India Air Quality Index Forecasting

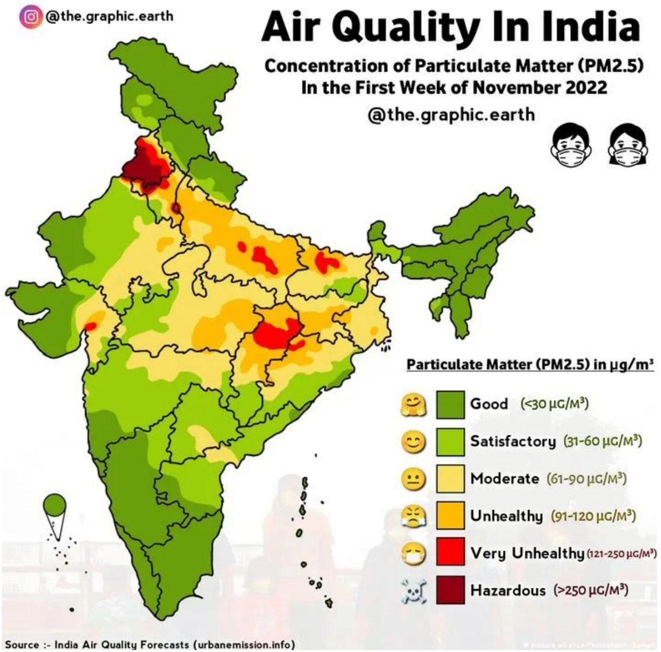
Team Forecaster: Katy Shih, Kitae Kim, Kshitiz Sahay, Welly Huang

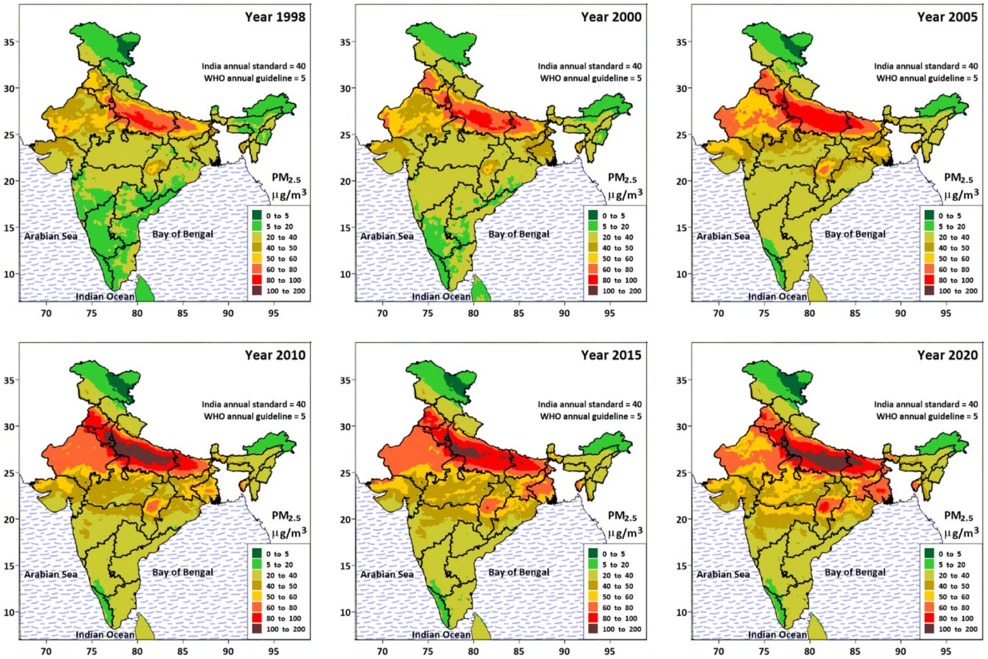
GitHub: https://github.com/k-kshitiz/india-air-quality-index-forecast/tree/main

Agenda

1. Business Problem
2. EDA
3. Analysis and Modeling
4. Model Performance Results
5. Future Work

Air quality in major Indian cities is hazardous





Rising Air Pollution Levels

- **Air quality** in many Indian cities is **deteriorating rapidly**.
- High levels of pollutants such as PM2.5, PM10, NO2, and SO2.

Potential Causes

- Factory emissions
- Vehicle emissions
- Waste burning
- Agricultural activities
- Fossil fuel
- Weak policies

Forecasting AQI would improve public health and economy

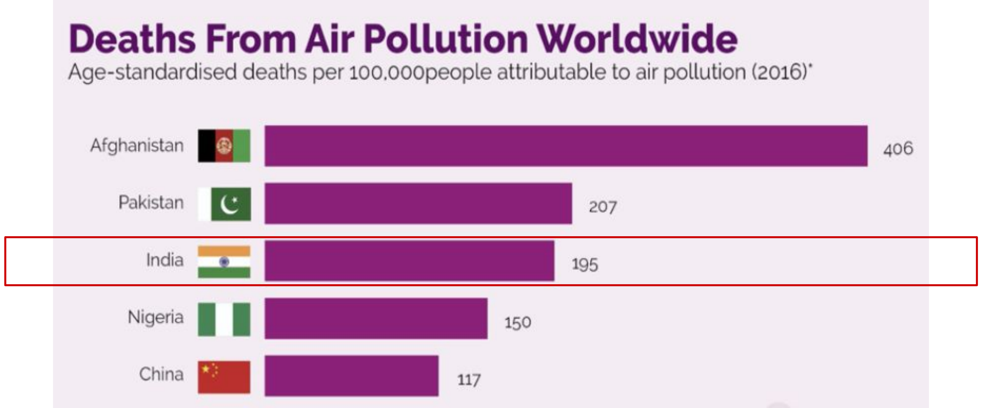
Objective

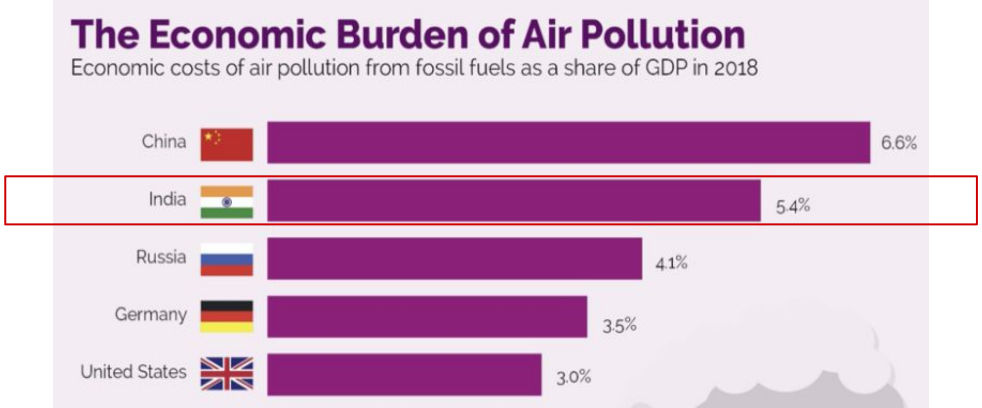
Forecast Air Quality Index (AQI) of major Indian cities to help mitigate the adverse effects of air pollution.

