Airguard
Analystics

AIR QUALITY FORECASTING MODEL

presented by :
CHRISTINE KIRIMI
JOHN ELVIS
OMARE BRYTONE

OVERVIEW

Air Quality Forecasting

Our assignment involves examining the air quality in the Gucheng area of China. The company has provided us with hourly air quality dataset spanning from March 1, 2013, to February 28, 2017. The objective is to develop a model capable of predicting the level of air pollution in Gucheng, specifically focusing on the PM2.5 feature within the mentioned timeframe.

BUSINESS PROBLEM

In rapidly industrializing regions air pollution has emerged as a critical environmental and public health issue.

The challenge faced by local authorities, environmental agencies, and public health organizations is the lack of predictive capability regarding air quality. This project aims to address this real-world problem by developing a predictive model for air pollution in Gucheng.

The model's forecasts can be used by these stakeholders to implement timely health advisories, pollution control measures, and urban planning strategies.

Major Sources of Air Pollution

Business Objective

The predictive model will empower stakeholders with foresight into air quality trends, enabling proactive measures rather than reactive responses. For instance, health advisories can be issued in advance, and industries can adjust operations in anticipation of high pollution periods. This approach not only mitigates the health impacts of air pollution but also aids in efficient resource allocation for pollution control.

Stakeholders:

Local Government and Environmental Agencies .

Public Health Organizations

General Public

Businesses and Industries

Methodologies, Techniques & Model

Exploratory Data Analysis (EDA): We conducted thorough EDA to understand the distribution, central tendencies, and variability of the PM2.5 data. Used statistical measures, histograms, and box plots to identify outliers and patterns in the data.

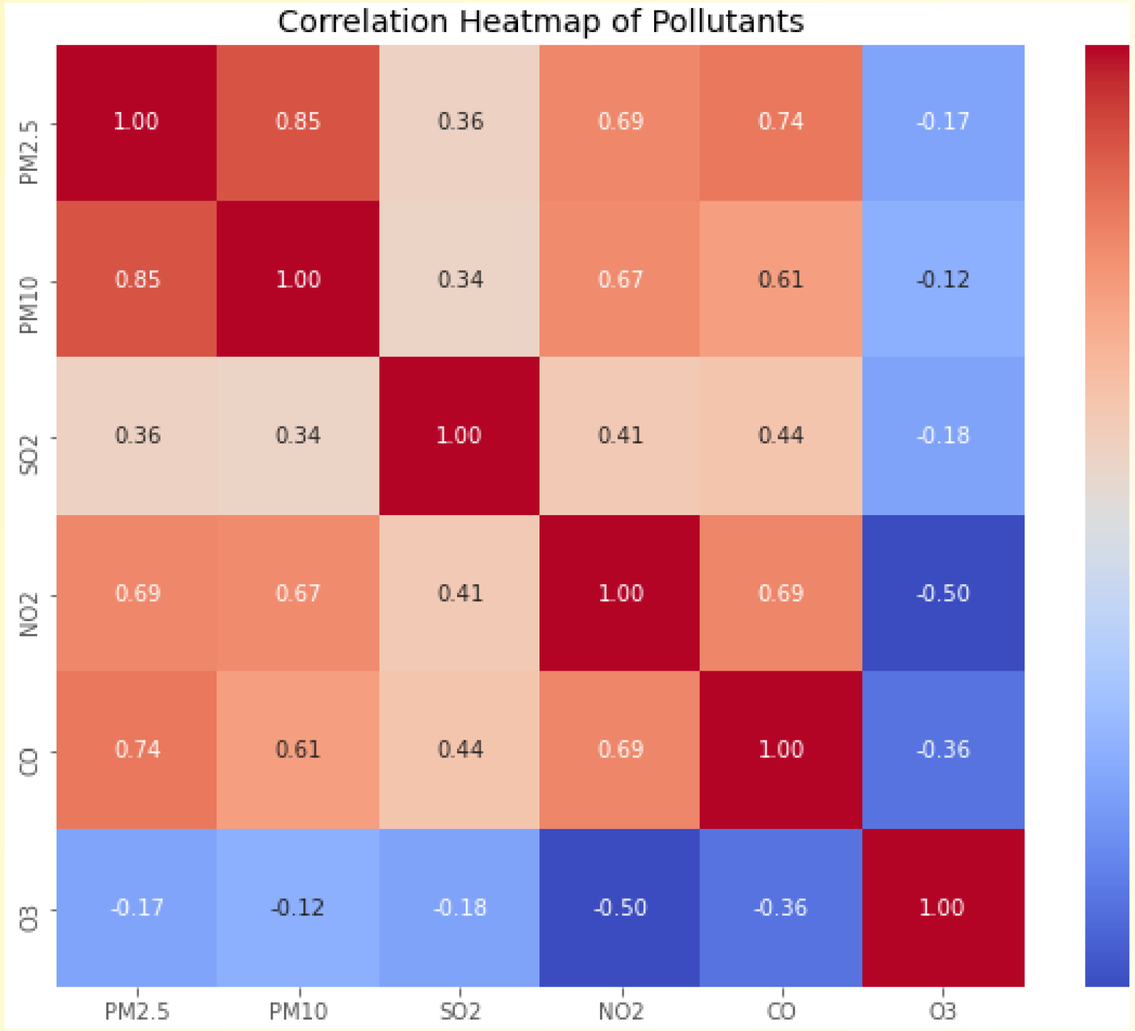
Time Series Analysis: Since the data involves a time component, perform time series analysis. Explore trends, seasonality, and cyclic patterns in the PM2.5 concentrations over the given time span. Use time series plots, autocorrelation, and decomposition techniques.

