Applying Analytic Models on the WHO Ambient Air Quality Database

Team Name

ETA Bros

Discussion Section

DIS-06

Team Members

Justin Wang: justin17@stanford.edu

Albert Tan: albert-tan@stanford.edu

Peter Fu: peter107@stanford.edu

Tianle Yao: tianle@stanford.edu

1. Introduction

Throughout the past century, rapid industrialization and growth has given mankind many marvels: the internet, air conditioning, rubber ducks, etc. However, these gifts come with a price, immense amounts of carbon emissions and industrial waste are released into our atmosphere, polluting our air and environments, presenting us with a novel and (potentially) lethal issue. Air pollutants from industrial zones have been linked to cardiovascular disease, breathing issues, lung cancer, and a myriad of other issues. However, we have grown aware of this growing issue and have already begun attempts to analyze the damage and mitigate it, creating multiple reports and organizations to reduce air pollution.

In this report, we will utilize one of these reports on global air quality to analyze the trends in the data to build a more complete data set. Collecting data for this set is a daunting task; not all regions in the world will have a detailed data record. Utilizing our knowledge of predictive modeling and model evaluation, we can attempt to predict missing data along with providing insights on trends in our data, creating a detailed report on our insights and models, helping others develop an improved understanding of air pollution.

We developed five main analytical questions to develop a better grasp of the dataset. We wanted to see the relationship between our pollutants, we wanted to find a way to predict the air quality of a region utilizing its background, we wanted to develop an accurate model for finding a country from only coordinates, we wanted to analyze the trend between the developmental status of a country and its air quality, and analyze the trend of air pollution in the past and utilize it to predict it in the future. By analyzing and solving these five questions, we can develop a more sophisticated understanding of our data along with its implications on our world.

2. Dataset Description

2.1 Main Dataset

WHO Ambient Air Quality Database
https://www.who.int/publications/m/item/who-ambient-air-quality-database-(update-jan-2024)

2.1.1 Data Provenance

Each of the 40,098 entries in the database represents the average air quality data observed in a specific location (city / town) on Earth in a specific year. The data dictionary is part of the first sheet of the spreadsheet. Useful information in the dataset includes concentration for three types of air pollutants (PM2.5, PM10, NO_2) along with facts about the location (geographic categories, types of stations in the location, geographic coordinates, etc). Some values about air pollutants are missing in particular rows; since the dataset is large enough, these rows are usually dropped.

2.1.2 Data Preparation

Since the data is structured clearly, not much preparation is necessary. The specific steps differ for each question, but mostly involve dropping rows with missing data and dropping columns with unnecessary information.

2.2 Supplementary Dataset

World Economic Outlook Groups and Aggregates Information
https://www.imf.org/external/pubs/ft/weo/2022/01/weodata/groups.htm (Original Source)
https://albertttan.github.io/static/sss/country_classification.json (Prepared Data)

2.2.1 Data Provenance

The webpage by the International Monetary Fund categorizes its 196 member countries and regions into two major groups, advanced economies along with emerging and developing economies, and further divides each group based on geographic location or membership of particular organizations. We focus on the large categorization as a brief indicator for the development status of a country. However, the data is not formatted in a machine-readable way, and the IMF's method of naming the countries differs from the WHO's.

2.2.2 Data Preparation

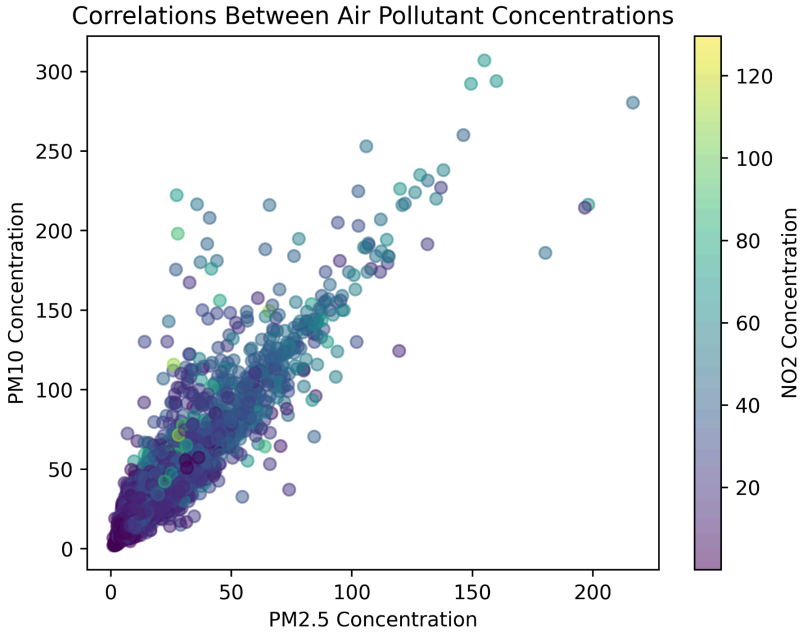
Since the original data is presented as a web page, we reformat it into a JSON file that groups the economies (see Prepared Data above). Group 2 represents advanced economies, while groups 0 and 1 together represent emerging and developing economies. Group 0 itself represents least developed countries listed by the United Nations, but is irrelevant in our analysis. The names of some economies

are manually edited so that they fit those in the WHO database. The few countries that are members of WHO but not members of IMF, like Cuba, are manually added to group 1.

