Predicting Air Pollution

Group 18

Name	Last name	Student Number
Błażej	Apanasik	2766691
Kacper	Chmielewski	2761683
Maciej	Wierzbicki	2743170
Mateusz	Wołodźko	2775842

Vrije Universiteit Amsterdam

30 June 2024

A report submitted for the course Project Big Data

Contents

1 Introduction

Air pollution presents a significant challenge to sustainable environmental conditions, especially as we progress into 21st century. It severely impacts health and living standards, particularly in densely populated urban areas. We focus on identifying the main natural and artificial contributors to increasing air pollution levels to gain more understanding of this issue. Our research will highlight a range of factors influencing air pollution, from industrial emissions and vehicle pollution to natural events such as forest fires and dust storms.

1.1 Problem Statement

The main purpose of this research is to delve into and identify the specific characteristics and elements that contribute significantly to the issue of air pollution. Upon identifying these factors, the information gathered will then be utilised to formulate and develop a reliable, scientifically-sound method of predicting future trends and patterns in air pollution. This prediction method will serve as a crucial tool for preemptive measures against environmental degradation.

In addition to this, the research pursues a secondary goal. This involves conducting an in-depth investigation into the effects, both direct and indirect, of the COVID-19 pandemic on the levels of air pollution within the United States. This aspect of the study is particularly important given the unprecedented global circumstances that the pandemic has brought forth and the potential implications it may have had on air pollution patterns.

2 Data Preparation

2.1 Data Overview

The data was collected from various sources, including the Air Quality Open Data Platform [1] and the US Environmental Protection Agency [2]. It covered a two-year period with daily measurements from 54 cities across USA. It encompasses several features related to anthropogenic and meteorological conditions, such as wind properties, temperature, humidity, travel miles, and 6 different pollution levels. For each feature there were three different explanatory values present: minimal level, maximal level and median from a given day. In our analysis we focused on median levels as we believe them to be most accurate representation of reality.

Pollutant	**Valid Samples**	**Number of unique cities**
PM 2.5	35134	54
PM 10	16965	29
O3	33950	54
NO2	14676	41
CO	23558	42
SO2	24538	39

Table 1: Pollutant Levels with Corresponding Values and Cities

2.2 Data Cleaning

2.2.1 Handling Missing Values

Missing values can introduce bias, particularly if the data gaps are correlated with factors like economic development. This can skew analysis results and lead to an under representation of pollution levels in less monitored areas. Integrating datasets from various sources requires sophisticated techniques to handle these gaps without significant bias.

In our dataset, the missing values scattered across pollution data can be attributed to different data sources having different scopes for data collection. For instance, some cities did not have sensors to measure specific pollutants.

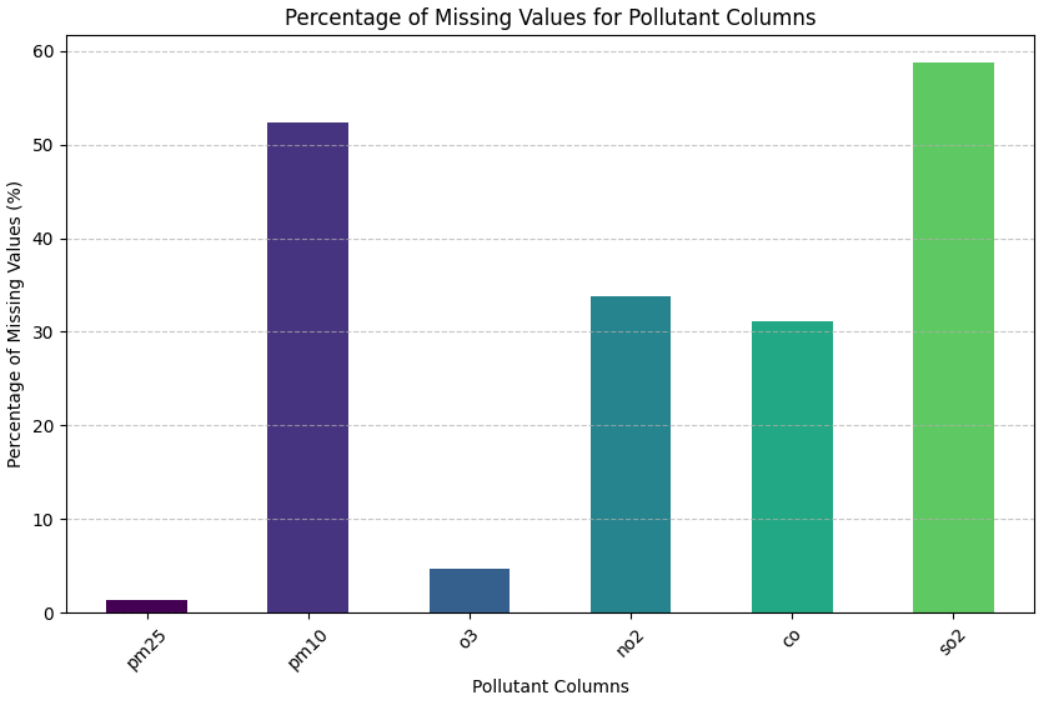

Figure 1: Missing values per pollutant

Initially, we considered dropping all rows with invalid input, but we quickly realized this approach would eliminate almost 90% of our dataset. Instead, we chose to remove NaN values only for the specific pollutant we were analyzing at any given time. This method allowed us to retain as much data as possible while still addressing the issue of missing values effectively.

2.2.2 Handling Outliers

For detecting outliers we have chosen IQR method as it is reliable and easy to implement. It firstly calculates Interquantile Range $IQR = Q3 - Q1$ and than classifies entries for which $value > Q3 + 1.5 * IQR \vee value < Q1 - 1.5 * IQR$ as outliers. In order to preserve as much data as possible we used similar approach as we did with missing values. Hence, we only removed outliers from the targeted column.

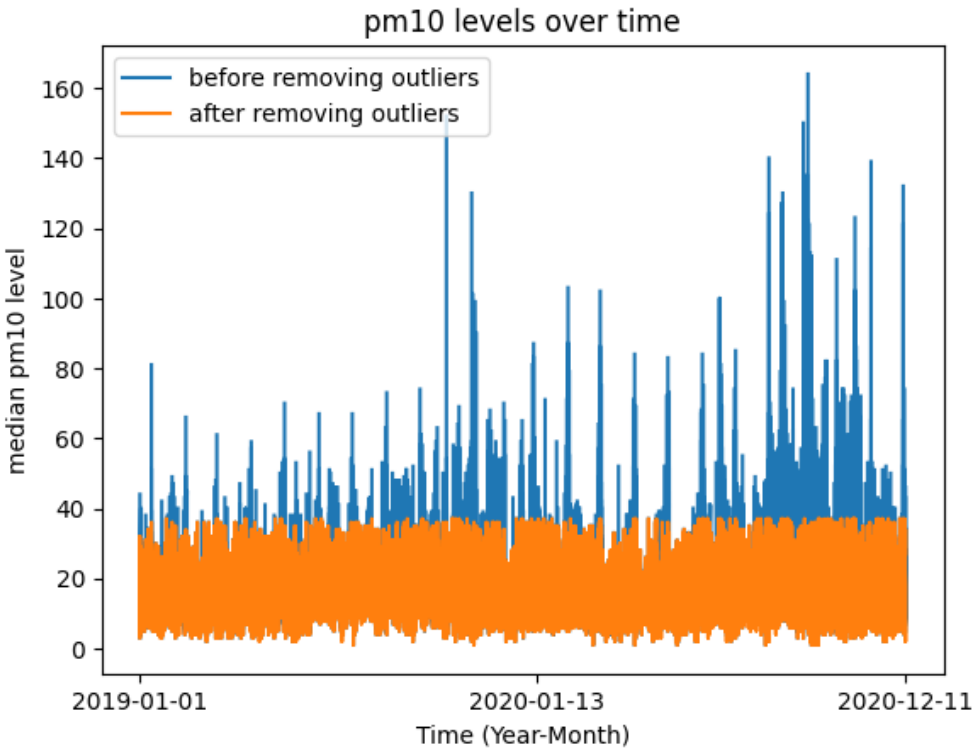

Figure 2: Visualization of data cleaning for pm10

2.2.3 Date format and Pre-/Post- COVID split

To split our dataset into pre- and post- COVID parts, we first converted the "Date" to 'pd.datetime' format to facilitate further analysis. Than we chose January 13^{th} 2020 as our cut-off point since it is the date when COVID 19 pandemic had begun in USA. To keep the comparison as precise as possible we decided to make those new datasets symmetric based on dates column. We did that by locating outliers and removing them from both datasets separately afterwards we removed all rows from 2019 where the day, month, and city were not present in the 2020 dataset, and vice versa.

3 Exploratory Data Analysis (EDA)

3.1 Descriptive Statistics

We computed descriptive statistics to understand our dataset's fundamental characteristics. This included calculating the mean, median, quantiles, min, max, and std of all features.

The dataset consists of 35,596 observations about population behavior and environmental factors. Findings show a lower average population staying at home, substantial variability in mobility, with geography spanning a vast range of latitudes and longitudes.

