CSE 6242 – Final Report – Team 20

Topic: Prediction and Visualization of Air Quality Index for US Regions

Team Members: Abhishek Gawande, Ayush Gordhan Agarwal, Ajay Krishna Manoj, Harshal Mittal, Nicholas Ramallo, Sandeep Sainath

1. Introduction

The aim of our project is to predict the concentration of different pollutants in the atmosphere, compute AQI across different US regions, and present the trends using interactive visualizations. We aim to provide dynamic, interactive visualization of AQI and pollutant concentrations with personal health recommendations directly to the user's inbox.

2. Problem Definition

Currently, there are dashboards that give the AQI or weather of a particular region but often do not provide any personal recommendations based on user location and sign-up. People do not proactively track AQI and weather on a regular basis and do not take appropriate actions.

3. Literature Survey

While COVID-19 has been the focus, there are more potential dangers that people should also be aware of. According to estimates from the American Lung Association, air pollution induced illnesses add $37 billion dollars to U.S. healthcare spending [1]. Air pollution can play a key role in health outcomes within a population. Roy et al. [2], analyze Kolkata, which along with many other cities in Asia has been undergoing rapid growth. This increase in population and urbanization has led to increased rates of air pollution-related illnesses. But, since our data is for the U.S., we will not see areas with the same level of rapid urbanization. Castelli et al. [1] discuss the impact of air pollution on healthcare spending. It is believed that within California, poor air quality leads to $15 billion in spending. Additionally, Pimpin et al. [3] simulate different scenarios in the U.K. around the concentrations of air pollutants and show that this will lead to billions of increased spending on the NHS and millions of added incidences of air pollution-induced diseases. However, the U.K. has considerable differences to the U.S. in terms of geography and demographics. Zhang et al. [4] explore how common pollutants (NO_x and VOCs) are affected by solar radiation and in turn become ozone. Also, Cao et al. [5] show that COVID-19 transmission is positively correlated with poor air quality, over both the short and long term. While this trend may be concerning, Torkmahalleh et al. [6] and Archer et al. [7] find that levels of NO_2 and $PM_{2.5}$ fell due to the reduction in human activity during COVID lockdowns. However, as was noted in [6], O_3 concentrations did see increases, especially in urban areas, because of a reduction in industrial activity. Air quality index (AQI) is a single metric to measure air quality with respect to human-health concerns and is measured differently in different countries [8]. In US, the maximum value out of the measured AQI values for individual air pollutants is reported as the final AQI [9][10]. While a single metric is good for providing the overall picture, every pollutant's level should also be considered while providing recommendations.

We explored several different ways to predict air quality and we aim to work on a novel Time Series approach that considers geospatial relationship between nearby regions to predict the pollutant concentration. Basic linear models like multiple linear regression were used in Rajput et al. [11] to build a model that can predict the air pollutants. However, in prediction, Singh et al. [12] found nonlinear models to capture complex nonlinearity in air quality data. Castelli et al. [1] propose using support vector regression (SVR) with a radial basis function (RBF) kernel to forecast pollutant and particulate levels to predict air quality in California. [13] uses single exponential smoothing (SES) of historic pollutant concentrations in a region to predict the air quality. The trend and seasonality components are however

not being captured. These serve as useful baseline models for time series forecasting on pollutant/particulate levels but have room for improvement through including data on more pollutants and attempting more advanced models. Cassard et al. [14] build on this approach by gaining data from low-cost sensors on both pollutants and temperature to predict air quality. They rely on a convolutional neural network with just one hidden layer, but do not perform robust hyperparameter tuning on their model. Janarthanan et al. [15] somewhat combine these approaches by using an ensemble of SVR and a Long Short-Term Memory model using various pollutants as features. However, it ignores the aforementioned time series approach in forecasting pollutant concentrations and only focuses on AQI. Feng et al. [16] also present an attempt to use deep learning in forecasting AQI in Wuhan, China by combining both predictions of pollutant features using time series models and meteorological data. While they can improve model accuracy through using models beyond just traditional multi-layer perceptron's (MLP), they do provide a framework of optimizing any given neural network approach to forecasting AQI.

