GOVERNMENT COLLEGE OF TECHNOLOGY, COIMBATORE-13

PROJECT NAME: Air Quality Assessment TN

Team Members:

1. Kokulavenkat K
2. Keerthana R
3. Kayalvizhi T
4. Karunasri A

Phase 2: Innovation

PROJECT DESCRIPTION:

Gather air quality data from monitoring stations across Tamil Nadu. This data may include measurements of pollutants like PM2.5, PM10, nitrogen dioxide (NO2), sulphur dioxide (SO2), ozone (O3), carbon monoxide (CO), and other relevant parameters.

Ensure the data is structured, accurate, and includes timestamps for each measurement. Aggregate data if multiple stations provide data for the same location. To visualise the data including line charts, bar graphs, heatmaps, and scatter plots and compare different pollutants. Geographic visualizations using tools like GIS software or Python libraries like Folium to map air quality across different regions in Tamil Nadu.

Consider using color-coding to represent air quality levels and station locations on the map. Calculate AQI based on the collected data and the relevant formula for Tamil Nadu. To visualise the air quality find how different pollutants relate to each other and apply time series analysis techniques, autoregressive integrated moving average (ARIMA) or prophet for forecasting air quality.

DATASET AND ITS DETAIL:

The dataset from https://tn.data.gov.in is available for various years, including 1998, with the latest version being from 2014. The dataset was originally published on October 17, 2016, and it was most recently updated on December 11, 2021.

DETAILS ABOUT COLUMNS:

- Stn Code: It serves as a unique code or identifier for different monitoring stations in Chennai, as mentioned. These codes, such as 38, 72, 71, 766, 765, and 764, are used to distinguish and reference specific monitoring stations. Each code corresponds to a particular monitoring station in the area, allowing for easy data management and differentiation between different locations where air quality measurements are taken.
- Sampling Date: This column represents the date on which air quality measurements were taken. It includes dates ranging from January to December in the year 2014.
- State: Indicates the state where the monitoring station is located, which is Tamil Nadu in this case.
- City: It appears that the monitoring stations are located in various areas, including "Kathivakkam, Municipal Kalyana Mandapam" in Chennai, as well as in other cities such as Coimbatore, Madurai, Salem, Thoothukudi, Trichy, and Mettur.
- Location of Monitoring Station: This field offers specific details about the monitoring station's exact placement, often with reference to municipal buildings. This information helps in precisely identifying where the monitoring station is situated and includes locations like Thiruvottiyur, Kathivakkam, and Thiyagaraja Nagar.
- Agency: Refers to the organization or agency responsible for conducting air quality monitoring. In this dataset, it's the "Tamilnadu State Pollution Control Board."
- Type of Location: This column categorizes the type of environment where the monitoring station is located, and in this case, it is classified as an "Industrial Area." This categorization provides context for the kind of air quality data collected.
- SO2 (Sulfur Dioxide): This column contains data on the concentration of sulfur dioxide (SO2) in the air, measured in unspecified units (possibly µg/m³ or ppm). SO2 is a common air pollutant resulting from industrial and combustion processes.
- NO2 (Nitrogen Dioxide): This column contains data on the concentration of nitrogen dioxide (NO2) in the air, measured in unspecified units. NO2 is another common air pollutant often associated with vehicle emissions and industrial activity.
- RSPM/PM10 (Respirable Suspended Particulate Matter/Particulate Matter 10): This column provides data on the concentration of particulate matter with a diameter of 10 micrometers or less. It is commonly expressed in units such as µg/m³.
- PM 2.5 (Particulate Matter 2.5): This column contains data on the concentration of finer particulate matter with a diameter of 2.5 micrometers or less, expressed in unspecified units. PM 2.5 is associated with various health and environmental

impacts.It's important to note that "NA" appears in the last column for each entry, indicating that there might be missing data for PM 2.5 in this dataset.

DETAILS OF LIBRARIES TO BE USED:

Here are some specific details about libraries that can be used for an air quality analysis project in Tamil Nadu:

- Pandas: Pandas is a fundamental library for data analysis and manipulation. You can use it to load, clean, and process air quality data in various formats such as CSV, Excel, or databases. This library is essential for data preprocessing.
- Matplotlib and Seaborn: Matplotlib and Seaborn are great choices for creating visualizations of air quality data. You can generate time series plots, scatter plots, bar charts, and maps to understand the trends and patterns in air quality over time and across locations in Tamil Nadu.
- Folium: For geospatial visualization, Folium is useful. You can create interactive maps that display air quality data by location. This can be particularly helpful for pinpointing areas with the worst air quality.
- NumPy: NumPy is essential for numerical operations. You can use it for various calculations, such as statistical analysis, averaging air quality values, and more.
- Scikit-Learn: If you want to develop predictive models for air quality forecasting or classification, Scikit-Learn can be helpful. You can train machine learning models to predict air quality levels based on various factors.
- Statsmodels: Statsmodels is useful for statistical analysis. You can perform hypothesis testing, regression analysis, and time series analysis to gain insights into the factors affecting air quality in Tamil Nadu.
- PyAirQuality: PyAirQuality is a Python library designed for air quality analysis. It may provide specific tools and functions tailored to air quality data analysis. You can check if it has features relevant to Tamil Nadu.
- Tamil Nadu Pollution Control Board (TNPCB) API: If TNPCB or other relevant agencies provide air quality data through APIs, you can use these to fetch real-time or historical data for your analysis.
- SQL and Databases: If you're dealing with large datasets, consider using a database like PostgreSQL or MySQL to store and manage the data efficiently.

WAY TO DOWNLOAD THE LIBRARIES:

1.Open a Command Prompt or Terminal: Depending on your operating system (Windows, macOS, or Linux), open a command-line interface.

