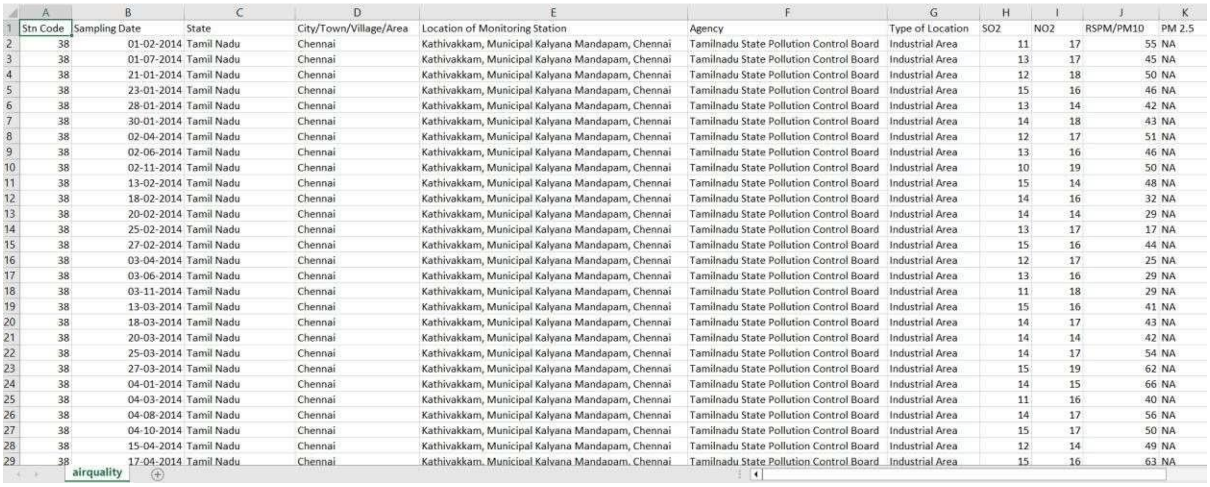
Air Quality Analysis Assessment in Tamil Nadu

Phase 5 : Final Documentation

Data set link : https://tn.data.gov.in/resource/location-wise-daily-ambient-airquality-tamil-nadu-year-2014



	A	B	C	D	E	F	G	H	I	J	K
1	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10	PM 2.5
2	38	01-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11	17	55	NA
3	38	01-07-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13	17	45	NA
4	38	21-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12	18	50	NA
5	38	23-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15	16	46	NA
6	38	28-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13	14	42	NA
7	38	30-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	14	18	43	NA
8	38	02-04-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12	17	51	NA
9	38	02-06-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13	16	46	NA
10	38	02-11-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	10	19	50	NA
11	38	13-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15	14	48	NA
12	38	18-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	14	16	32	NA
13	38	20-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	14	14	29	NA
14	38	25-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13	17	17	NA
15	38	27-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15	16	44	NA
16	38	03-04-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12	17	25	NA
17	38	03-06-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13	16	29	NA
18	38	03-11-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11	18	29	NA
19	38	13-03-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15	16	41	NA
20	38	18-03-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	14	17	43	NA
21	38	20-03-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	14	14	42	NA
22	38	25-03-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	14	17	54	NA
23	38	27-03-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15	19	62	NA
24	38	04-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	14	15	66	NA
25	38	04-03-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11	16	40	NA
26	38	04-08-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	14	17	56	NA
27	38	04-10-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15	17	50	NA
28	38	15-04-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12	14	49	NA
29	38	17-04-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15	16	63	NA

INTRODUCTION:

I can provide a general overview of how you might approach such a project, but I can't provide specific outputs or code implementation due to the complex and detailed nature of such a project. Here's a high-level overview:

Project Objectives:

The objective of the project is to analyze air pollution trends and pollution levels in Tamil Nadu, using historical air quality data. Key goals may include identifying pollution hotspots, understanding temporal trends, and assessing the impact on public health.