Key Questions We Want to Answer

1. What are the future trends in air quality across India?
2. Can accurate forecasting **help policy making** and **public health**?
3. What seasonal variations affect air quality?

Business Impact of AQI Forecast Model	
Improved public health	Implementing better policies
Reduced economic costs	Raising awareness



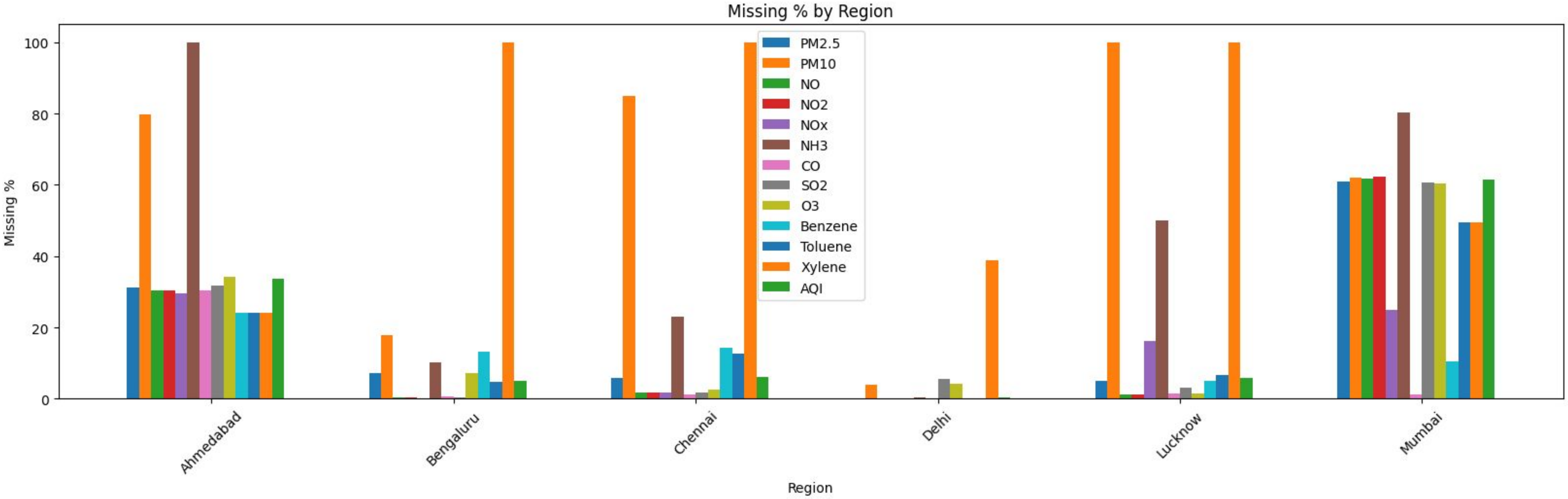


Assumptions/Hypotheses

1. The historical data is accurate and consistent
2. The time series data is stationary
3. AQI patterns exhibit seasonal variations

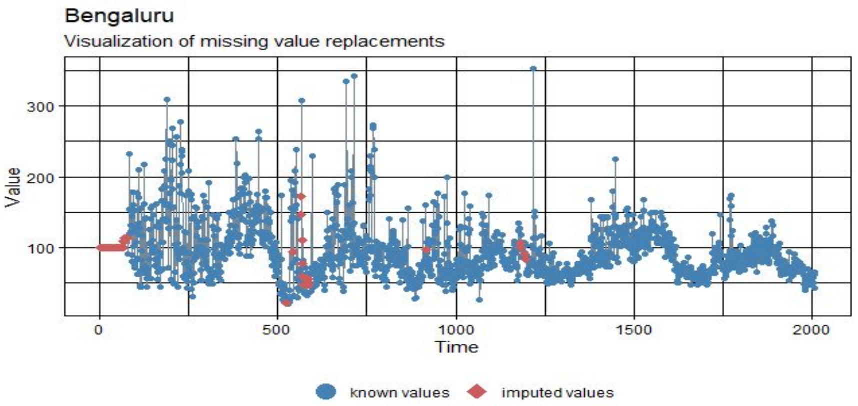
Data Processing

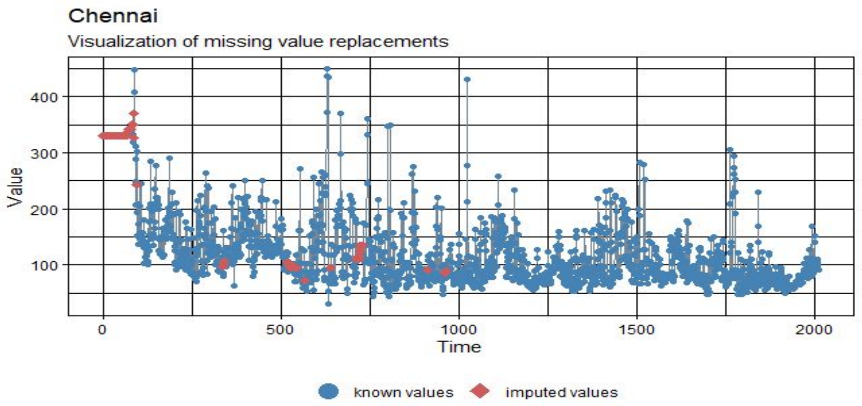
- Original dataset has 29531 rows and contains data from 26 regions
- Only 6 regions has full-length data from 2015-01-01 ~ 2020-07-01
- Ahmedabad and Mumbai has too many missing data across all columns = dropped