2. Install Python: Make sure you have Python installed on your system. You can download Python from the official website (https://www.python.org/downloads/) and follow the installation instructions.

3.Install Libraries with Pip: Use `pip` to install the required libraries. For example, to install Pandas, you would type

```
pip install pandas
```

To install other libraries, replace "pandas" with the name of the library you want to install. For example

```
pip install matplotlib
pip install seaborn
pip install folium
pip install numpy
pip install scikit-learn
pip install statsmodels
```

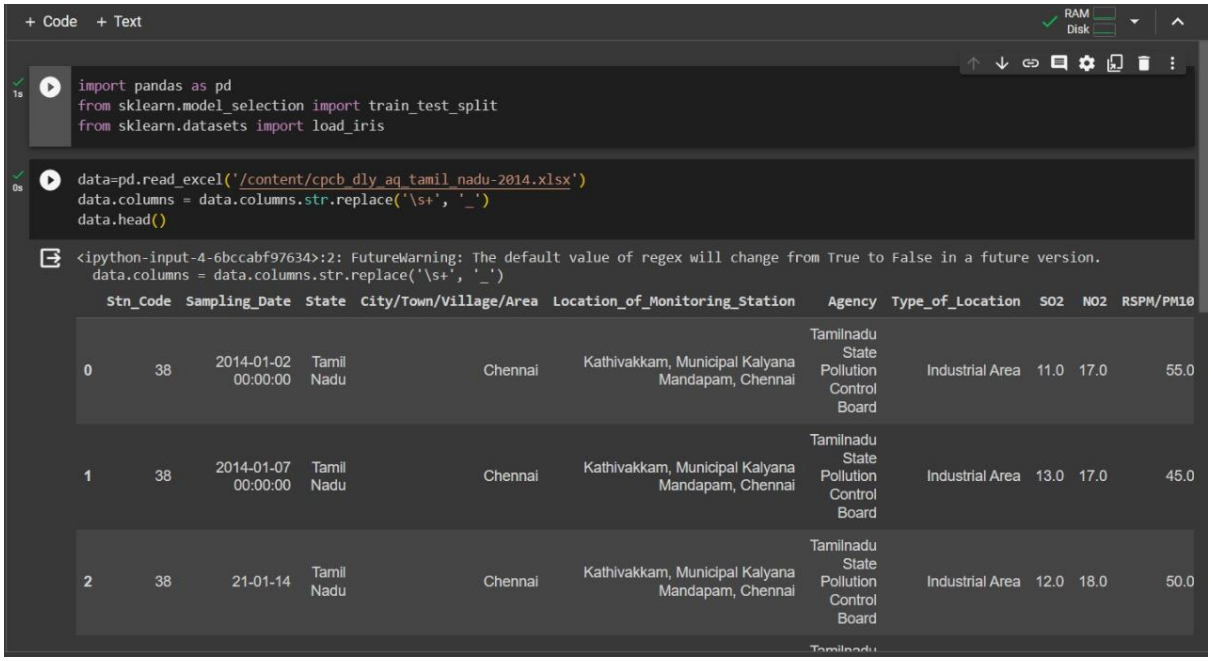
4.Additional Considerations:

- Some libraries might have dependencies, and `pip` will automatically install them for you.
- If you're working on a specific project, consider creating a virtual environment to manage the library dependencies for that project. You can use `venv` or a tool like `conda` if you're using Anaconda.

TRAIN AND TEST:

As we work with datasets, a machine learning algorithm works in two stages. We usually split the data around 20%-80% between testing and training stages. Under supervised learning, we split a dataset into a training data and test data in Python ML.

Prerequisites for Train and Test Data and loading the dataset



```
+ Code  + Text                                                    RAM Disk

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.datasets import load_iris

data=pd.read_excel('/content/cpcb_dly_aq_tamil_nadu-2014.xlsx')
data.columns = data.columns.str.replace('\s+', '_')
data.head()

<ipython-input-4-6bccabf97634>:2: FutureWarning: The default value of regex will change from True to False in a future version.
  data.columns = data.columns.str.replace('\s+', '_')
```

	Stn_Code	Sampling_Date	State	City/Town/Village/Area	Location_of_Monitoring_Station	Agency	Type_of_Location	SO2	NO2	RSPM/PM10
0	38	2014-01-02 00:00:00	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0
1	38	2014-01-07 00:00:00	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0
2	38	21-01-14	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0

ANALYSING AND VISUALISATION:

Creating a heatmap for air quality data can provide a clear visualization of the variations in SO2 and NO2 content across different cities and sampling dates. Since as mentioned that the city column has repeated city names, we may want to aggregate the data before creating the heatmap. Here's a step-by-step guide:

Aggregate Data:

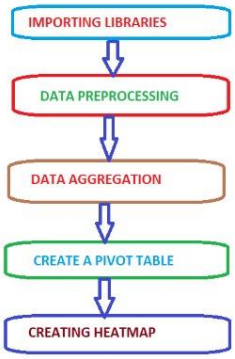
Calculate the average SO2 and NO2 content for each city on each sampling date. This will help in reducing repeated city names and provide a representative value for each city-date combination.

Pivot Table:

Create a pivot table where cities are along the rows, sampling dates along the columns, and the values are either average SO2 or NO2 content.

Heatmap:

Use a heatmap to visualize the data. You can use tools like Python with libraries such as Matplotlib and Seaborn or other data visualization tools like Tableau or Excel.



The code preprocesses air quality data by converting dates and creating a formatted date column. It then aggregates the data by calculating mean 'SO2' and 'NO2' values for different locations and dates. Finally, it visualizes the 'SO2' and 'NO2' levels across locations and dates using a heatmap with color-coded values and a color bar, making air quality trends easily interpretable.