The data reveals significant variation in air quality measurements for ozone and particulate matter (PM10). Wind statistics show large differences in gusts, with the PP feature indicating diverse conditions. Overall, the dataset shows significant variability in population movement, environmental conditions, and air quality.

3.2 Pollutants Distribution

Our six pollutants varied in scale and range, hence some trends are hardly visible, such as SO2 which is nearly a flat line (see Figure 3). In terms of pollutant levels, PM2.5 consistently shows the highest levels, ranging mostly between 25-40 parts per billion (ppb). PM10 and O3 form a middle tier, typically between 15-25 ppb. NO2 and CO occupy lower range, generally below 15 ppb, while SO2 cannot be reliably ranked due to scaling issues in the provided graph. Furthermore, the variability of pollutants is diverse. O3 shows the most pronounced fluctuations, which may be caused by higher than usual temperatures in certain parts of the United States. PM2.5 and PM10 display moderate variability. Some pollutants, such as O3, PM2.5 and NO2 suggest potential seasonal cycles. PM2.5 and PM10 often move in tandem, while other pollutants show less clear relationships. The distribution pattern in Figure 3 suggests a complex air quality picture with varying levels of data reliability across pollutants, emphasizing the need for cautious interpretation, especially regarding the SO2.

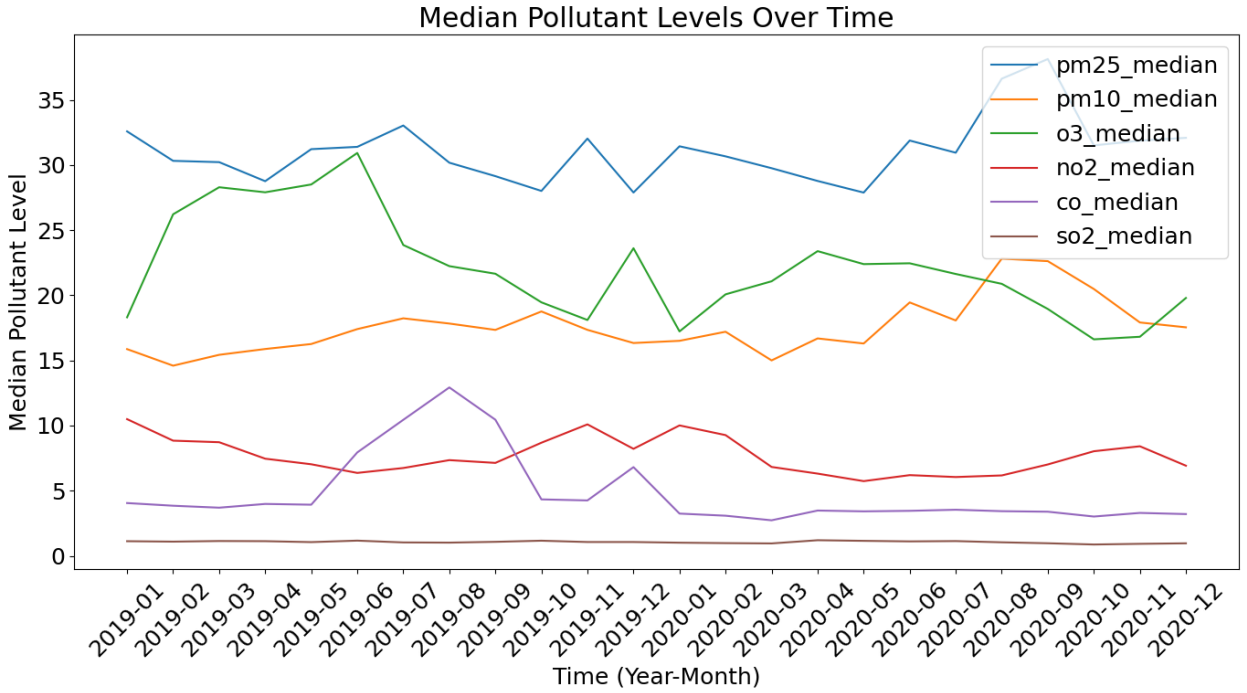

Figure 3: Median pollutant levels over time

Figure 4 shows the median NO2 levels over period of two years, revealing general decrease with notable fluctuations. A seasonal pattern with peaks and downs suggests varying NO2 levels. Peaks occurred in the November and December 2019 with levels around 10 ppb, while lows were observed mid-2019 and 2020, dropping to about 6 ppb after a mid-2020 low occurred. Levels rose gradually by year-end, then dropped sharply. NO2 levels varied between 6 and 10.5 ppb over the period. A significant dip in 2020 is most likely linked to COVID-19 lockdowns' reduced emissions, and the 2019 peak might be due to weather conditions or increased heating usage during winter. The NO2 pollutant displays useful seasonal characteristics which might suggest easier and more informative model predictions.

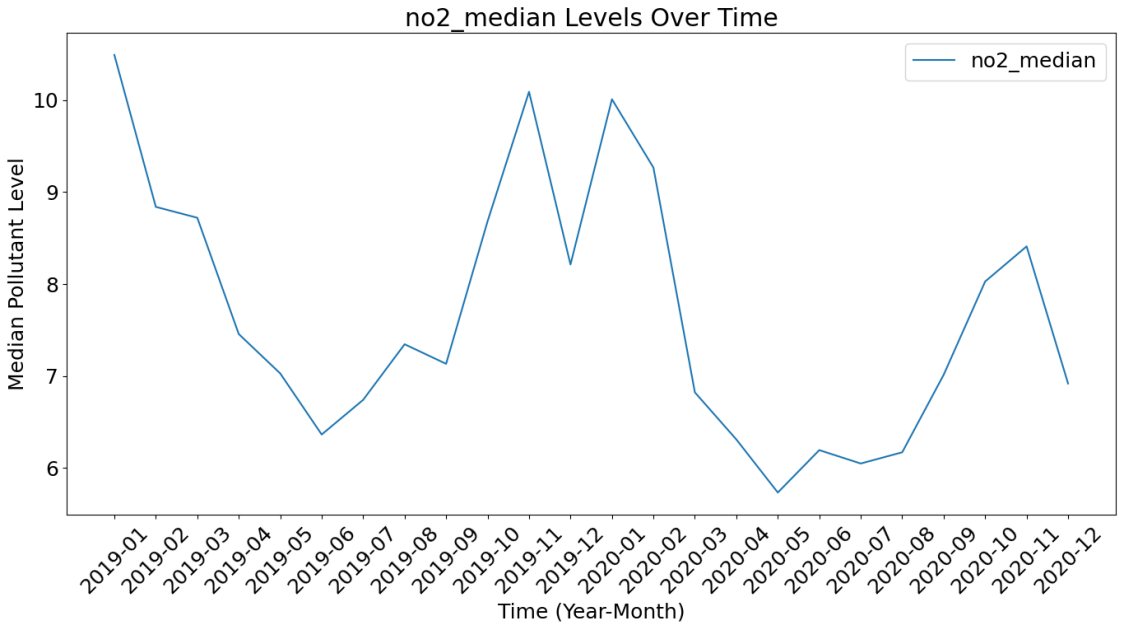

Figure 4: NO2 levels over time

Figure 5 shows significant fluctuations in SO2 levels over the same period of time. A peak was observed in December 2019, followed by a drastic surge around June 2020, and a slight increase towards late 2020. Seasonal variations and reduced emissions can be explained by COVID-19 lockdowns.
These insights into pollution trends highlight the influence of seasonal factors, policy changes, and major events on air quality.

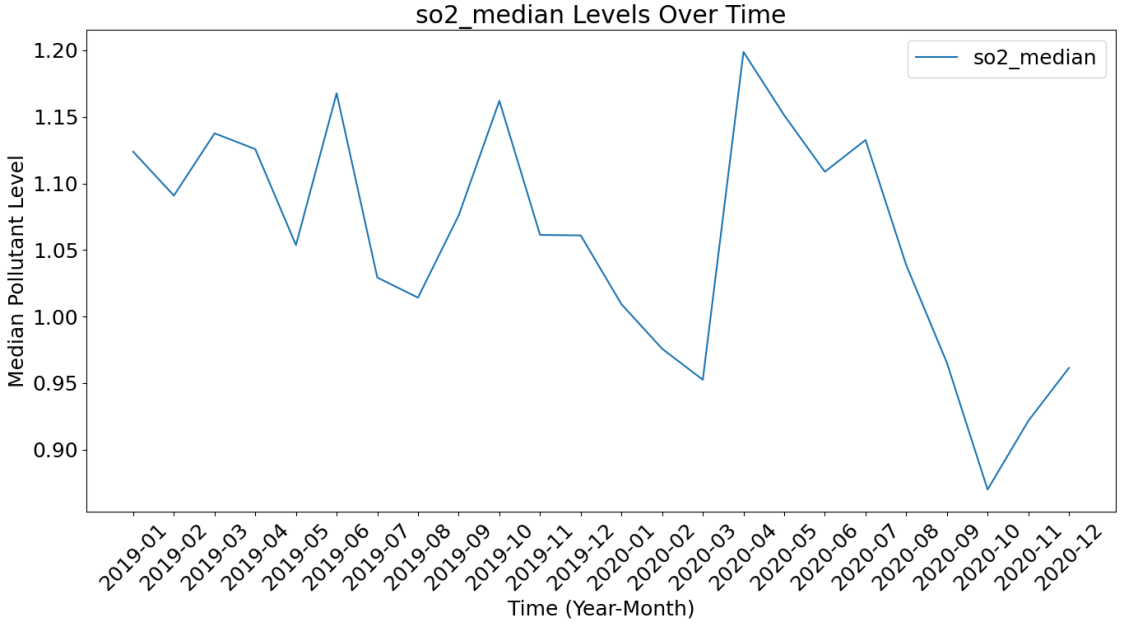

Figure 5: SO2 levels over time

3.3 Correlation Analysis

Correlation analysis was conducted to identify relationships between different pollutants and meteorological factors. We later use those finding while choosing features to be used in regression models.