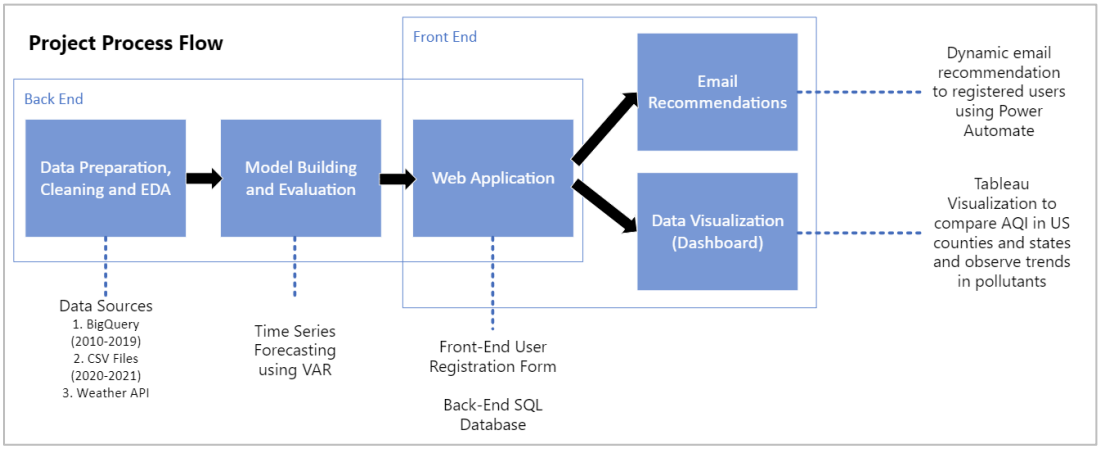
A critical issue is creating awareness amongst the general public about air pollution and promoting public participation. A study in China states that about 49% of the respondents are aware that air pollution is a looming threat [17]. Cromar et al. [19] found that AQI communication was the most effective predictor of respiratory risk in California. Visualization is an effective method for sharing information. Li et al. [18] mentions the limitations of scatter plots and other simple techniques when dealing with complex data and advocates the use of more intuitive multi-perspective graphs and geo-visualizations.

4. Methodology

4.1 Intuition/Motivation

Our tool will provide an in-depth analysis of the impact of air quality in a selected region and comparative study of different regions and the effects of each pollutant on the AQI. This would both help people with health conditions and families with small kids and government and healthcare agencies to plan the urban development and restrict certain industries based on the pollutant levels, while tracking long term impacts of AQI on the health. We see two main benefits of this: users can take preventive measures to stay healthy (prevent allergy and asthma attacks) based on adequate preparation through our recommendations, and government bodies can take effective and quicker decisions based on predicted air quality and pollutants in the atmosphere.

4.2 Approach



4.2.1 Data Preparation

For this project, we are using two main sources of data to compile the last 11 years of data for six key atmospheric pollutants. The bulk of the data comes from a BigQuery database (2010-2019) for which we

set up a client in Google Cloud, to allow our data queries to be authorized. The BigQuery dataset had data until the end of 2019, so to supplement the recent data, we found another official data source (2020-11). While we intend to conduct our analysis at the county level, the data is given at the site (measurement station within a county) level, which requires extra adjustments when creating and merging our data sources using python. To ensure that our analysis could be done and would be meaningful, we set some criteria on the quality of each stream of data. We wanted at least 750 data points (out of 868) from 2019 onward, and that we had at least 10 years of consistent data over the last 11 (3650 data points minimum). To fill in any data missing for specific dates, we interpolated the values using time series estimation. Due to the lack of consistent data from Dec 2020 onwards, our modeling approach is based on predicting AQI in Dec 2020, since the data for that is more robust. Additionally, to access weather forecasts for each location we are modelling, we are utilizing a weather API, OpenWeatherMap. This will provide us with the current day (as of API call) and a seven-day daily forecast. We plan to use the forecast to enhance the visualization and email aspects of the project.

4.2.2 Time Series Forecasting

Our aim is to forecast the concentration of atmospheric pollutants in each county for the next few weeks based on the past concentration levels. There are 6 atmospheric pollutants we are considering, namely: CO, O_3, PM_{10}, $PM_{2.5}$, NO_2, and SO_2. From our dataset, we have 100 counties for which we need to forecast the pollutant concentration and their corresponding AQI value. For the prediction of each pollutant in a single county, we considered 11+ years of historical pollutant concentration, temperature, pressure, relative humidity, and windspeed in that region. As pollutant concentrations and weather are highly influenced by geospatial locations, factoring in the concentrations of nearby counties would improve our prediction. This novel approach required us to investigate models that can be used on multivariate data for 100 plus regions and 6 pollutants. We used data from Jan 2010 to Nov 2020 for training and validating the models, and we have provided forecast from Dec 2020 onwards.

