Air Quality Index Prediction in Phoenix, AZ - Report

Abstract

The goal of this project is to determine how well different pollutants are able to forecast changes in the Air Quality Index (AQI) in Phoenix, AZ, a vital indicator of air quality that has an impact on the environment and public health. Through the use of an extensive dataset from the Environmental Protection Act, containing concentrations of various pollutants, such as PM10, PM2.5, CO, Ozone, NO2, and SO2, we implemented complex data preprocessing methods to guarantee data accuracy and relevance. We showed the significant impact of specific pollutants on air quality by identifying important patterns and correlations between pollutants and AQI values through exploratory data analysis. By utilizing the XGBoost machine learning technique, which is optimized for both prediction accuracy and efficiency, we were able to create a model that can accurately estimate AQI. The results highlight the model's potential to assist environmental agencies and policymakers in developing strategies for pollution control and air quality management. This study adds to the expanding body of research on environmental monitoring and predictive analytics by providing a solid framework for other investigations that seek to improve air quality forecasts.

Introduction

Air quality prediction has become a crucial task for environmental scientists and public health officials, as it directly impacts human health and the environment. The Air Quality Index (AQI) serves as a standard measure to communicate how polluted the air currently is or might become in the future. Given the complex nature of air pollution, which involves multiple pollutants and variable weather conditions, predicting AQI accurately presents a significant challenge. Recent advancements in machine learning (ML) have offered new avenues for enhancing AQI prediction accuracy. For instance, XGBoost, a gradient boosting framework, has gained popularity for its effectiveness in dealing with non-linear data and its ability to handle missing values, making it a strong candidate for air quality forecasting. Studies such as those by T Madan (2020) and Zhang et al. (2018) have demonstrated the potential of ML models, including neural networks and decision trees, in predicting AQI with considerable accuracy. These studies underscore the importance of incorporating various pollutants as predictors and highlight the need for comprehensive data preprocessing to improve model performance. However, the literature also reveals controversies and challenges in AQI prediction, such as the selection of relevant features, dealing with spatial and temporal data variability, and the interpretability of ML models. The debate continues on the best practices for model training and validation, especially in the context of changing environmental regulations and the emergence of new pollutants. This project aims to contribute to the ongoing efforts in air quality prediction by

employing XGBoost to forecast AQI values based on historical pollutant data. By conducting a thorough literature review, we identify gaps in current methodologies and propose a novel approach to address these challenges. This study not only seeks to enhance the accuracy of AQI predictions but also aims to provide insights into the relative importance of different pollutants, thereby informing more effective pollution control policies.

Methods

This study aimed to predict the Air Quality Index (AQI) using data from various pollutants by leveraging the capabilities of XGBoost, a machine learning algorithm known for its efficiency and accuracy. The dataset utilized in this research was sourced from the Environmental Protection Agency (EPA, link in references), specifically their outdoor air quality data collection available at EPA's outdoor air quality data. This dataset includes daily measurements of pollutants such as Carbon Monoxide (CO), Nitrogen Dioxide (NO2), Ozone (O3), Particulate Matter 10 micrometers or smaller (PM10), Particulate Matter 2.5 micrometers or smaller (PM2.5), and Sulfur Dioxide (SO2).