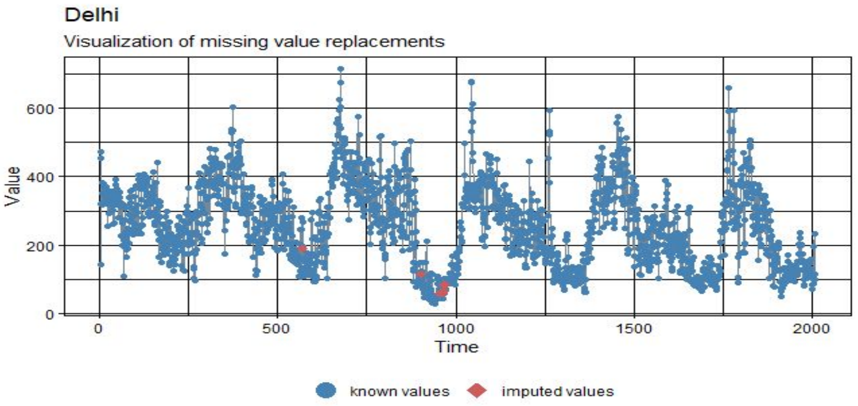


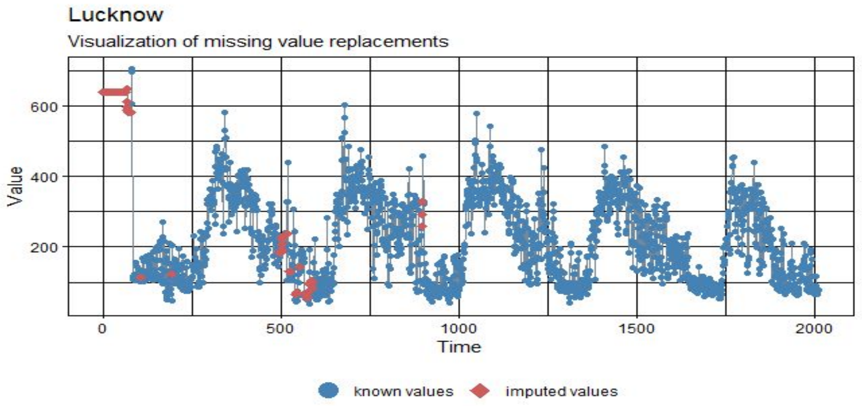
- Interpolation was made using na_ma from imputeTS package
- Moving Average = na_ma(data, k = 15, weighting = "exponential")

Data Imputations





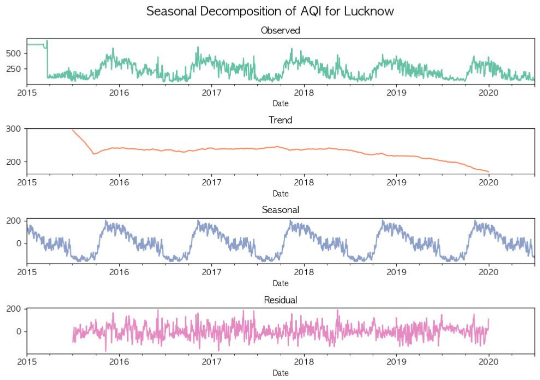




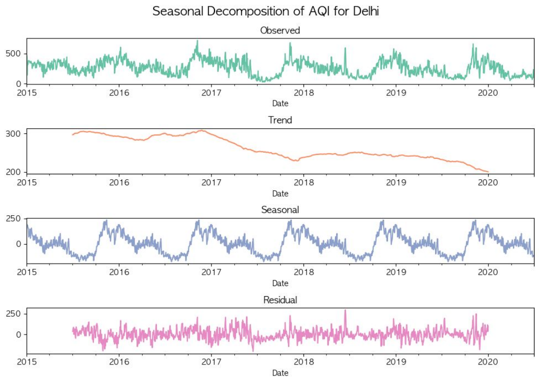
- After interpolation, 4 regions$AQI each contains 2009 data points.

Data Properties/EDA

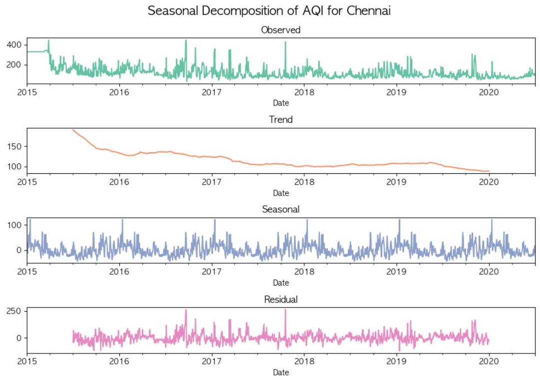
Lucknow



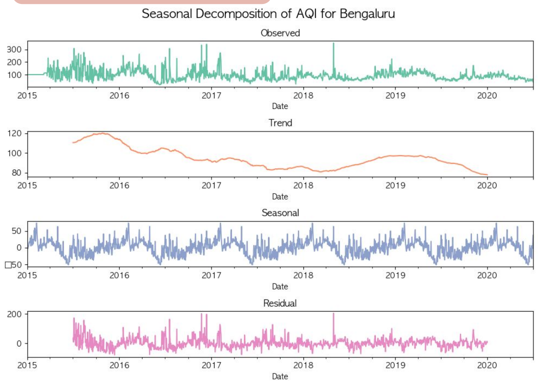
Delhi



Chennai



Bengaluru



Commonalities

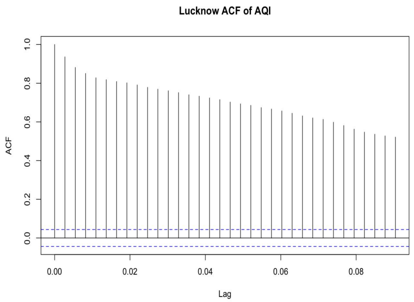
- **Seasonal Presence**: All cities exhibit some level of seasonality in AQI, indicating a cyclic pattern that repeats annually

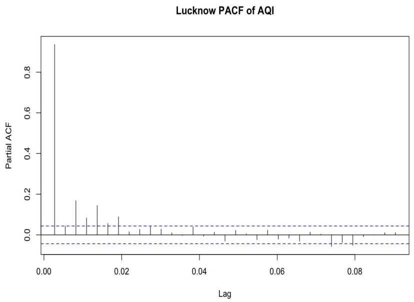
Differences

- **Lucknow**: Strong seasonal fluctuations, peak in late winter.
- **Delhi**: Moderate seasonality, noticeable peaks in winter months.
- **Chennai**: Weak seasonality, relatively stable throughout the year
- **Bengaluru**: Very mild seasonal changes, minimal fluctuation.