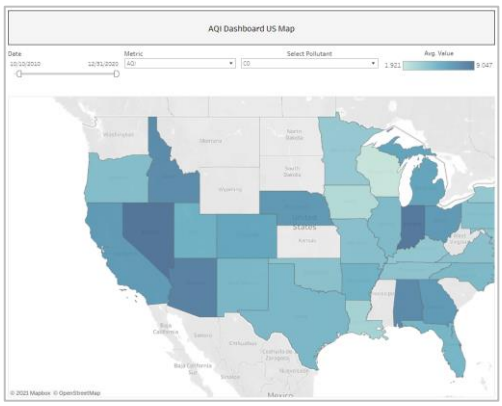
The model we have built is a Vector Auto Regression (VAR) using weather and pollutant concentration along with concentrations for 5 nearest counties (distance calculated using spatial coordinates and the number decided by distance threshold and data availability). VAR is a method that considers multiple times series and how they all relate to one another. Without knowing the scale at which the underlying pattern in the data would be best modeled, we considered methods that looked at the trends as a whole and others that consider certain series as separate, but still allowed for the potential influence of other series on to each other. We have also modelled using ARIMA as a baseline for each pollutant concentration for a given county.

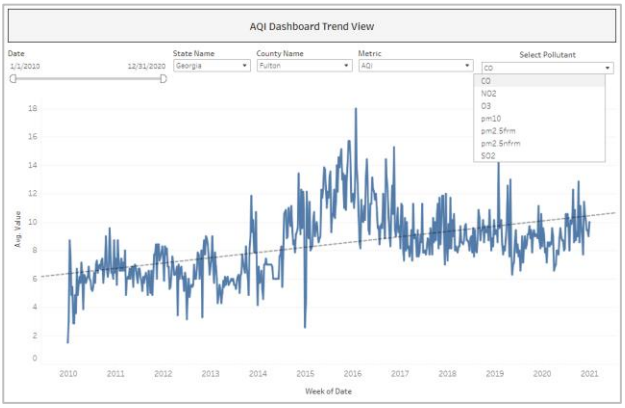
4.2.3 Web App for Home Page and User Registration

To create the web app and the web page within it, we are using Flask as our framework. Flask is a simple, Pythonic, and easy-to-deploy solution that fit our needs of hosting a basic web page with a registration form. In addition, Flask integrates easily with SQLite, which is our current database solution, through SQL Alchemy, a database toolkit for Python. In addition, Flask and SQLite each are easy and scalable solutions and offer many options for future customization. The web app consists of one page containing the registration form and a “Register” button prompting the user to fill in their registration information and answering some questions for personalized recommendations via email updates. After a user registers, the page refreshes to the same form for additional user registration with a message indicating that they have successfully registered.

4.2.4 Data Visualization

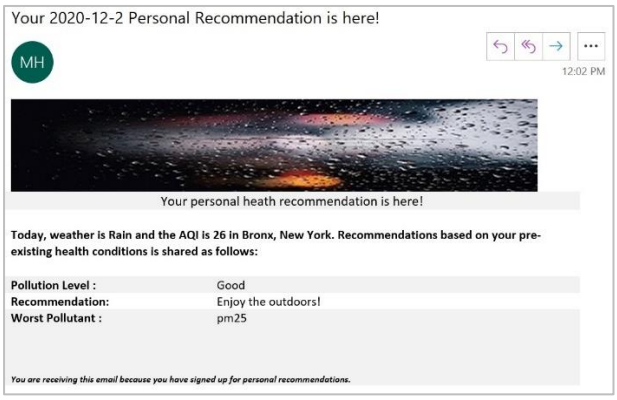
We are using Tableau to build a dashboard that shows the pollutant levels across the states as well as the predicted values and trends of pollutants across the years. The user can filter the pollutant, the date range and county name as per the requirement. The dashboard is published and hosted on Tableau Public and the link is embedded into the web application.