Correlation Analysis: Investigate the relationships between PM2.5 and other features. Calculate correlation coefficients to identify variables that are strongly correlated or inversely correlated with PM2.5. This analysis helps in feature selection.

Feature Engineering: Create new features that may enhance the model's predictive power. For example, derive features such as time of day, day of the week, and seasonal indicators from the timestamp. Experiment with lag features to capture temporal dependencies.

Data Preprocessing: Handle missing data, outliers, and any anomalies in the dataset. Impute missing values using appropriate techniques, and consider normalization or scaling to ensure uniformity in feature magnitudes.

Model Selection: Choose appropriate modeling techniques for time series forecasting. Common approaches include autoregressive models (ARIMA), machine learning algorithms (such as linear regression, decision trees, and ensemble methods), and deep learning models like Long Short-Term Memory (LSTM) networks.

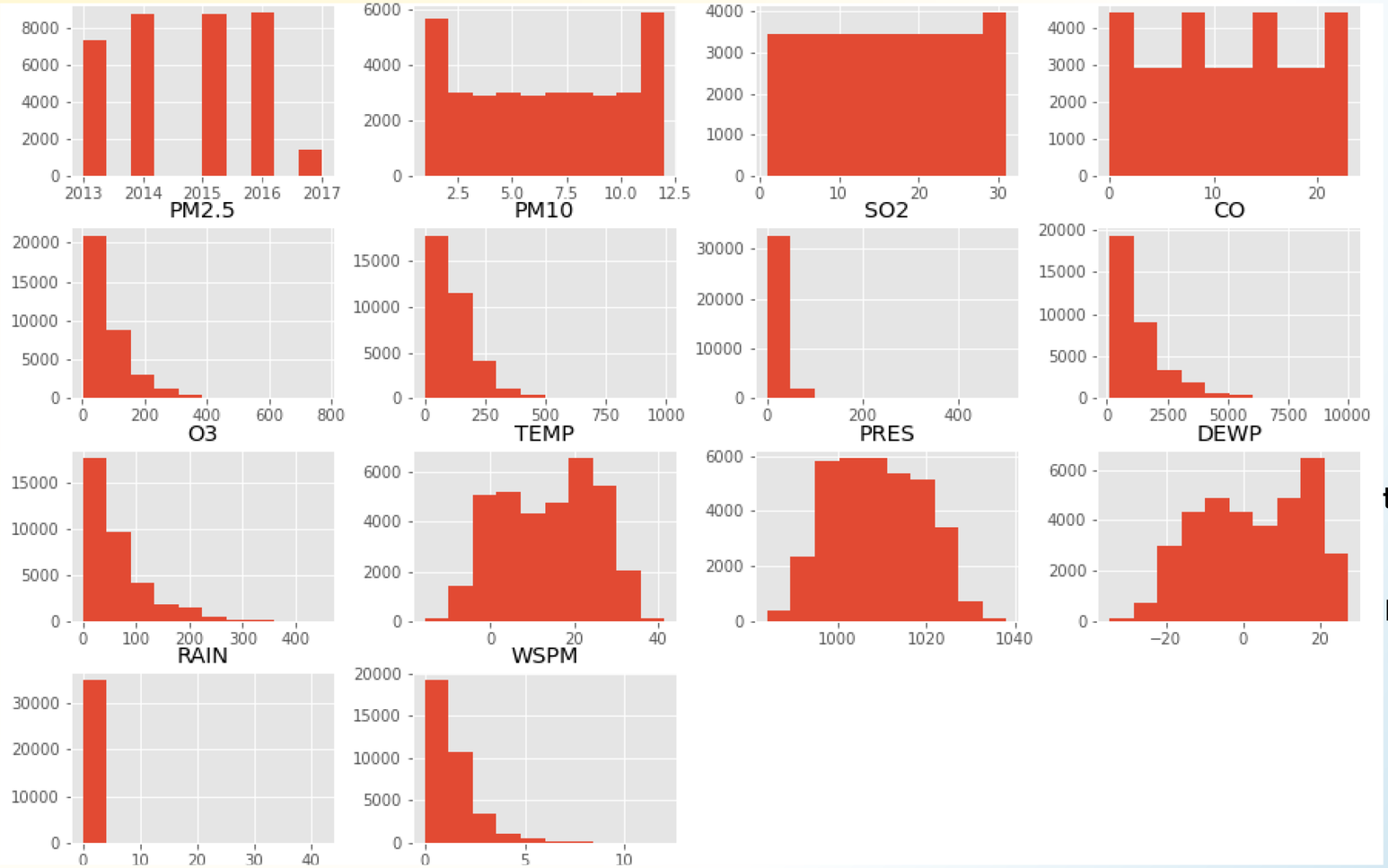


The heatmap illustrates the correlations between different air pollutants. Notable observations include:

- PM2.5 and PM10: There is a strong positive correlation between PM2.5 and PM10. This is expected as they are both particulate matters, albeit of different sizes.

- Correlations with NO2 and CO: Both PM2.5 and PM10 show significant positive correlations with NO2 and CO, suggesting common sources such as vehicle emissions or industrial activities.

Data Exploration/Visualization



Histograms of the various Pollutant features

Summary of the columns observed

Consideration of various meteorological features involved provides a comprehensive understanding of the interplay between weather and air quality. This highlights the importance of considering multiple factors when analyzing air quality data.

- Analyzing temperature patterns helps identify seasons with potentially higher pollution levels. Understanding how air quality varies with the seasons is crucial for environmental planning.

<u>Wind</u>:
Higher wind speeds contribute to better air quality by dispersing pollutants. Periods of low wind speed may lead to the accumulation of pollutants, impacting air quality negatively.