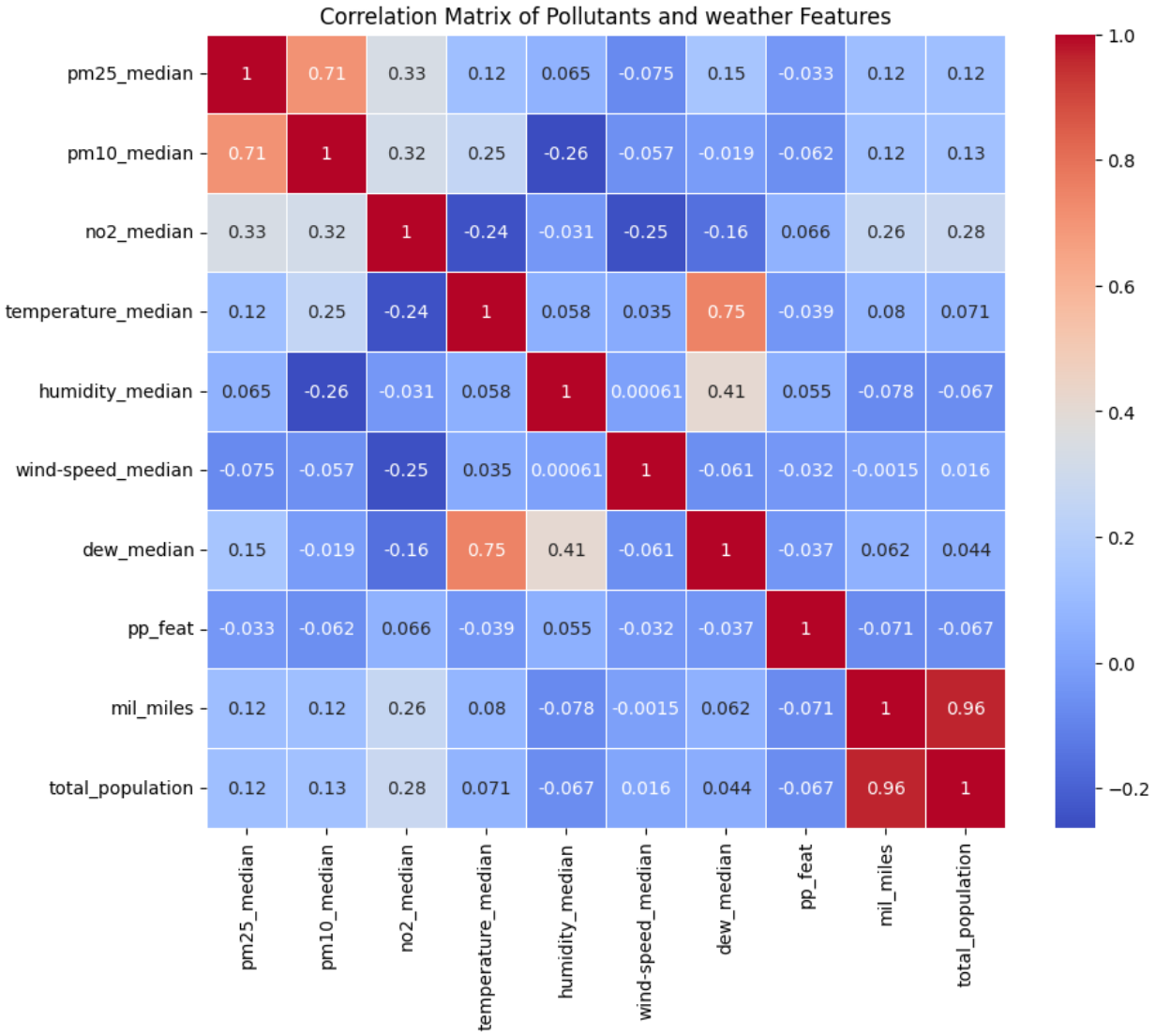

Figure 6: Matrix of strongest correlations

Figure 6 presents matrix of strongest correlations meaning it only contains features for which at least one correlation $corr_{feature}$ had property: $|corr_{feature}| > 0.2$. An exception to this was the pp_feat, which we included as this variable was added by the authors of paper *Deciphering Environmental Air Pollution with Large Scale City Data* [1]. It was the only feature expressing impact of power plants for given entry. In spite of our high hopes, its correlations were very low (see Figure 6). As for other features we can see some promising outcomes for example between NO2 and mil_miles which is average distance covered in car per citizen from given city. Some correlations were more obvious yet equally interesting such as strong correlation between PM10 and PM2.5 which both come from the same sources. NO2 also shares some sources with PM pollutants which is reflected on the matrix.

3.4 COVID Influence Analysis

The COVID-19 pandemic has had varying impact on different types of pollutants. While some pollutants like NO2 and O3 have seen a reduction in emissions, others like PM2.5 and PM10 have shown increases, possibly due to some environmental regulations, transportation patterns or industrial stagnation during the pandemic.

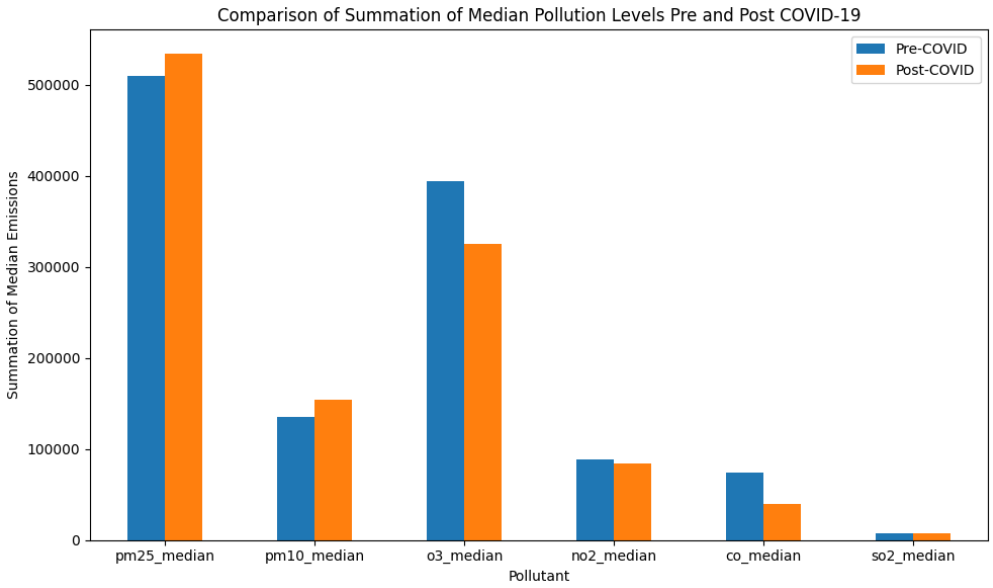

Figure 7: Comparison of sums of pollution levels across USA

3.4.1 Statistical Testing

To check whether pre-COVID dataset was stochastically smaller than post-COVID dataset we have conducted Two Sample Kolmogorov-Smirnov Test with "less" alternative hypothesis. The results of this test can be found in table 2.

	PM2.5	PM10	O3	NO2	CO	SO2
p-value (with outliers)	0.947	0.001	≈ 0	≈ 0	≈ 0	≈ 0
p-value (no outliers)	0.032	0.985	≈ 0	≈ 0	≈ 0	≈ 0

Table 2: KS test results for original and cleaned data

In the result we can observe that distribution of most pollution levels was indeed stochastically smaller before COVID pandemic have started. The interesting behavior can be observed for PM10 and PM2.5 p-values. For those pollutants removing outliers caused changes to the test outcome. This is possible since KS test is very sensitive to outliers and could mean that there are some non-trivial trends present in those distributions.

3.5 Geolocation Analysis

We plotted our data over the map of USA and found the following: Apart from PM10 (see Figure 8) and PM2.5 (see Appendix, Figure 22), for which the maps we made suggest that they tend to occur in the western part of the country, most pollutants do not exhibit obvious trends or patterns (see Appendix, Figures 23-26). We identified some local extremities or case-specific high emissions that require further investigation to reach definitive conclusions.

