**Project:** Analyzing Environmental Air Pollution in Cities

**Student:** Leonardo Cruz de Souza, NEPTUN: PUL5SX

## OBJECTIVE

The objective of this project is to analyse environmental air pollution data from various cities, identify patterns, and explore relationships between pollution levels and various factors.

## HYPOTHESIS

(H1) There is a positive correlation between vehicle travel distance and pollution levels.

## DATASET

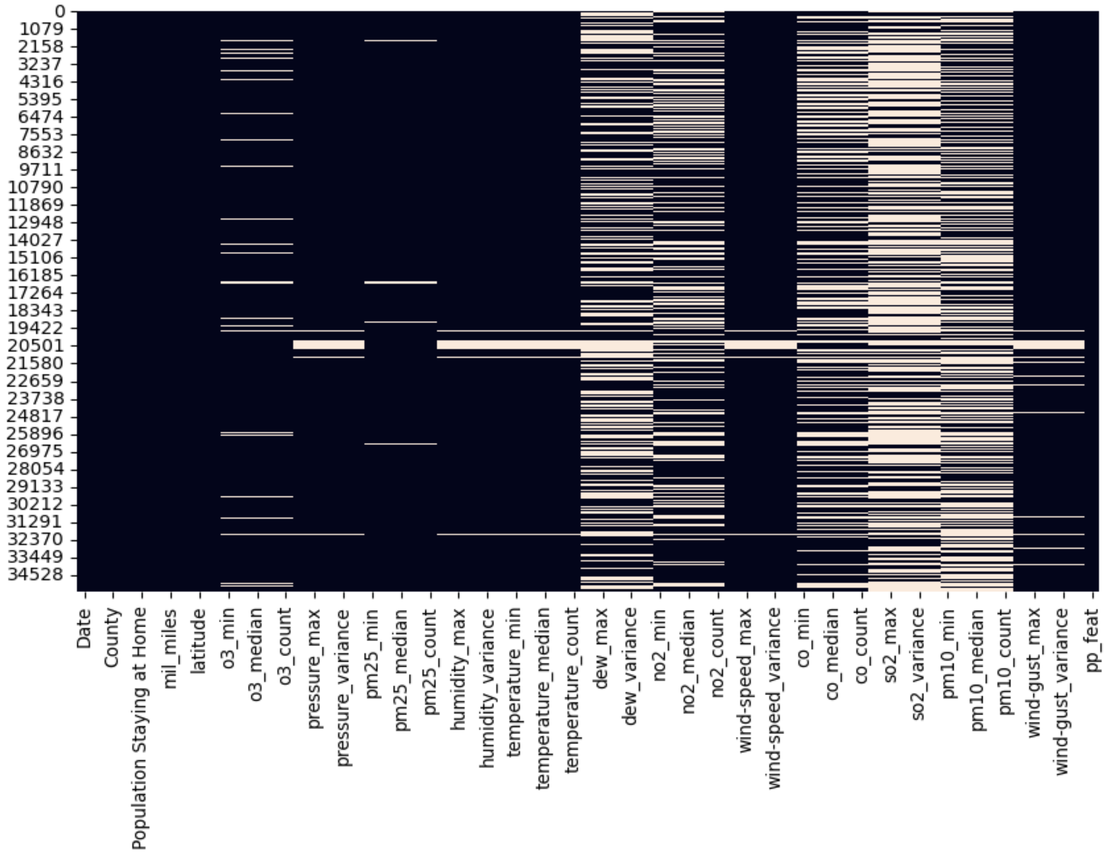
The dataset used for this project is from Bhattacharyya et al. (2022) and contains comprehensive information related to environmental air quality and various factors potentially influencing pollution levels in multiple cities.