Analysis Approach:

1. **Data Collection:**
   Gather historical air quality data for Tamil Nadu, including parameters like PM2.5, PM10, CO, NO2, SO2, O3, and meteorological data.
2. **Data Preprocessing:**
   Clean the data, handle missing values, and standardize formats.
3. **Exploratory Data Analysis (EDA):**
   Conduct EDA to understand the data's distribution, correlations, and outliers.
4. **Temporal Analysis:**
   Analyze pollution trends over time, possibly using time series analysis or trend analysis techniques.
5. **Spatial Analysis:**
   Utilize geographic information systems (GIS) to visualize spatial distribution of pollutants.
6. **Health Impact Assessment:**
   Assess the potential health impacts using available health data.

Visualization Techniques:

1. **Time Series Plots:**
   Visualize pollution trends over time using line graphs.
2. **Heatmaps:**

Show spatial distribution of pollutants on a map of Tamil Nadu.

3. **Bar Charts:**
   Compare pollution levels across cities or regions.
4. **Correlation Matrix:**
   Visualize relationships between different pollutants and meteorological factors.
5. **Geospatial Maps:**
   Use GIS tools to create interactive maps showing pollution levels.

Code Implementation:
For code implementation, you would use programming languages like Python or R. Libraries such as Pandas, Matplotlib, Seaborn, Plotly, and geospatial libraries (e.g., Folium) would be useful. Data analysis and visualization would involve writing scripts to preprocess data, generate plots, and create interactive visualizations.

Example Outputs:
- A time series plot showing the increase in PM2.5 levels over the last decade.
- A heatmap illustrating the spatial distribution of NO2 concentrations in different regions of Tamil Nadu.
- Bar charts comparing air quality indexes of major cities in the state.
- A correlation matrix highlighting relationships between pollutants and weather variables.
- An interactive geospatial map showing pollution levels across different districts.

Insights into Air Pollution Trends:
The analysis would provide insights into how air pollution has changed over time, whether it's increasing or decreasing. You could identify regions with consistently high pollution levels and potential factors contributing to this. Moreover, the correlation analysis may reveal which pollutants are interrelated and how meteorological conditions impact pollution.

Pollution Levels in Tamil Nadu:
The project would help in understanding the current pollution levels in various parts of Tamil Nadu, enabling policymakers and the public to take informed actions to mitigate the health and environmental risks associated with air pollution.

Phase 1 : Problem definition and Design Thinking

Project Description:

Air quality is a critical concern in Tamil Nadu, given its impact on public health, the environment, and overall quality of life. This project aims to develop an advanced system for air quality analytics in Tamil Nadu. By collecting, analyzing, and visualizing air quality data, this project will provide valuable insights to government agencies, researchers, and the public to better understand and manage air pollution in the region.

Design Thinking Approach:

Design thinking is a human-centered problem-solving approach that emphasizes empathy, creativity, and iterative prototyping. Here's a design thinking framework for the "TN Marginal Workers Assessment" project:

1. Empathize:

-Understand the current air quality situation in Tamil Nadu, including sources of pollution, pollution hotspots, and the impact on public health.

-Engage with residents, environmental activists, health experts, and government officials to gather insights into their concerns and priorities related to air quality.

-Conduct surveys and interviews to empathize with the needs and expectations of the community.

2. Define:

- Define the problem by identifying specific challenges and opportunities related to air quality analytics in Tamil Nadu.

- Prioritize the most critical issues, such as inadequate monitoring infrastructure, data accessibility, or public awareness.

- Create user personas for different stakeholders, including policymakers, researchers, and the general public.

3. Ideate:

- Generate creative ideas for improving air quality analytics and addressing the identified challenges.

- Organize ideation workshops and brainstorming sessions with experts in environmental science, data analytics, and technology.

- Encourage cross-disciplinary collaboration to explore innovative solutions.

4. Prototype:

- Develop prototypes or mockups of potential solutions, such as air quality monitoring apps, data visualization tools, or educational materials.
- Create a prototype of an improved monitoring network that covers critical areas across Tamil Nadu.
- Build a user-friendly interface for accessing real-time air quality data.