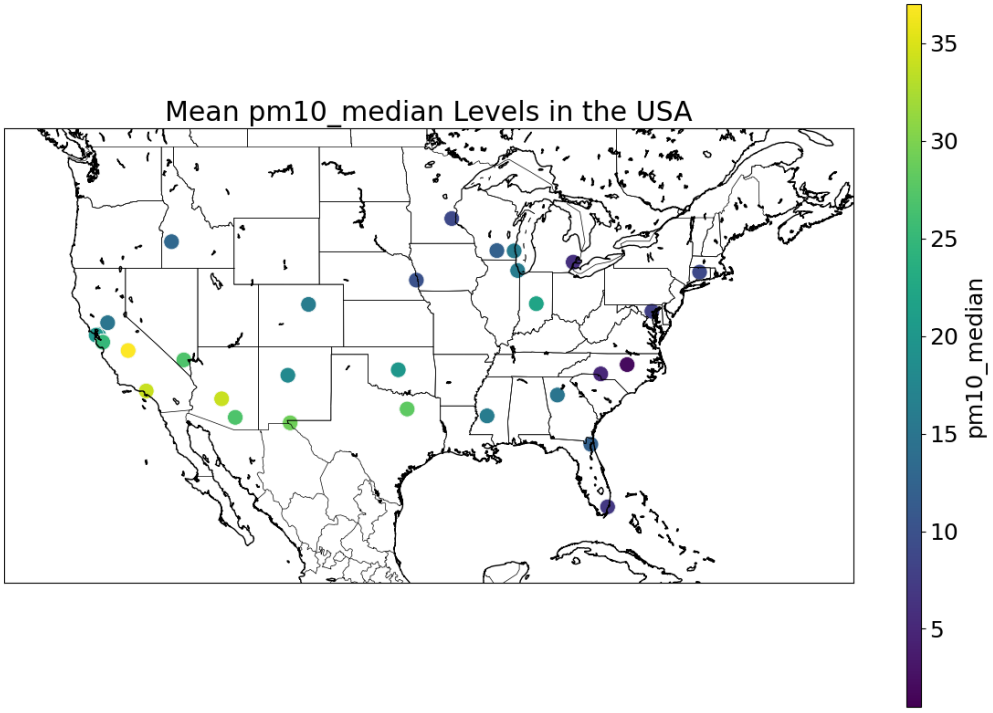

Figure 8: Map of PM10 emissions across USA

4 Methodology - Predictive Models

4.1 Modeling Approach

The goal of our prediction analysis was to create models that could accurately predict pollution levels, measured in the specific cities. As mentioned above we had six pollutants in our dataset from which we decided to use only five as targets for our models. The pollutant we dropped after first iteration of our models was SO2 since we have basis to believe that data was improperly collected resulting in unpredictable behaviour while used in regression models as it can be seen on Figure 9. Furthermore, we opted to utilize the method of cross-validation in our study. This approach is not only easy to understand, but also straightforward to implement, making it an effective and accessible tool for our research. One of the key benefits of cross-validation is its reduced bias compared to other models, leading to more reliable and robust outcomes. During our training phase, we employed a data division ratio of 80/20. This implies that 80% of the data was used as a training set to develop and fine-tune our model, while the remaining 20% was reserved as a test set. For cross validation we used five-set validation. This involved splitting our training set into five subsets. Each of these subsets was used once as validation set and remaining four for training. This helped to ensure none of our models could generalize well to new data.

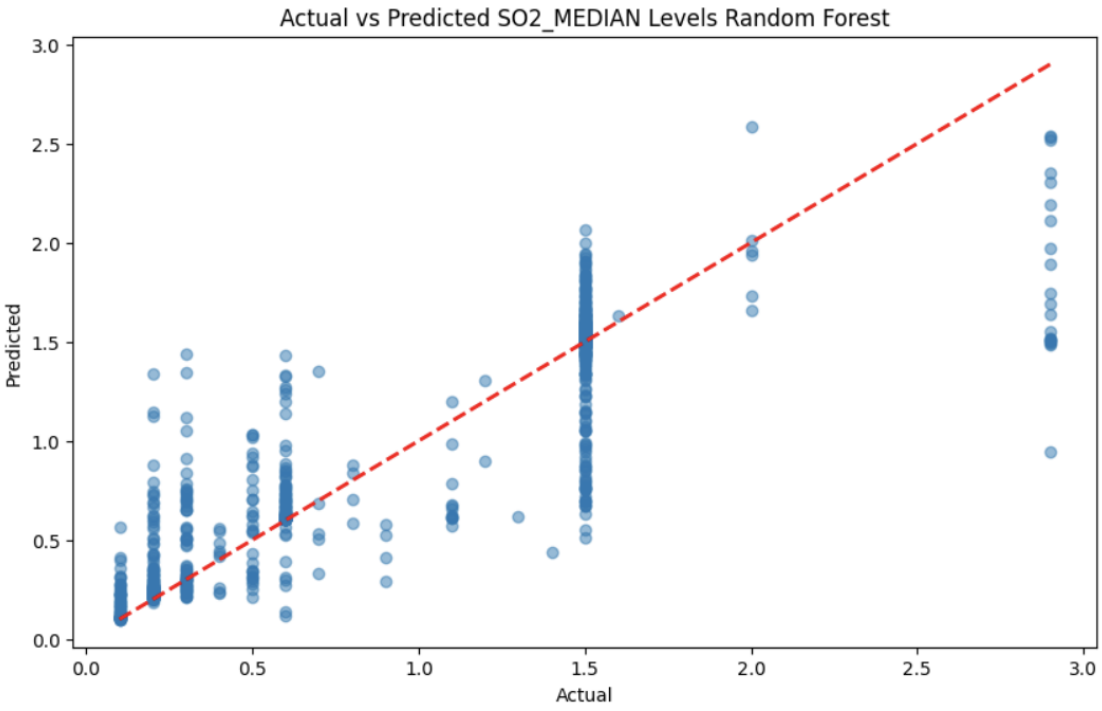

Figure 9: Actual vs Predicted SO2 Median Levels

4.2 Random Forest Regressor

The Random Forest Regressor was selected as our initial choice based on our team's familiarity with its application. This method constructs multiple decision trees trained on random subsets of data, each providing individual predictions. The resulting predictions are then averaged to produce the final prediction, leveraging the collective insights generated. By averaging predictions from multiple trees the model tends to generalize well which improves its accuracy with new data. Additionally, Random Forest technique is relatively straightforward to implement and tune making it an effective benchmark model.

4.3 XGBoost

XGBoost or Extreme Gradient Boosting is a powerful model known for its high performance and efficiency with regression tasks. This model is based on gradient boosting framework, which sequentially builds new models by correcting errors of the previous ones. This approach helps to reduce bias and variance resulting in more accurate and robust predictions.

4.4 Support Vector Regression

The Support Vector Regression is a type of machine learning algorithm based on the concept of Support Vector Machine. The goal of the SVR is to predict a continuous output in the regression tasks. The principle of SVR is to find a function that has at most epsilon deviation from the actually obtained targets for all the training data. It tries to fit the error within a certain threshold.

5 Methodology - Deep Learning

5.1 Data Preparation

5.1.1 Initial Preparation

Several assumptions and decisions were made to streamline the analysis, such as treating certain data as negligible or using specific time zones. Data preparation involved organizing the data into a suitable format for analysis and modeling. The dataset includes various parameters: geographic, temporal, demographic, and pollution metrics. Temporal features such as day of the week, month, and year were extracted to enrich the dataset, enabling models to recognize seasonal and weekly patterns in pollution levels. Irrelevant and redundant columns, which do not contribute to the prediction of pollution levels, were identified and removed. This step prepared the dataset for effective modeling.

5.1.2 Handling Missing Values and Outliers

Our approach addressed missing values by replacing them with column means, ensuring that models are trained on complete data without discarding valuable information. Outlier detection and capping were implemented, using interquartile range(IQR) measures to minimize the negative impact of extreme values on the performance of our model.

5.2 Feature Selection and Standardization

Recursive Feature Elimination (RFE) with RandomForestClassifier was utilized for identifying the most predictive features, optimizing the model's focus on relevant variables.

5.3 Deep Learning Model Structure and Evaluation

The Sequential model architecture from TensorFlow's Keras package was created, comprising dense layers with various activation functions to detect complex data relationships. Dropout layers were introduced to prevent overfitting. Data standardization was accomplished through RobustScaler, mitigating the influence of outliers and improving overall model's accuracy The model was optimized using the Adam optimizer, targeting Mean Squared Error (MSE) as the loss function suitable for regression tasks, punishing large differences between predicted and actual values Interactive training function was implemented, letting us evaluate models performance in real time. Based on interim results we were able to decide whether further training on prospective parameters was sensible. Automatic early stopping mechanisms were also put in place, judging the model based on validation loss and stopping training when no progress was being made.

6 Results

6.1 Results Overview

In the course of our research, we made a conscious decision to utilize PM10 data to display our results. We found that this particular data for the pollutant yielded more representative errors. The outcomes when compared to the data collected by the researchers presented more accurate data, which made it an optimal choice for our purposes.

