- ## Data Acquisition :

  As an API I used meersens that contains Air Data Quality API which is designed to let you retrieve and integrate environmental data seemlessly into your own services and products.

  The air quality API lets you request air quality data for a specific location. Retrieved information includes pollutant concentration levels, data confidence, comparison with W.H.O applicable standards & norms and generic health recommendations.
  Major pollutants are available, including:

  CO: Carbon Monoxide (identifier co)
  NO2: Nitrogen Dioxide (identifier no2)
  O3: Ozone (identifier o3)
  PM10: Fine Particulate Matter 10 (identifier pm10)
  PM2.5: Fine Particulate Matter 2.5 (identifier pm25)
  SO2: Sulphur Dioxide (identifier so2)

- ## Data Cleaning

➔ **Median for Numerical Values :**
I used the median for :

Resilience to Outliers: The median is the middle value in a sorted list of numbers, which means it is less affected by outliers than the mean (average). Outliers can skew the mean, making it an unreliable measure for central tendency when your data contains extreme values. The median, on the other hand, remains stable because it is not influenced by the magnitude of outliers.

Robust Central Tendency: When imputing missing values, you want to replace them with a value that represents the central tendency of your data. The median provides a robust estimate of this central tendency, especially in skewed distributions where the mean may not accurately represent the majority of the data.

Maintains Distribution Shape: Imputing missing values with the median helps to preserve the shape of the distribution of your data. This is important for ensuring that statistical analyses and machine learning models trained on this data are not biased by extreme values.

➔ **Mode (Most Frequent Value) for Categorical Values**
I used Most Frequent Value for :
Most Common Value: Categorical data represents distinct groups or categories, and the mode is the value that occurs most frequently in the dataset. Using the mode to fill missing values ensures that the imputed value is the one that is most

representative of the existing data, maintaining consistency with the observed categories.

No Arithmetic Mean: Unlike numerical data, categorical data does not have a meaningful arithmetic mean. The mode provides a practical solution for filling missing values by selecting the most common category, which aligns with the nature of categorical data.

Preserving Distribution: By imputing with the mode, you preserve the distribution of categorical data, avoiding the introduction of new categories or skewing the existing distribution. This helps in maintaining the representativeness of the data and ensuring that categorical variables remain informative for subsequent analyses.

- ## Data Transformation:

  I use some aggregation like (total_value which means the total value of pollutants in each moroccan city …)

  ETL pipeline : as pipeline to automate the processes I use airflow which make to run codes on a certain time for our case I make it to run with a daily method in this format :

  Retrieve Data from meersens API **—-->** cleaning and transformation **—----->** Loading in mysql database in relational tables