**AIR QUALITY ANALYSIS AND PREDICTION IN TAMIL NADU**

**INTRODUCTION:**

Technological advancements lead to the emissions of air pollutants over the decades. Major concerns in industrial cities which experience air pollution, can be harmful not only for the environment but also for human health. Due to this urban resident are more likely to live in less polluted neighbourhood to avoid the health impact of air pollution. Atmospheric pollution can be classified into three types based on the source mobile, stationery and area sources. Mobile sources are due to the motor vehicles, airplanes, locomotives and other engines and equipment that are able to move to different locations. Stationary sources include foundries, fossil fuel burning, food processing plants, power plants, refineries and other industrial sources. Area sources is caused by certain local actions. Air pollution can be caused due to the pollutants which are emitted directly from a source or which are not directly emitted as such. It can result in the degradation of ambient air quality in the industrial cities also daily exposure of people to air pollution results in diseases like asthma, wheezing, and bronchitis.

**DATASET:**

**the link used for dataset**

**https://tn.data.gov.in/resource/location-wise-daily-ambient-air**

**quality-tamil-nadu-year-2014**

**COLUMNS USED:**

1. Standard code
2. Sampling Date
3. State
4. City/Town/Village
5. Location of agency
6. Type of location
7. SO2
8. NO2
9. RSPM/PM10

10.PM2.5

**Libraries :**

The essential libraries used in this project are :

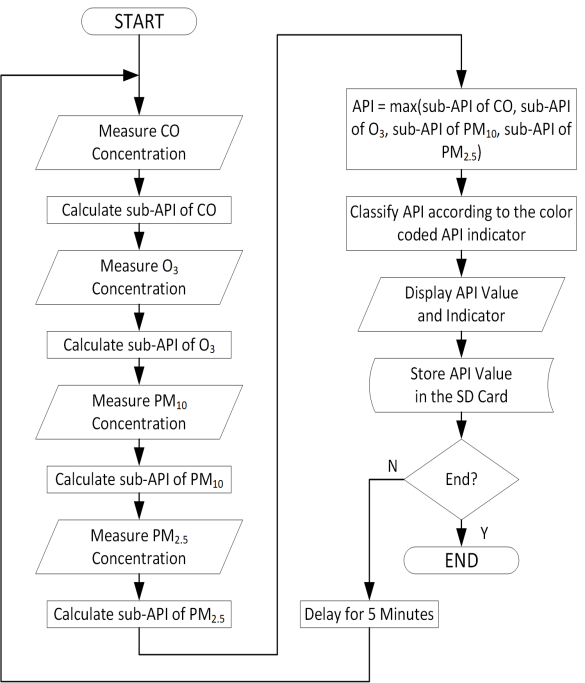
1. Importing OS
2. numpy and Pandas libraries
3. Matplotlib
4. Seaborn

**Visualization:** Visualize the customer segments using techniques like scatter plots, bar

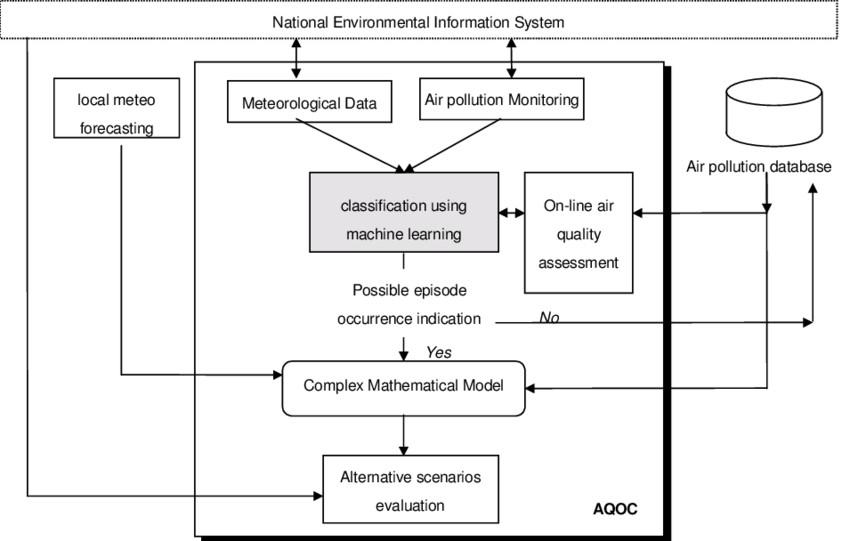
charts, and heatmaps. Interpretation: Analyze and interpret the characteristics of each customer

segment to derive actionable insights for marketing strategies.