6.2 Random Forest Regressor (RFR)

The Random Forest Regressor achieved a mean RMSE of 5.5195 across five cross-validation sets with mean MAPE of 34.1% . The plot below presents a strong linear relationship between the actual and predicted values for PM10 levels. However, there is a significant spread, which is to be expected due to inherent noise in our data. The residuals show nearly normal distribution that is slightly skewed to the right indicating little to no bias.

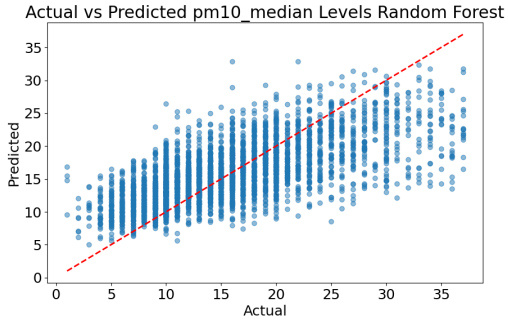

Figure 10: Actual vs Predicted PM10 Median Levels

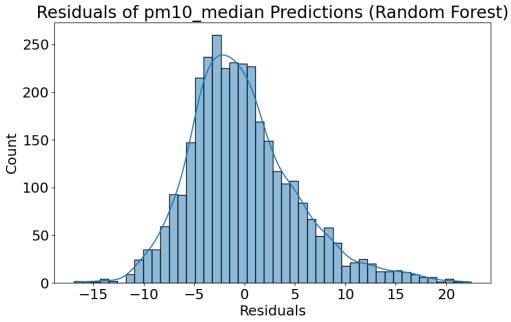

Figure 11: Residuals of PM10 Median Predictions

6.3 XGBoost (XGB)

The XGBoost model achieved a mean RMSE of 5.4799 across five cross validation sets. The MAPE for this model was 33.44%. The actual vs predicted plot (Figure 12) reveals firm relationship for PM10 levels, though there is a noticeable spread reflecting some inherent prediction error. The residuals plot (Figure 13) displays a near-normal distribution centered around zero, indicating that the model's predictions are generally unbiased, with slight skewness to the right.

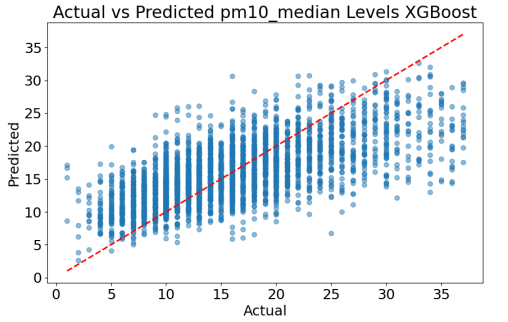

Figure 12: Actual vs Predicted PM10 Median Levels

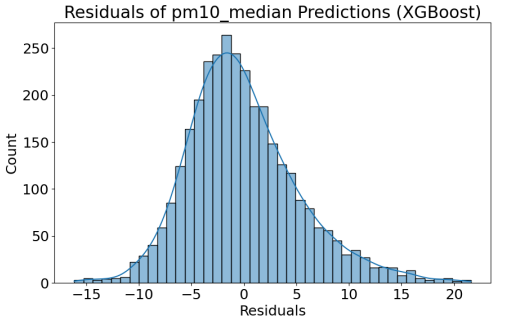

Figure 13: Residuals of PM10 Median Predictions

6.4 Support Vector Regression (SVR)

The SVR model scored a mean RMSE of 5.9188. The MAPE for this model was 34.09%. Figure 14 shows the actual vs predicted plot and it suggests the model captures overall data trend. However, the predicted values show considerable spread around the red dashed line, indicating some prediction error. The model tends to underestimate higher pm10_median levels and overestimate lower levels. The residuals in Figure 15 are centered around zero, indicates no systemic bias. It appears to roughly follow a normal distribution with slight skewness to the right. Hence, suggesting that the model's errors are randomly distributed and the model is capturing most of the systemic patterns in the data.

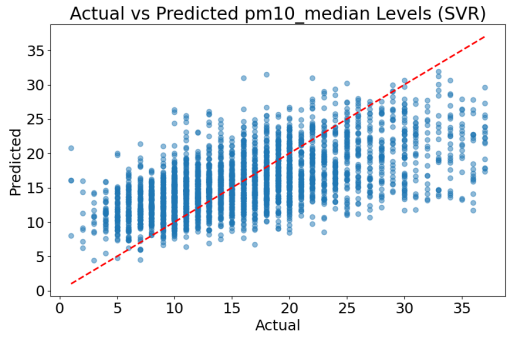

Figure 14: Actual vs Predicted PM10 Median Levels

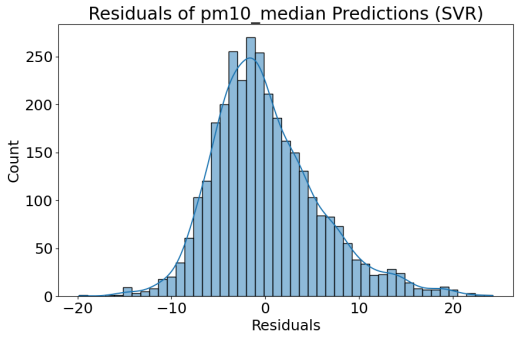

Figure 15: Residuals of PM10 Median Predictions

6.5 Deep Learning

6.5.1 Limitations of Deep Learning

Deep Learning approach had high potential of outperforming our simple models, however we have came across a few obstacles. The biggest one being the dataset size. Depending on the target variable, our dataset contained up to 35 thousand entries. It is considered medium sized dataset, while for deep learning desired are large datasets - above million entries. Another issue arose from high computational power required - with a large number of small correlations between our variables, for our neural network to detect the complex relationship between them it was required to use multiple layers of thousands of neurons. It soon became infeasible for our use.

6.5.2 Deep Learning Predictions

Despite all that we were still able to make some interesting observations. We were able to predict CO Median values with an MSE of 0.959 despite all previous limitations.

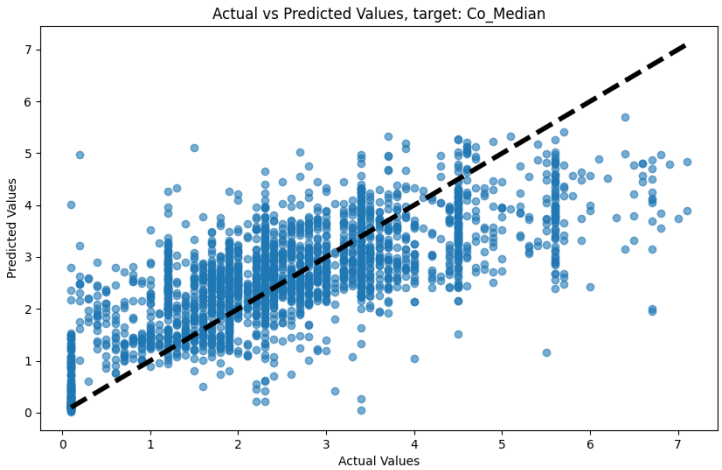

Figure 16: Actual vs Predicted CO Median Levels Using Deep Learning

On the other hand, predicting PM 2.5 - which had the largest dataset - was not that successful, achieving roughly 127 test MSE - unsatisfactory result considering that variable range was between 0 and 50. What's noteworthy is that it was the highest error margin amongst all pollutants despite the largest training set.

It leads us to believe that insufficient correlations and neural network not complex enough is the main cause of our error. Another finding of ours was, that including information about other pollutants into our predictions, rewarded us with better accuracy - most likely because of moderate correlation between pollutants, higher than with environmental features. For PM2.5 it still is not accurate enough for precise predictions, but gives an insight of what can be achieved with a bigger dataset - MSE 82.299

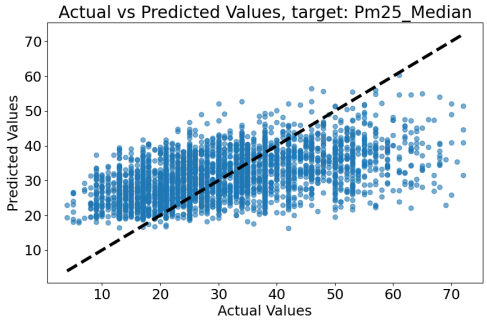

Figure 17: Actual vs Predicted PM2.5 Median Levels: other pollutants not included

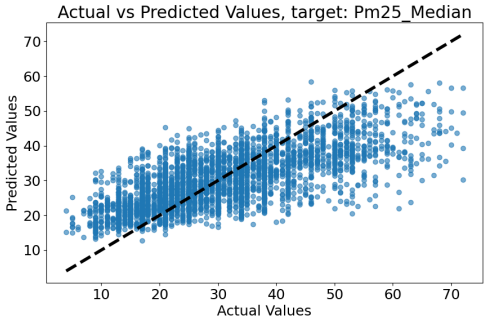

Figure 18: Residuals of PM2.5 Median Predictions: other pollutants included

Another example would be PM10 - highly correlated with PM2.5, making its prediction while including PM2.5 data quite accurate. Here we can see predictions with other pollutants and without them (MSE = 17.998 vs MSE = 38.5948) - error was basically cut in half.

