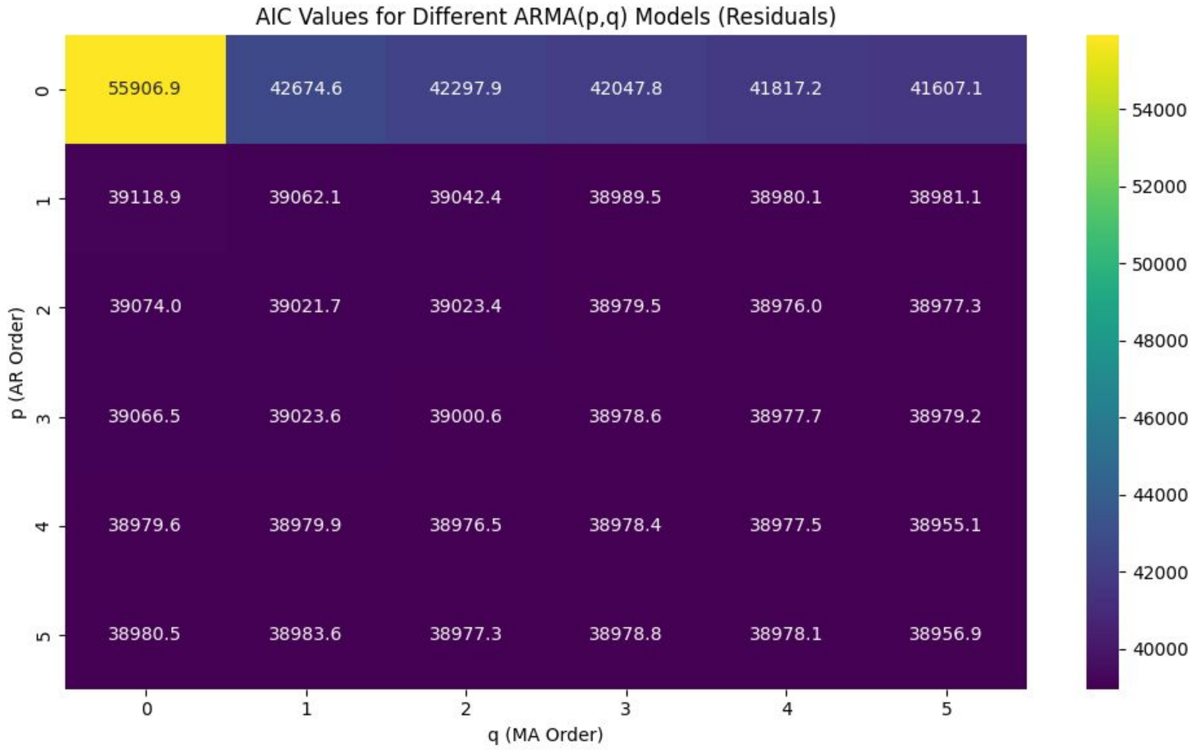


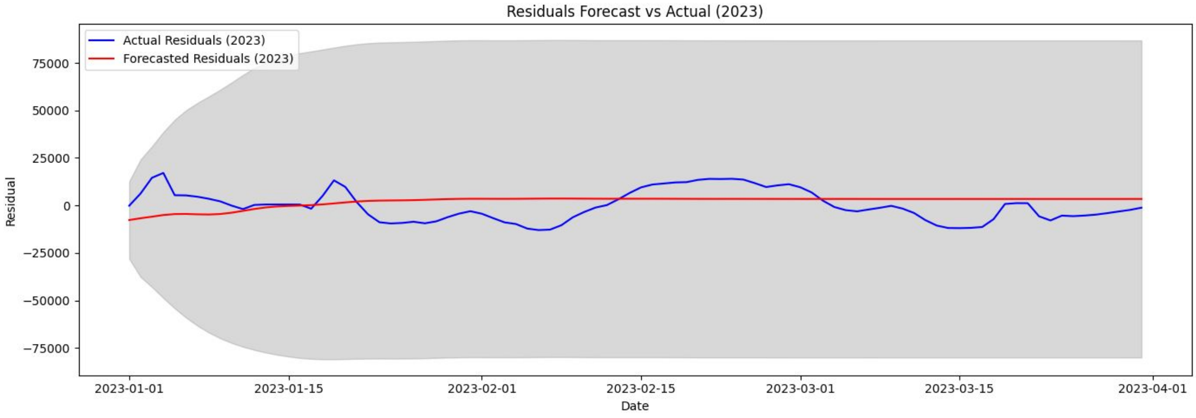
Figure 4: ACF and PACF of residuals during the training phase

Based on the plots, the ACF appears to be exponentially decaying and the PACF cuts off after lag 1, suggesting AR(1) as a possible initial guess. However, to determine the optimal model more systematically, a grid search was conducted over p and q values in the range [0, 5], selecting the model with the lowest AIC. The optimal residual model is found to be ARMA(4, 5), with an AIC of approximately 38,955.