3. Exploratory Data Analysis

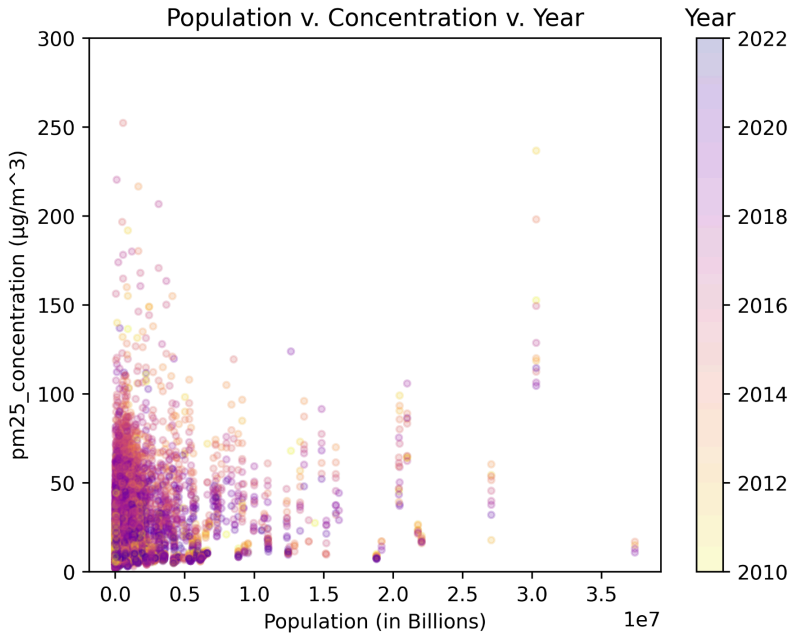
3.1 Question 1

A scatterplot is produced to show the correlation between the concentration of PM2.5, PM10, and NO_2. There is a clear positive correlation between the three variables, perhaps except for some outliers with high NO_2 concentration where the other two pollutants are less concentrated. This gives us confidence to proceed to our prediction of PM10 concentration based on the other two values given.



3.2 Question 2

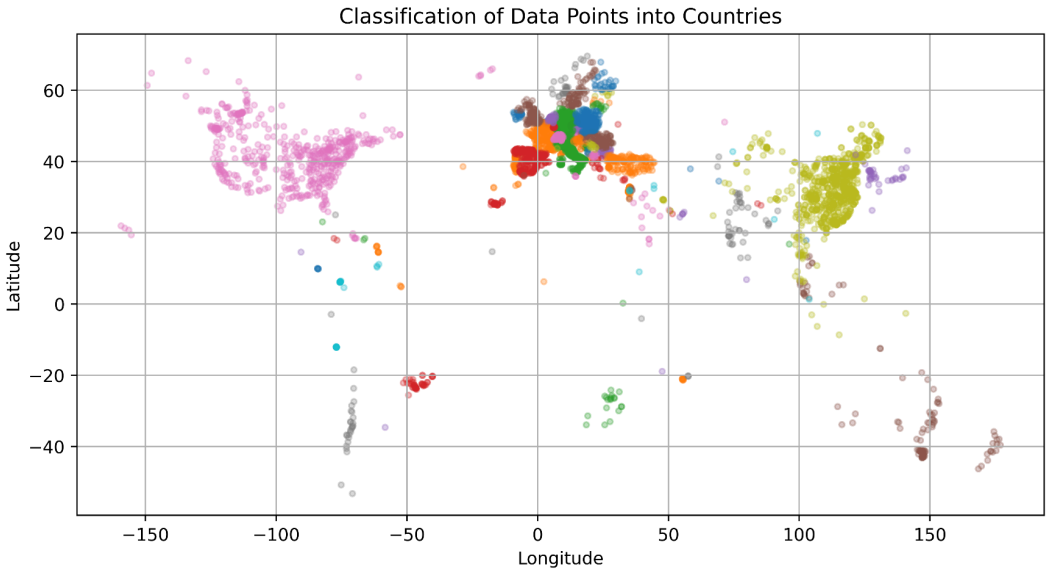
A scatterplot will be made based on the population (in billions) of the region, the PM2.5 concentration (in $\mu g/m^3$), and the year the data was collected. We avoided some of the other variables for sake of clarity. This allows us to examine the relationship between the variables to make hypotheses about which model is a better fit for our data.



As we can see, the data is heavily clustered in the bottom right of the graph, which indicates the majority of measurements have low population and low PM2.5 concentration. This visualization leads to our hypothesis that since our data is heavily clustered, a K Neighbors Regression would be the best fit model for our data. We can now move on to validate our claim.

