Air Quality Analysis and Prediction in Tamil Nadu

Phase 3

17-10-2023

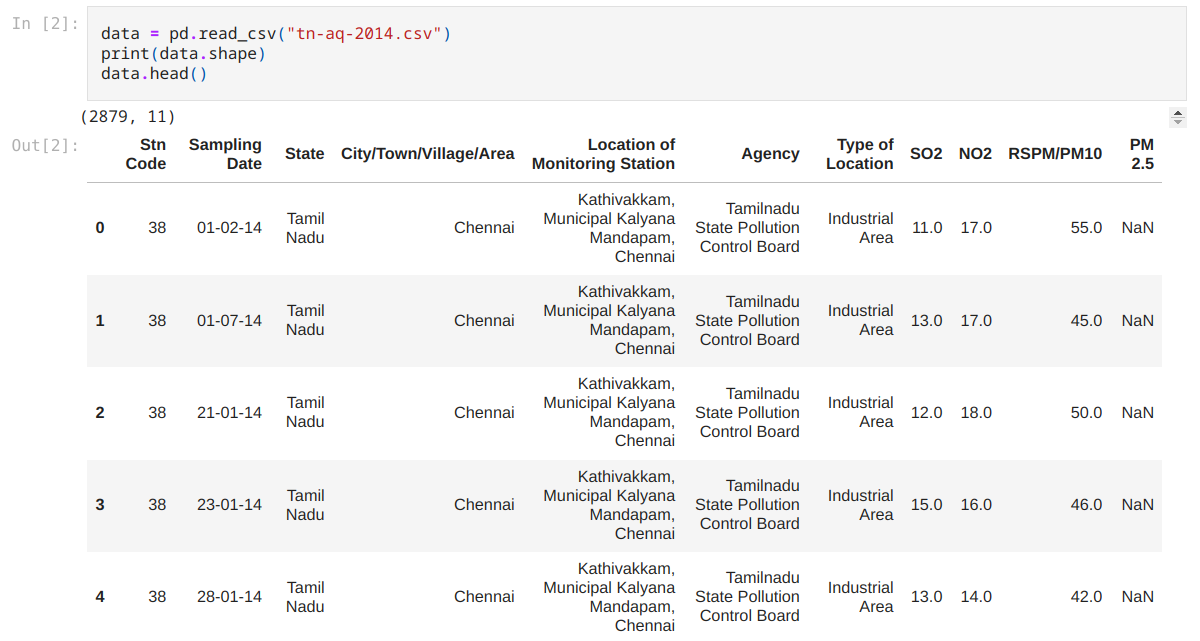
Problem Definition:

The project aims to analyze and visualize air quality data from monitoring stations in Tamil Nadu. The objective is to gain insights into air pollution trends, identify areas with high pollution levels, and develop a predictive model to estimate RSPM/PM10 levels based on SO2 and NO2 levels.

**Data Preprocess:**

**1) Data Loading and Initial Exploration:**

* Load the dataset from the CSV file using the Pandas library.
* Display the shape of the dataset, showing the number of rows and columns.
* Display the first few rows of the dataset to inspect its structure



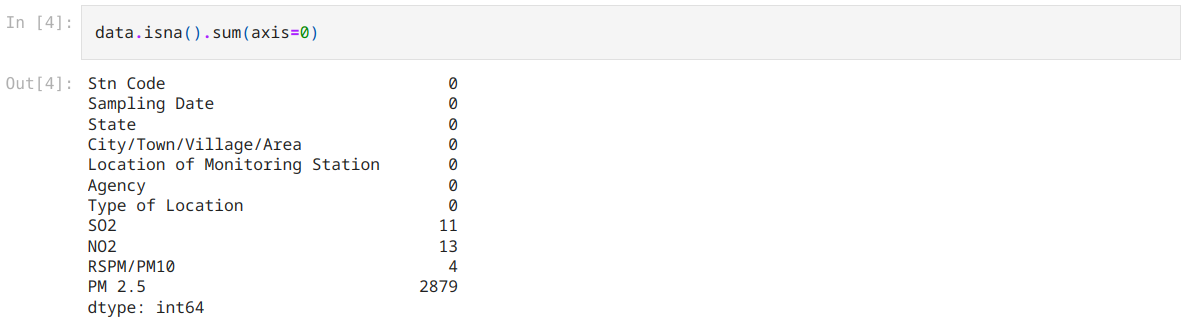
**2) Creating a DataFrame for Monitoring Stations:**

* Create a new DataFrame `stn\_df` by selecting and dropping duplicates of the columns "Stn Code" and "Location of Monitoring Station" from the original dataset.
* Rename the columns in stn\_df as "code" and "name."
* Adding station coordinates ("coord") using Bard AI with Google maps extension. By just giving the list of address of the station from the dataset to the Bard AI and prompting a python list format we can obtain the coordinates of the locations.



**3) Checking Missing Values:**

Checking for missing values in the dataset and display the count of missing values for each column.

****

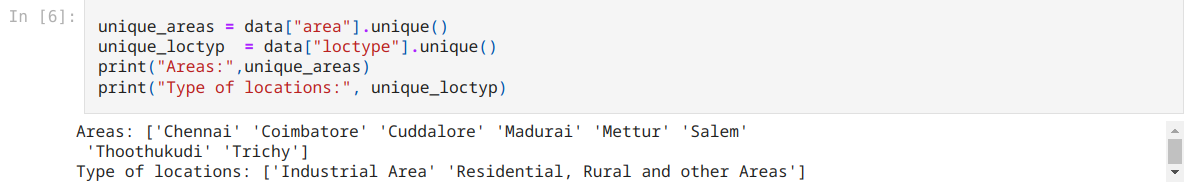
**4) Cleaning and Formating the Dataframe:**

* Removing the columns "Stn Code," "State," "Agency," and "PM 2.5" from the dataset as they may not be needed for the analysis.
* Remove rows with missing values (NaN) in any column.
* Convert the "Sampling Date" column to a datetime data type.
* Rename the columns to more descriptive names.



**5) Exploration of Unique Areas and Location Types:**

Identify the unique values in the "area" and "loctype" (Type of Location) columns and print them.



**6) Data Grouping:**

Now that we identified unique areas, We grouping the data by the "area" column, creating a dictionary where each key is an area name, and each value is a DataFrame containing data specific to that area. This can later used to analyse every single area individually.

  
Full code:

data\_by\_area = {area\_name : area\_data.drop("area",axis=1).sort\_values("date") for area\_name, area\_data in data.groupby("area") }

**Conclusion:**

In this data preprocessing step, we successfully prepared the "tn-aq-2014" dataset for further analysis.