**Physical data flow diagram:**



**Data flow diagram:**



**AQI:**

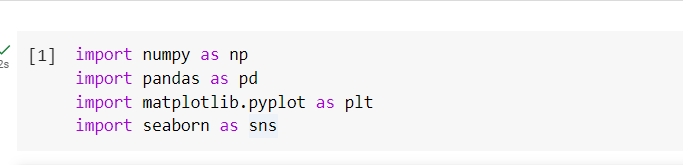
The air quality index is an index for reporting air quality on a daily basis. In other words, it is a measure of how air pollution affects one’s health within a short time period. The AQI is calculated based on the average concentration of a particular pollutant measured over a standard time interval. Generally, the time interval is 24 hours for most pollutants, and 8 hours for carbon monoxide and ozone. We can see how air pollution is by looking at the AQI.



**Importing the libraries:**

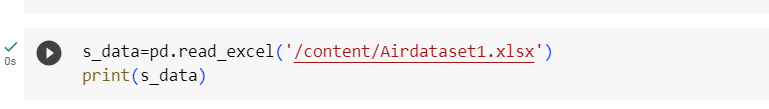
Import three basic libraries which are very common in machine learning and will be used every time you train a model

* **NumPy:** it is a library that allows us to work with arrays and as most machine learning models work on arrays NumPy makes it easier
* **matplotlib:** this library helps in plotting graphs and charts, which are very useful while showing the result of your model
* **Pandas:** pandas allows us to import our dataset and also creates a matrix of features containing the dependent and independent variable.
* **Seaborn :** seaborn helps in visualization