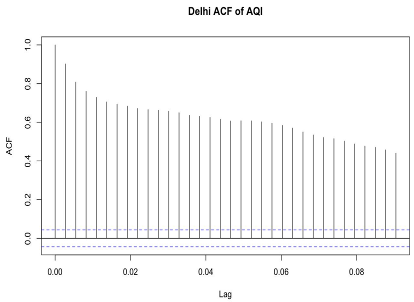
Data Properties/EDA

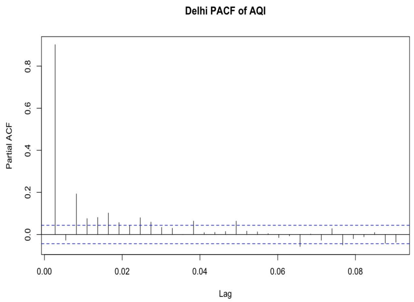
Lucknow



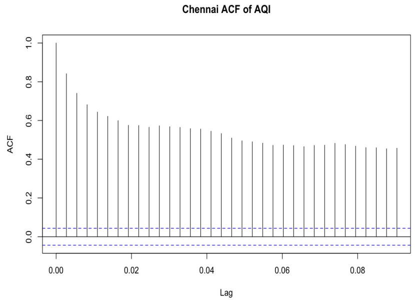


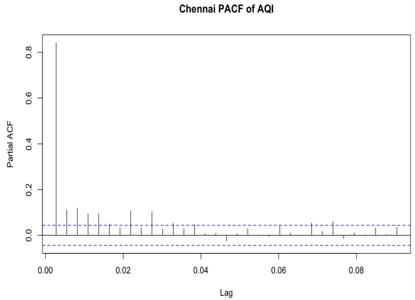
Delhi



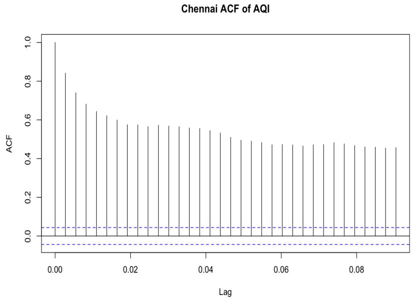


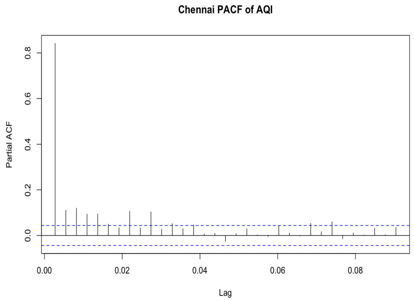
Chennai





Bengaluru

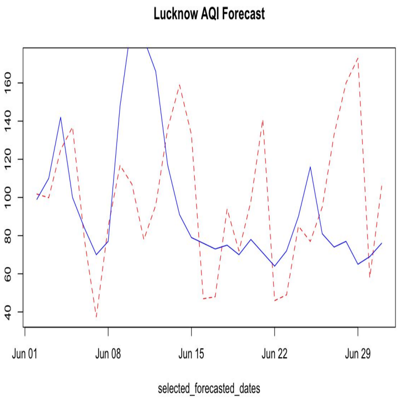


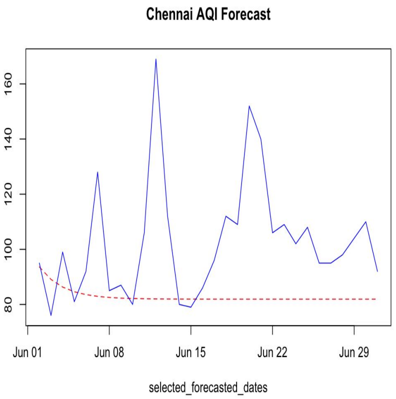


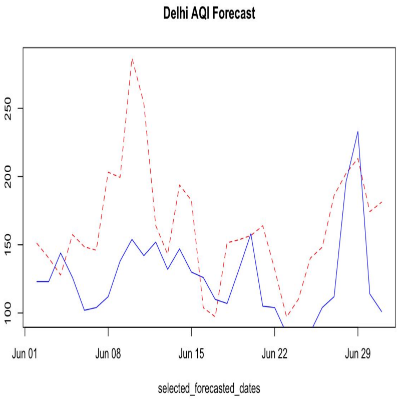
- Data stationarity check using ACF and PACF analysis
- Data from four regions show instability; adjustments are needed
- Using Auto ARIMA to find optimal parameters for P D Q
- This method enhances the robustness of our time series forecasting

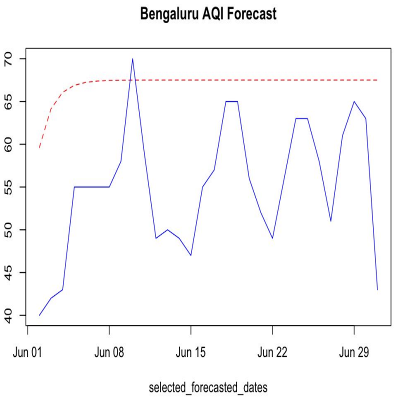
ARIMA

❖ Prediction Result









Metrics for Modeling

City	RMSE	MAE	MAPE	AMAPE
Lucknow	47.57	36.71	39.98	38.16
Chennai	29.47	21.42	18.34	20.85
Delhi	52.78	42.68	36.48	33.93
Bengaluru	13.88	12.24	24.40	22.27

- **Bengaluru** shows the highest forecast accuracy with the lowest RMSE and MAE values.
- **Delhi** displays the highest error rates, indicating the least accurate predictions.
- **Chennai** has notably low MAPE and AMAPE, suggesting consistent and reliable forecasts.
- **Lucknow** experiences moderate forecasting errors, positioned between the best and worst performing cities.

About ARMA-GARCH

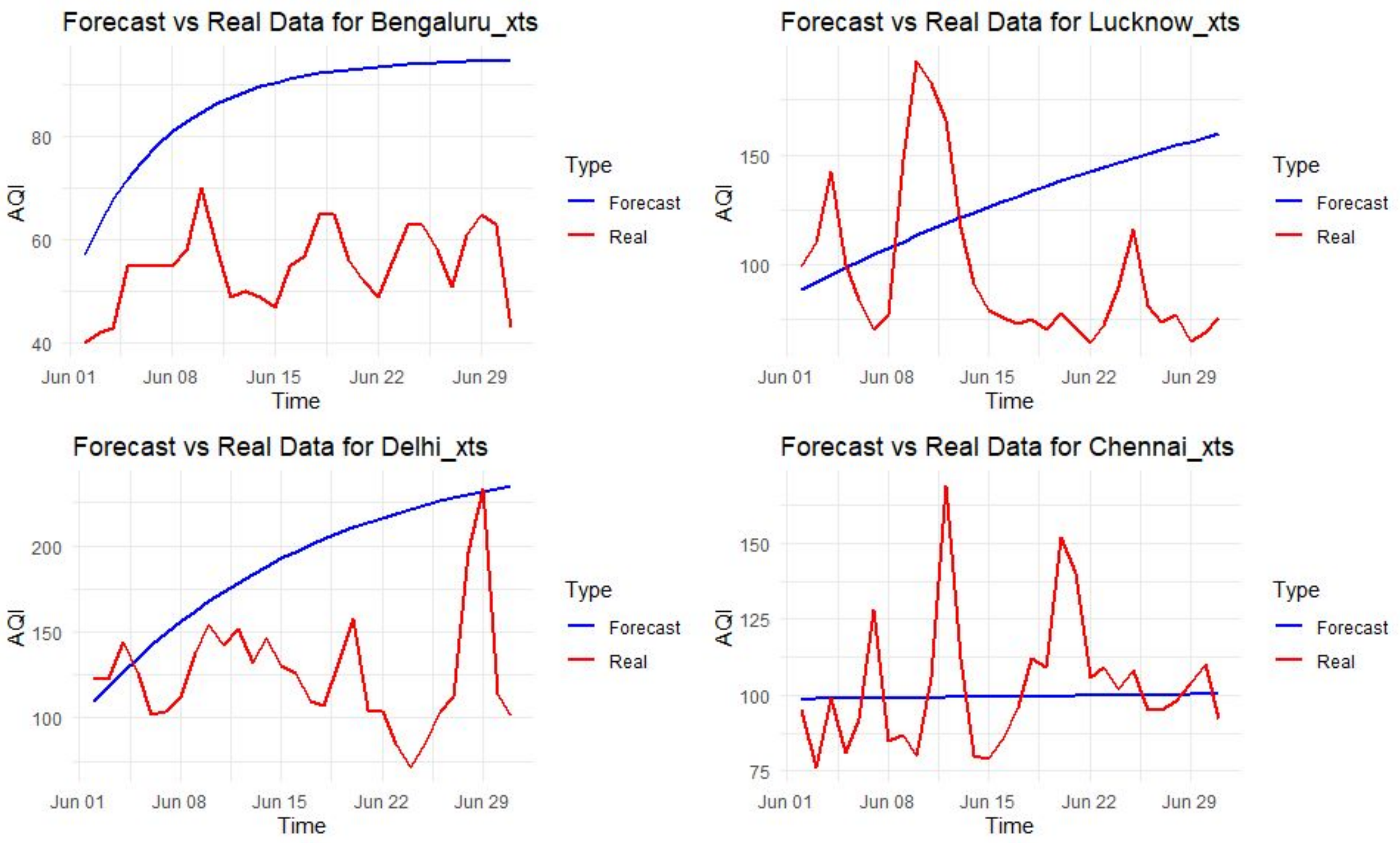
- ARMA model for the linear dependencies + GARCH model for the conditional heteroskedasticity
- Suitable for modeling and forecasting time series data with changing volatility over time.

