1. **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. I can access it using python Pandas Library.**
2. **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**
3. # Column Non-Null Count Dtype
4. **--- ------ -------------- -----**
5. 0 Unnamed: 0 665414 non-null int64
6. 1 Date 665414 non-null object
7. 2 Address 665414 non-null object
8. 3 State 665414 non-null object
9. 4 County 665414 non-null object
10. 5 City 665414 non-null object
11. 6 O3 Mean 665414 non-null float64
12. 7 O3 1st Max Value 665414 non-null float64
13. 8 O3 1st Max Hour 665414 non-null int64
14. 9 O3 AQI 665414 non-null int64
15. 10 CO Mean 665414 non-null float64
16. 11 CO 1st Max Value 665414 non-null float64
17. 12 CO 1st Max Hour 665414 non-null int64
18. 13 CO AQI 665414 non-null float64
19. 14 SO2 Mean 665414 non-null float64
20. 15 SO2 1st Max Value 665414 non-null float64
21. 16 SO2 1st Max Hour 665414 non-null int64
22. 17 SO2 AQI 665414 non-null float64
23. 18 NO2 Mean 665414 non-null float64
24. 19 NO2 1st Max Value 665414 non-null float64
25. 20 NO2 1st Max Hour 665414 non-null int64
26. 21 NO2 AQI 665414 non-null int64
27. 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** |
| --- |
|  |

****