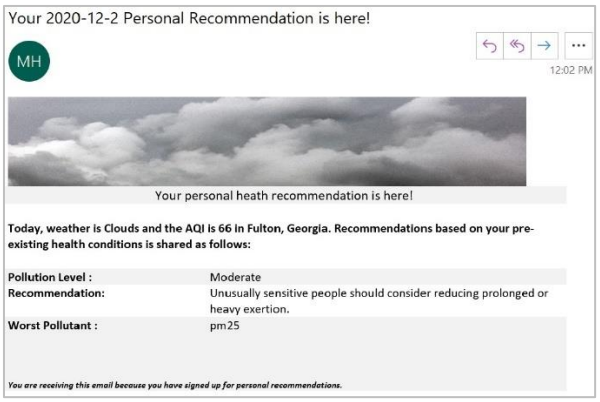




4.2.5 Automated Email System

We are leveraging the Microsoft Power Automate platform to send out daily recurring emails to users based on the location they have signed up for and the existing health conditions they have. The structured data used for the emails comes from the predicted AQI dataset, recommendations mapping and the user database from the web application. We have also used the weather API to get the data about the weather in terms of whether it is sunny, cloudy, windy, rainy, or snowy. The merged file contains the predicted AQI and personal recommendations for each user who has registered with us. The email content itself is very dynamic with guidelines and visuals in the emails changing based on weather (sunny, cloudy, and rainy) and whether the AQI level is hazardous or not. The design of the email has been developed with HTML and CSS, and it has been incorporated into the Power Automate email stage. A workflow has been setup in Power Automate which checks for several parameters and fetches the relevant visuals to display in the email.





4.2.6 List of Innovations

1. End-to-end web app/visualization product that deliver personalized recommendations to individuals.
2. Time series forecasting model that uses K-nearest neighbors to understand geospatial effects in conjunction with time series modeling on pollutant concentrations to predict AQI for each region.

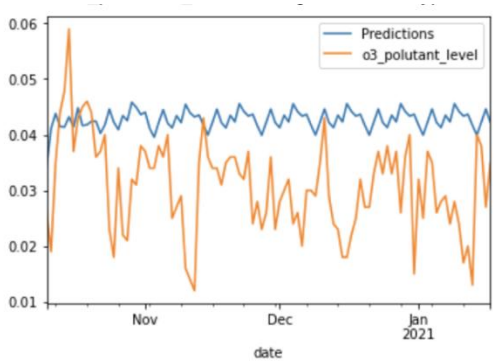
5. Distribution of Team Member Effort

Ajay, Ayush, and Nicholas worked on gathering data from the sources stated earlier, and the preparation and cleaning of the collected data. Sandeep and Abhishek worked on developing the web app using Flask and SQLite to set up the registration form via database solution and hosting the visualizations on the platform. Harshal developed the design of the emails using HTML and CSS scripts and configured the workflow for personalized recommendation emails using Power Automate. Abhishek and Ajay worked on building and validating the models for prediction. Ajay, Ayush and Harshal also worked on creating the AQI dashboard with the pollutant level heat-map and pollutant trends. All team members worked on integrating the different components into a system that works in tandem.

6. Experiments/Evaluation

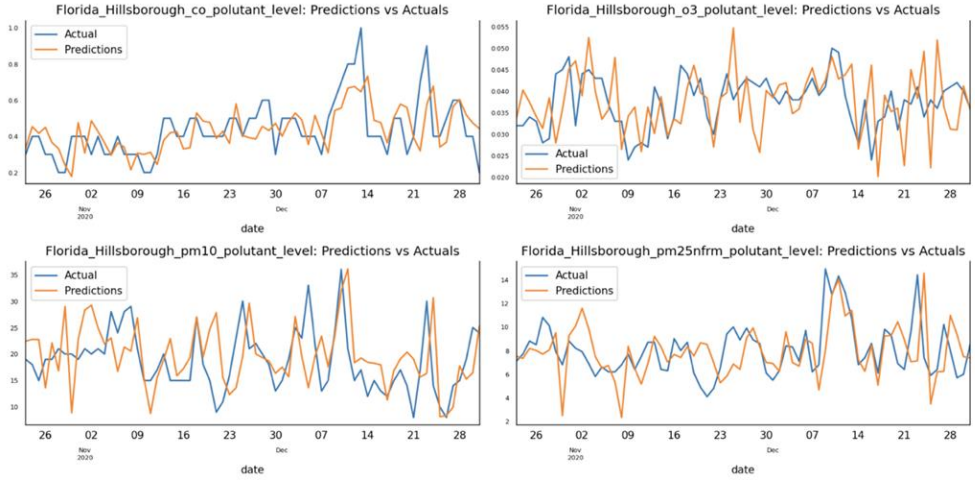
Our experiments primarily involve training and hyperparameter tuning of our time series-based machine learning model. We investigated different models which can incorporate the seasonality we see in our data, but also take advantage of new techniques which are effective on non-linear paths. We deterministically calculate AQI based on individual pollutant concentrations as per the U.S EPA guidelines. To evaluate our visualization, we used best practices learned in class to ensure a simple, accessible dashboard that is easy to navigate and highly interactive. To evaluate our web application, we manually tested the input registration information from several different computers simultaneously to test the scalability and robustness of our backend, as well as perform this action over several days to ensure our email recommendations are based on the current day.