**Relevant Columns:**

1. **Date:** Date of the sample - This column provides a timestamp for each data point, allowing for time-based analysis.
2. **City:** City of the sample - Identifies the city from which the data was collected, enabling city-specific insights.
3. **X_median/min/max/variance:** Median/min/max/variance value of the pollutant/meteorological feature X for the day.
4. **mil_miles:** Total vehicle travel distance for the sample - Provides data on the total distance traveled by vehicles, which can indicate the impact of transportation on air quality.
5. **pp_feat:** Calculated feature for the influence of neighboring power plants - This feature assesses the potential influence of nearby power plants on pollution levels.
6. **Population Staying at Home:** Used as a measure of domestic emissions.
7. **Pollutants:** This dataset covers various pollutants, including PM2.5, PM10, NO2 (Nitrogen Dioxide), O3 (Ozone), CO (Carbon Monoxide), and SO2 (Sulfur Dioxide).
8. **Meteorological Features:** The dataset also includes meteorological data, encompassing temperature, pressure, humidity, dew point, wind speed, and wind gust.

## DATA EXPLORATION

The dataset comprises daily environmental and pollution data for 54 cities in the United States, covering the time period from January 1, 2019, to November 12, 2020. In total, the dataset contains 35,596 unique observations. Notably, the dataset exhibited a significant number of missing values, particularly in the case of pollutants such as NO2, CO, SO2, and PM10, as illustrated below (white lines represents cells with missing values).



## DATA PREPROCESSING

First, the commas from the "Population Staying at Home" column and converted it to a numerical format. Additionally, a new "weekday" column was introduced to categorize data points by the day of the week. Furthermore, only columns containing the median value was selected for the analyses, excluding the min, max and variance ones.

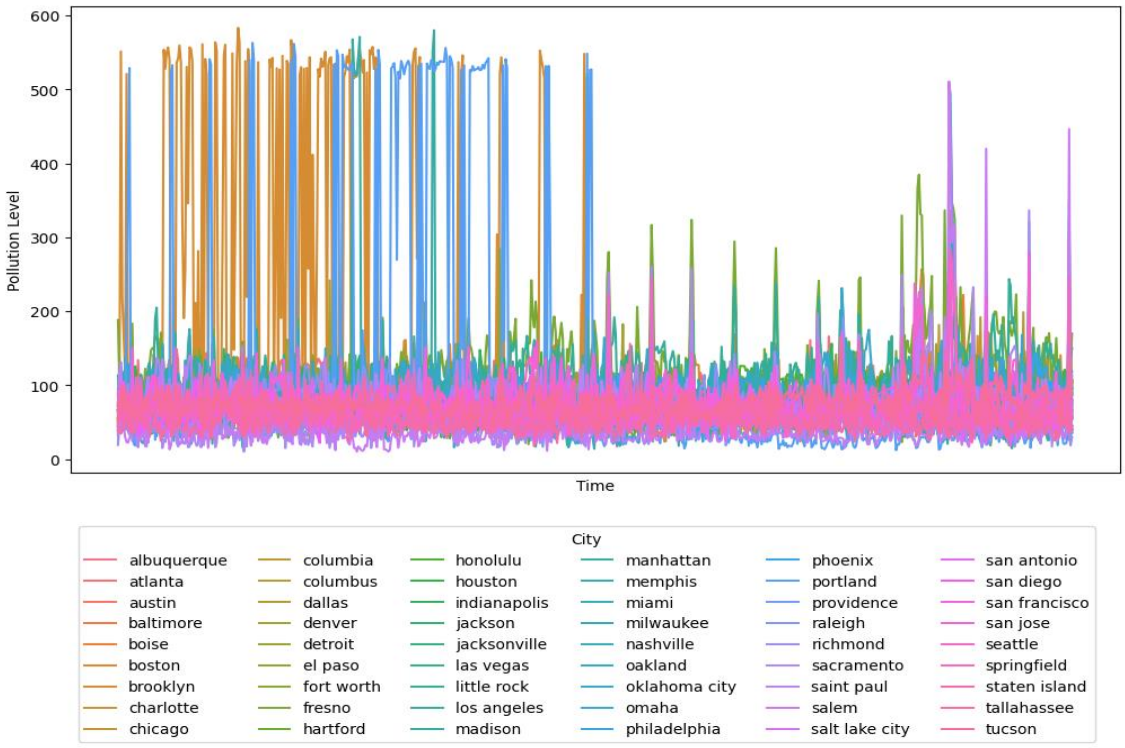
As a second step, I addressed the missing values within the dataset. To accomplish this, the data is grouped by both 'City' and 'weekday' to calculate the group-specific means, and the NaN values are replaced with these values. In cases where a group contains exclusively NaN values, the NaN values were filled with zeros.