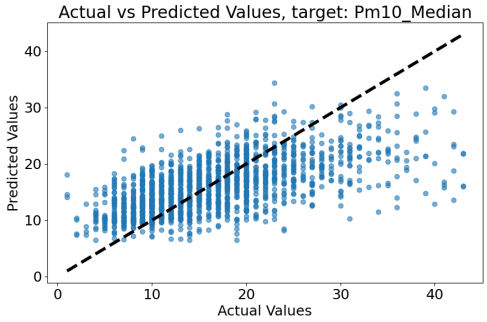

Figure 19: Actual vs Predicted PM10 Median Levels: other pollutants not included

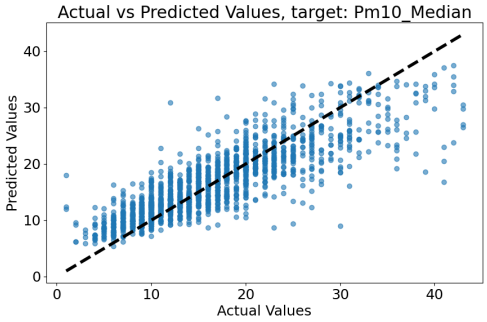

Figure 20: Residuals of PM10 Median Predictions: other pollutants included

7 Conclusion & Discussion

7.1 Summary of Findings

The Random Forest Regressor showed a strong linear relationship between actual and predicted PM10 levels, despite significant spread due to inherent data noise. XGBoost performed slightly better, displaying a firmer relationship between actual and predicted values. The Support Vector Regression (SVR) model captured the overall trend but exhibited considerable spread, resulting in higher root mean square error values compared to the other models. The residuals plot for SVR suggests no systematic bias.
Despite the dataset size limitations, deep learning showed potential. Including data from other pollutants significantly improved prediction accuracy, highlighting moderate correlations between pollutants. For PM10, the Mean Squared Error (MSE) improved from 38.5948 to 17.998, and for PM2.5, the MSE improved from 127 to 82.299.
The medium size of this dataset likely limited our deep learning model's predictive capabilities, necessitating larger datasets for optimal performance. High computational demands were an issue, not only with our

deep learning model but also with the Random Forest Regressor. This issue was mitigated by utilizing the XGBoost model, which displayed much better performance. Despite computational power constraints, our models have outperformed those presented in the paper on *Deciphering Environmental Air Pollution with Large-Scale City Data* [1]. We have achieved lower Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) for nearly every pollutant.

Method	**RMSE**					**MAPE (%)**				
	PM2.5	**PM10**	**NO_2**	**O_3**	**CO**	**PM2.5**	**PM10**	**NO_2**	**O_3**	**CO**
RFR	10.35	5.33	2.92	5.49	0.76	33.4	34.1	47.8	118.1	37.2
XGB	10.25	5.33	2.88	5.44	0.77	32.7	33.6	46.9	114.0	41.7
SVR	11.26	5.8	3.44	7.19	1.08	33.9	34.7	56	429.3	107.3

Table 3: Comparison of Predictive Models

7.2 Future Work

This area of research is crucial to combat a number of health hazards caused by pollutants and improve quality of life for people all around the world. If researchers will be able to crate a very accurate prediction models and further analyse datasets such as this one we will be able to locate areas prone to dangerous pollution levels and help policymakers drive data informed legislation. Our study yielded very promising results especially given very limited time. Additionally, it could be beneficial to further pursue predicting one pollutant based on other's levels, making it easier and more efficient to control air quality levels in less populated areas with fewer detectors. We are confident that given more time and resources we would be able to create even better prediction models.

References

[1] Bhattacharyya, M., Nag, S., & Ghosh, U. (2022). Deciphering environmental air pollution with large scale city data.