3.4 Forecasting Revenue with Residual Model

The residual ARMA(4, 5) model is trained on data up to 2022. Forecasts for 2023 are generated, and then combined with trend and seasonal components to reconstruct the complete revenue forecast for 2023.



The final forecast is reconstructed as:

$$\hat{X}_t = \hat{T}_t + \hat{S}_t + \hat{R}_t$$

Where:

- $\hat{T}_t$: Trend component forecast, extrapolated from the last known trend values.
- $\hat{S}_t$: Seasonal component forecast, repeating the last full seasonal cycle.
- $\hat{R}_t$: Residual component forecast, from the fitted ARMA(4, 5) model.

Confidence intervals at 0.95 are computed for the residual forecast and propagated into the final revenue forecasting.

These bounds allow us to construct a reliable envelope of uncertainty for each day's forecast. A comparison of predicted revenue and actual revenue in 2023 shows improved alignment, with most observed values falling within the confidence bounds.

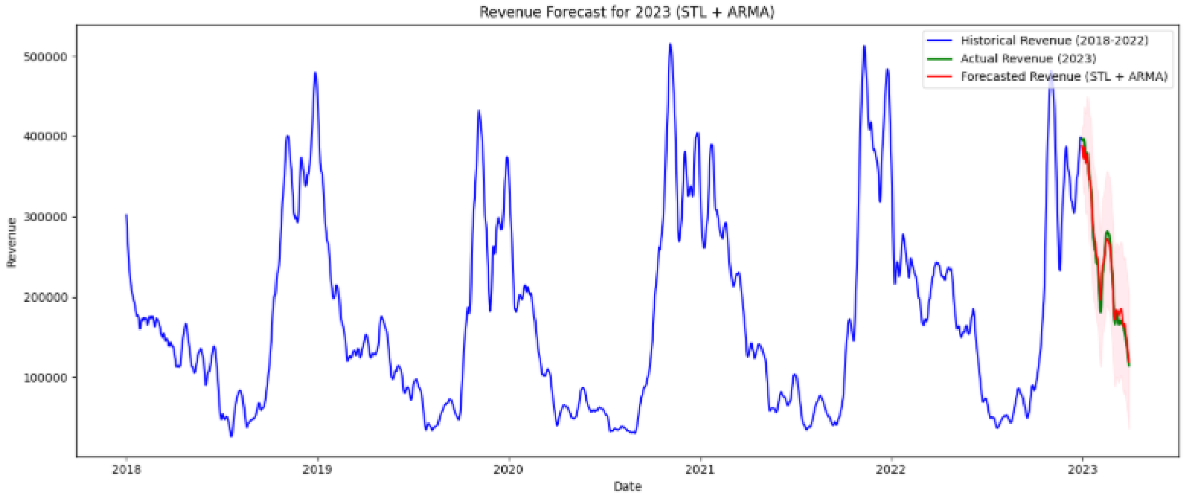


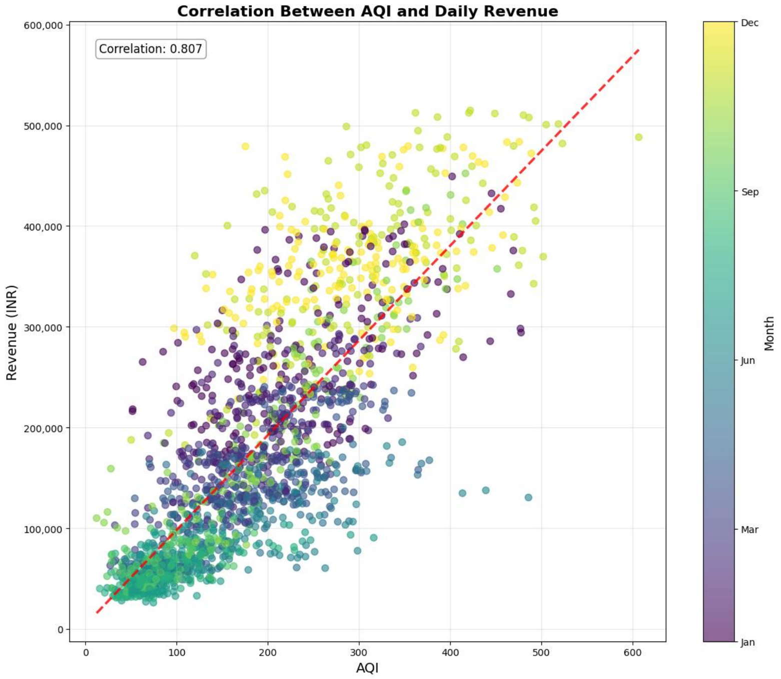
Figure 5: Revenue Forecast for 2023 (STL + ARMA)

4 Final Synthesis: Strategic Insight from Forecasting

This section consolidates our key findings by integrating **statistical modeling outcomes** with **consumer behavior insights**, providing a holistic understanding of how Delhi's deteriorating air quality translates into commercial opportunities and strategic decisions for air purifier companies like Blue Star.