6.1 Time Series Forecasting Model Experiments



ARIMA output through visual inspection and error analysis gives poorer predictions than VAR. Though it is simpler and more explainable, it is unable to capture the complexities of pollutant concentrations time series data.

Below are some of the predictions achieved by VAR models for Hillsborough County for November and December months. Best VAR order was selected according to Akaike Information Criterion.



Results from the Tuned VAR Model (Averaged across all counties):

Pollutants	Mean Error (ME)	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
CO	0.17265066	0.68773626	0.84060875
NO2	3.79870041	22.3608868	28.3761837
O3	-0.0036609	0.02023488	0.02484914
PM10	19.4364637	83.7446786	102.176048
PM2.5frm	12.3982723	17.9873591	22.289607
PM2.5nfrm	9.54910864	24.2619868	29.4094337
SO2	0.73753354	5.13394351	7.07410957

Our model predicts the pollutant concentration and AQI accurately for most of the counties and across various pollutants. As shown above, PM10 has a higher error rate but this is unevenly distributed across various counties. The errors are in acceptable range for majority of the counties. Due to lack of available data for some counties and the variations in the scale of the pollutant concentrations, the overall error and average of values are on the higher end for a few pollutants.

6.2 Web App Evaluation

To evaluate our web app, we first attempted user registration locally to test whether the user input was saved properly to a database file (.db). In addition, to evaluate personalization of our web app, we add three questions inquiring users about any existing respiratory and heart conditions that they may have as well as if they have any children or elderly people around them. These questions target any relevant medical conditions about the user as well as any vulnerable populations that they may be surrounded by. Once local development was completed, we use Render, a cloud application hosting platform, to deploy our web app as a web service (link: https://dva-app-test.onrender.com) to ensure that individuals would be able to access our site. In the deployment process through storing our web app as a GitHub repository, Render automatically verifies that the web app is ready for deployment since it logs deployment attempts as failures if it is not. Once the site became live, we tested user registration on each of our project members' computers as a simple scalability and robustness check. We use the free plan on Render that provides 256 MB of RAM and a shared CPU as a baseline for our framework.

7. Conclusion

This project provides an end-to-end, robust, and scalable framework for a health recommendation system based on novel multivariate time series forecasting of AQI. This system delivers personalized health recommendations to users who register for updates on a simple web application based on their region of residence and relevant health factors. Finally, the web application also includes interactive Tableau dashboards to visualize AQI forecasts and trends by region to further enforce the urgency and importance of monitoring air quality.

While our project is limited in scope due to the lack of recent AQI and accurate data, the main goal was to provide an interpretable and scalable framework that can be generalized by health experts as necessary. We believe that with this framework, health experts with more domain knowledge in the health and weather domain can use this recommendation system to deliver interactive visualizations and personalized recommendations based on accurate air quality forecasts to assist individuals across the U.S. in safeguarding against bad and harmful weather.

Acknowledgements
We sincerely thank Prof. Nicoleta Serban and Prof. Joel Sokol from the ISYE department for their guidance in approaching time series forecasting for our project.