## FEATURE ENGINEERING

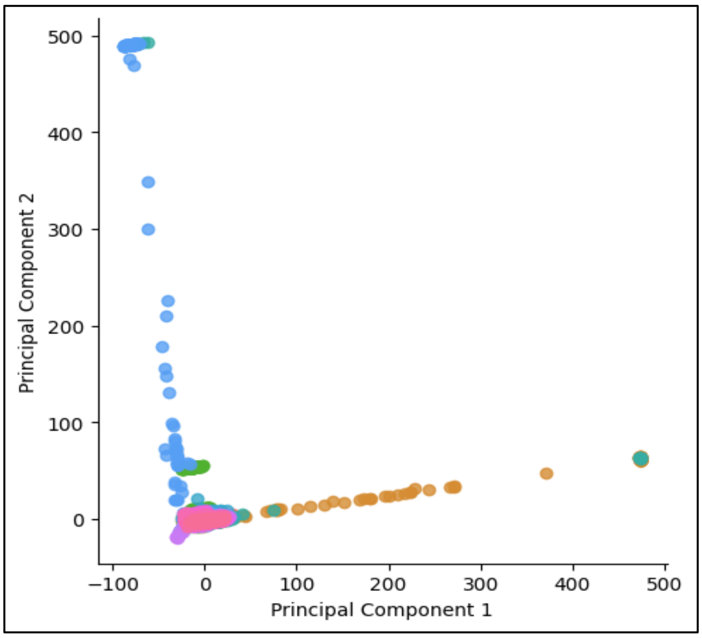
A new composite pollution column is created by summing the values of various individual pollutant columns, including 'o3_median', 'pm25_median', 'no2_median', 'co_median', 'so2_median', and 'pm10_median'. This new 'pollution_level' column represents the cumulative pollution level for each data point. To further categorize pollution levels, a 'pollution_class' column is introduced. This categorization is based on quartiles, which divide the data into four parts. Data points are classified as 'Low' if their pollution level is below the first quartile, 'Moderate' if falling between the first and third quartiles, and 'High' if exceeding the third quartile.

## EXPLORATORY DATA ANALISES (EDA)

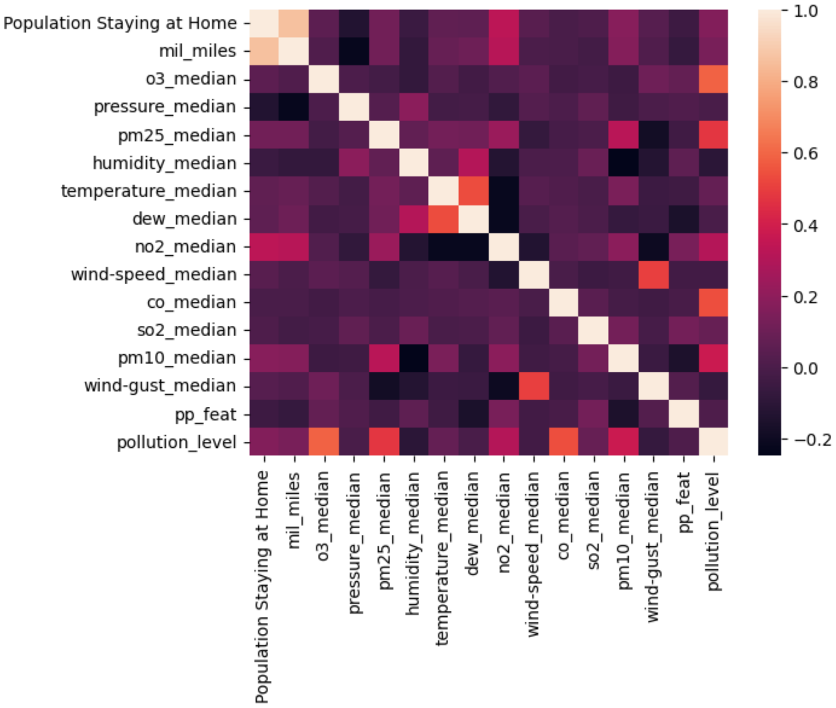
1. **The pollution level over time per city.** Most of the cities in the dataset show a similar trend and levels of pollution over time. Notably, Brooklyn (brown line) and Portland (blue line) show a very high level of pollution in 2019 but a significant drop in the beginning of 2020.