Bibliography

[1] Air Quality Open Data Platform - *https://aqicn.org/*

[2] US Enviromental Protection Agency - *https://www.epa.gov/*

[3] US Census - *https://www.census.gov/*

[4] SciKit-learn documentation *https://scikit-learn.org/stable/index.html*

[5] TensorFlow documentation *https://www.tensorflow.org/api$_d$ocs/python/tf*

[6] Grammarly - text correction *https://www.grammarly.com/*

[7] ChatGPT - debugging *https://chatgpt.com/*

[8] Pandas documentation *https://pandas.pydata.org/docs/*

[9] Cartopy - *https://scitools.org.uk/cartopy/docs/latest/*

Appendices

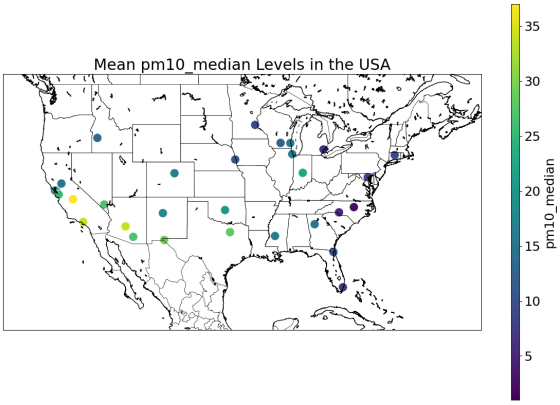

Figure 21: Geo-Analysis for PM10

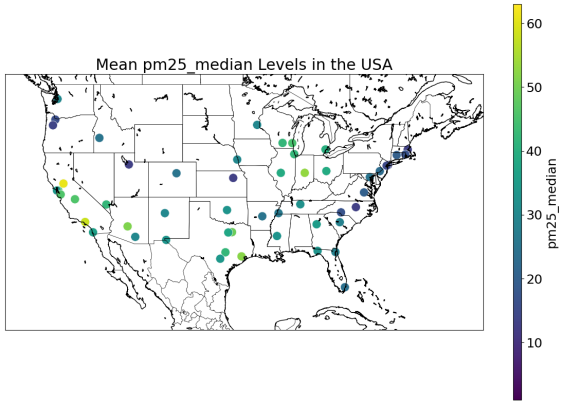

Figure 22: Geo-Analysis for PM2.5

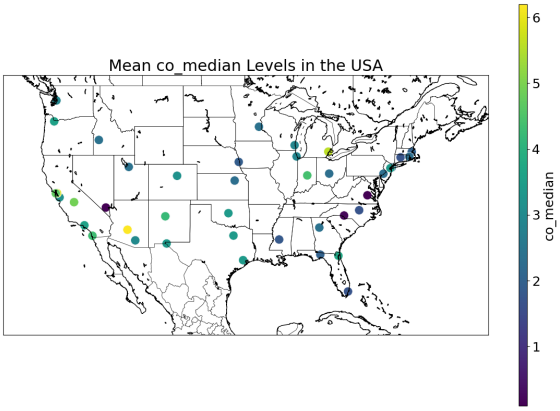

Figure 23: Geo-Analysis for CO

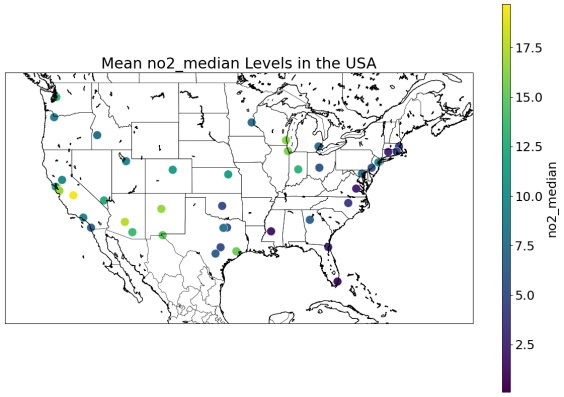

Figure 24: Geo-Analysis for NO2

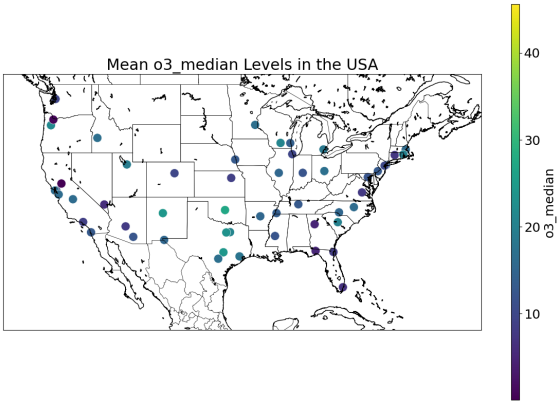

Figure 25: Geo-Analysis for O3

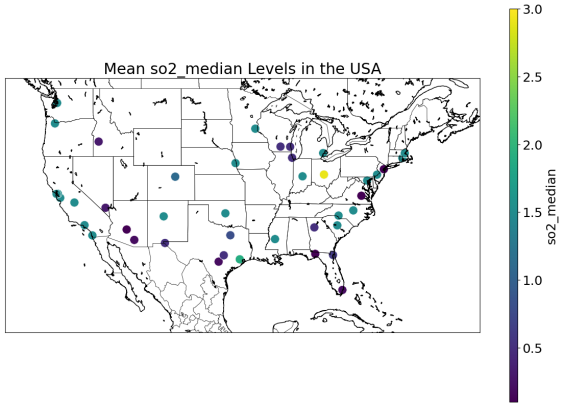

Figure 26: Geo-Analysis for SO2

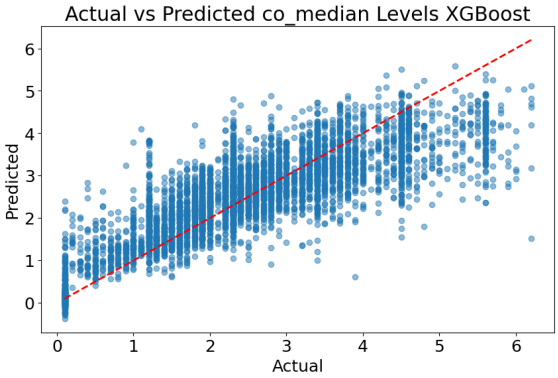

Figure 27: Predicted vs Actual CO

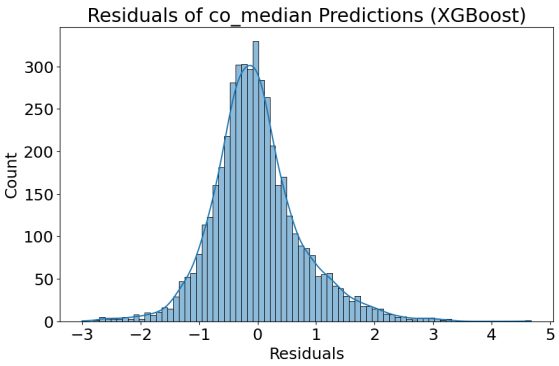

Figure 28: Prediction residuals CO

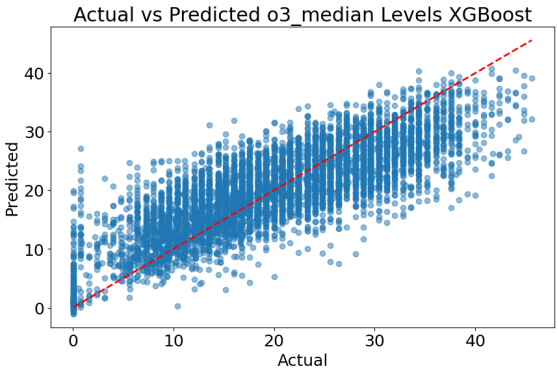

Figure 29: Predicted vs Actual O3

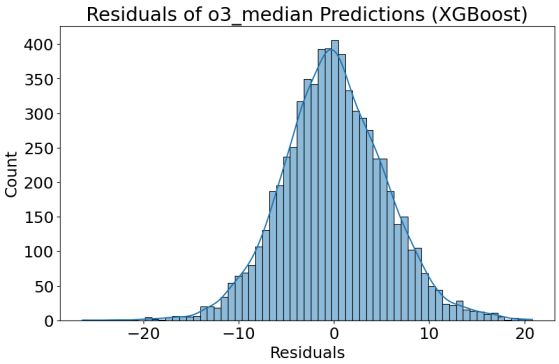

Figure 30: Prediction residuals O3

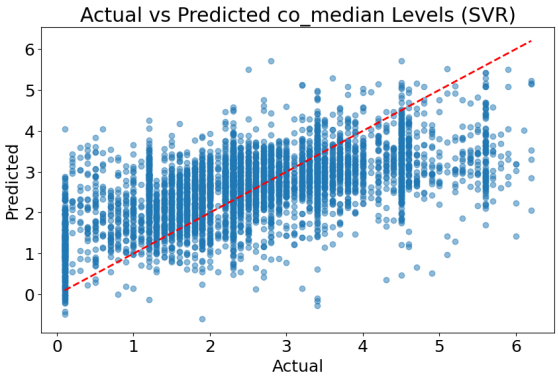

Figure 31: Predicted vs Actual NO2

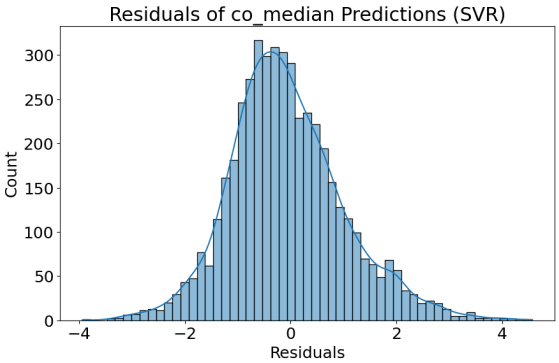