3.3 Question 3

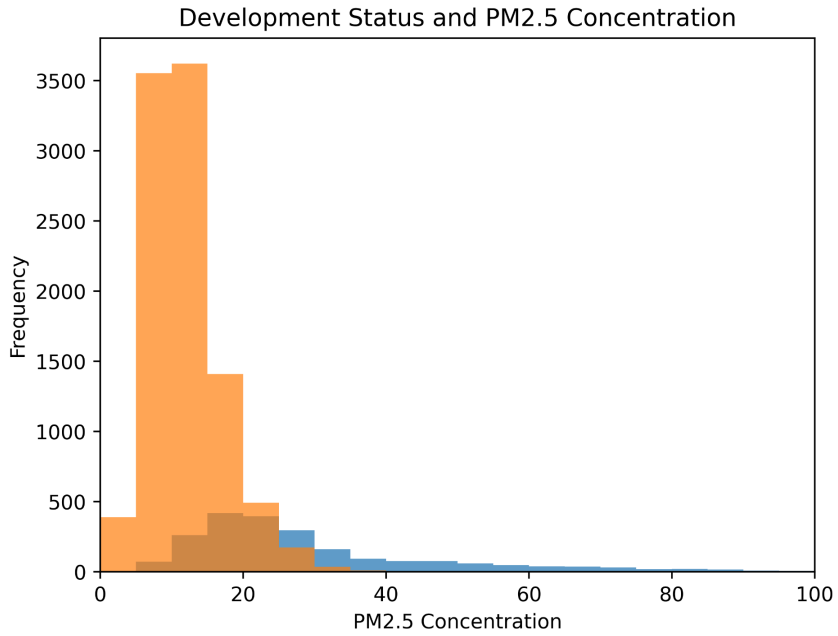
A scatterplot is produced based on the latitudes and longitudes of each available data entry; the color of each dot represents the country the entry belongs to. This allows us to visualize the distribution of data across different countries along with where potential errors might occur.



We discover that most entries are located in Europe, followed by North America and East Asia. Africa, South America, and North Asia have the least amount of data. Our model might thus be more effective for classifying coordinates in Europe compared to the parts sparsely populated by existing entries.

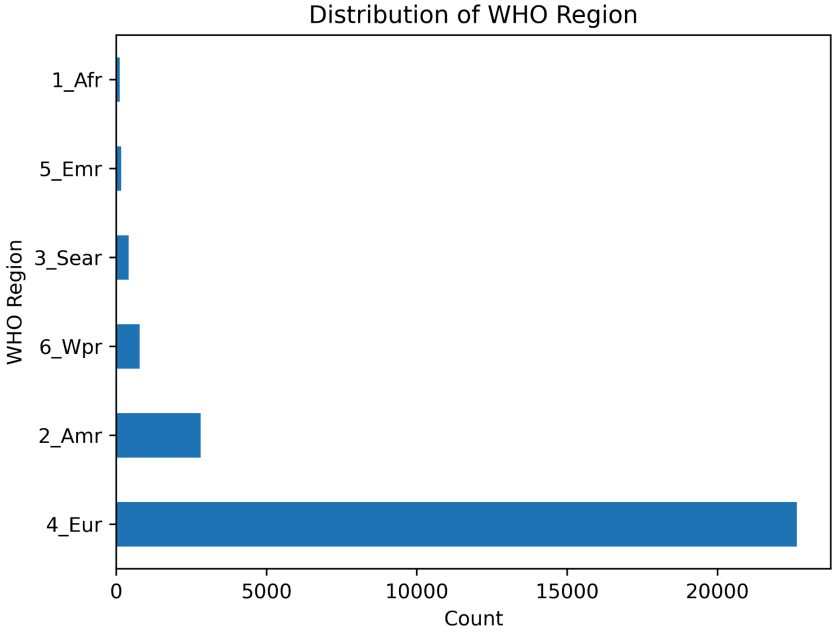
3.4 Question 4

A histogram is produced to visualize the distribution of data for developed and developing countries after entries with empty values are removed. Only PM2.5 concentration is used in the visualization, but our actual analysis will incorporate all three air pollutants. There is much more data from developed countries (shown in orange) than developing ones (shown in blue). Developed countries tend to have lower air pollution levels in general. We also see a clear overlap between the two color blocks, which shows that the task can be potentially challenging.

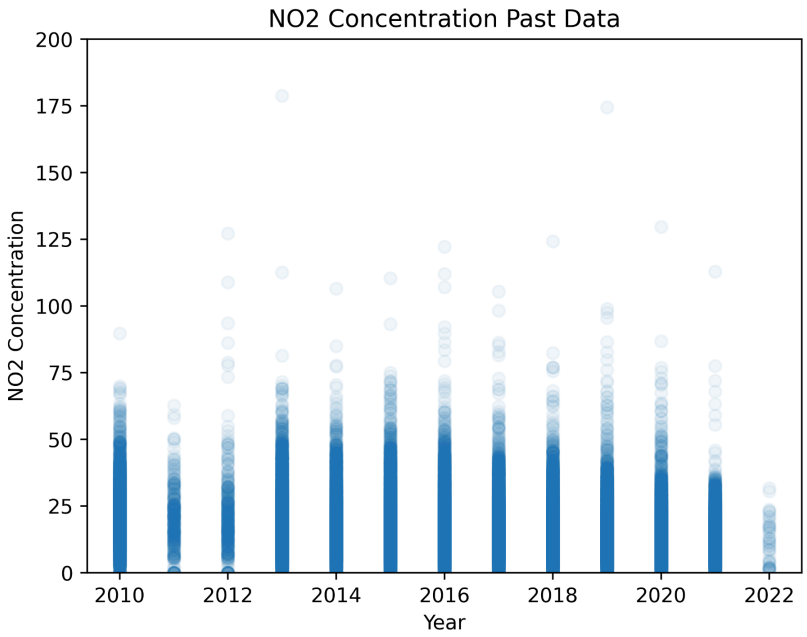


3.5 Question 5

The data question 5 used is a collection of entries with year values between 2010 and 2022 inclusive. We first visualize the number of entries produced in each WHO region to show their relative weights.