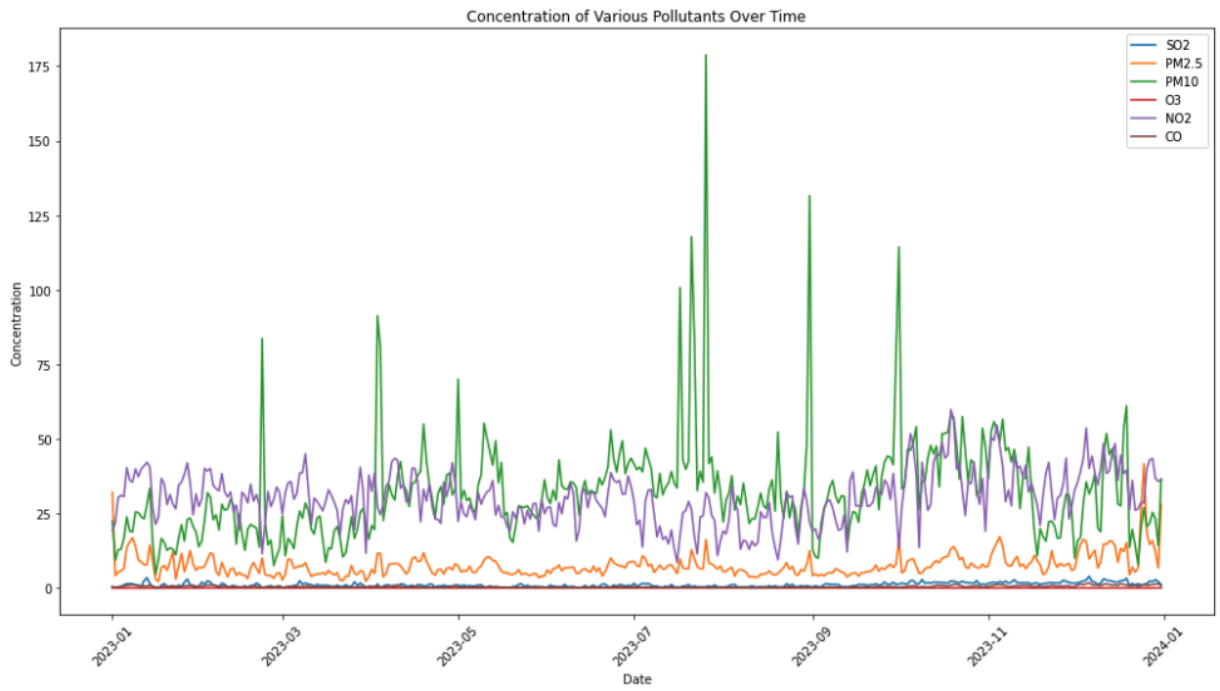
Data Cleaning and Manipulation
Initially, the data for each pollutant was loaded into separate Pandas DataFrames from their respective CSV files. The data cleaning process involved removing rows with missing values, normalizing the data formats, and ensuring consistency across datasets. We then merged these individual pollutant datasets into a single DataFrame based on 'Date' and 'Site Name' to facilitate a comprehensive analysis. This merged dataset underwent further cleaning to remove any remaining inconsistencies or outliers, ensuring the data's reliability for model training.

Spatial Processing
Given the geographic diversity of the dataset, we implemented spatial processing to account for variations in air quality across different locations. This involved aggregating data points by their geographical coordinates (latitude and longitude) to analyze AQI trends across regions.

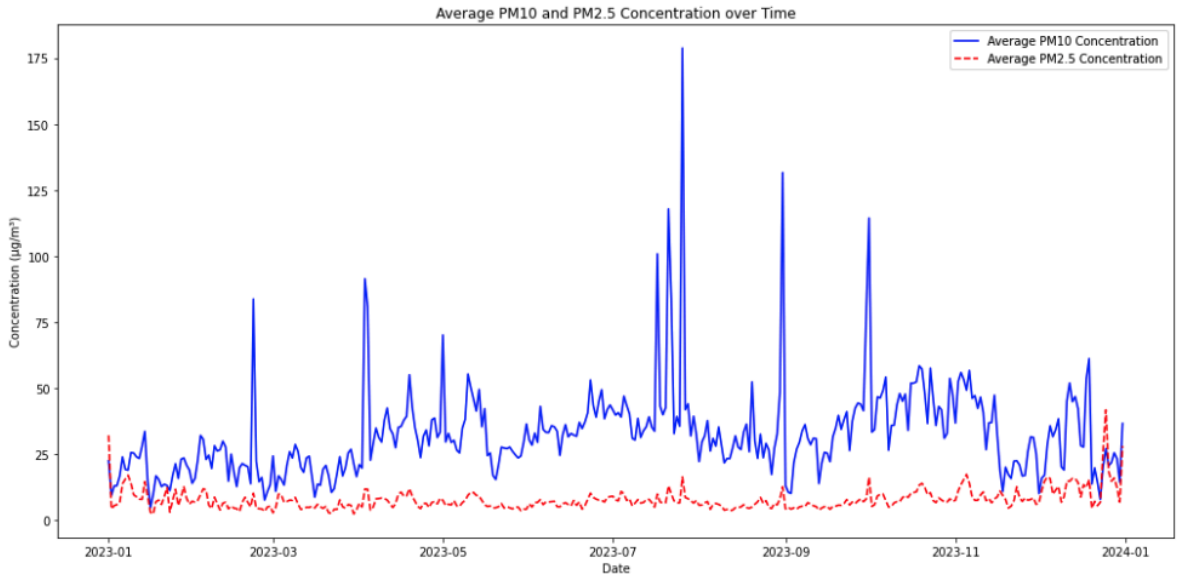
Machine Learning Application
For predicting AQI values, we employed the XGBoost regressor due to its proven track record in handling regression tasks efficiently. The model was trained using features derived from the pollutants' concentration levels and their respective AQI values on historical dates. The dataset was split into training and testing sets, with 90% used for training the model and the remaining 10% reserved for evaluation. We applied feature engineering to enhance model performance, including selecting relevant features that significantly impact AQI and conducting hyperparameter tuning to optimize the model's accuracy.

