Air Quality Analysis- Project Team 3

Problem Statement:

Air pollution is a critical environmental threat impacting urban populations, particularly in New York City (NYC). This project aims to analyze NYC air quality data to understand pollutant levels, their distribution across neighborhoods, temporal trends, and spatial patterns. The analysis will focus on pollutants such as PM 2.5 and SO2, providing insights into their impacts on public health.

**Analysis Questions:**

**Pollutant-Specific Trends (Krishna)**:

* How have Nitrogen dioxide (NO2) levels changed over time across different neighborhoods?
* What are the trends in Ozone (O3) levels in various neighborhoods over the years?

**Health Impact Analysis (Michael)**:

* How do asthma emergency department visits correlate with Ozone levels in different neighborhoods?
* What is the relationship between PM 2.5 levels and cardiovascular hospitalizations (age 40+) across NYC?
* Analyze the correlation between respiratory hospitalizations (age 20+) and PM 2.5 levels.

**Seasonal Patterns (Halima)**:

* Do NO2 levels exhibit any significant seasonal patterns?
* What are the seasonal variations in O3 and PM 2.5 levels?

**Spatial Analysis of Emissions (Nicole)**:

* Create a spatial distribution map of SO2 emissions from boiler emissions across neighborhoods.
* Identify neighborhoods with consistently high levels of boiler emissions (NOx, PM 2.5, SO2).

**Toxic Air Pollutants**:

* Analyze the spatial and temporal distribution of outdoor air toxics like Benzene and Formaldehyde.
* What neighborhoods show higher concentrations of these toxic air pollutants?

August 7th: Finish the analysis questions

August 8th- Solution Outline  
August 12th – Visulization plan and deployment plan

August 14th: Presentation

**Data Source:**

* **Dataset**: NYC Air Quality Surveillance Data (Uploaded as Air\_Quality\_20240801.csv)

**Solution Outline:**  
**1. EDA Plan:**

* **Objective**: Summarize and visualize the distribution and basic statistics of the air quality data.
* **Steps**:
* Calculate descriptive statistics (mean, median, standard deviation) for PM 2.5 across neighborhoods.
* Identify outliers in PM 2.5 measurements.
* Plot the distribution of SO2 emissions across neighborhoods to identify areas with significantly different emission levels.

**2. Data Storage Plan:**

* **Local Storage**: Store the initial dataset and processed files locally for analysis.
* **Cloud Storage**: Use Amazon S3 for storing data files to ensure scalability and accessibility.

**3. Modelling Plan:**

* **Steps**:
* Use time series analysis to study trends in SO2 emissions in specific neighborhoods (e.g., Southeast Queens and Bensonhurst - Bay Ridge).
* Apply seasonal decomposition to identify patterns in PM 2.5 data.

**4. Model Optimization Plan:**

* **Steps**:
* Tune parameters of time series models (e.g., ARIMA) for better accuracy.
* Validate models using cross-validation techniques.

**5. Model Evaluation Plan:**

* **Metrics**: Evaluate models based on accuracy, mean squared error, and seasonal pattern identification.
* **Steps**:
* Compare actual vs. predicted values in time series analysis.
* Assess model performance using residual plots and error metrics.

**6. Visualization Plan:**

* **Tools**: Matplotlib, Plotly, Tableau
* **Visuals**:
* Use Matplotlib for initial EDA visualizations (e.g., histograms, scatter plots).
* Create interactive maps using Plotly to visualize spatial distribution of pollutants.
* Develop a dashboard using Tableau to present key metrics and trends.

**7. Deployment Plan:**

* **Tools**: AWS Sagemaker, Databricks, S3
* **Steps**:
* Deploy time series models on AWS Sagemaker for real-time predictions.
* Use Databricks for processing large-scale data and performing advanced analytics.
* Store the model and data files on S3 for easy access and management.