Then, if we look at the scatterplot of last year's NO2 Concentration, we can see that there are several outliers in the graph. But since each data point corresponds to the NO2 Concentration of a specific city and is averaged over the entire dataset, we do not exclude them.



4. Project Design, Implementation, and Results

4.1 Question 1

Predict `pm10_concentration` from `pm25_concentration` and `no2_concentration`.

4.1.1 Formulation of Question

Air pollution is a significant global issue; various pollutants contribute to the degradation of air quality and pose serious health risks. Among these pollutants, PM10, PM2.5, and NO_2 are critical due to their harmful effects on respiratory and cardiovascular health. Understanding the relationship between these pollutants can help in developing better strategies for monitoring and controlling air pollution.

By accurately predicting PM10 concentrations, we can fill in missing data, identify areas at high risk of pollution, and provide crucial information for policymakers and health organizations. This will contribute to better air quality management and allows the public to be more informed.

4.1.2 Design

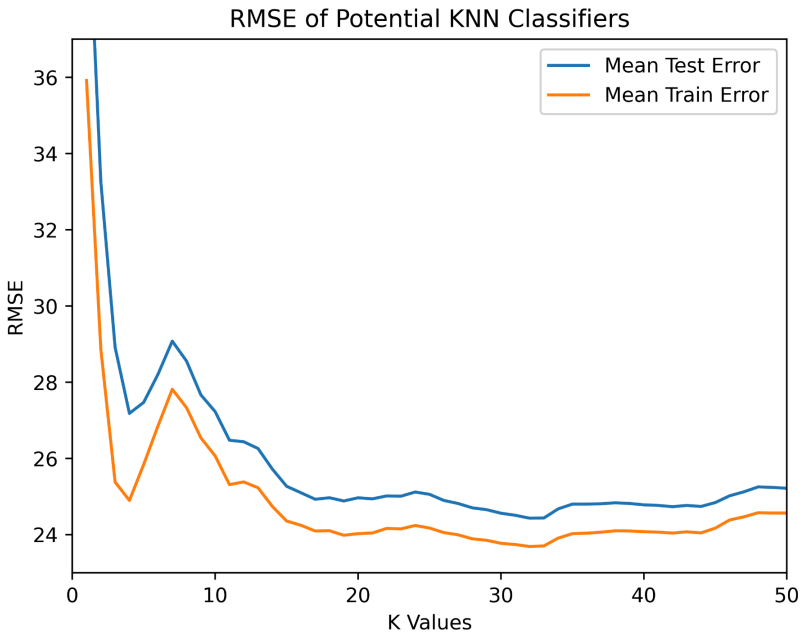
The type of problem we are facing is a regressor problem. So to address this question, we propose to train two predictive models: Linear Regression and K Nearest Neighbors (KNN). For our data, we used data from columns `pm25_concentration` and `no2_concentration` in our main dataset. In each case, we will standardize the dataset before training them on the models. The dataset contains tens of thousands of inputs to be trained so the inputs should provide a decent result.

4.1.3 Implementation

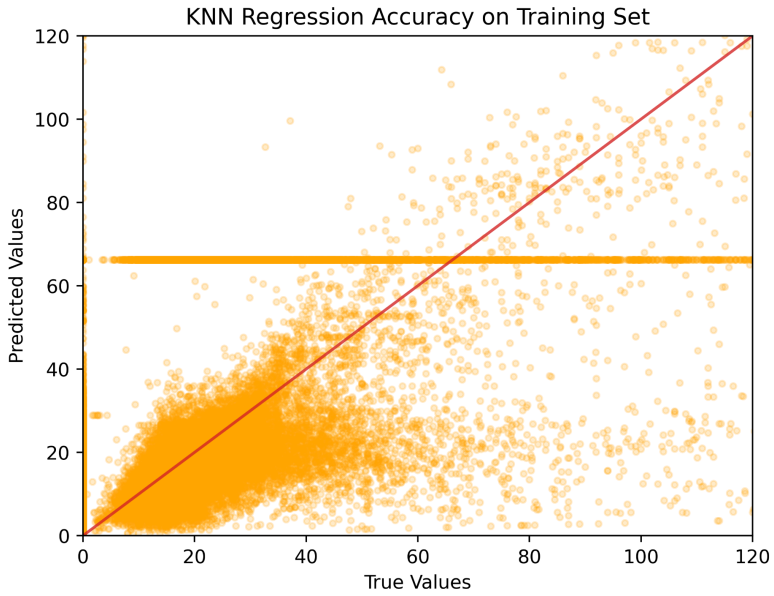
For the implementation, we will first find the best hyperparameters for KNN by utilizing GridSearchCV to find the best n_neighbors and distance metric. After that, we will use cross-validation on both models with 5 folds. After determining the most optimal parameters for K Nearest Neighbors and conducting cross-validations on the models, we will display the RMSE value from KNN with the RMSE from our regression model in a data frame and provide visualizations of the training results.

4.1.4 Results

The best K value for the K Nearest Neighbors regressor is 32.



Linear regression yields an RMSE of 28.34, while K Nearest Neighbors yields an RMSE of 24.84. Reapplying the model on the training set and visually inspecting the results also shows that K Nearest Neighbors seems to be the more accurate model.



4.2 Question 2

What is the best model in predicting `pm25_concentration` from `country`, `year`, `type_of_stations`, and `population`?