Results and Discussion

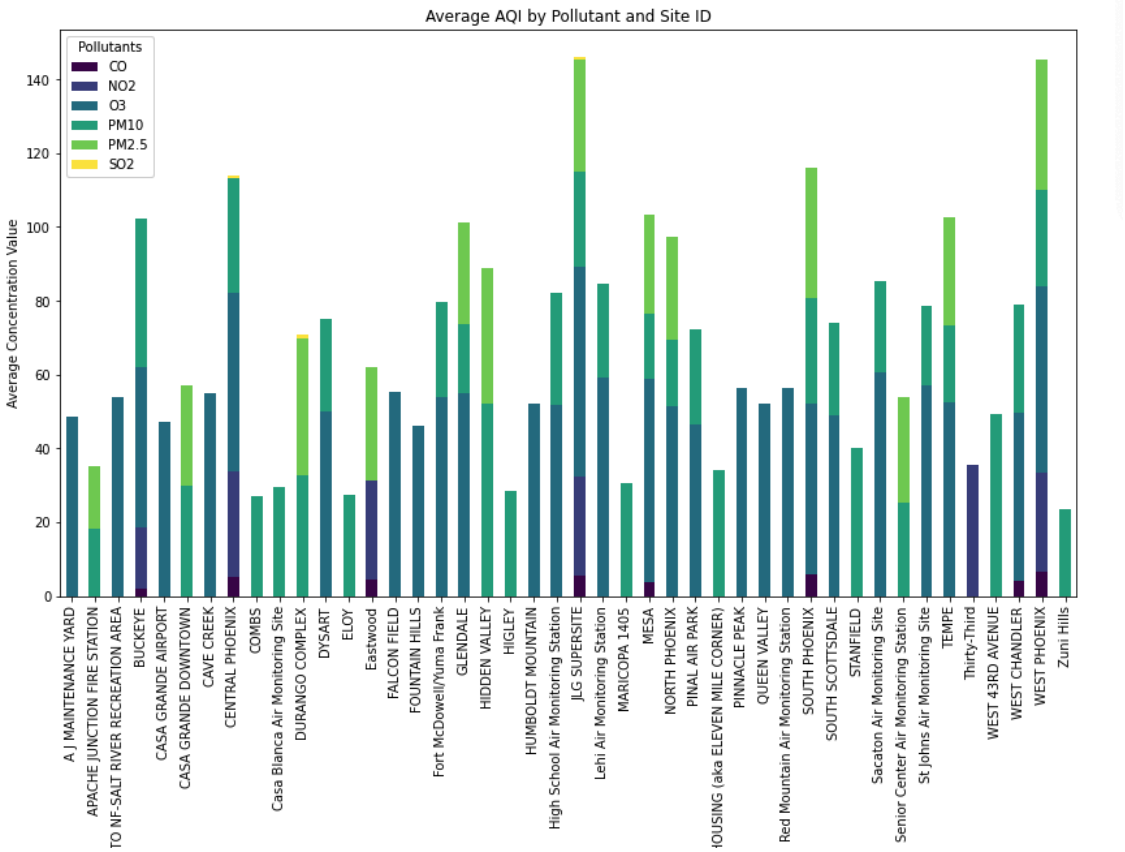
The exploratory data analysis (EDA) conducted in this study revealed significant insights into the relationship between various pollutants and the Air Quality Index (AQI).