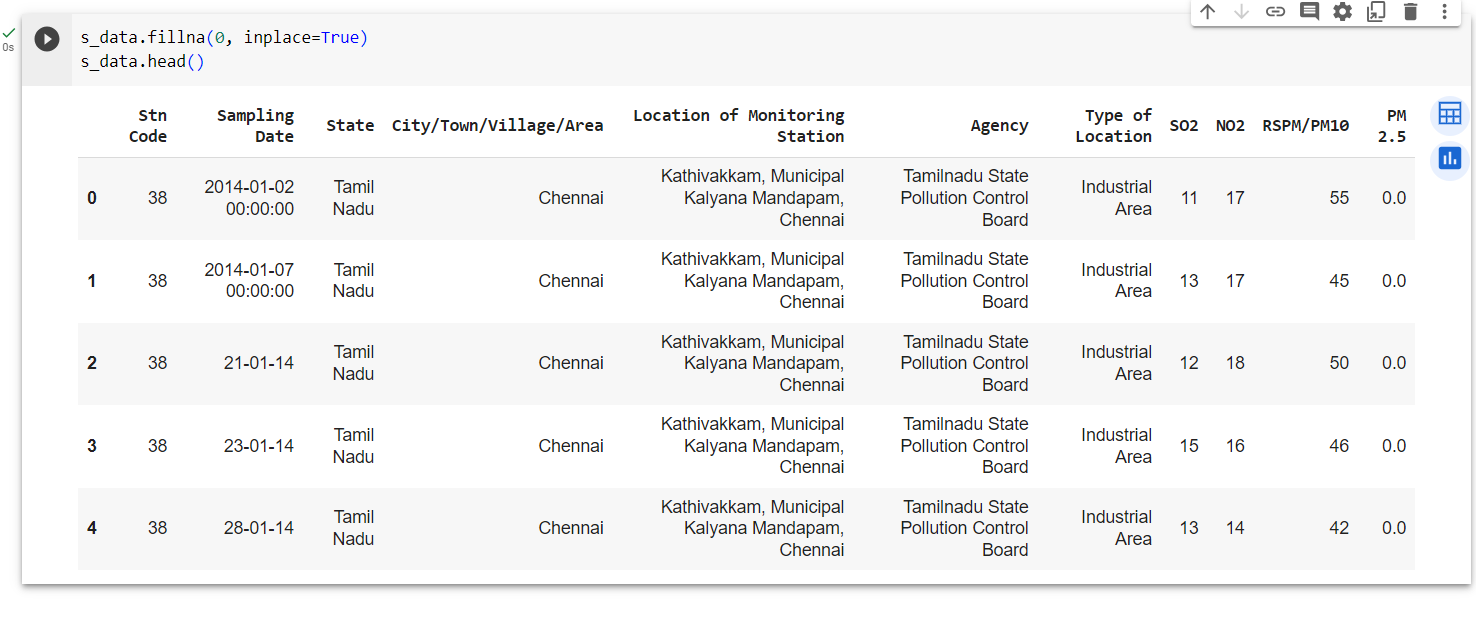
**Load the dataset: (DATASET 1)**

* Data sets are available in .csv format. A CSV file stores tabular data in plain text.
* Each line of the file is a data record. We use the read.csv method of the pandas library to read a local CSV file as a data frame.
* Load our customer data from the CSV file



**Head() Function:**

* The head() function is used to get the first n rows.
* This function returns the first n rows for the object based on position.
* It is useful for quickly testing if your object has the right type of data in it.
* If the value of the n is not assigned it returns a default value of first 5 rows

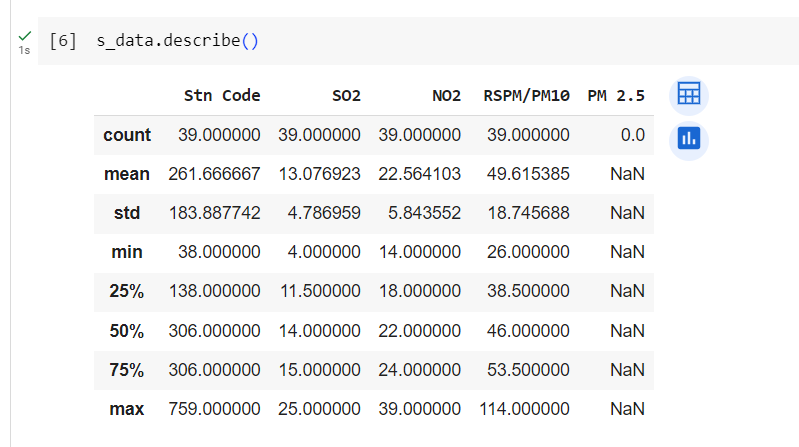


**Describe Function:**

* The describe() function in pandas, a popular Python data analysis library, is used to generate summary statistics of a Series.

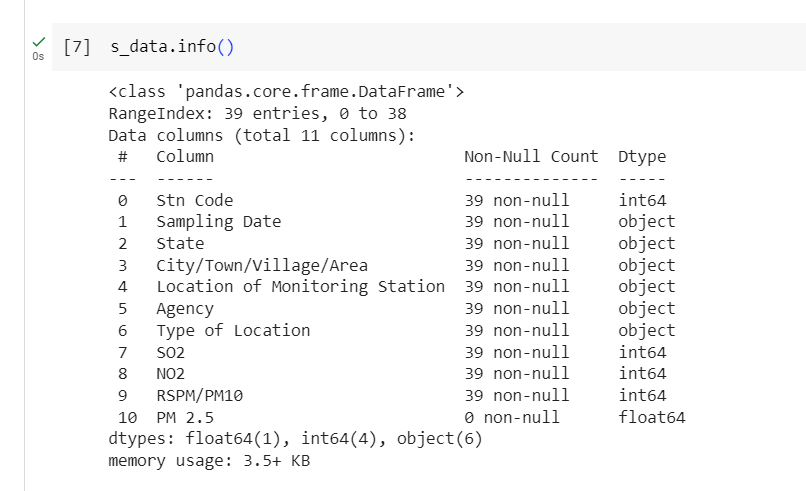
It provides a quick overview of the key statistics for numerical data in the dataset, including:

