**Air Q Assessment TN**

**Phase 2 Submission Document.**

**Project Title:** Air Q Assessment TN

**Project Definition:**

The project aims to analyse 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. This project involves defining objectives, designing the analysis approach, selecting visualization techniques, and creating a predictive model using Python and relevant libraries.

**1.Project Objective:**

* Analyze air quality data from monitoring stations in Tamil Nadu.
* Gain insights into air pollution trends.
* Identify areas with high pollution levels.
* Develop a predictive model to estimate RSPM/PM10 levels based on SO2 and NO2 levels.

**2.Analysis Approach:**

* Data Collection: Specify how you will gather air quality data, including sources, frequency, and data formats.
* Data Preprocessing: Outline steps for cleaning, handling missing data, and ensuring data quality.
* Exploratory Data Analysis (EDA): Describe techniques for exploring data to uncover trends and patterns.
* Feature Engineering: Detail how you’ll create relevant features for modeling.
* Model Selection: Decide on machine learning algorithms for building the predictive model.y
* Evaluation Metrics: Define how you’ll measure the model’s performance.
* Cross-validation: Plan for validating the model’s performance.
* Hyperparameter Tuning: Discuss methods for optimizing model parameters.

**3.Visualization Techniques:**

* Select appropriate visualization techniques based on the nature of the data:
* Line charts for time series analysis of air quality trends.
* Heatmaps or spatial maps for pinpointing pollution hotspots.
* Scatter plots or correlation matrices to understand relationships between variables.
* Ensure that the visualizations are user-friendly and insightful for stakeholders.

**4.Python and Libraries:**

* Specify the Python libraries you plan to use for data analysis and modeling (e.g., pandas, NumPy, scikit-learn).
* Mention any specific data visualization libraries (e.g., Matplotlib, Seaborn) that will be employed.
* Include any additional libraries required for geospatial analysis if applicable.

**Contents for Project Phase 2:**

To incorporate machine learning algorithms to improve the accuracy of a predictive model, you can follow these steps:

**Data Pre-processing:**

* Collect and clean your data, handling missing values and outliers.
* Encode categorical variables and scale numerical features if necessary.

**Split the Data:**

* Divide your dataset into training, validation, and test sets.
* Select machine learning algorithms suitable for your problem (e.g., regression, classification, clustering).

**Data source:**

Dataset Link:

(<https://tn.data.gov.in/resource/location-wise-daily-ambient-air-quality-tamil-nadu-year-2014>)

**Program:**

import numpy as np # linear algebra

import pandas as pd # data processing, CSV file I/O (e.g. pd.read\_csv)

import os

for dirname, \_, filenames in os.walk(‘/input/cpcb\_dly\_aq\_tamil\_nadu\_2014.csc’):

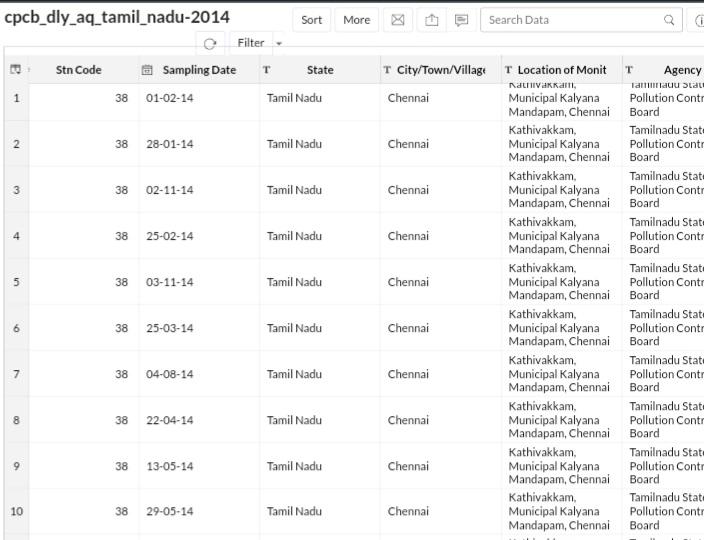
for filename in filenames:

print(os.path.join(dirname, filename))

df=pd.read\_csv(‘../input/input/cpcb\_dly\_aq\_tamil\_nadu\_2014.csc’,encoding=’cp1252’)

df.head()

**Out:**



df.shape

**Out:**

(2879, 11)

df[df[‘state’]==’Tamil Nadu’][‘location’].unique()

