**Predicting Air Quality Index (AQI) in Mumbai Using Linear Regression**

1. Why This Topic

Air quality is a pressing public health concern in urban centers like Mumbai, where high pollution levels from traffic, industry, and dense populations frequently exceed safe thresholds. Predicting AQI enables authorities to issue timely health warnings, enforce pollution controls, and protect vulnerable groups, such as children and the elderly, from respiratory conditions like asthma or bronchitis. I chose this topic because it blends data science with real-world impact—using linear regression to forecast AQI can guide environmental policies and raise awareness about air quality challenges in Mumbai. This project also aligns with global efforts to combat pollution and supports Sustainable Development Goal 11 (Sustainable Cities and Communities).

2. About the topic

This project develops a linear regression model to predict daily AQI in Mumbai using key air pollutants as predictors. AQI is a composite index that aggregates concentrations of pollutants—PM2.5, PM10, NO2, CO, and O3—into a single value from 0 to 500, where higher values indicate worse air quality. The model incorporates polynomial features (e.g., PM2.5², PM2.5·PM10) to capture non-linear relationships and month-based dummy variables to account for seasonal patterns, such as increased pollution in winter. By training on historical data, the model aims to provide accurate, interpretable AQI forecasts to help anticipate pollution spikes and inform mitigation strategies. The use of polynomial features enhances the model’s ability to reflect the complex interactions inherent in AQI calculations, while the time-series approach (no shuffling in data splits) respects the temporal nature of the data.

3. Data Source Used for the Model

The dataset is sourced from the “Air Quality Data in India” collection, specifically the city\_day.csv file, available on platforms like Kaggle (e.g., “Air Quality Data in India (2015-2020)”). It contains daily air quality measurements for multiple Indian cities, including Mumbai, spanning 2015 to 2020. The dataset includes columns for Date, City, PM2.5, PM10, NO2, CO, O3, and AQI, which were filtered for Mumbai to focus on local patterns.

4. Why This Data Source

This dataset was chosen for its reliability, comprehensiveness, and relevance to India’s air quality context. Collected from official monitoring stations (likely managed by the Central Pollution Control Board or state authorities), it ensures accuracy and includes key pollutants directly tied to AQI calculations. Its multi-year coverage captures seasonal and long-term trends, ideal for modeling Mumbai’s diverse pollution sources (e.g., vehicles, industries). The open access via Kaggle made it practical for a student project. Alternatives, such as global datasets (e.g., WHO air quality data), lack daily granularity or city-specific focus, while real-time APIs were unsuitable for historical analysis.

5. Important Variables Used in This Model

The model uses five primary predictor variables:

* PM2.5: Fine particulate matter (≤2.5 micrometers), a major contributor to AQI due to its severe health impacts.
* PM10: Coarser particulate matter (≤10 micrometers), prevalent in urban pollution.
* NO2: Nitrogen dioxide, emitted from vehicles and linked to smog formation.
* CO: Carbon monoxide, resulting from incomplete combustion.
* O3: Ozone, a secondary pollutant formed in sunlight, affecting AQI.

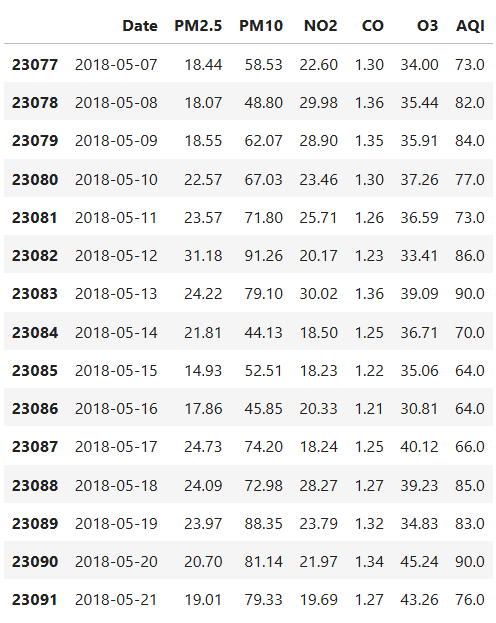
Additional features include:

Month Dummies: Binary variables (e.g., Month\_2 for February) to capture seasonal variations.

These variables are critical because India’s AQI calculation uses sub-indices based on these pollutants, selecting the maximum value. The model’s polynomial terms reflect the non-linear nature of these sub-indices, while month dummies address seasonal trends like winter pollution peaks.

6. Sample of the Dataset

Below is a sample of 15 rows from the Mumbai AQI dataset, showing Date, PM2.5, PM10, NO2, CO, O3, and AQI.



7. Descriptive Stats of the Data Set

Below are the summary statistics for the key variables (PM2.5, PM10, NO2, CO, O3, and AQI) in the Mumbai dataset, including mean, median, minimum, maximum, and standard deviation.

A table with numbers and letters

AI-generated content may be incorrect.