Why ARMA-GARCH for AQI forecasting?

- **Heteroskedasticity:** ACF plots of squared original data show heteroskedasticity.
- **Starting Orders:** arma(1,1) + garch(1,1)

City	RMSE	MAE	MAPE	AMAPE
Lucknow	57.9988	52.2579	62.74	62.7362
Chennai	21.2329	14.8511	13.66	13.6629
Delhi	80.1953	65.9042	62.35	62.3521
Bengaluru	33.0371	31.7739	5944	59.4431

ARMA-GARCH: Initial Model



ARMA-GARCH: Improvement

	Bengaluru_xts <dbl>	Lucknow_xts <dbl>	Delhi_xts <dbl>	Chennai_xts <dbl>
Jarque-Bera (R) pv	0.0000	0.0000	0.0000	0.0000
Shapiro-Wilk (R) pv	0.0000	0.0000	0.0000	0.0000
Ljung_Box (R, Q = 10) pv	0.0003	0.0000	0.0000	0.0000
Ljung_Box (R, Q = 15) pv	0.0000	0.0000	0.0000	0.0000
Ljung_Box (R, Q = 20) pv	0.0000	0.0000	0.0000	0.0000
Ljung_Box (R^2, Q = 10) pv	0.9993	0.0980	0.9345	0.4468
Ljung_Box (R^2, Q = 15) pv	1.0000	0.3013	0.9747	0.5054
Ljung_Box (R^2, Q = 20) pv	1.0000	0.5300	0.9873	0.6099
LM Arch pv	0.9996	0.1498	0.8825	0.4209

Test on the standardized residuals

- Normality Tests (Jarque-Bera and Shapiro-Wilk):
  Not normally distributed for any of the datasets
- Autocorrelation Tests (Ljung-Box for residuals):
  Significant autocorrelation, indicating that the model has not fully captured the time-series dynamics
- Autocorrelation Tests for Squared Residuals (Ljung-Box for R^2):
  No significant autocorrelation, suggesting that the GARCH model is effectively capturing the volatility
- LM Arch Test:
  The GARCH model adequately captures the autoregressive conditional heteroskedasticity in the data.

ARMA-GARCH: Improvement

Used grid search to find the optimal ARMA components

- Bengaluru: (1, 1, 1)
- Lucknow: (0, 1, 4)
- Delhi: (1, 1, 2)
- Chennai: (1, 1, 1)

Differentiate the time series by d = 1 and run ~ arma(p, q) + garch(1, 1)

City	RMSE	MAE	MAPE	AMAPE
Lucknow	101.8298	95.6396	99.53	99.5317
Chennai	105.5568	103.4924	100.8	100.7989
Delhi	129.1922	125.1313	99.50	99.4955
Bengaluru	55.0340	54.3562	98.57	98.5669

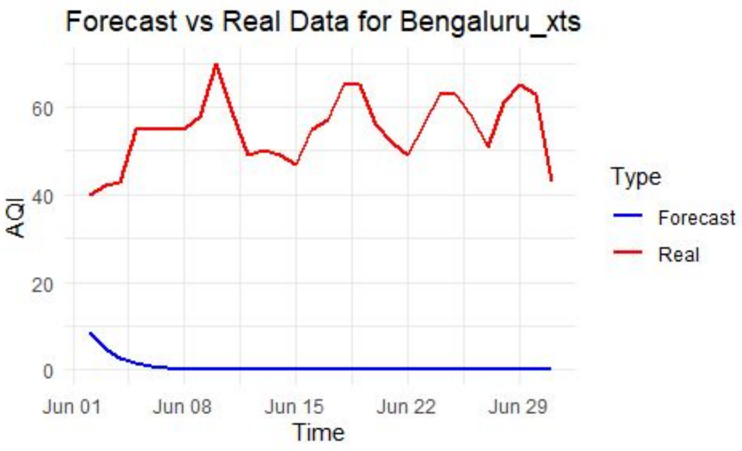
ARMA-GARCH: Improvement

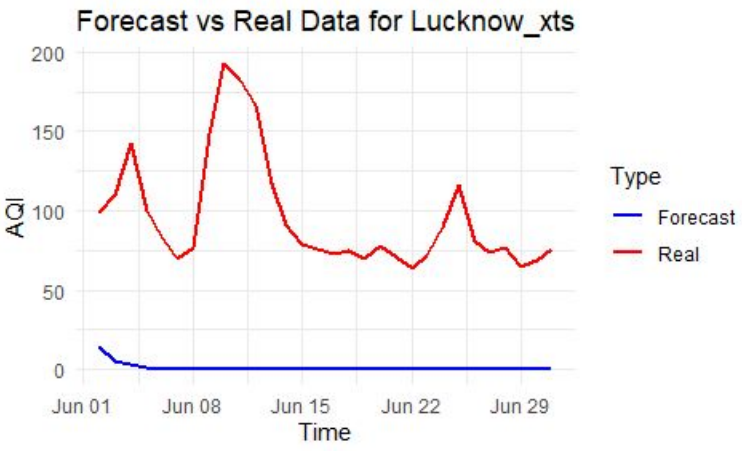
	Bengaluru_xts <dbl>	Lucknow_xts <dbl>	Delhi_xts <dbl>	Chennai_xts <dbl>
Jarque-Bera (R) pv	0.0000	0.0000	0.0000	0.0000
Shapiro-Wilk (R) pv	0.0000	0.0000	0.0000	0.0000
Ljung_Box (R, Q = 10) pv	0.0246	0.7637	0.0493	0.6210
Ljung_Box (R, Q = 15) pv	0.1172	0.8192	0.1222	0.8187
Ljung_Box (R, Q = 20) pv	0.1475	0.5996	0.0875	0.6769
Ljung_Box (R^2, Q = 10) pv	0.9992	0.2134	0.5818	0.3169
Ljung_Box (R^2, Q = 15) pv	1.0000	0.5315	0.5563	0.4711
Ljung_Box (R^2, Q = 20) pv	1.0000	0.7687	0.6027	0.6418
LM Arch pv	0.9995	0.3151	0.3712	0.3517