-<u>Atmospheric Conditions</u>
Monitoring air pressure and humidity alongside pollution levels provides a holistic view of atmospheric conditions. Changes in these parameters can be correlated with variations in air quality.
<u>PM2.5 Spikes</u>:
- Examining spikes in PM2.5 concentrations helps identify potential pollution events. Investigating the causes of these spikes can guide pollution control strategies. Holistic Approach:

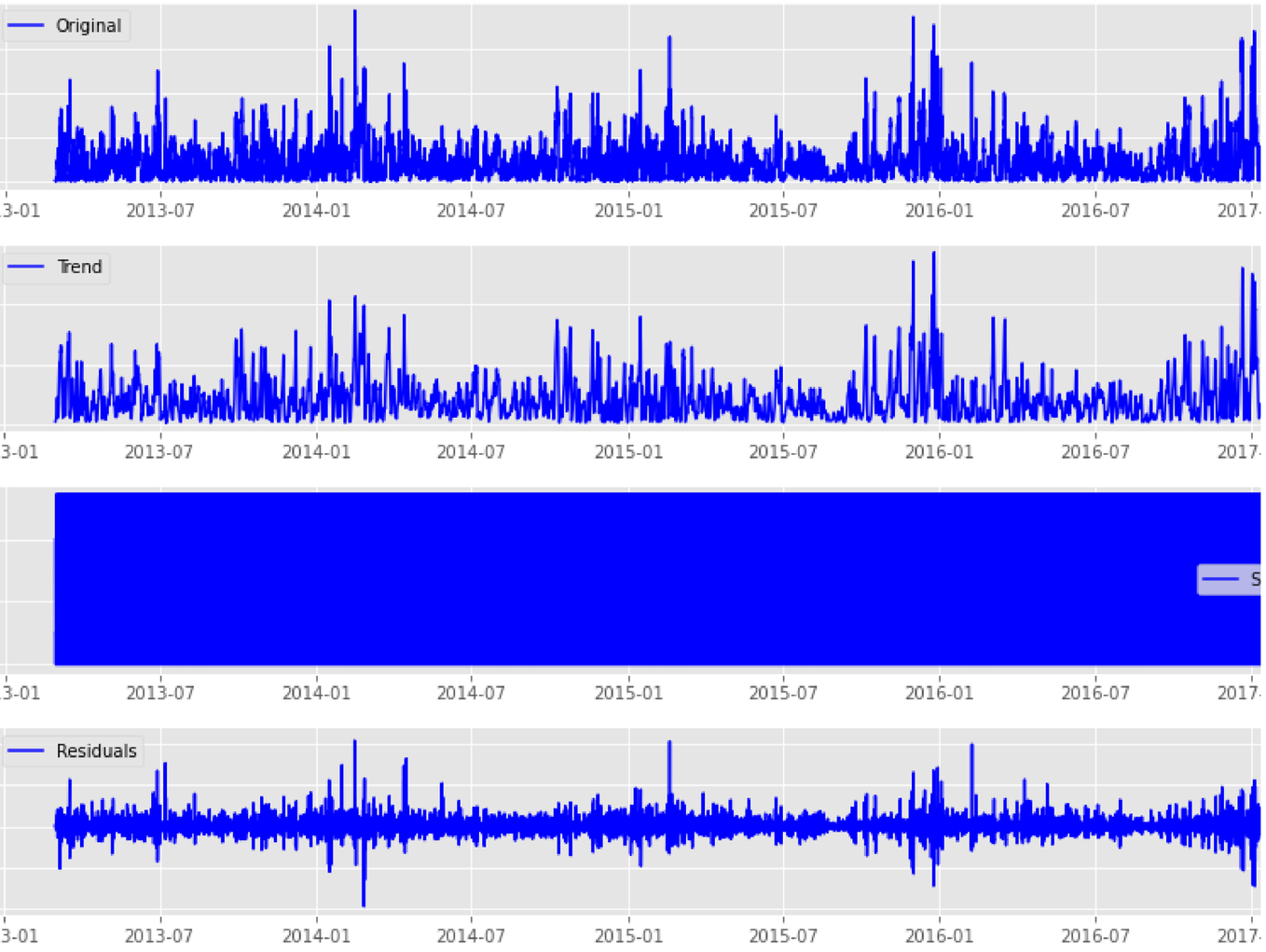
Temporal Analysis for PM2.5 levels over the years

This visualization analyzes air quality trends over time, emphasizing the severity and duration of different pollution levels.

