25th Feb 25

Khyathi M R- ENG22CT0009

Data Engineering - Assignment 1

1. Dataset Overview

- **Dataset Name:** Real-time Air Quality Index from Various Locations
- **Description:** This dataset contains information regarding air quality measurements from various locations across India, including pollutants, their levels, and geographic details.
- **Purpose of Collection:** To monitor and assess air quality in real-time, helping to inform public health and environmental policies.

2. Data Collection Details

- **Source of Data:** Public dataset - collected from air quality monitoring stations across India.
- **Collection Method:** Automated system using API
- **Time Period:** 24-02-2025
- **Frequency of Collection:** Once in then minutes.
- **Sampling Method:** Continuous monitoring at predefined locations.

3. Dataset Structure

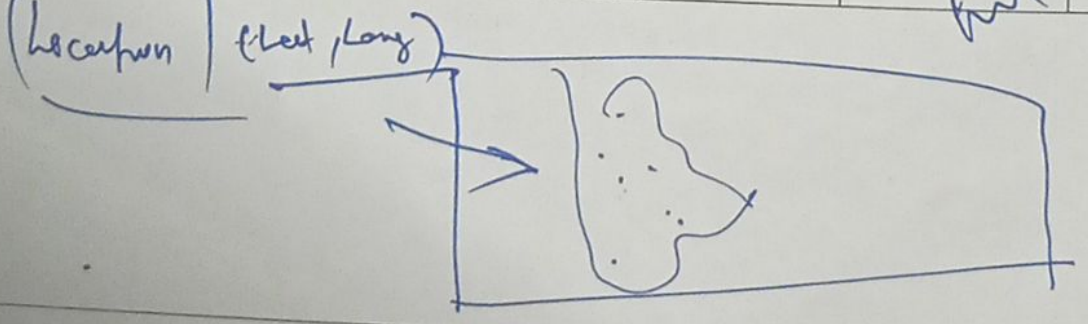
- **Number of Rows:** 3,222 records
- **Number of Columns:** 13 columns
- **File Format:** CSV (common for such datasets).
- **Storage Location:** Typically stored locally in a Jupyter environment (could be on a local server or cloud storage).

4. Column Descriptions

Column Name	Description	Data Type	Range of Values	Distribution	Missing Values	Notes
Country	The country of the monitoring station	String	India	Uniform distribution across locations	None reported	N/A

Datetime

State	The state within India where the station is located	String	Various Indian states	Varies by region	None reported	N/A
City	The city where the monitoring station is based	String	Various Indian states	Varies by region	None reported	N/A
Station	Name of the monitoring station	String	Various names	Unique names per location	None reported	N/A
Last Update	Timestamp of the last update	DateTime	24-02-2025	Single date	None reported	Timestamp format: dd-mm-yyyy hh:mm:ss
Latitude	Latitude coordinate of the station	Float	Range of latitudes in India	Varies between -8.0 to 37.0	None reported	N/A
Longitude	Longitude coordinate of the station	Float	Range of longitudes in India	Varies between 68.0 to 97.0	None reported	N/A
Pollutant Id	Identifier for the pollutant	String	CO, PM2.5, OZONE, PM10, NH3	Categorical distribution	None reported	N/A
Min Value	Minimum measured value for the pollutant	Integer	Varies by pollutant	Varies by pollutant	None reported	N/A
Max Value	Maximum measured value for the pollutant	Integer	Varies by pollutant	Varies by pollutant	None reported	N/A
Avg Value	Average value of the pollutant	Integer	Varies by pollutant	Varies by pollutant	None reported	N/A



5. Data Quality and Cleaning

- **Handling Missing Values:** No missing values reported in the preview.
- **Outlier Treatment:** No specific method ; typically involves statistical techniques.
- **Data Transformation:** Likely raw data.
- **Data Validation:** Assumed checks for accuracy during automated data collection.

6. Limitations and Assumptions

- **Potential Biases:** Potential bias in sampling locations; some areas may have more stations than others.
- **Limitations:** Lacks historical context or trends.
- **Assumptions:** Assumes all monitoring stations provide consistent and accurate readings.

7. Data Usage and Privacy

- **Data Sensitivity:** Does not appear to contain sensitive personal information.
- **Privacy Considerations:** No personal identifiers present, so anonymization not particularly relevant.
- **Usage Restrictions:** Not specified; assume public domain unless otherwise indicated.