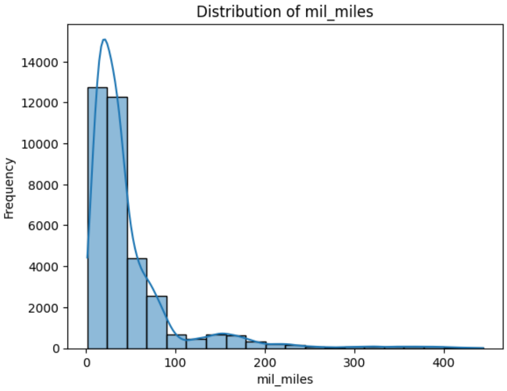
2. **Pollutants composition per city.** A PCA analysis, using the combination of every individual pollutant, was performed to visualize if the cities differed among themselves. The result show that Brooklyn and Portland are not only different from the other cities, but also between them.

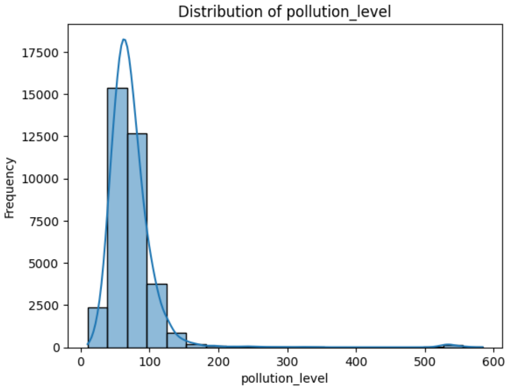


3. **Correlation between features.** All features had low correlation among them, except between population staying at home and vehicle travel distance, wind gust and wind speed. Therefore, population staying at home and wind gust were dropped. In addition, the first hypothesis was rejected as there was close to no correlation between vehicle travel distance and pollution level.



4. **Distribution of variables.** Most features and pollutants showed a non-normal and long-tailed distribution, as exemplified bellow.





## DATA MODELING

The data was split into train (80%) and test (20%), evenly distributed between cities and weekdays. Both sets were normalized.

To model the data I have used four machine learning algorithms: Linear regression, Random Forest, K-Nearest Neighbors and Support Vector Machine.

## MODEL TRAINING

First, I trained the models to predict the pollution level (sum of all pollutants) using all the other features (mil_miles, pressure_median, humidity_median, temperature_median, dew_median, wind-speed_median and pp_feat). Then, I used the same machine learning algorithms but a with multi-output response, which included all individual pollutants. The predictor variables remained the same. Lastly, I used a Random Forest classifier to model the pollution class categorical variables.