The areas are filled with different colors and labels to represent varying air quality categories, each defined by specific PM2.5 index ranges:

- 'Good' (green)
- 'Moderate' (yellow)
- 'Unhealthy For Sensitive Group' (orange)
- 'Unhealthy' (red)
- 'Very Unhealthy' (dark red)
- 'Hazardous' (Grey)

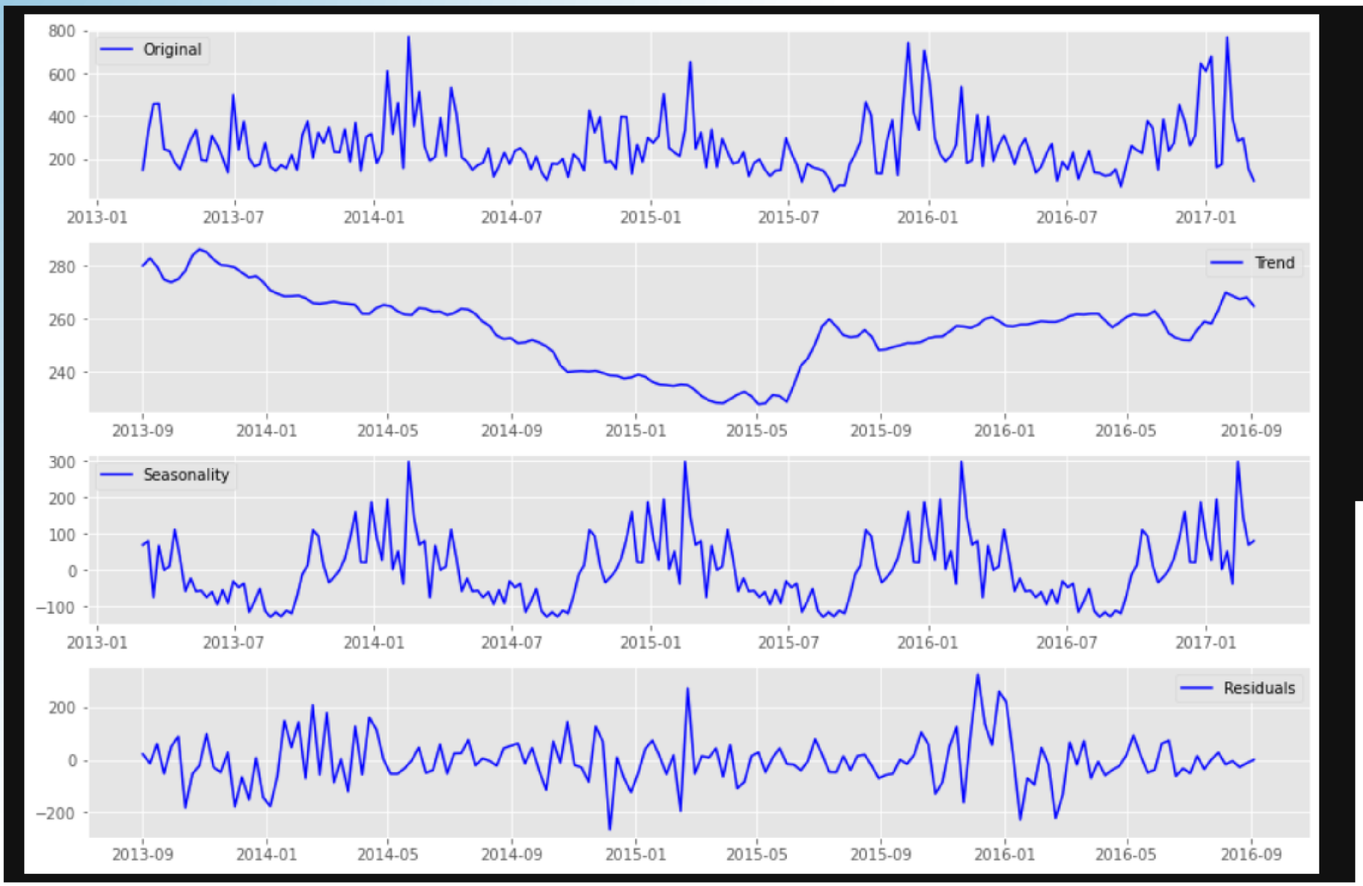
Time Series Decomposition : Hourly Analysis of Pollution trends




```
Results of Dickey-Fuller Test:

Test Statistic                  -18.847174
p-value                           0.000000
#Lags Used                       51.000000
Number of Observations Used    35012.000000
Critical Value (1%)               -3.430537
Critical Value (5%)               -2.861623
Critical Value (10%)              -2.566814
```