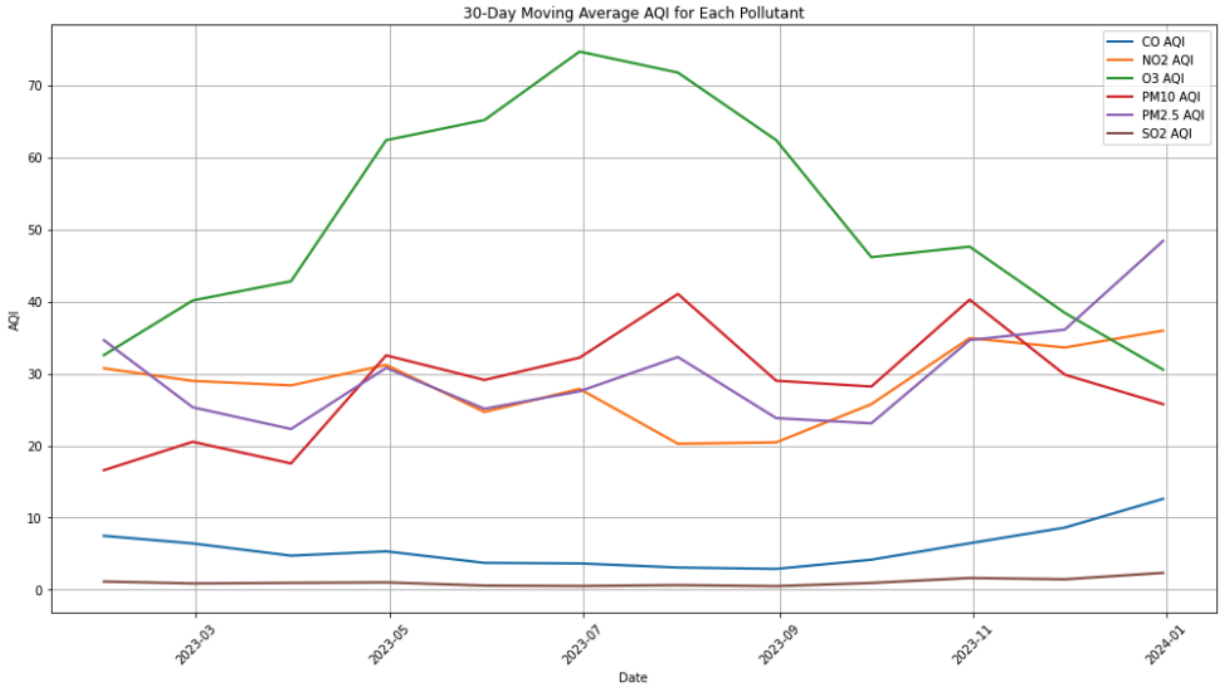
Initial visualizations showed the fluctuating concentrations of SO2, PM2.5, PM10, O3, NO2, and CO over time, highlighting PM2.5 and PM10 as major pollutants affecting AQI. This was consistent with the literature, which identifies particulate matter as a significant contributor to air quality degradation.



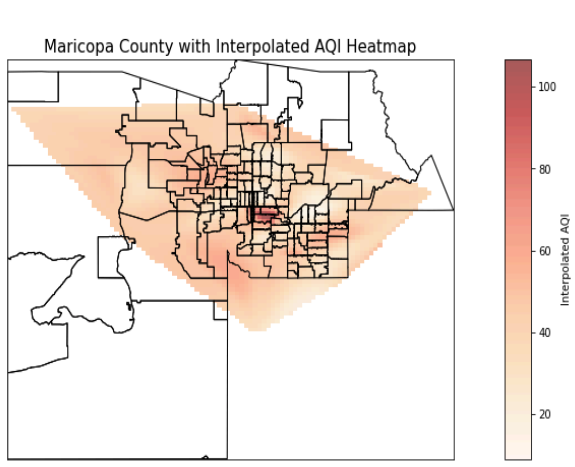
A deeper analysis with a focus on PM10 and PM2.5 concentrations over time further underscored their predominant impact on AQI levels. It became evident that PM10, in particular, had a more pronounced effect on air quality, aligning with findings from various weather data sites. This insight is crucial for targeting specific pollution control measures.

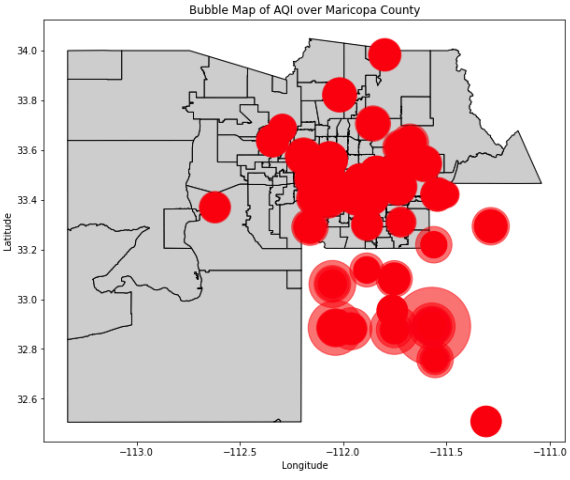


We also utilized stacked bar plots to visually represent the average AQI by pollutant and site ID, offering a clear understanding of which pollutants predominantly affect specific sites. This spatial analysis was instrumental in identifying pollution hotspots and could inform localized air quality management strategies.