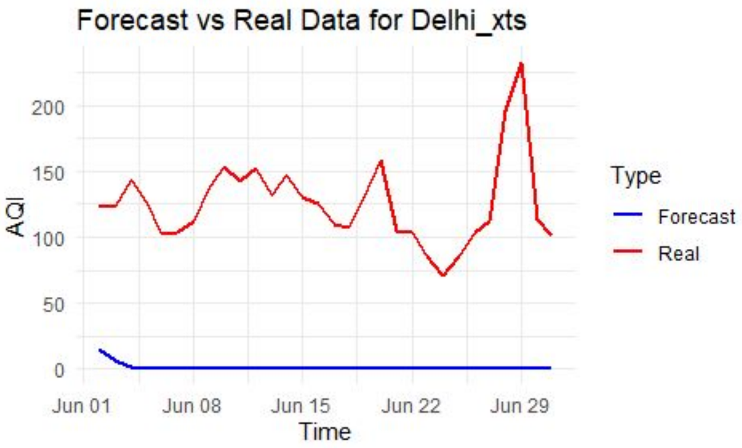
Test on the standardized residuals

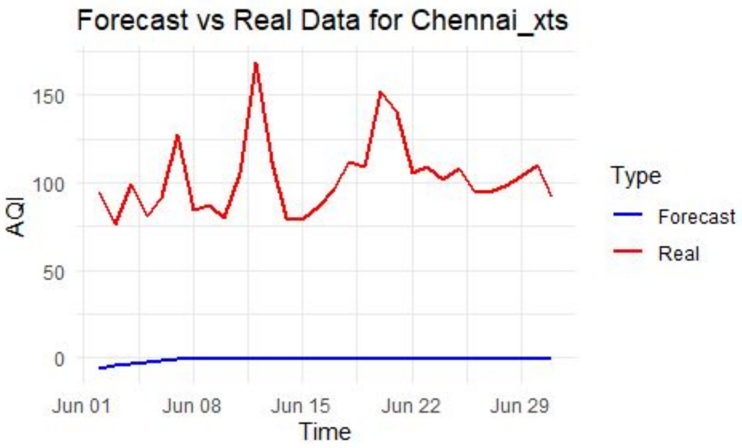
- Normality Tests (Jarque-Bera and Shapiro-Wilk):
  Still not normally distributed for any of the datasets
- Autocorrelation Tests (Ljung-Box for residuals):
  Reduced significance for autocorrelation, indicating that the model has captured the time-series dynamics
- Autocorrelation Tests for Squared Residuals (Ljung-Box for R^2):
  No significant autocorrelation, suggesting that the GARCH model is effectively capturing the volatility
- LM Arch Test:
  The GARCH model adequately captures the autoregressive conditional heteroskedasticity in the data.

ARMA-GARCH: Improvement









VARIMA: Model Overview & Results

About VARIMA

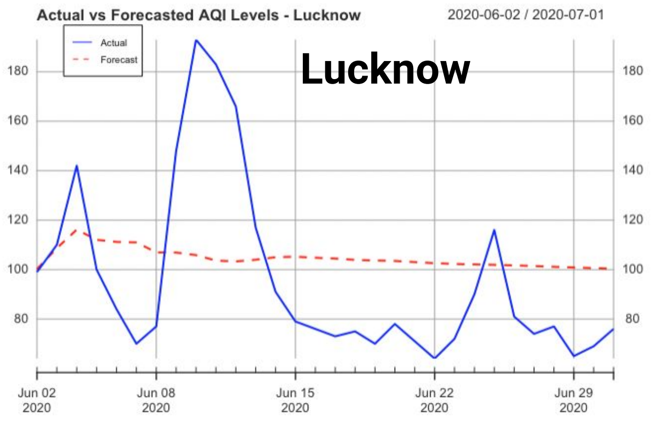
1. Multivariate time series model
2. Forecasting by past observations of itself and of the other variables within the data set

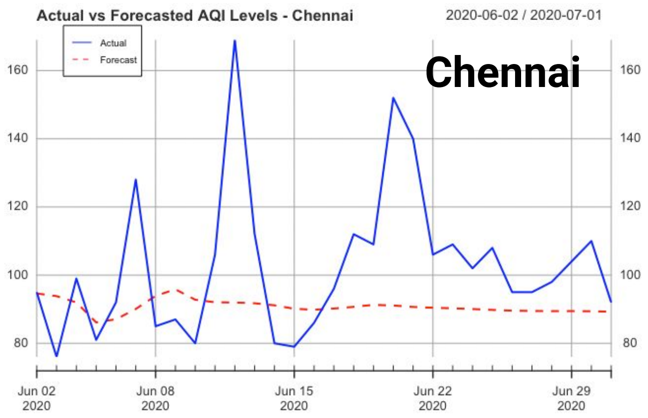
Why VARIMA for AQI forecasting?

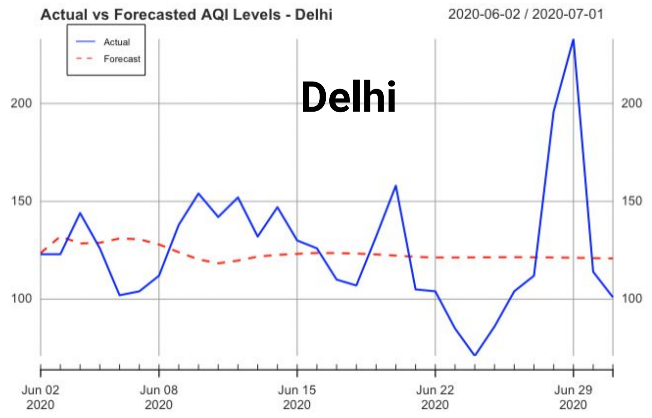
- **Spatial Interdependencies:** Regions in India may exhibit spatial interdependencies in terms of air quality.
- **Integration and Differencing:** Air quality data may be non-stationary, exhibiting trends and seasonality.