From the Dickey-Fuller test, the obtained p-value is 0.000000, leading to the inference that the time series exhibits stationarity. This suggests an extensive volume of available data.

Weekly Maximum Data: Time Series Decomposition



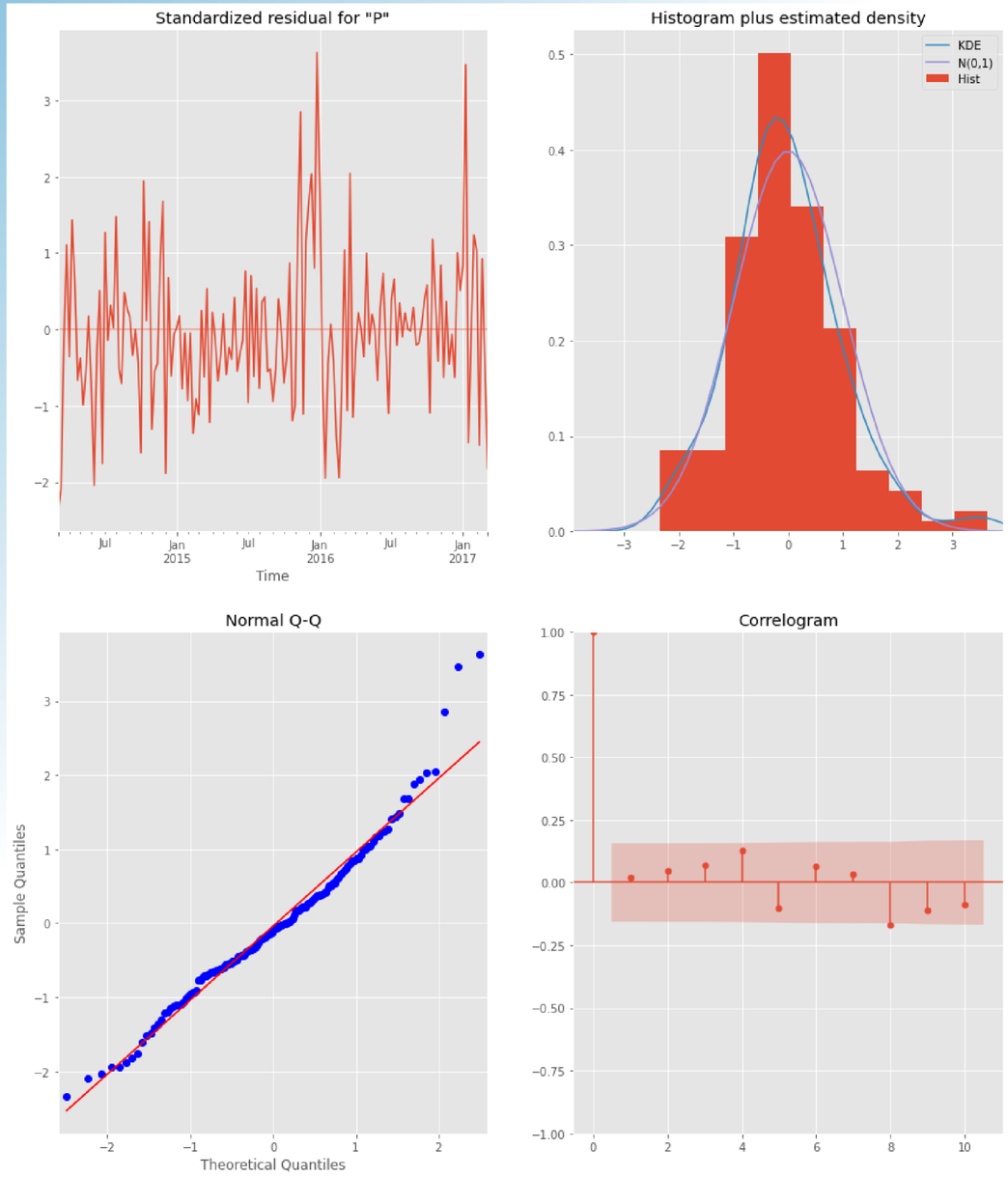

```
Results of Dickey-Fuller Test:

Test Statistic                  -4.567230
p-value                          0.000148
#Lags Used                       3.000000
Number of Observations Used    206.000000
Critical Value (1%)             -3.462499
Critical Value (5%)             -2.875675
Critical Value (10%)            -2.574304
```