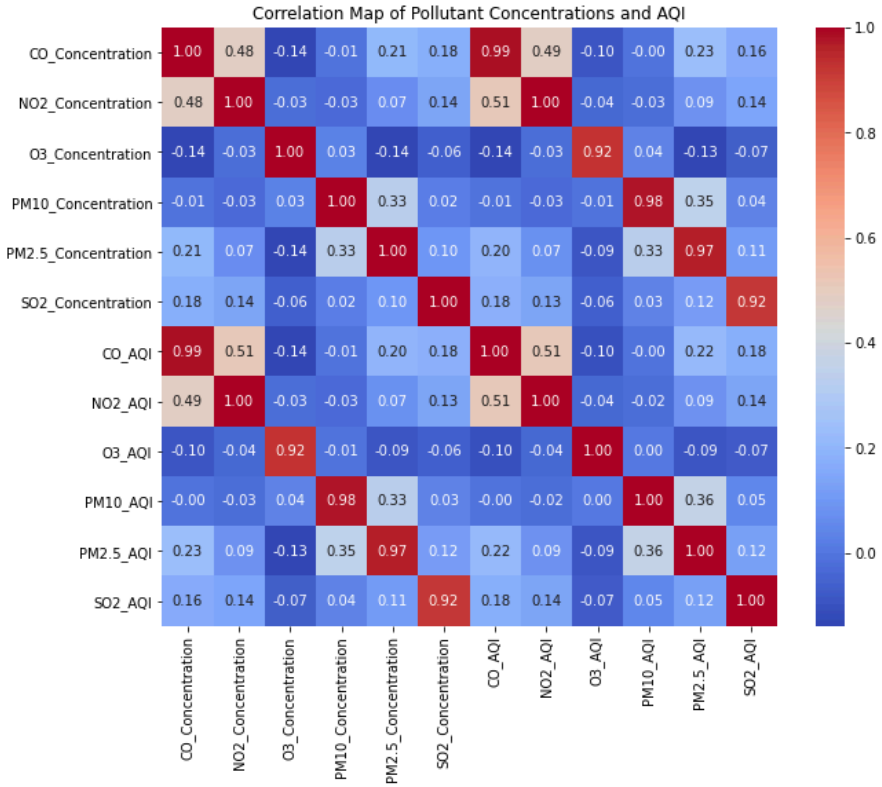


Moving average plots of AQI for each pollutant provided a smoothed overview of air quality trends, indicating that while PM10 levels were consistently high, Ozone also significantly impacted air quality. This dual focus on particulate matter and Ozone is essential for comprehensive air quality management.