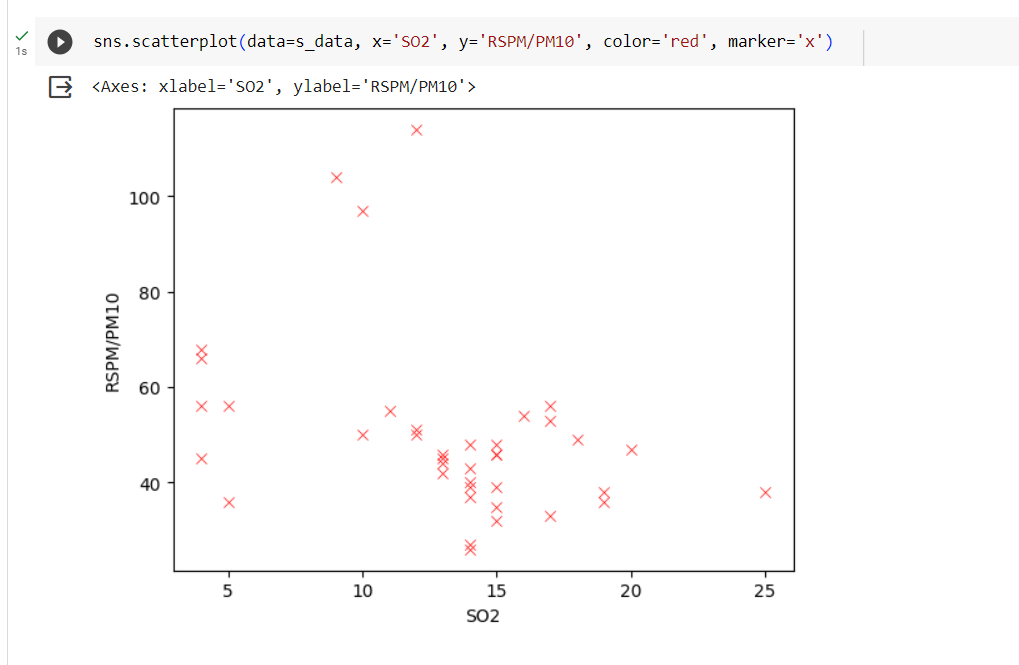
* **Count:** The number of non-null values.
* **Mean:** The average of the values.
* **Standard Deviation (std):** A measure of the spread or dispersion of the data.
* **Minimum:** The minimum value in the dataset.
* **25th Percentile (25%):** The value below which 25% of the data falls (the first quartile).
* **Median (50% or the 2nd quartile):** The middle value when the data is sorted.
* **75th Percentile (75%):** The value below which 75% of the data falls (the third quartile).
* **Maximum:** The maximum value in the dataset.



**Info() Function:**

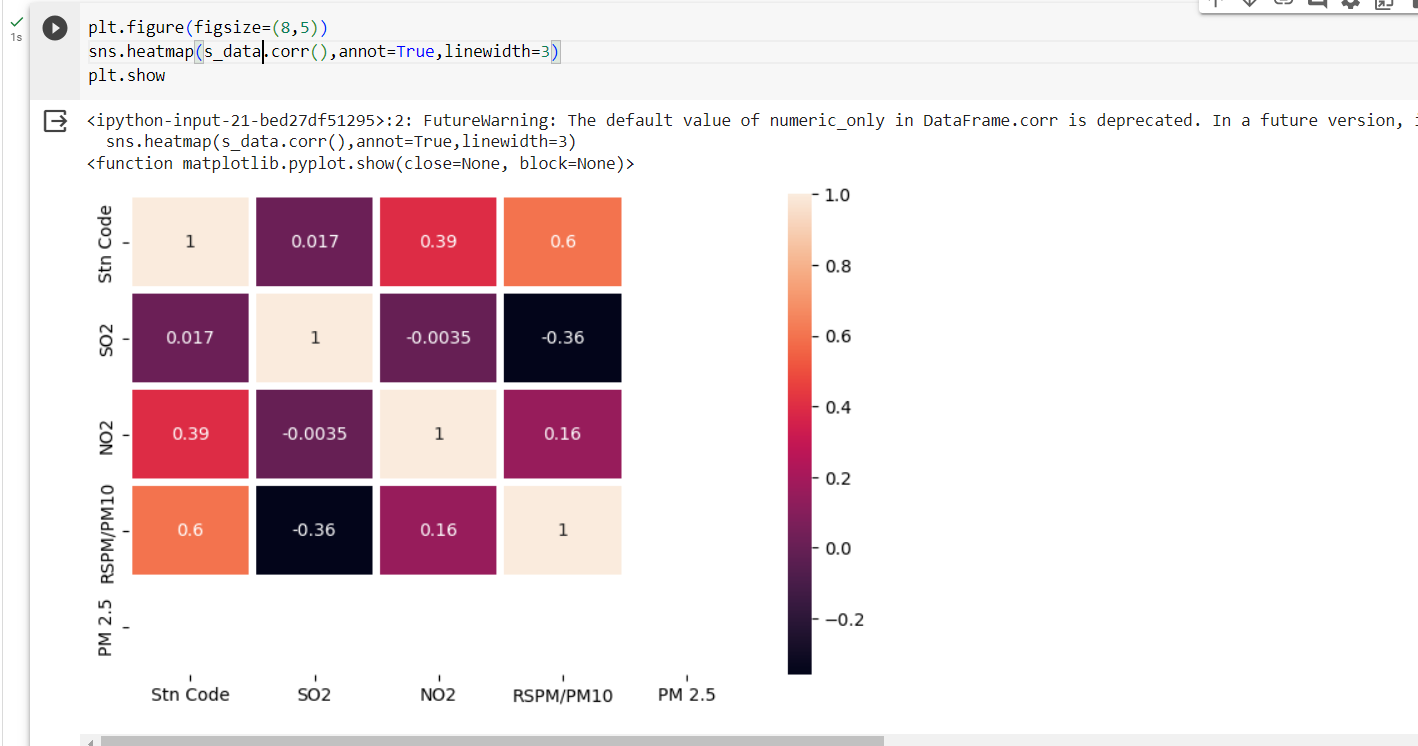
* The info() method prints information about the data frame.
* The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values).