**Out:**

array([‘Madras’, ‘Turicorin’, ‘Tuticorin’, ‘Coimbatore’, ‘Madurai’,

‘Salem’, ‘Chennai’, ‘Thoothukudi’, ‘Trichy’, ‘Mettur’, ‘Cuddalore’],

Dtype=object)

df[‘location’]=df[‘location’].replace((‘Madras’, ‘Turicorin’, ‘Thoothukudi’, ‘Mettur’),(‘Chennai’, ‘Tuticorin’, ‘Tuticorin’, ‘Salem’))

df[df[‘state’]==’Tamil Nadu’][[‘location’, ‘rspm’]].groupby([‘location’]).agg(‘mean’).sort\_values(‘rspm’,

ascending=False).style.background\_gradient(cmap=’cool’)

**Out:**

 #Let’s check which type of area causing more pollution in tuticorin

df[df[‘location’]==’Tuticorin’][[‘type’, ‘rspm’]].groupby([‘type’]).agg(‘mean’).sort\_values(‘rspm’,

ascending=False).style.background\_gradient(cmap=’inferno ’)

**Out:**

****

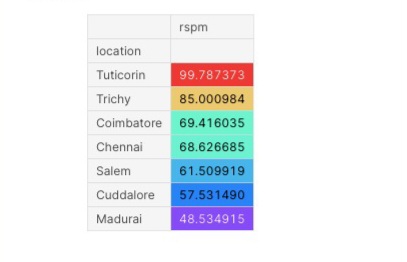
#As we know Industrial areas causing more pollution let’s check which industrial areas causing more pollution in tamilnadu

df[(df[‘type’]==’industrial’)|(df[‘state’]==’Tamil Nadu’)][[‘location’,

‘rspm’]].groupby([‘location’]).agg(‘mean’).sort\_values(‘rspm’,

ascending=False).style.background\_gradient(cmap=’rainbow’)

**Out:**



**Data visualization :**

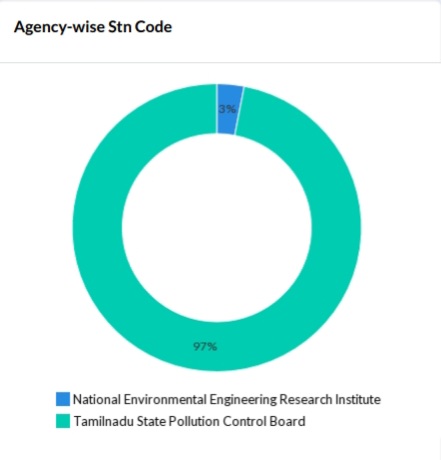
Data visualization is a crucial aspect of data analysis and communication. It involves creating graphical representations of data to help people understand patterns, trends, and insights in the data.

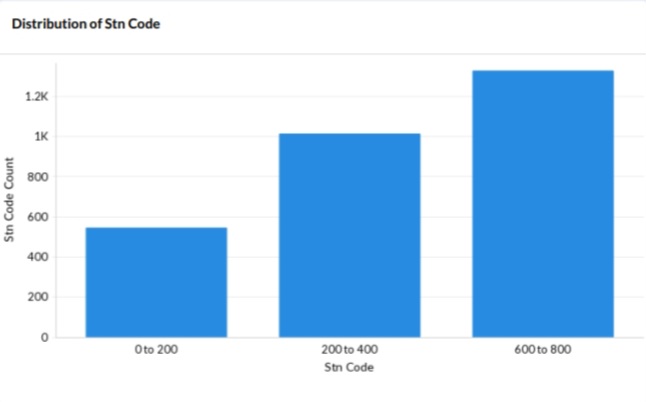
**Types of Visualizations:**

There are various types of data visualizations, including bar charts, line graphs, scatter plots, heatmaps, pie charts, histograms, and more. The choice of visualization depends on the data and the insights you want to convey.

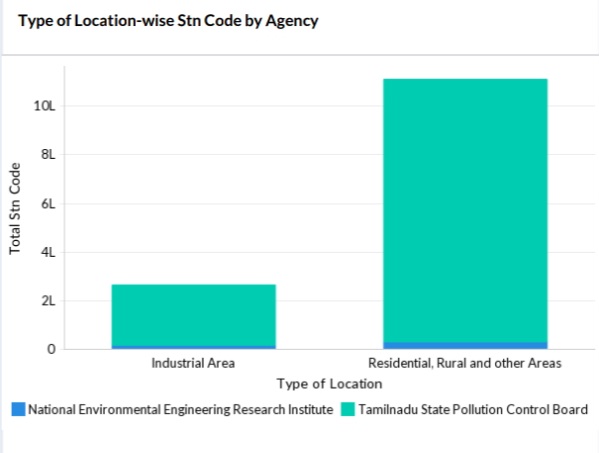
**Air Q Assessment TN- Data Visualization:**