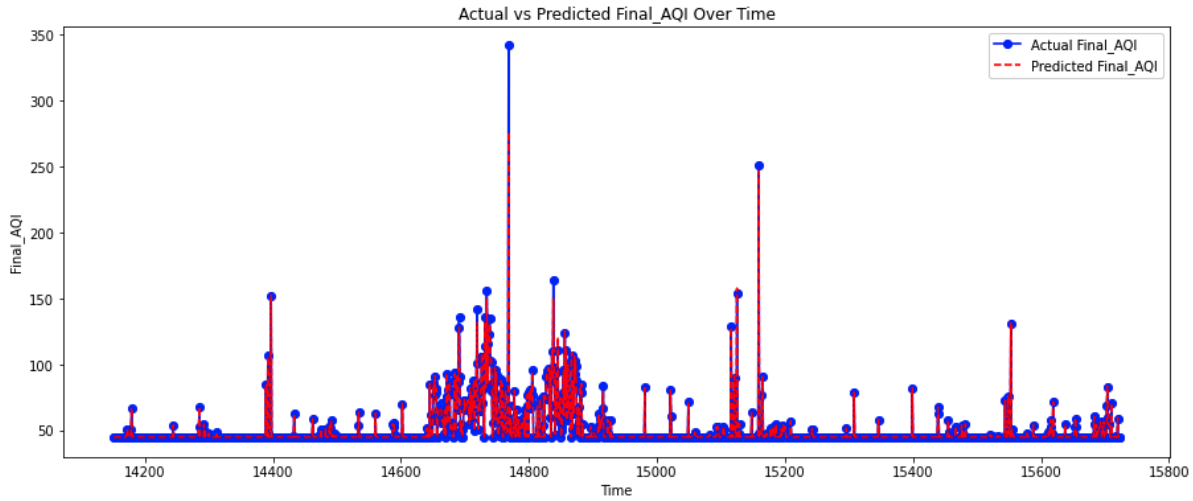


Several non-spatial plots were also generated to explore the temporal trends of pollutants and their AQI values. For instance, time series plots demonstrated seasonal variations in pollutant levels, with higher concentrations observed during specific periods of the year. This seasonal trend is crucial for understanding the temporal dynamics of air pollution and can aid in the development of time-specific pollution management strategies.



One of the key visualizations created was a heatmap of the correlations between different pollutants (CO, NO2, O3, PM10, PM2.5, and SO2) and the AQI. The heatmap illustrated a varied strength of correlation between these pollutants and AQI, with PM2.5 and O3 showing the strongest positive correlations. This suggests that these pollutants have a more significant impact on the AQI, highlighting their importance in air quality assessment and the need for targeted pollution control measures in areas where these pollutants are prevalent.

The application of the XGBoost regressor model, both in its base form and with hyperparameter tuning, yielded promising results in predicting AQI. The model achieved a root mean square error (RMSE) of 1.80 and a mean absolute error (MAE) of 0.187 for the base model, which improved to an RMSE of 1.31 and an MAE of 0.447 with tuning. These metrics indicate a high degree of accuracy in AQI prediction, showcasing the effectiveness of XGBoost in handling complex environmental data.



Visual comparison of actual vs. predicted AQI values over time demonstrated the model's precision, with the predicted values closely mirroring the actual AQI measurements. This accuracy is especially noteworthy, considering the inherent variability in air quality data and the complexities of accurately forecasting AQI.

In summary, the findings from this study highlight the significant impact of PM10 and Ozone on AQI, underscoring the value of advanced machine learning techniques, like XGBoost, in accurately predicting air quality. These insights have implications for environmental monitoring, public health, and policy-making, providing a data-driven foundation for targeted interventions to improve air quality.

Conclusion

The final analysis of the Air Quality Index (AQI) prediction project, utilizing the XGBoost algorithm, demonstrates a significant capability in forecasting AQI values from environmental pollutant data. The base XGBoost model achieved commendable performance, with a Root Mean Square Error (RMSE) of 1.80 for the training set and an impressive 0.19 for the testing set. Through hyperparameter tuning, the model's performance was further enhanced, reducing the training RMSE to 1.31 and slightly increasing the testing RMSE to 0.45. This improvement shows the model's enhanced generalization and its effectiveness in accurately predicting AQI values. The project successfully illustrated the potential of machine learning techniques, particularly XGBoost, in environmental science applications, such as air quality prediction. By meticulously cleaning and preparing the dataset for analysis, applying an advanced machine learning algorithm, and evaluating the model's performance, this study contributes valuable insights into the prediction of air quality levels. The findings underscore the importance of continuous improvement and adaptation of predictive models to enhance their accuracy and reliability in real-world applications. This research not only aids in the advancement of air quality monitoring and prediction strategies but also supports policymakers and environmental agencies in devising more effective pollution control and public health initiatives.

References

1. Huabing, Ke & Gong, Sunling & He, Jianjun & Zhang, Lei & Mo, Jingyue. (2022). A hybrid XGBoost-SMOTE model for optimization of operational air quality numerical model forecasts. Frontiers in Environmental Science. 10. 1007530. 10.3389/fenvs.2022.1007530.

2. Wang, J., Li, X., Jin, L., Li, J., Sun, Q., & Wang, H. (2022). An air quality index prediction model based on CNN-ILSTM. *Scientific Reports*, *12*(1), 8373. https://doi.org/10.1038/s41598-022-12355-6

3. Zhao, Z., Wu, J., Cai, F., Zhang, S., & Wang, Y.-G. (2023). A hybrid deep learning framework for air quality prediction with spatial autocorrelation during the COVID-19 pandemic. *Scientific Reports*, *13*(1), 1015. https://doi.org/10.1038/s41598-023-28287-8

4. US EPA,OAR. (2016, August 18). *Download Daily Data | US EPA*. US EPA. https://www.epa.gov/outdoor-air-quality-data/download-daily-data