4.2.1 Formulation of Question

We wanted to formulate a robust method in predicting the `pm25_concentration` from our data as it can help provide hypothetical data in areas that we overlooked, allowing for a more complete measure of global air pollution. This column represents pm2.5 concentration, which is a collection of fine particles. These particles can be absorbed deep into the lungs and into the bloodstream, causing a variety of issues such as triggering asthma attacks, bronchitis, or leading to cardiovascular disease. By predicting the concentration in areas without collected data, we can take appropriate action and prevent more harm.

In order to account for both numerical and categorical variables and our training and test set, we will utilize a column transformer to prepare our data for our models.

4.2.2 Design

For our training set, we will utilize the source country of each data point, the year of the data point (2010-2022), the population (numerical and discrete), and the type of stations (which is a combination of urban, suburban, rural, residential and commercial area, urban traffic, etc..)

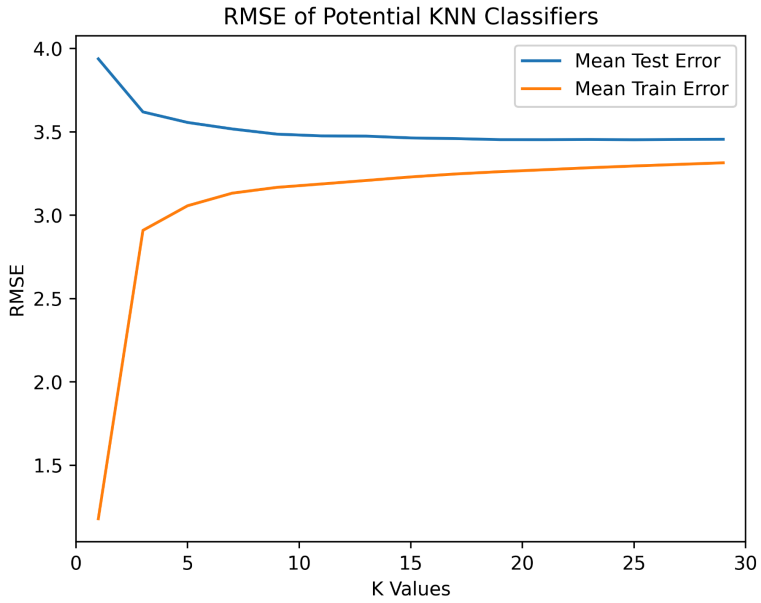
We plan to train two models and compare them to one another. We will first do feature engineering to make `type_of_stations` provide more useful values for predicting. Instead, we will make new columns: `rural`, `suburban`, and `urban`, where each column contains the percentage of the respective station in the current region. We will also account for `Residential and Commercial Area`, `Urban Traffic` / `Residential and Commercial Area`, and `Urban Traffic`. These are not all the possible stations, but for the sake of simplicity and feasibility, we'll keep it limited to these six columns. We will train a K Nearest Neighbors model and a Linear Regression Model.

We will utilize a 5-fold cross validation to find our RMSE. After determining the most optimal parameters for K Nearest Neighbors using `GridSearchCV`, we will compare the RMSE value from our result with the RMSE from our regression model gained through cross validation.

4.2.3 Implementation

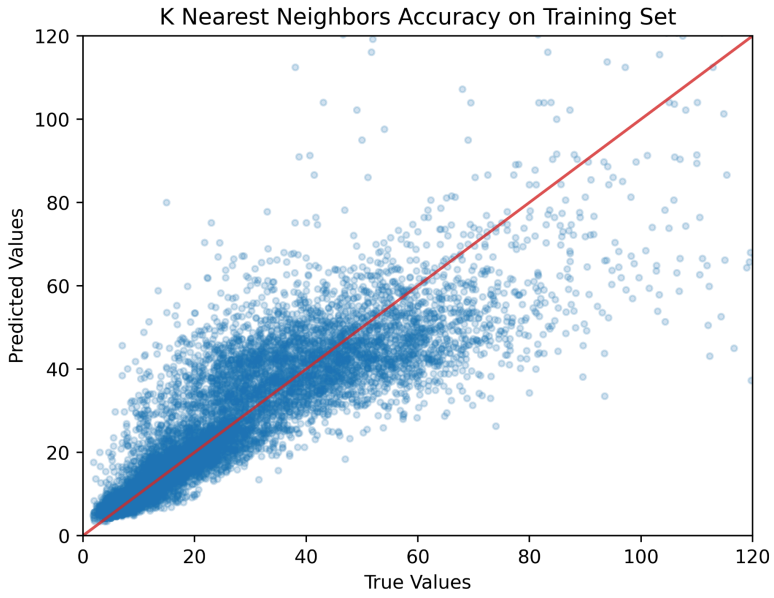
In the process of modifying `type_of_stations` into more viable data, there were a number of challenges and complications, such as that it wasn't logical to account for all the differing stations, so only six columns were made: `rural`, `suburban`, `urban`, `Residential and Commercial Area`, `Urban Traffic / Residential and Commercial Area`, and `Urban Traffic`. In order to do this, we converted each value in the column to a list through splitting the string, creating six columns that contained the proportion of their respective station through `value_counts`.

In order to construct our models, we used column transformers and pipelines to transform, fit, and predict our data. For our linear model, we utilized a 5-fold cross validation and calculated the average RMSE. For our K Nearest Neighbors model, grid search is used to tune our hyperparameters.