**Corelation:**

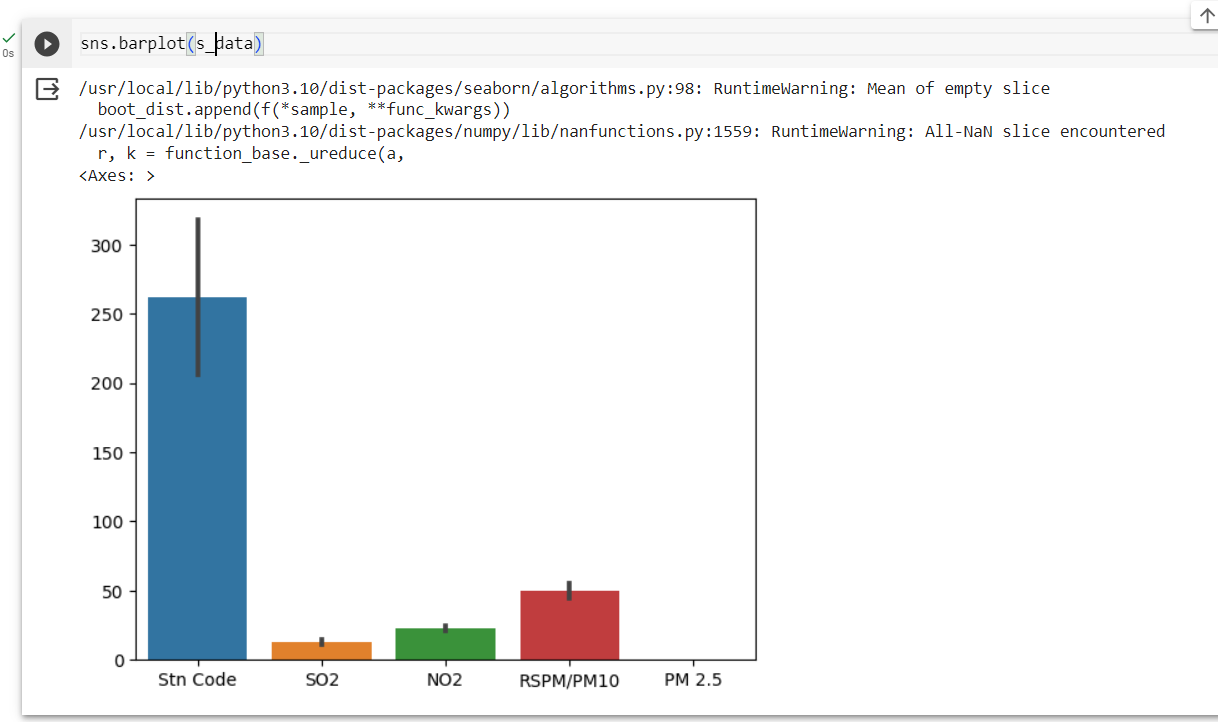
* Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate in relation to each other.
* Correlation describes the relationship between variables. It can be described as either strong or weak, and as either positive or negative.



**Dataset columns :**

For example the Grades column, we could simply use the loc function and specify the name of the column in order to retrieve it.





**Missingno Function:**

* "missing no" is a Python library used for visualizing and analyze missing data in a dataset.
* It provides various visualization tools to quickly understand and identify missing values in your data, allowing you to make informed decisions on how to handle or impute missing data.