5. Test:

- Pilot test the prototypes with a diverse group of stakeholders, including citizens, government officials, and researchers.
- Gather feedback on usability, accessibility, and effectiveness.
- Observe how users interact with the prototypes and identify pain points

6. Implement:

- Based on feedback and insights from testing, refine and implement the most promising solutions.
- Collaborate with relevant government agencies, environmental organizations, and tech companies to ensure successful implementation.
- Ensure the solutions are scalable and adaptable to different regions within Tamil Nadu.

7. Evaluate:

- Establish key performance indicators (KPIs) to measure the impact of the implemented solutions.
- Monitor air quality improvements, increased data accessibility, and heightened public awareness.
- Collect data on user engagement and satisfaction.

8. Feedback and Iterate:

- Maintain an open feedback loop with users and stakeholders.
- Continuously gather feedback to make iterative improvements to the solutions.

- Adapt to changing environmental conditions and emerging technologies.

9. Educate and Advocate:

- Develop educational materials and campaigns to raise public awareness about air quality issues and their health impacts.

- Advocate for policy changes based on the data and insights generated through improved air quality analytics.

- Engage with local communities to empower them to take action to improve air quality.

10. Impact Assessment:

- Regularly assess the long-term impact of the improved air quality analytics on public health and environmental quality.

- Share success stories and best practices to inspire similar efforts in other regions.

- Collaborate with research institutions to publish findings and contribute to the broader scientific community.

By applying the design thinking approach to air quality analytics in Tamil Nadu, stakeholders can develop user-centric solutions that address the specific needs of the community, enhance data-driven decision-making, and contribute to a healthier and more sustainable environment for the region.

Visualization:

Creating visualizations for air quality analysis in Cognos, a business intelligence and data visualization tool, to analyze air quality data in Tamil Nadu would involve several steps. Below, I'll outline a general process for creating such visualizations:

1. Data Preparation:

- Obtain air quality data for Tamil Nadu from reliable sources, such as government agencies or environmental organizations. Ensure that the data is clean and well-structured.

- Import the data into Cognos. You can use data connectors, import from databases,
or import flat files, depending on the source of your data.

2. Data Modeling:

- Create a data model or package in Cognos to define the structure of your data. This includes defining dimensions (e.g., time, location) and measures (e.g., air pollutant levels).

3. Create Reports:

- Build reports in Cognos that will serve as the basis for your visualizations. Reports can include tables, cross-tabs, and lists to display raw data.

4. Visualization Types:

- Select appropriate visualization types for your air quality analysis. Common types for this type of analysis might include:

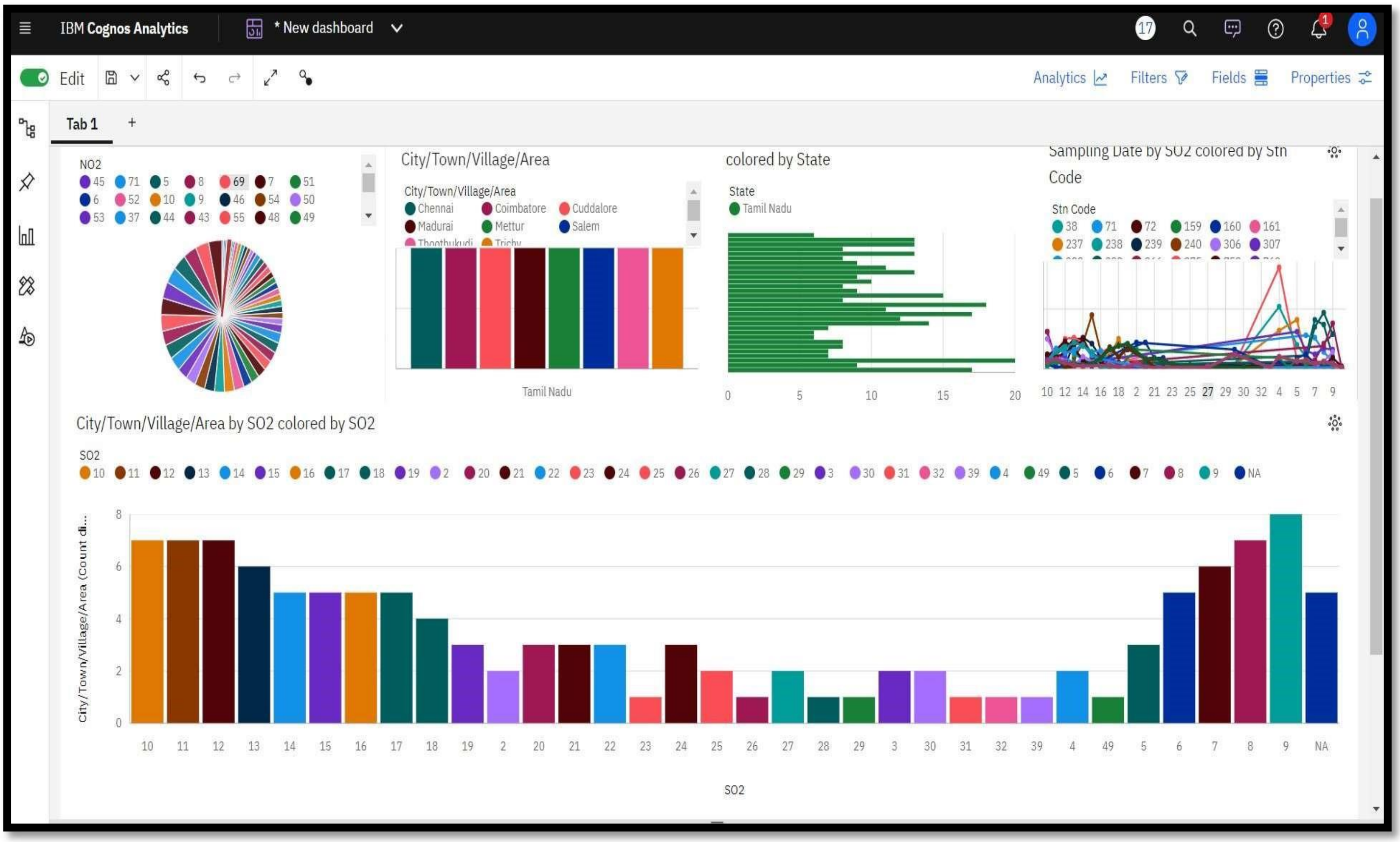
□ Line Charts: To show trends in air quality over time.

□ Maps: To display geographical variations in air quality across Tamil Nadu.

□ Bar Charts or Heatmaps: To compare air quality across different cities or regions.

□ Gauges: To display current air quality levels against predefined thresholds.

□ Dashboard: Combine various visualizations into a single dashboard for a comprehensive view.



5. Data Filtering and Slicing:

- Allow users to filter data by date, location, or pollutant type, so they can focus on specific aspects of air quality.

6. Color Coding and Legends:

- Use appropriate color coding to represent air quality levels (e.g., green for good, red for poor) and provide legends to explain the color scheme.

7. Interactivity:

- Add interactivity to your visualizations, such as drill-down capabilities, tooltips, and parameterized reports, to enable users to explore the data more deeply.

8. Annotations and Labels:

- Include labels, titles, and annotations to provide context and explanation for the visualizations.

9. Performance Optimization:

- Optimize the performance of your reports and visualizations, especially if dealing with large datasets, by tuning queries and caching results.

Phase 2 : Innovation

Consider incorporating machine learning algorithms to improve the accuracy of the predictive model

Finding pattern in Air quality analysis

Analyzing air quality data involves finding patterns and trends in the data to gain insights into the quality of the air in a specific location or over a period of time. Here are some steps and techniques for finding patterns in air quality data:

1. Data Collection:

    - Gather air quality data from various sources such as government monitoring stations, weather stations, or sensors.
    - Ensure the data includes relevant parameters like particulate matter (PM2.5, PM10), ozone (O3), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), temperature, humidity, and wind speed.

2. Data Preprocessing:

    - Clean the data by handling missing values and outliers.
    - Convert timestamps into a consistent format and set them as the index if applicable.
    - Normalize or scale the data if the parameters have different units or scales.

3. Exploratory Data Analysis (EDA):

   - Visualize the data using plots like time series plots, histograms, box plots, and scatter plots.
   - Look for seasonal, daily, or hourly patterns in the data.
   - Calculate basic statistics and correlations between different variables.