4.2.4 Results

Using grid search, we found that the best parameters for K Nearest Neighbors is Manhattan distance with K = 25. After applying these estimators to our data and calculating the RMSE, we found that the RMSE for our K Nearest Neighbors model is approximately 11.78 and the RMSE for our Linear Regression model is approximately 11.94. With the smallest of the RMSE values, K Nearest Neighbors is the best model, although the results are extremely close. In addition, these RMSE error values are fairly small in context and display that both models are valid predictors for `pm25_concentration`.



4.3 Question 3

Classify the locations into the countries they belong to (`country_name`) based on the geographic coordinates (`latitude` and `longitude`).

4.3.1 Formulation of Question

It can be confusing when a random geographic coordinate is given: without a reliable map, it is impossible to make any sense from two decimal numbers. This question aims to solve the problem by constructing a predictor that receives a set of coordinates as input and returns a country name, providing an intuitive sense of where an unknown location is. This question is not related to the air pollutant concentration data; instead, it is a creative exploration attempting to utilize and connect subsidiary columns in the main dataset.

4.3.2 Design

We pick out all rows in the year 2019 from the dataset to proceed the analysis in the question. From the selected data, we continue to drop the rows whose countries have less than 3 entries in total, since the limited number of available neighbors makes it hard, if not impossible, to predict these points correctly.

A single classifier, K Nearest Neighbors, is utilized, since it is intuitively reasonable to classify and predict a location's `country_name`, a geographic categorical value, from the geographic neighbors of the location. Grid search is used to find out the best hyperparameter K for the classifier, with 10 cross validation rounds.

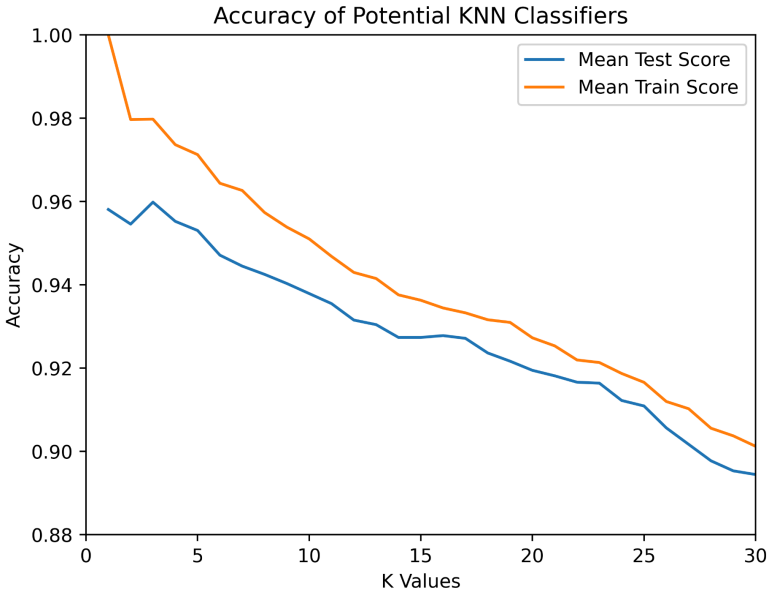
4.3.3 Implementation

In our dataset, the same location might give multiple entries in different years. We pick only one year to conduct our analysis, since otherwise, the nearest neighbors of a location might be another entry of the location itself. 2019 has the second highest number of entries and the most number of unique countries and is thus selected.

Accuracy is used to evaluate the K Nearest Neighbors classifier since there are many potential categories, none of which takes up a majority of the data.

4.3.4 Results

Grid search shows that the best K value for the classifier is K = 3, yielding an accuracy of 0.9598.



The results can be visualized when the model is applied back to the training set. It is clear to see that all the errors occur near the borders of countries, particularly inside Europe, between Canada and the US, and around the Levant. Locations where data are sparser tend to perform better since the borders are better defined and entries belonging to different countries are further apart.

4.4 Question 4

Predict and classify the development status of a country based on `pm10_concentration`, `pm25_concentration`, and `no2_concentration`.

4.4.1 Formulation of Question

There has been a heated international debate about whether developing countries are obliged to follow the same environmental regulation compared to developed countries. The basis of the debate is the assumption that developing countries are not as capable at controlling pollutants, including air pollutants. This question explores whether the difference in pollutant concentration is great enough to successfully classify entries into those from developed countries and developing countries.

4.4.2 Design

We first drop all rows where there are empty values for any air pollutant concentration. We then join the supplementary dataset to the main dataframe by mapping the group 2 into the developed category and the rest into developing. This is stored as a separate column, development. We also standardize the three air pollutant concentration rows to make sure they are on the same scale.

We try different methods, including supervised classifiers and unsupervised K-Means clustering, to find out which yields the best result.

Supervised classifiers used include Logistic Regression, K Nearest Neighbors, and Support Vector Machine. For K Nearest Neighbors and Support Vector Machine, grid search is used to find out the best values for K and C, respectively. All three resulting models are applied to 6 potential combinations of the 3 different air pollutant metrics to determine the best classifier.

K-Means clustering with 2 clusters is used to divide the data into two groups based on all three air quality indicators without supervision. The resulting labels are analyzed to see if they correspond well with the 2 categories of developed and developing countries.