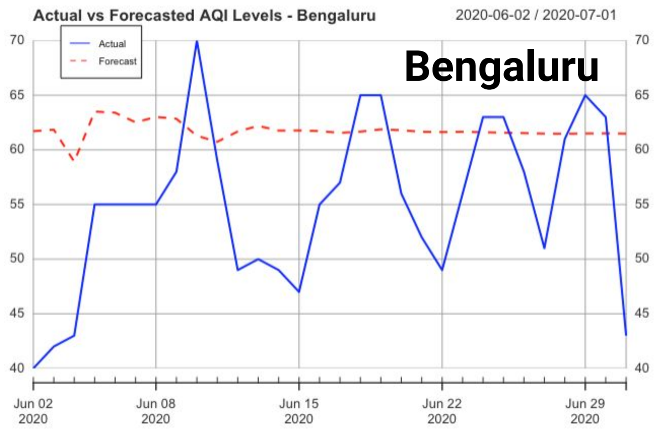
City	RMSE	MAE	MAPE	AMAPE
Lucknow	35.37972	30.05236	32.55386	29.52032
Chennai	24.33768	17.26720	15.04081	16.54072
Delhi	32.77869	23.93517	19.07055	18.68528
Bengaluru	35.37972	30.05236	32.55386	29.52032

VARIMA: Actual vs Forecast









Prophet: Model Overview & Results

About Prophet

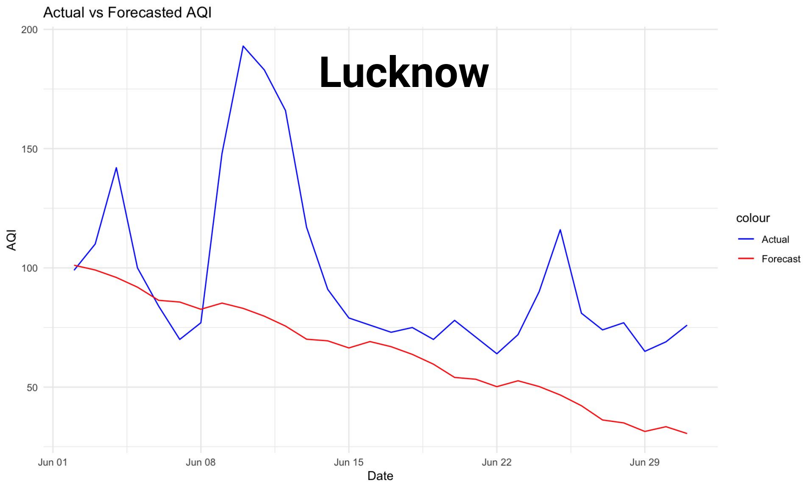
1. Developed by Facebook
2. Automatic detect trends and seasonality with minimal manual training
3. Intuitive and user-friendly

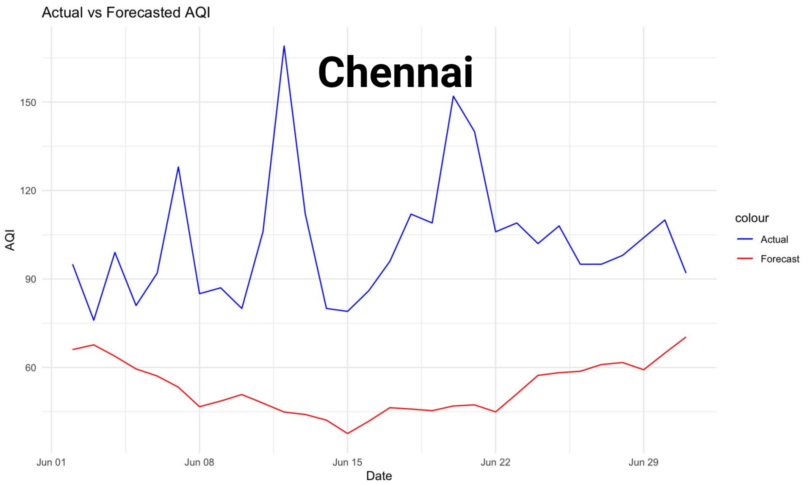
Why Prophet for AQI forecasting?

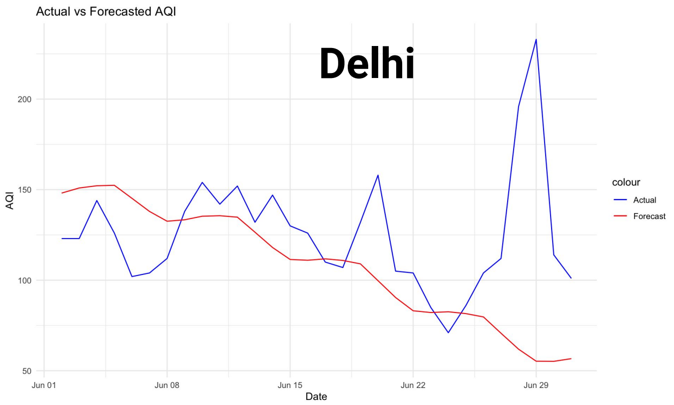
Prophet excels at capturing daily, weekly, monthly, and yearly seasonal effects and can incorporate holidays or special events.

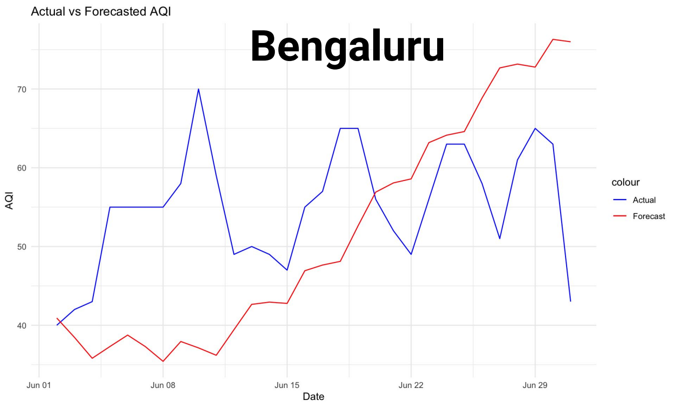
City	RMSE	MAE	MAPE	AMAPE
Lucknow	43.8092	32.99207	0.3155502	40.46366
Chennai	55.41889	49.7491	0.4633163	62.37414
Delhi	48.23874	30.73455	0.2210615	26.18271
Bengaluru	14.58552	11.92744	0.2166511	22.76304

Prophet: Actual vs Forecast









Results

Lucknow

Model	RMSE	MAE	MAPE %	AMAPE %
ARIMA	47.57	36.71	39.98	38.16
ARMA GARCH	58.00	52.26	63.00	62.74
VARIMA	35.38	30.05	32.55	29.52
Prophet	43.8	32.9	31.50	40.4

Delhi

Model	RMSE	MAE	MAPE %	AMAPE %
ARIMA	52.78	42.68	36.48	33.93
ARMA GARCH	80.20	65.90	62.00	62.35
VARIMA	32.78	23.94	19.07	18.69
Prophet	48.2	30.7	22.00	26.2

Chennai

Model	RMSE	MAE	MAPE %	AMAPE %
ARIMA	29.47	21.42	18.34	20.85
ARMA GARCH	21.23	14.85	14.00	13.66
VARIMA	24.34	17.27	15.04	16.54
Prophet	55.4	49.7	46.3	62.4