4.1 Key Insights and Contributions

- **High Correlation:** There is a high positive correlation between AQI levels and Blue Star's revenue, notably during the months of November to January. This indicates that AQI acts as a leading indicator for sales, which can inform marketing and inventory strategies.



- **Seasonal Patterns:** Revenue peaks are distinct during winter months when AQI exceeds 300, enabling predictable demand cycles and supply chain optimization.

- **Consumer Behavior:** There is a sharp increase in purchases during pollution spikes, with up to a 70% sales surge, highlighting health-driven consumption and the market's responsiveness to environmental changes.

- **COVID-19 Impact:** The year 2020 showed anomalies due to lockdowns, resulting in the lowest AQI and revenue. This validates the model's sensitivity to external shocks.

- **Post-2020 Growth:** There has been a sharp rebound in AQI and sales from 2021 to 2023, with record values in 2023, indicating rising health awareness and increasing market penetration.

- **Modeling Performance:** The use of STL + ARMA reduced AQI forecast RMSE by **41%**, and revenue forecasts effectively captured seasonality, demonstrating the power of advanced time series techniques in real-world forecasting.

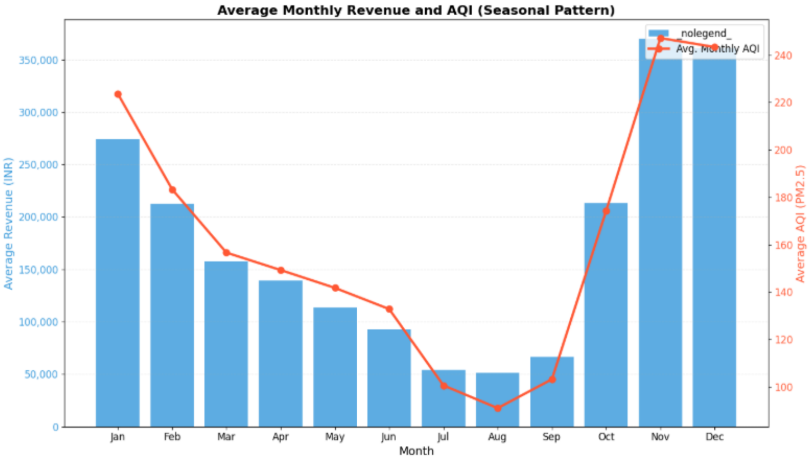
4.2 Broader Implications

- **Environmental Analytics Meets Market Strategy:** Our approach showcases how classical statistical models (ADF, ARMA) combined with STL decomposition can deliver both environmental insight and commercial foresight.

- **Business Sensitivity to Climate Indicators:** Firms in pollution-sensitive sectors can benefit from predictive analytics by treating AQI as a macroeconomic driver.

- **Policy and Corporate Synergy:** Government regulations like GRAP and CAQM influence consumer behavior and therefore must be integrated into business forecasts.

We will now demonstrate how similar both curves are using a graphic.



Aspect	**Observation**	**Implication**
Correlation	High positive correlation between AQI levels and Blue Star's revenue (notably in Nov–Jan)	AQI acts as a leading indicator for sales—can inform marketing and inventory strategies
Seasonality	Distinct revenue peaks during winter months when AQI exceeds 300	Predictable demand cycles enable supply chain optimization
Consumer Behavior	Sharp increase in purchases during pollution spikes (up to 70% sales surge)	Health-driven consumption underlines the market's responsiveness to environmental changes
COVID-19 Impact	2020 showed anomalies due to lockdowns, with lowest AQI and revenue	Validates the model's sensitivity to external shocks
Post-2020 Growth	Sharp rebound in AQI and sales in 2021–2023, reaching record values in 2023	Indicates rising health awareness and increasing market penetration
Modeling Performance	STL + ARMA reduced AQI forecast RMSE by **41%**, revenue forecast captured seasonality effectively	Demonstrates power of advanced time series techniques in real-world forecasting

Table 4: Key Statistical and Behavioral Insights

Forecasting Output	**Business Application**
Accurate AQI Prediction	Public health alerts, policy intervention timing
Revenue Forecasting	Inventory planning, workforce allocation, supply chain readiness
Seasonal Decomposition	Tailored promotional campaigns aligned with high-demand months
Residual Modeling Accuracy	Risk-adjusted business planning using confidence intervals
Long-Term Trend Analysis	Product development and pricing strategies in anticipation of future pollution patterns

Table 5: Practical Business Applications of Our Forecasting Approach

5 Conclusion

This study demonstrates how classical time series models, when complemented with STL decomposition, can be used effectively for modeling environmental and economic indicators. The AQI data benefits substantially from decomposition and residual modelling, showing a dramatic reduction in forecasting error. Revenue data, while more volatile and complex, also sees improvement with the STL + ARMA approach, although not to the same extent. Both models are validated with strong diagnostics, including ADF tests, RMSE metrics, and confidence interval analysis. Forecasts from both series are not only point predictions but are also wrapped in probabilistic bounds, allowing for transparent uncertainty communication.