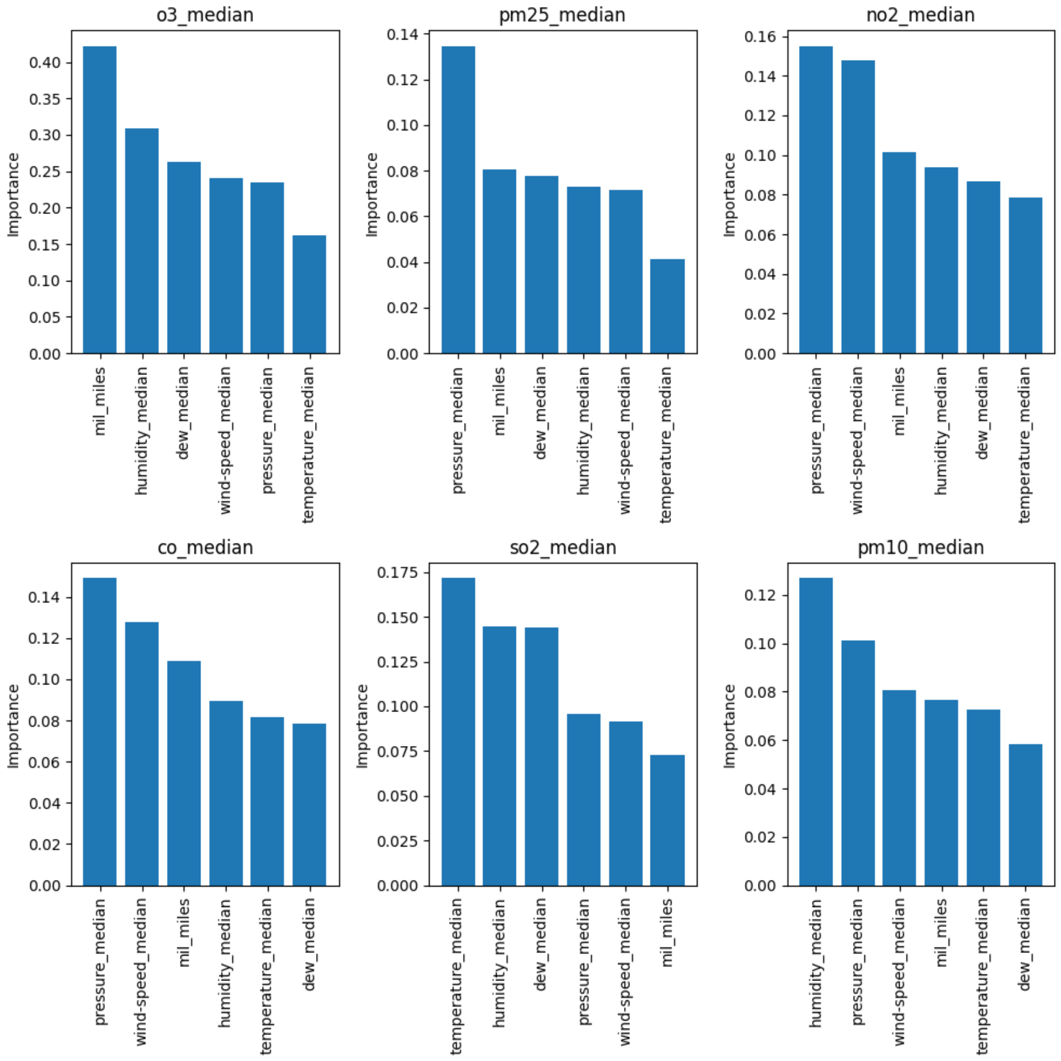
## MODEL EVALUATION

All regression models were evaluated using the Mean Squared Error (MSE) and $R^2$, whereas the classification one using accuracy. The results are shown below, ‘M’ stand for multi-output.

| Model | MSE | $R^2$ |
|---|---|---|
| Linear regression | 2699.54 | 0.01 |
| Linear regression (M) | 456.51 | 0.06 |
| Random Forest | 1641.43 | 0.40 |
| **Random Forest (M)** | **269.86** | **0.46** |
| K-Nearest Neighbor | 2363.86 | 0.14 |
| K-Nearest Neighbor (M) | 384.06 | 0.25 |
| Support Vector Machine | 2674.38 | 0.02 |
| Support Vector Machine (M) | 448.94 | 0.11 |

The Random Forest classifier had an accuracy of 62%.

Therefore, the multi-output Random Forest was selected as the best model, taking into account that it can predict every pollutant individually reasonably. The variable importance plots are shown below.



## INSIGHTS AND CONCLUSIONS

The temporal trends in pollution levels and pollutant composition exhibited remarkable consistency among the majority of the studied cities, suggesting a common pattern for pollution management. This finding implies that similar strategies may be effective for most cities, which is a favorable outcome, despite the distinct dynamics among individual pollutants, as discussed below.

Although the initial hypothesis (H1) was rejected, the vehicle travel distance was important in most cases in the multi-output Random Forest, particularly in predicting O3 levels. Furthermore, the analysis revealed that the studied pollutants display diverse relationships with various influencing factors, necessitating distinct strategies based on the specific pollutant of interest. These valuable insights are also provided by the modeling and guide the development of targeted measures for controlling pollutants on an individual basis. For example, the study uncovered a strong association between O3 levels and the intensity of vehicle travel. Additionally, the results underscore the effectiveness of multi-output models in accurately forecasting pollutant levels.

## CITATION

Bhattacharyya, M., Nag, S., & Ghosh, U. (2022). Deciphering Environmental Air Pollution with Large Scale City Data. In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. International Joint Conferences on Artificial Intelligence Organization.