Bengaluru

Model	RMSE	MAE	MAPE %	AMAPE %
ARIMA	52.78	12.24	24.40	22.27
ARMA GARCH	33.03	31.77	59.00	59.44
VARIMA	35.38	30.05	32.55	29.52
Prophet	14.5	11.9	21.00	22.7

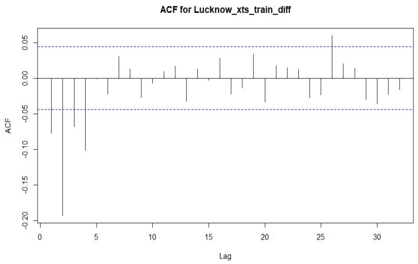
Future Work

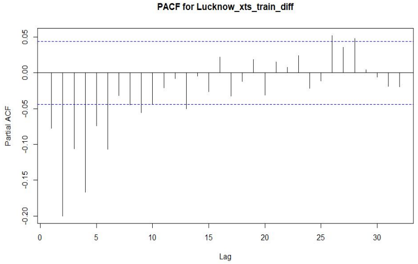
1. Incorporate additional exogenous variables like weather data, traffic data, industrial emissions, and social events
2. Explore more advanced models like LSTM, GRU, and hybrid models
3. Account for spatial dependencies between different regions using techniques
4. Incorporate holiday effects, COVID impact, and other seasonal events
5. Real-time forecasting

Appendix

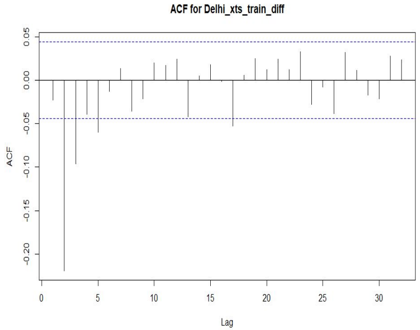
ACF & PACF for Differentiated Data (d = 1)

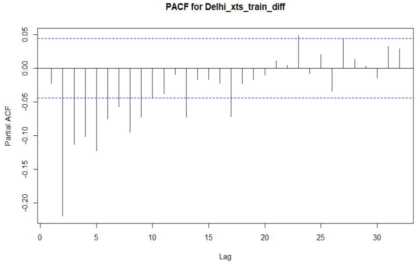
Lucknow



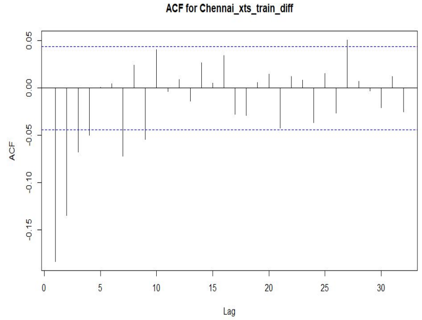


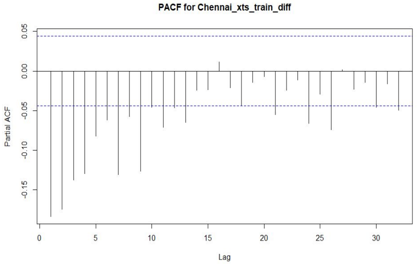
Delhi



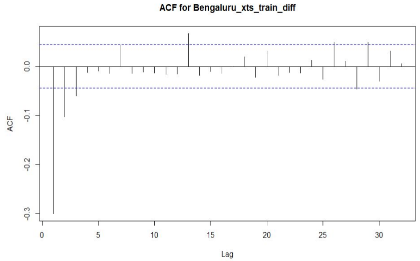


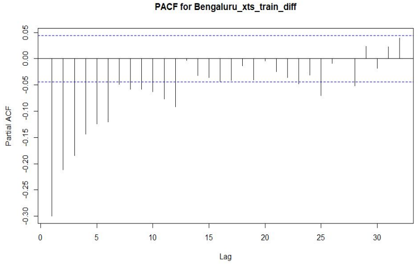
Chennai





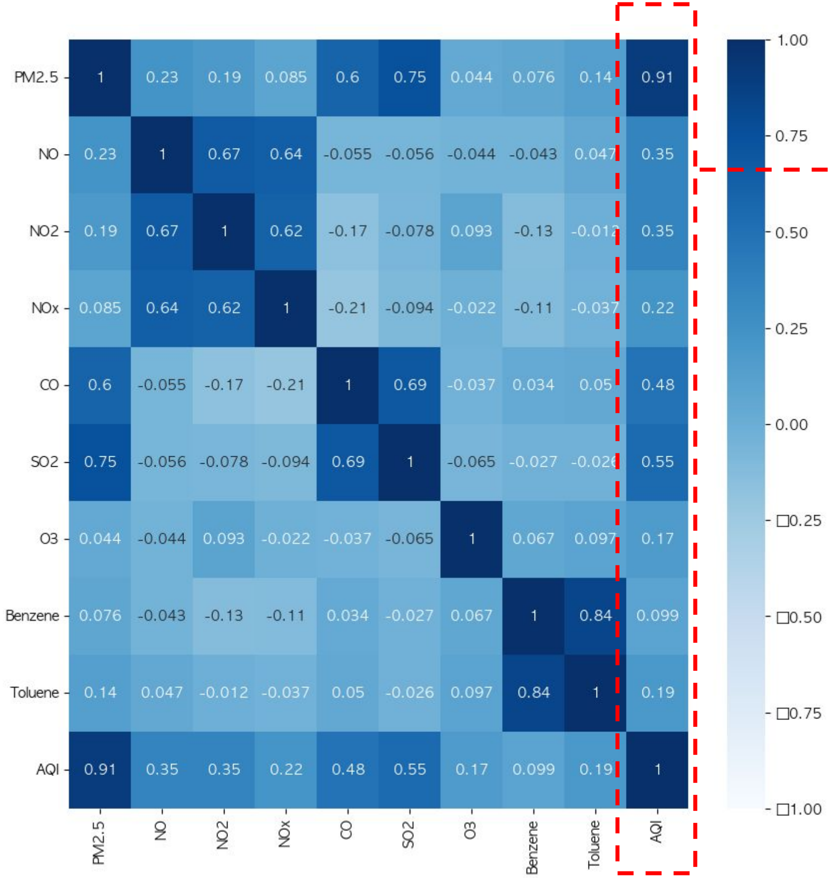
Bengaluru





- Differencing removes trends and cycles, stabilizing mean across the series.
- Reduces dependence among values, enhancing model's predictive accuracy and stability.

AQI Check



All features are needed to do analysis?

$$AQI = \frac{I_{high} - I_{low}}{C_{high} - C_{low}} \times (C - C_{low}) + I_{low}$$

- **The concentration of each pollutant is converted into a sub-index**
- **and the highest sub-index value among the pollutants is taken as the final AQI**
- **India's Air Quality Index (AQI) considers the highest sub-index among various pollutants as the final AQI**
- **This means that the AQI is determined based on the pollutant with the highest sub-index value, reflecting the most severe impact on air quality**