References
1. Mauro Castelli, Fabiana Martins Clemente, Aleš Popovič, Sara Silva, Leonardo Vanneschi, "A Machine Learning Approach to Predict Air Quality in California", Complexity, vol. 2020, Article ID 8049504, 23 pages, 2020. https://doi.org/10.1155/2020/8049504.
2. Roy, Sandeep, et al. Smart and Healthy City Protecting from Carcinogenic Pollutants . 2017, https://www.ripublication.com/ijaes17/ijaesv12n9_04.pdf.
3. Pimpin, Laura, et al. "Estimating the Costs of Air Pollution to the National Health Service and Social Care: An Assessment and Forecast up to 2035." PLOS Medicine, Public Library of Science, https://journals.plos.org/plosmedicine/article?id=10.1371%2Fjournal.pmed.1002602.
4. Zhang, Junfeng (Jim), et al. "Ozone Pollution: A Major Health Hazard Worldwide." Frontiers, Frontiers, https://www.frontiersin.org/articles/10.3389/fimmu.2019.02518/full.
5. Cao, Ru, et al. "Estimating Short- and Long-Term Associations between Air Quality Index and Covid-19 Transmission: Evidence from 257 Chinese Cities." *Frontiers*, Frontiers, 1 July 2021, https://www.ssph-journal.org/articles/10.3389/ijph.2021.1604215/full.
6. Torkmahalleh, Mehdi Amouei, et al. "Global Air Quality and Covid-19 Pandemic: Do We Breathe Cleaner Air?" *Aerosol and Air Quality Research*, Taiwan Association for Aerosol Research, 8 Feb. 2021, https://aaqr.org/articles/aaqr-20-09-covid-0567.
7. Archer, Cristina L., et al. "Changes in Air Quality and Human Mobility in the USA during the COVID-19 Pandemic." *Bulletin of Atmospheric Science and Technology*, Springer International Publishing, 26 Oct. 2020, https://link.springer.com/article/10.1007/s42865-020-00019-0.
8. Tan, Xiaorui, et al. "A Review of Current Air Quality Indexes and Improvements under the Multi-Contaminant Air Pollution Exposure." Journal of Environmental Management, Academic Press, 13 Dec. 2020, https://www.sciencedirect.com/science/article/pii/S0301479720316066?casa_token=mDK3JSa5vSwAAAAA%3AgdTng2oktHYcEOmy-wC0xKiNpB5VjownjjTOOFk07ETYek9K3Cd4ercSBgx5XCc1Q6ufsWZy.
9. Bishoi, Biswanath, et al. "A Comparative Study of Air Quality Index Based on Factor Analysis and US-EPA Methods for an Urban Environment." *Aerosol and Air Quality Research*, Taiwan Association for Aerosol Research, 28 Feb. 2009, https://aaqr.org/articles/aaqr-08-02-oa-0007.
10. "Technical Assistance Document for the Reporting of Daily Air Quality – the Air Quality Index (AQI)." U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Air Quality Assessment Division, Sept. 2018. Publication No. EPA-454/B-18-007.
11. Rajput, Tripti Singh, and Nandini Sharma. "Multivariate Regression Analysis of Air Quality Index for Hyderabad City: Forecasting Model with Hourly Frequency." *International Journal of Applied Research*, 2017.
12. Singh, Kunwar P., et al. "Linear and Nonlinear Modeling Approaches for Urban Air Quality Prediction." *Science of The Total Environment*, Elsevier, 25 Apr. 2012, https://www.sciencedirect.com/science/article/pii/S0048969712004809?casa_token=8AVHQ4A85KoAAAAA%3AlBtt-427V549fupsO3kKS08po3xnBycPzCGC2BWTNZxu0Poq4_3yTg4N62-H4hbZRLRtXd7q8oQ.
13. Roy, Sandip, et al. "Time Series Forecasting Using Exponential Smoothing to Predict the Major Atmospheric Pollutants." *IEEE Xplore*, 2018, https://ieeexplore.ieee.org/abstract/document/8748326/.
14. Cassard, Thibaut, et al. "High-Resolution Air Quality Prediction Using Low-Cost Sensors." *Arxiv.org*, Plume Labs, 2020, https://arxiv.org/pdf/2006.12092.pdf.

15. Janarthanan, R., et al. “A Deep Learning Approach for Prediction of Air Quality Index in a Metropolitan City.” *Sustainable Cities and Society*, Elsevier, 20 Jan. 2021, https://www.sciencedirect.com/science/article/pii/S2210670721000159.
16. Feng, Qi, et al. “Improving Neural Network Prediction Accuracy for PM10 Individual Air Quality Index Pollution Levels.” *Environmental Engineering Science*, Mary Ann Liebert, Inc., 1 Dec. 2013, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3875204/.
17. Xu, Zhihua, et al. “Extending the Theory of Planned Behavior to Predict Public Participation Behavior in Air Pollution Control: Beijing, China.” *Taylor & Francis*, May 2019, https://www.tandfonline.com/doi/full/10.1080/09640568.2019.1603821?scroll=top&needAccess=true&.
18. Li, Huan, et al. “A Visualization Approach to Air Pollution Data Exploration-A Case Study of Air Quality Index (PM2.5) in Beijing, China.” *MDPI*, Multidisciplinary Digital Publishing Institute, 29 Feb. 2016, https://www.mdpi.com/2073-4433/7/3/35/htm.
19. Cromar, Kevin R, et al. “Evaluating the U.S. Air Quality Index as a Risk Communication Tool: Comparing Associations of Index Values with Respiratory Morbidity among Adults in California.” *PLOS ONE*, Public Library of Science, Nov. 2020, https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0242031.