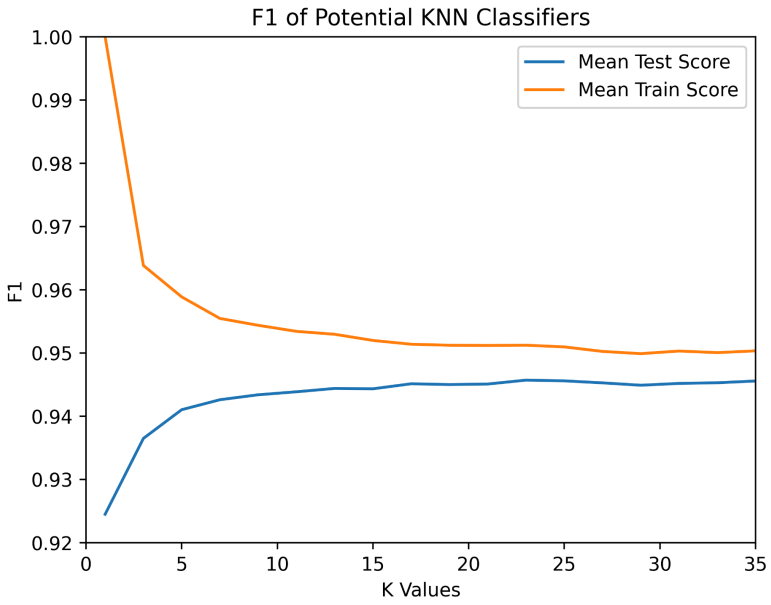
4.4.3 Implementation

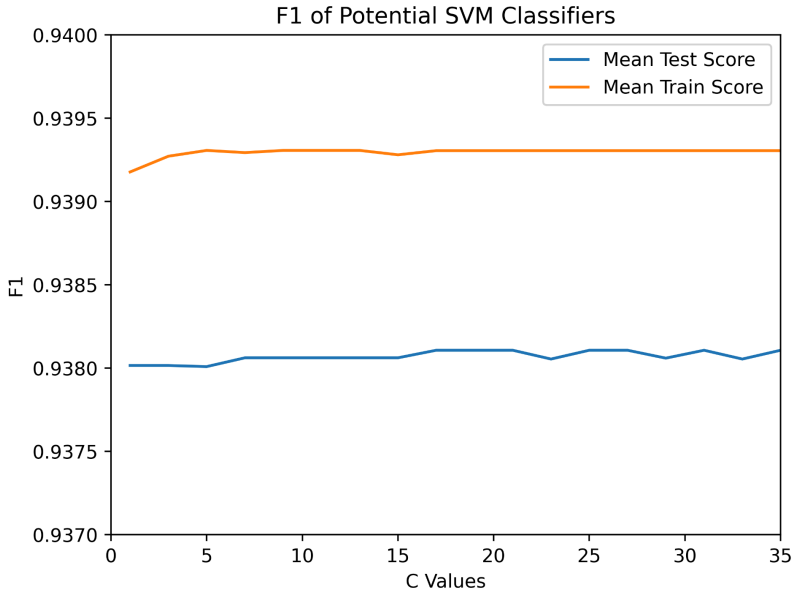
Dropping the rows with empty data removes most entries from developing countries, making it possible that a seemingly accurate model significantly underpredicts developing countries.

The scoring method used to evaluate the supervised classifiers is F1 since more data entries belong to the developed countries, so recall should be considered along with precision. We also use 5 folds for grid search since it takes too long to run an evaluation with 10 grids.

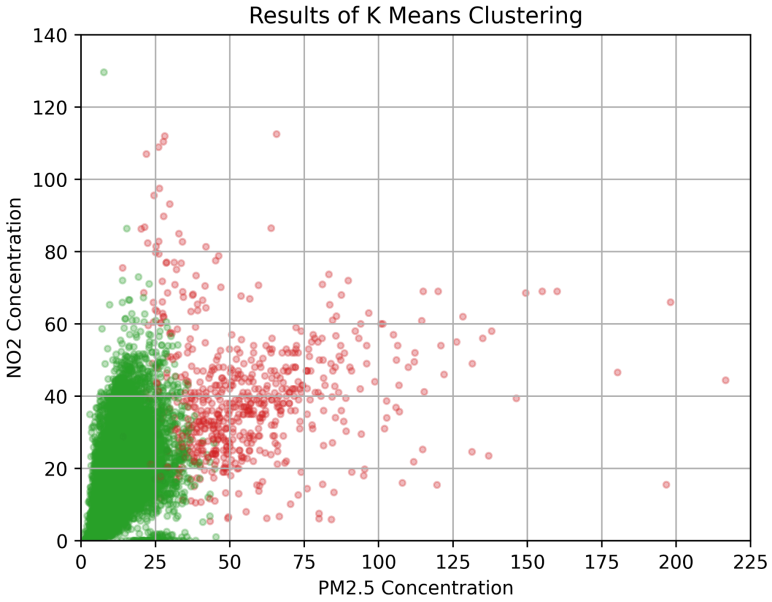
4.4.4 Results

The best K Neighbors Regressor has K = 23, while the best Support Vector Machine has C = 17. Out of all the supervised classifiers, the best one is K Neighbors Classifier on all three potential air pollutant indicators, reaching an F1 score of 0.9457; the second best is Logistic Regression on PM10 only, reaching an F1 score of 0.9401.





The K Means model has a relatively high cluster purity score of 0.8700, indicating good correlation with the development status of a country. However, it also yields a terrible V Measure of 0.2830, suggesting that it does not act as a good classifier.