Figure 32: Prediction residuals NO2

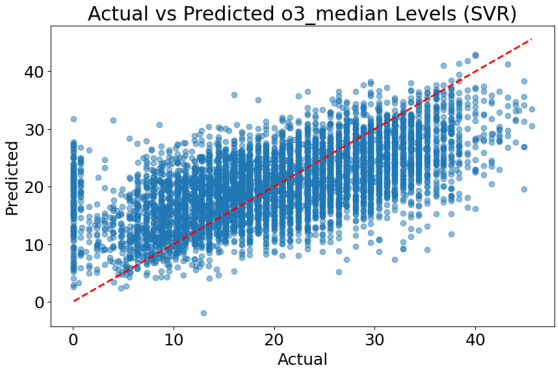

Figure 33: Predicted vs Actual PM10

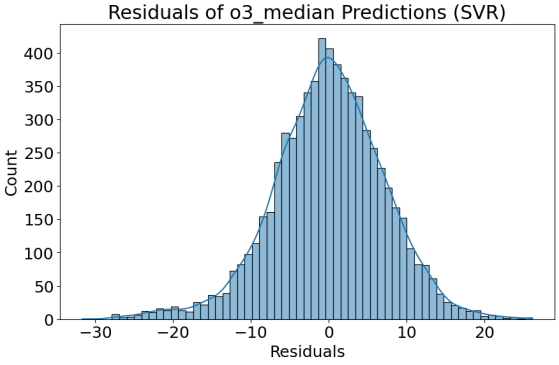

Figure 34: Prediction Residuals PM10

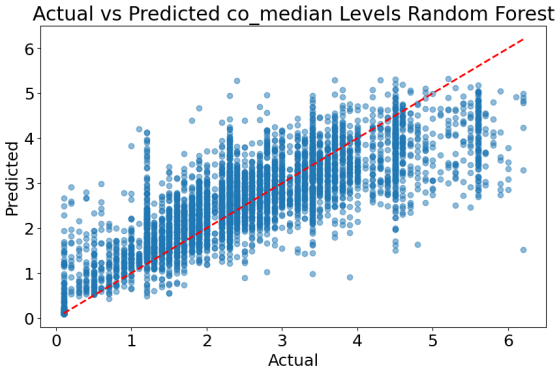

Figure 35: Original vs Cleaned PM2.5

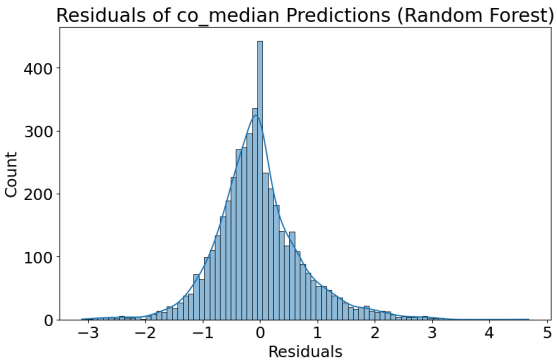

Figure 36: Prediction residuals SO2

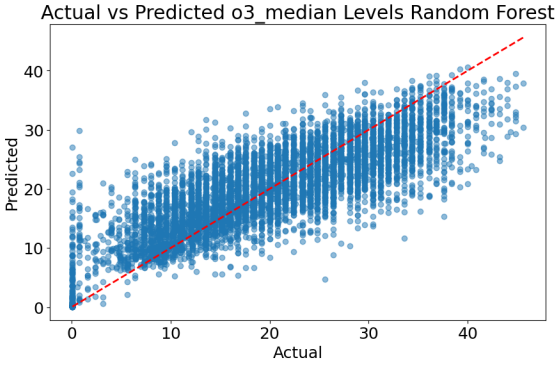

Figure 37: Predicted vs Actual PM2.5

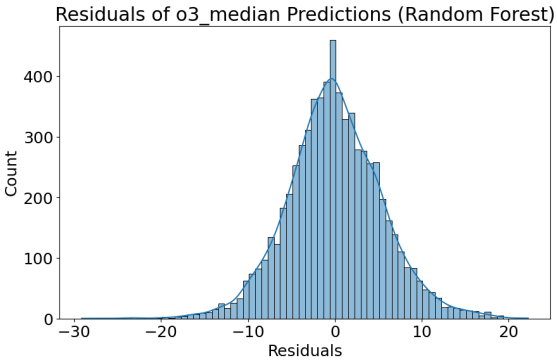

Figure 38: Prediction Residuals PM2.5

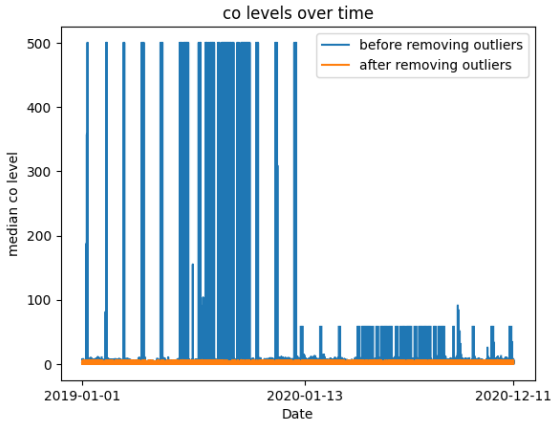

Figure 39: Original vs Cleaned CO

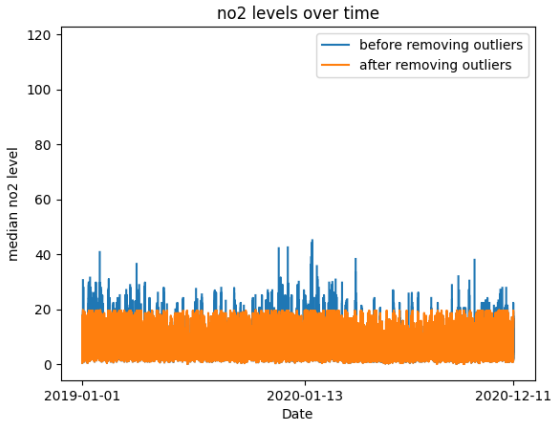

Figure 40: Original vs Cleaned CO

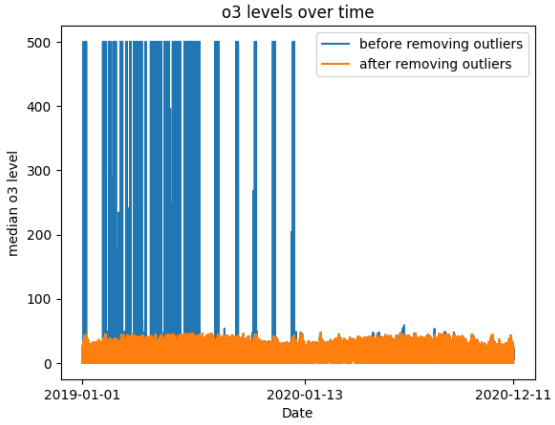

Figure 41: Original vs Cleaned O3

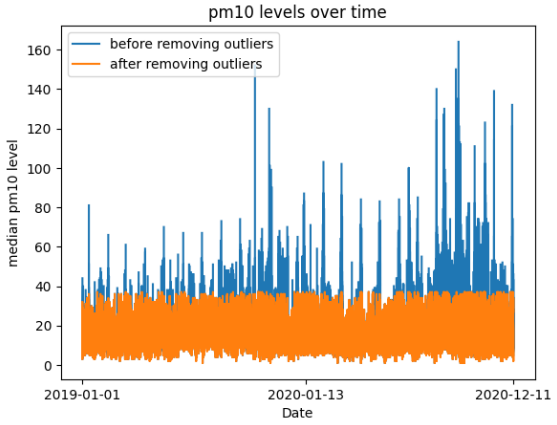

Figure 42: Original vs Cleaned PM.10

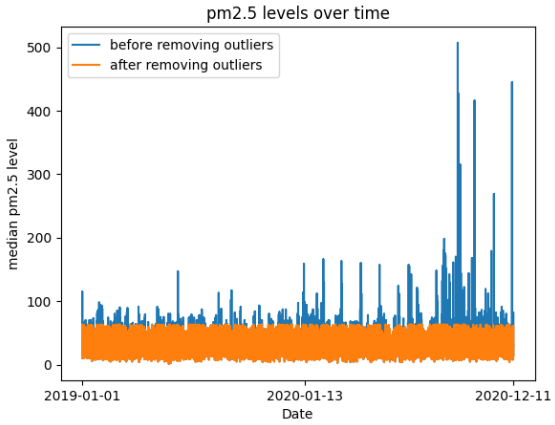

Figure 43: Original vs Cleaned PM2.5

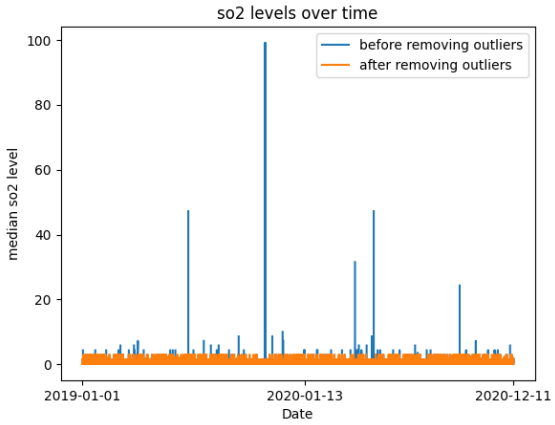

Figure 44: Prediction residuals SO2

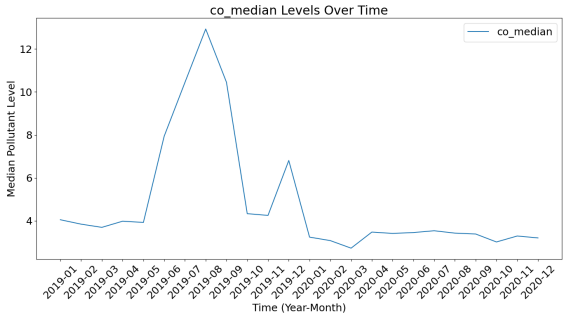

Figure 45: CO Levels Over Time

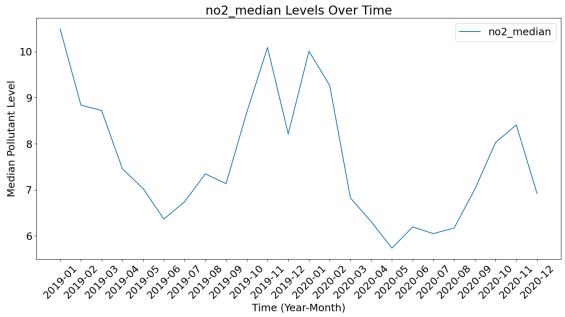

Figure 46: NO2 Levels Over Time

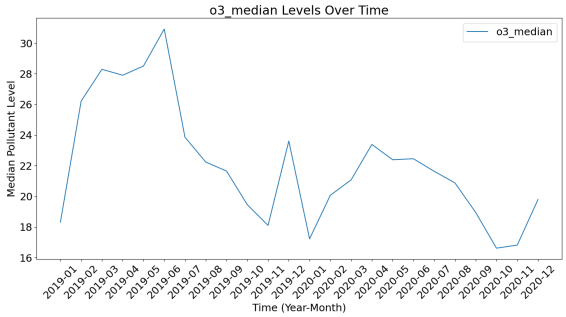

Figure 47: O3 Levels Over Time

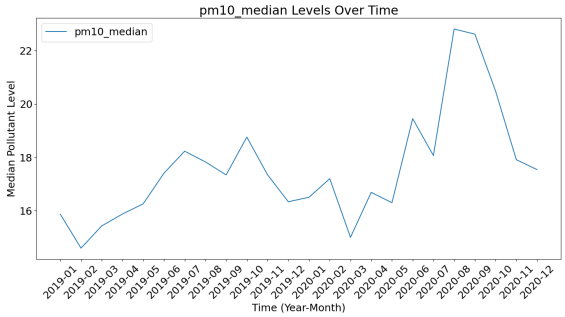

Figure 48: PM10 Levels Over Time

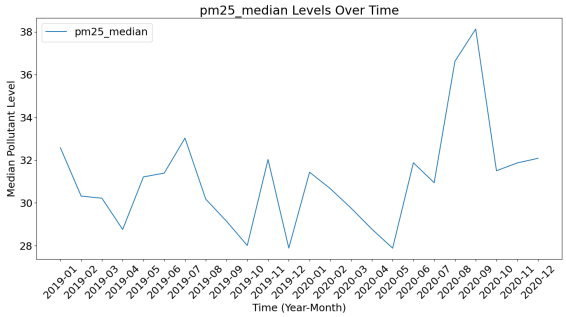

Figure 49: PM2.5 Levels Over Time

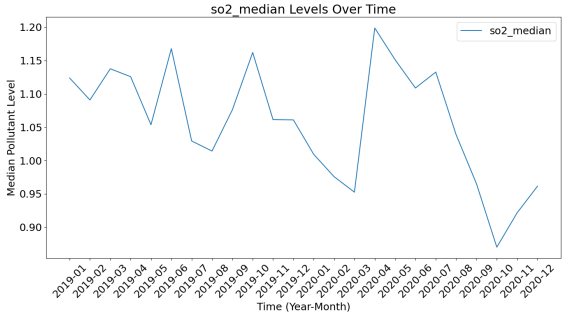

Figure 50: SO2 Levels Over Time