**Agency wise Stn-Code(Station-Code):**

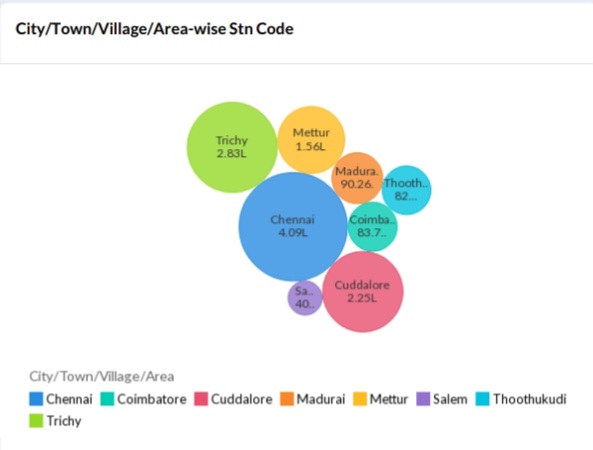
****

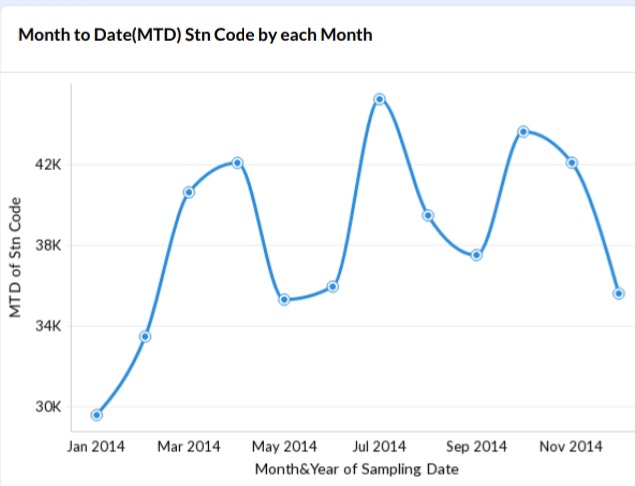
**Distribution of Stn-Code:**

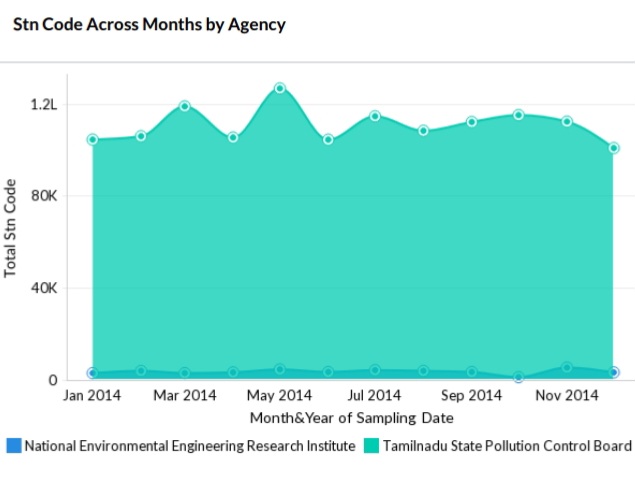
**Type of location wise Stn-Code by Agency:**

****

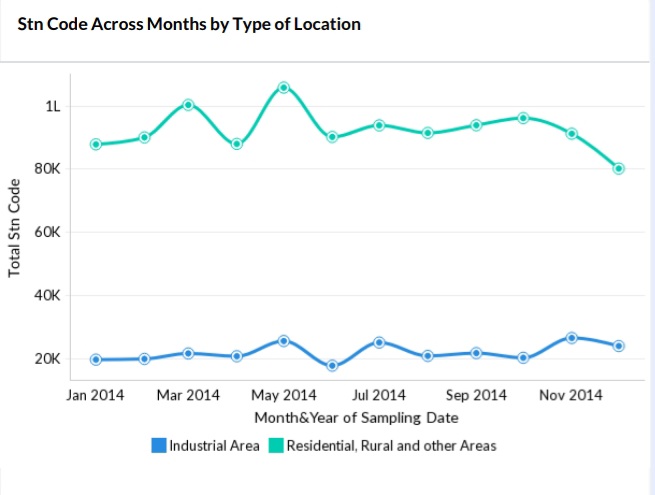
**City/Town/Village/Area-wise Stn-Code:**