- For the Weekly Max Data , The Dickey-Fuller test results suggest that the time series is stationary after removing the trend, which means that the PM2.5 levels fluctuate around a stable mean and variance over time.
- For forecasting models, it's critical to account for both the trend and seasonality components identified in the time series decomposition. The stationary nature of the time series, indicated by the Dickey-Fuller test, means that models like ARIMA could be well-suited for forecasting these PM2.5 levels.

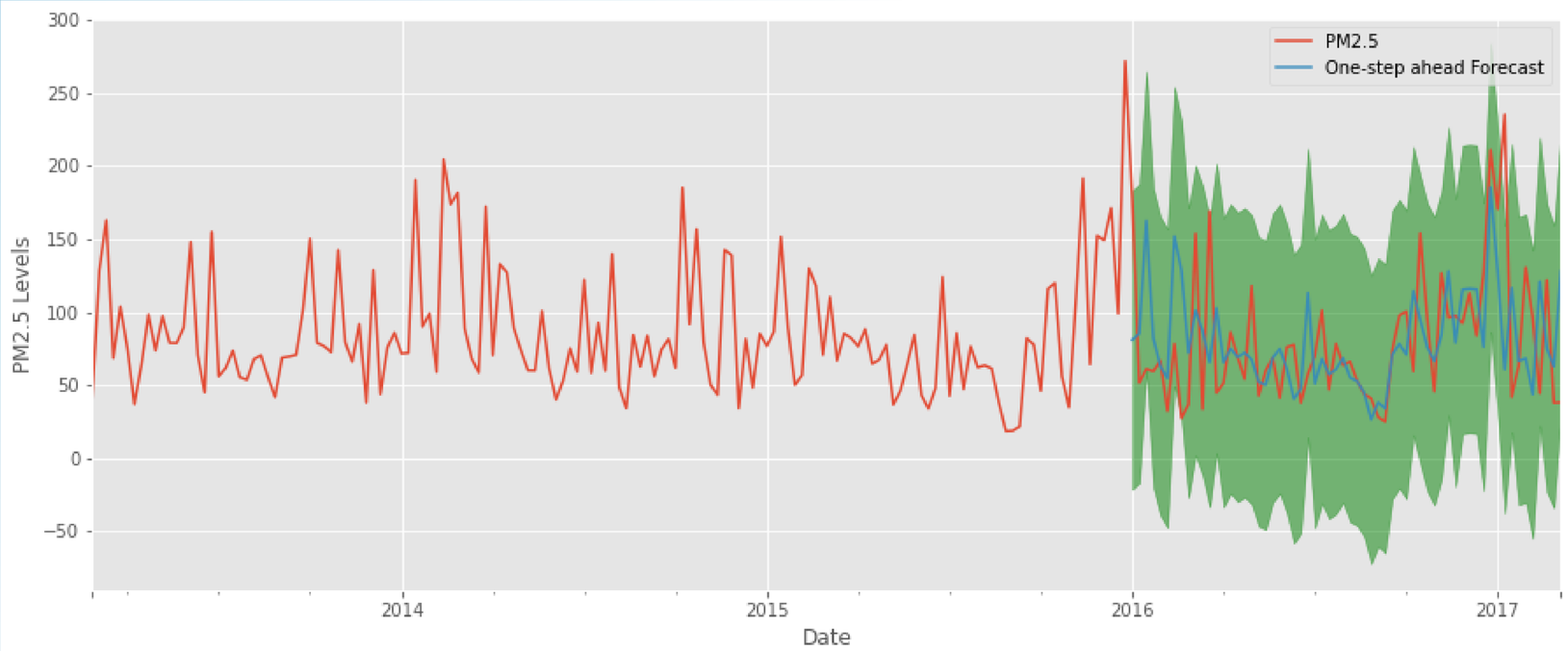
Modelling

SARIMA Model

- SARIMA Parameters: Selecting the SARIMA parameters, including autoregressive, differencing, moving average, and seasonal components, is crucial for capturing the complex patterns in air quality data.
- Training: Model training focuses on fitting the SARIMA model to historical data, identifying the best configuration to achieve accurate predictions of pollutant PM2.5 concentrations.



- The residuals of the SARIMA model appear to be white noise, which is good, indicating that the model captures the data's patterns well.
- The slight skew in the histogram and the deviations in the Q-Q plot's tails suggest that there may be some outliers or non-normal aspects in the data that could be explored further.
- The absence of significant autocorrelation in the residuals is a positive sign that the model has accounted for the time series' autocorrelation.
- The occasional spikes in the residuals may warrant further investigation — they could represent unusual or extreme air quality events that the model cannot predict.
- Future work could explore more complex models, outlier detection, or additional variables that could explain these extreme values.



The validation graph shows the actual observed values of PM2.5 levels (in red) and the one-step-ahead forecasts from the SARIMA model (in green), along with confidence intervals for these forecasts (the shaded area)

