**Preeti Gupta**

**ML Pollution AQI PROJECT**

**NOTE 2000-2022 DATA NOT AVAILABLE SO I USED 2000-2023 DATA**

**Understand the Dataset Structure:-It is a csv data set name pollution\_2000\_2023. Data organize in row and columns. No of column are 665414 Rows and 22 Columns. we can access it using python Pandas Library.**

1. **There a 22 columns in dataset in which 5 are text based column and 17 numeric based column. Name and data types are given below**
2. # Column Non-Null Count Dtype
3. **--- ------ -------------- -----**
4. 0 Unnamed: 0 665414 non-null int64
5. 1 Date 665414 non-null object
6. 2 Address 665414 non-null object
7. 3 State 665414 non-null object
8. 4 County 665414 non-null object
9. 5 City 665414 non-null object
10. 6 O3 Mean 665414 non-null float64
11. 7 O3 1st Max Value 665414 non-null float64
12. 8 O3 1st Max Hour 665414 non-null int64
13. 9 O3 AQI 665414 non-null int64
14. 10 CO Mean 665414 non-null float64
15. 11 CO 1st Max Value 665414 non-null float64
16. 12 CO 1st Max Hour 665414 non-null int64
17. 13 CO AQI 665414 non-null float64
18. 14 SO2 Mean 665414 non-null float64
19. 15 SO2 1st Max Value 665414 non-null float64
20. 16 SO2 1st Max Hour 665414 non-null int64
21. 17 SO2 AQI 665414 non-null float64
22. 18 NO2 Mean 665414 non-null float64
23. 19 NO2 1st Max Value 665414 non-null float64
24. 20 NO2 1st Max Hour 665414 non-null int64
25. 21 NO2 AQI 665414 non-null int64
26. dtypes: float64(10), int64(7), object(5)

**Check Meta Data**

**Data Dictionary:- there is variable x ,x1,x2,x3,x4,x5,x6,x7 to access the mean median mode and sum and etc. from the dataset.**

**Missing Data:- There is no missing data in pollution\_2000\_2023 dataset**

**Data Source And Information:- Data Available on Kaggle Site publicly and there is no restriction to access the data set**

**Initial Data Exploration:-**

** Sample Data Inspection:- Load the dataset (using commands like read\_csv() in Python with pandas library) This helps you spot any irregularities or patterns early.**

**Descriptive Statistics:- numeric data involve mean median and mode to access different data.**

* **Data Types and Consistency: It Ensure that the data types are appropriate for the kind of analysis you intend to perform (e.g., dates should be in date format, categorical data should be properly labeled).**

**. Look for Data Quality Issue**

* **There are some outlier in numeric columns like AQI**

**. Examine Data Provenance and Licensing**

* ** Provenance:-the data come from Kaggle site, and it is available publicly This can help you assess the credibility and limitations of the dataset.**

**MECE Framework**

* **Define data set(columns detail)**
* # Column Non-Null Count Dtype
* **--- ------ -------------- -----**
* 0 Unnamed: 0 665414 non-null int64
* 1 Date 665414 non-null object
* 2 Address 665414 non-null object
* 3 State 665414 non-null object
* 4 County 665414 non-null object
* 5 City 665414 non-null object
* 6 O3 Mean 665414 non-null float64
* 7 O3 1st Max Value 665414 non-null float64
* 8 O3 1st Max Hour 665414 non-null int64
* 9 O3 AQI 665414 non-null int64
* 10 CO Mean 665414 non-null float64
* 11 CO 1st Max Value 665414 non-null float64
* 12 CO 1st Max Hour 665414 non-null int64
* 13 CO AQI 665414 non-null float64
* 14 SO2 Mean 665414 non-null float64
* 15 SO2 1st Max Value 665414 non-null float64
* 16 SO2 1st Max Hour 665414 non-null int64
* 17 SO2 AQI 665414 non-null float64
* 18 NO2 Mean 665414 non-null float64
* 19 NO2 1st Max Value 665414 non-null float64
* 20 NO2 1st Max Hour 665414 non-null int64
* 21 NO2 AQI 665414 non-null int64
* dtypes: float64(10), int64(7), object(5)

**MECE BREAKDOWN for Pollution\_2001\_2023**

**Unnamed:-S. no**

**Date:- Year-wise , Month-wise**

**Address:- State, Country, City**

**O3 Mean**

**O3 1st Max Value**

**O3 1st Max Hour**

**O3 AQI**

**CO 1st Max Hour**

**CO AQI**

**SO2 Mean**

**SO2 1st Max Value**

**SO2 1st Max Hour**

**SO2 AQI**

**NO2 Mean**

**NO2 1st Max Value**

**NO2 1st Max Hour**

| **NO2 AQI** |
| --- |
|  |

**MECE CHART**

**Unnamed Date Address**

**Month-wise Year-Wise State Country City**

**O3 CO SO NO**

**Mean Max Mean Mean**

**Max Value AQI Max Value Max Value**

**Max Hour Max Hour Max Hour**

**AQI AQI AQI**

**Based on MECE breakdown 15 problem statement that encapsulate the key issue.**

**Certainly! Here's a set of 15 problem statements based on a MECE (Mutually Exclusive, Collectively Exhaustive) breakdown, which covers a range of issues or areas for improvement in a dataset:**

**1. Data Completeness**

* **The dataset contains significant missing values across key variables, leading to potential biases in analysis.**

**2. Data Accuracy**

* **Inaccurate or inconsistent data entries are present, causing discrepancies in critical data points such as timestamps and numeric values.**

**3. Data Consistency**

* **Multiple representations of the same data exist (e.g., different formats for dates or addresses), leading to difficulties in data merging and analysis.**

**4. Data Redundancy**

* **There are multiple instances of duplicated records in the dataset, affecting data integrity and increasing processing time.**

**5. Data Granularity**

* **The dataset lacks sufficient detail for certain variables, limiting the ability to conduct granular analysis on specific subgroups.**

**6. Data Formatting**

* **Several variables have inconsistent formatting, such as mixed data types (e.g., numbers stored as text), making analysis and integration more challenging.**

**7. Data Labeling**

* **The dataset includes incorrectly or ambiguously labeled categories or variables, making it difficult to interpret or classify the data accurately.**

**8. Outlier Detection**

* **Extreme outliers in certain fields are not properly flagged, potentially skewing analysis and leading to incorrect conclusions.**

**9. Variable Relevance**

* **Several variables in the dataset are irrelevant or redundant, which adds unnecessary complexity and noise to analysis.**

**10. Data Timeliness**

* **The dataset is outdated or lacks real-time information, reducing its relevance for current decision-making or predictions.**

**11. Data Completeness in Key Variables**

* **Key variables necessary for predictive modeling or business insights have low completion rates, impacting model accuracy and decision-making.**

**12. Data Alignment**

* **Data from different sources or systems are not aligned correctly, leading to misalignment between corresponding variables and making integration efforts inefficient.**

**13. Bias in Sampling**

* **The dataset exhibits bias due to an imbalanced sample, such as overrepresentation or underrepresentation of certain groups, affecting generalizability.**

**14. Data Accessibility**

* **The dataset is stored in formats or systems that are difficult to access or integrate, reducing the effectiveness of analysis and reporting.**

**15. Data Scalability**

* **As the dataset grows, issues related to storage and computational efficiency arise, leading to slower query responses and data processing challenges.**

**These problem statements cover a wide range of issues related to data quality, consistency, and utility, offering a comprehensive view of potential areas for improvement in a dataset.**

****