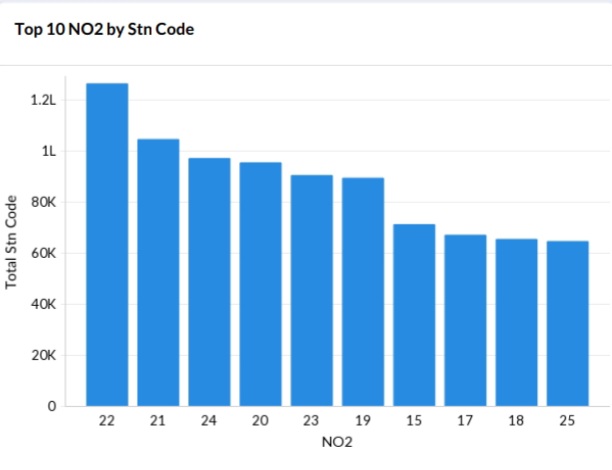


**Month to Date(MTD)Stn-Code by Each Month:**

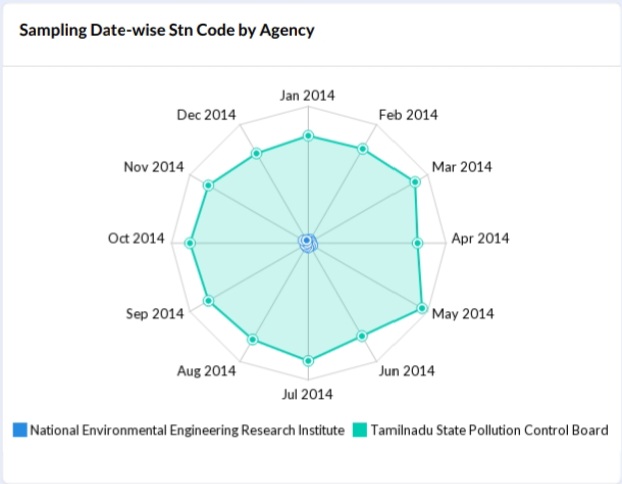
**Stn-Code Across Month by Agency:**

**Stn-Code Across Months by Type of Location:**

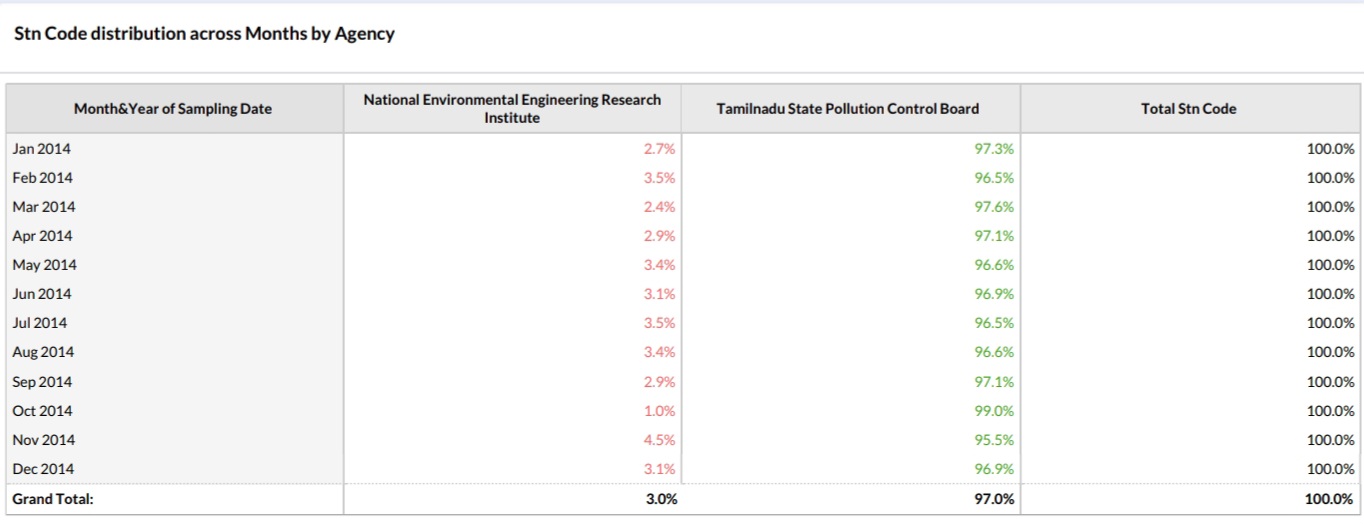


**Top 10 NO2 by Stn-Code:**

**Sampling Data wise Stn-Code by Agency:**



**Stn-Code Distribution Across Months by Agency:**



**Conclusion of Phase-2 Project:**

In the Phase 2 conclusion, we will summarize the data analysis and data visualization of **Air Q Assessment of Tamilnadu** by using the given dataset.

From this project, we innovativly analysis the air quality of Tamilnadu. In future we will develop the project for the given dataset.