- The SARIMA model is performing well in terms of one-step-ahead forecasting for the air quality dataset
- The model's parameters should be periodically re-evaluated to ensure they remain appropriate as new data becomes available
- The occasional spikes and the width of the confidence intervals highlight the need for continuous monitoring and model updating. Collaboration between data scientists, environmental experts, and public health officials is crucial to leverage the full potential of the model for societal benefit.

How this model can be utilized:

- **Anticipating High Pollution Events**- The model's forecasts can be integrated into early warning systems to alert the public and relevant authorities of impending high pollution days. This can facilitate preemptive measures to mitigate health risks
- **Public Health Measures**- The model can inform the timing of health advisories, recommending when to stay indoors, reduce outdoor exercise, or use air filtration systems to minimize exposure to high PM2.5 levels.
- **Urban Planning and Infrastructure** i.e The model can guide the planning and usage of public spaces, such as parks and outdoor recreational facilities

Limitations

Reactivity to New Data: The model's ability to update and react to new data is limited by the need for re-estimation of parameters. In rapidly changing situations, the model may lag behind the current state until it is updated.

The model's accuracy is contingent on the quality and representativeness of the training data. Any biases or gaps in the data may impact the model's generalization to real-world scenarios.

Conclusion

In conclusion, the development of air quality prediction models yields critical insights and paves the way for future advancements. By summarizing key learnings and exploring opportunities for further improvement, we can steer the evolution of air quality prediction towards even greater accuracy, reliability, and impact on public health and well-being.

The End