4.5 Question 5

Predict future `no2_concentration` by modeling the trend from 2010 to 2022.

4.5.1 Formulation of Question

NO_2 is a significant air pollutant with adverse effects on human health and the environment. High concentrations can cause respiratory problems and contribute to the formation of other harmful pollutants such as ozone and particulate matter. We want to know how `no2_concentration` will change in the next decade and see if the trend is good or not.

4.5.2 Design

The 14 entries from non-member states, although classified as a region, clearly do not have as much impact as a real continent or geocultural area. These entries are first dropped.

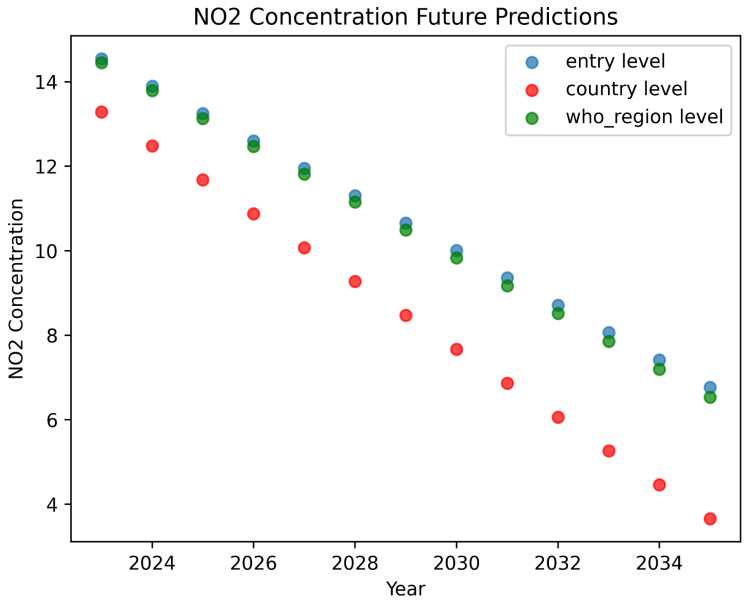
The question is focused on three levels, the first level is that each entry in the dataset contributes equally to the prediction; the second level is that entries from country that appear a lot in the dataset tend to be more important and given higher weight; the third level is that entries from who_region that appear a lot in the dataset tend to be more important and given higher weight.

4.5.3 Implementation

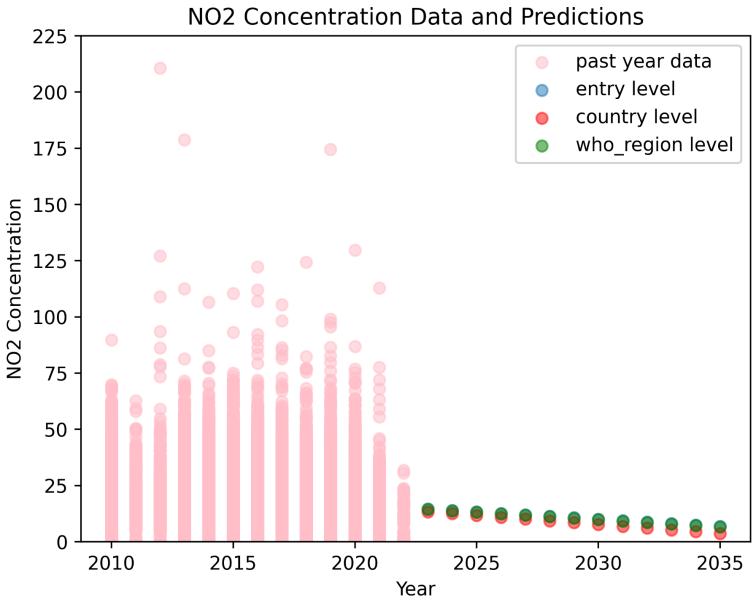
When getting the weights of each country or who_region, we use their normalized value_counts() in the whole dataset as their assigned weights; and then add the two weights as new columns in the dataframe. After that, when we fit into the linear regression model, we use the sample_weight parameter of the linear regression model to fit into our calculated weights in the process of model training.

4.5.4 Results

The final results do not show a big difference in the predictions when we predict at different levels, and all three predictions show a common decreasing trend of the `no2_concentration`.



However, when using weights calculated from the country level, the prediction is slightly smaller than the prediction of the other two levels. When using weights calculated from the country level, the prediction is significantly smaller than the predictions of the other two levels, and entry level and who_region level almost collide on the graph drawn.



5. Discussion and Conclusions

Our goal for this project was to help develop models and methods to predict and evaluate trends involving global air pollutants. From our results, we can state that we have successfully developed multiple models of valuable use in helping analyze the data along with making accurate predictions. We helped predict missing gaps in data for PM10 concentrations, analyze the trend for NO_2 concentrations, analyze the relationships between the three pollutants, find countries based on coordinates, and analyze the trend between the development status of a country and their air quality. These improvements have matched our goal in creating a more concise interpretation of the earlier data, helping visualize details regarding air pollution and its patterns: patterns such as lowering no2 concentrations and pm2.5 particles in our air for the previous years. Information is crucial in combating any issue, and with a sophisticated understanding and control of data, we can use it to understand the extent of the issue and develop methods to solve it. Data has grown to be an instrumental tool in the modern world; through our report and many others, we can gather more information about air pollution, its impact, along with other specifics. Through this, we can begin the first step of many towards building a cleaner, safer world.