1. Introduction

The dataset analyzed contains air quality measurements from various cities, focusing on pollutants such as PM2.5, PM10, NO2, NOx, CO, O3, benzene, toluene, xylene, and AQI (Air Quality Index). The primary objective of this analysis is to understand the relationships between pollutants, identify key drivers of poor air quality, and provide actionable insights for improving air quality. The analysis includes data cleaning, exploratory data analysis (EDA), correlation analysis, and visualizations to uncover patterns and trends.

2. Data Loading and Initial Inspection

- **Dataset Structure**: The dataset contains 29,531 entries with 16 columns, including city names, dates, pollutant levels, AQI, and AQI categories.
- **Initial Observations**: The dataset has missing values, particularly for pollutants like Xylene (61.3% missing) and NH3 (35% missing). The first few rows reveal that PM2.5 and PM10 levels are often missing, while NO, NO2, and CO are more consistently recorded.
- **Data Types**: Most columns are numerical (e.g., PM2.5, NO2, AQI), while categorical columns include `City` and `AQI_Bucket`.

3. Data Cleaning

- **Handling Missing Values**: Missing numerical values were imputed using the median, while categorical values were filled with the mode. Rows with remaining missing values were dropped.
- **Duplicate Records**: No duplicate rows were found in the dataset.
- **Outlier Treatment**: Outliers were detected using the Interquartile Range (IQR) method and capped to the upper and lower bounds to reduce their impact on the analysis.
- **Standardization**: Categorical values like `City` and `AQI_Bucket` were standardized to ensure consistency.

4. Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

- **Summary Statistics**:
    - **PM2.5**: Mean ≈ 67.45 µg/m³, Median ≈ 48.57 µg/m³. Positively skewed, indicating occasional extreme pollution events.
    - **NO2**: Mean ≈ 28.56 µg/m³, Median ≈ 21.69 µg/m³. Also positively skewed.
    - **AQI**: Mean ≈ 166.46, Median ≈ 118. Positively skewed, with extreme values reaching up to 2049.

- **Frequency Distributions**:
    - **Cities**: Ahmedabad, Delhi, and Mumbai dominate the dataset, suggesting these are major pollution hotspots.
    - **AQI_Bucket**: Most records fall into "Moderate" (45.7%) and "Poor" (15.5%) categories, indicating suboptimal air quality.
- **Visualizations**:
    - **Histograms**: PM2.5 and NO2 distributions are positively skewed, with long tails indicating extreme pollution events.
    - **Box Plots**: PM2.5 and NO2 show wide interquartile ranges and many outliers, confirming variability and extreme events.

4.2 Bivariate Analysis

- **Correlation Matrix**:
    - **Strong Positive Correlations**:
        - PM2.5 and PM10 (0.93): Both are particulate matter from similar sources.
        - NO2 and NOx (0.97): NO2 is a component of NOx.
        - Benzene and Toluene (0.85): Both are VOCs emitted together.
    - **Moderate Positive Correlations**:
        - CO and PM2.5 (0.65): Both emitted from combustion processes.
        - O3 and NO2 (0.55): Ozone formation depends on NOx and sunlight.
    - **Weak or No Correlations**:
        - NH3 and SO2 show weak correlations, suggesting distinct sources (e.g., agriculture for NH3, coal combustion for SO2).
- **Scatter Plots**:
    - **PM2.5 vs PM10**: Strong linear relationship, confirming high correlation.
    - **NO2 vs NOx**: Strong linear relationship, as NO2 is a component of NOx.
    - **CO vs O3**: Weak positive relationship, with some outliers.
    - **Benzene vs Toluene**: Strong positive relationship, as both are VOCs.
- **Bar Plots, Violin Plots, and Box Plots**:
    - **AQI by AQI_Bucket**: Higher AQI values correspond to worse air quality categories (e.g., "Severe" or "Very Poor").
    - **PM2.5 by AQI_Bucket**: PM2.5 levels are significantly higher in worse air quality categories.
    - **NO2 by AQI_Bucket**: NO2 levels are higher in worse air quality categories, with more outliers.

5.Key Findings

1. **PM2.5 and PM10 Dominance**:
    - Strong correlation (0.93) between PM2.5 and PM10, indicating shared sources like vehicles and industries.
    - High mean values (PM2.5: 67.45 µg/m³, PM10: 118.13 µg/m³) and extreme outliers (PM2.5 > 949.99 µg/m³, PM10 > 1000 µg/m³) highlight severe particulate pollution.
2. **NO2 and NOx Relationship**:
    - Strong correlation (0.97) between NO2 and NOx, with NO2 being a component of NOx.
    - Higher NO2 levels in "Severe" and "Very Poor" AQI categories, indicating significant contribution to poor air quality.
3. **VOCs (Benzene and Toluene)**:
    - Strong correlation (0.85) between benzene and toluene, both emitted from vehicles and industries.
    - Outliers in benzene levels suggest localized industrial pollution sources.
4. **NH3 and SO2**:
    - Weak correlations with other pollutants, indicating distinct sources (agriculture for NH3, coal combustion for SO2).
    - Lower variability and fewer outliers compared to other pollutants.
5. **AQI Trends**:
    - AQI distribution is positively skewed, with most values in "Moderate" to "Poor" range (AQI: 100–200).
    - Extreme AQI values (> 400) are linked to high PM2.5, PM10, and NO2 levels.
6. **City-Specific Insights**:
    - Ahmedabad, Delhi, and Mumbai dominate the dataset with the highest pollution levels and extreme events.
    - Smaller cities like Aizawl and Shillong show lower pollutant levels, possibly due to less monitoring.
7. **Extreme Pollution Events**:
    - Outliers in PM2.5, PM10, and NO2 levels are common in "Severe" and "Very Poor" categories.
    - Likely caused by stubble burning, industrial emissions, and adverse meteorological conditions.
8. **Ozone (O3) Formation**:
    - Moderate correlation (0.55) between NO2 and O3, indicating NOx's role in ozone formation.
    - O3 levels are more stable compared to other pollutants, with fewer outliers.

9. **Health Risks**:
    - Extreme pollution events pose significant health risks, especially in cities with frequent "Severe" AQI categories.

6. Conclusion

The analysis reveals that air quality is significantly impacted by pollutants like PM2.5, PM10, NO2, and NOx, which are strongly correlated and originate from common sources such as vehicles and industries. Extreme pollution events are more frequent in worse air quality categories, particularly in cities like Ahmedabad, Delhi, and Mumbai. Addressing baseline pollution and extreme events through targeted measures and improved monitoring is essential for improving air quality.

This report provides a comprehensive understanding of the dataset and actionable insights for policymakers and stakeholders to mitigate air pollution effectively. Further analysis could include time-series analysis to identify trends and seasonal variations in air quality.