4. Time Series Analysis:

   - Decompose the time series data into its components (trend, seasonality, and residual) using techniques like seasonal decomposition of time series (STL).
   - Use autocorrelation and partial autocorrelation plots to identify lagged relationships in the time series.

5. Data Clustering:

   - Apply clustering algorithms (e.g., k-means, hierarchical clustering) to group similar air quality patterns.
   - Clustering can help identify areas with similar air quality characteristics.

6. Machine Learning Models:

   - Utilize machine learning models for prediction and classification tasks.
   - Train models to predict air quality based on historical data and meteorological variables.
   - Use classification models to categorize air quality as good, moderate, unhealthy, etc.

7. Spatial Analysis:

   - If you have data from multiple monitoring stations, perform spatial analysis to identify geographical patterns.
   - Use geospatial tools to create maps showing air quality variations across different locations.

8. Anomaly Detection:

   - Implement anomaly detection techniques to identify unusual events or spikes in air quality data.
   - These anomalies could indicate pollution events or measurement errors.

9. Feature Engineering:

- Create new features or transformations of existing features that may reveal hidden patterns.
- Feature engineering might involve lagged variables, moving averages, or time- based features.

10. Statistical Tests:

- Conduct statistical tests (e.g., t-tests, ANOVA) to compare air quality under different conditions or in different locations.
- Test hypotheses related to the effects of various factors on air quality.

11. Time Series Forecasting:

- Build time series forecasting models (e.g., ARIMA, LSTM) to predict future air quality levels.
- These models can help in planning and decision-making for air quality management.

12. Visualization:

 Use advanced data visualization techniques, such as heatmaps, contour plots, and geographic information system (GIS) tools, to represent patterns spatially and temporally.

Identifying outliers and anomalies in Air quality analysis

1. Data Preprocessing:

Begin by loading and preprocessing your air quality data. Handle missing values appropriately, as they can affect the accuracy of outlier detection methods.

2. Data Visualization:

Visualize the data to get an initial understanding of its distribution and potential outliers. Box plots, scatter plots, and histograms can be helpful for this purpose.

3. Z-Score Method:

Calculate the Z-score for each data point, which measures how many standard deviations a data point is from the mean. Data points with high absolute Z-scores (typically greater than a threshold like 3 or -3) are considered outliers.

4. IQR Method:

Calculate the Interquartile Range (IQR) and use it to identify outliers. Data points that fall below `Q1 - 1.5 * IQR` or above `Q3 + 1.5 * IQR` are considered outliers.

5. Visual Inspection:

Manually inspect data points that are identified as outliers or anomalies to determine whether they are genuine anomalies or data errors. Sometimes, visual inspection is necessary to understand the context of the data.

6. Anomaly Detection Models (Optional):

Consider using more advanced anomaly detection techniques like Isolation Forest, One-Class SVM, or autoencoders if the data has complex patterns that are challenging to detect with simple statistical methods.

These methods should help you identify outliers and anomalies in your air quality data. Keep in mind that the choice of method and threshold may depend on the characteristics of your specific dataset and the objectives of your analysis.

Making better predictions in Air quality analysis

Making better predictions in air quality analysis requires a combination of data preparation, feature engineering, model selection, and evaluation. Here are steps and techniques to improve the accuracy and reliability of predictions in air quality analysis:

1. High-Quality Data:

- Ensure your data is accurate, complete, and representative of the target area or region.
- Address missing values and outliers through appropriate data preprocessing techniques.
- Consider collecting additional relevant data such as meteorological information (e.g., temperature, wind speed, humidity) to enhance predictions.

2. Feature Engineering:

- □ Create informative features by incorporating domain knowledge and considering the physics of air pollution.
- □ Lagged variables: Include historical air quality values as features, as past air quality can influence the current state.
- □ Time-based features: Incorporate time-related features such as time of day, day of the week, or seasonality.
- □ Meteorological data: Use weather-related features that may affect air quality, such as temperature inversions or wind direction.

3. Model Selection:

- Experiment with different machine learning models suited for regression or time series forecasting tasks.

- Common models for air quality prediction include linear regression, decision trees, random forests, support vector machines, gradient boosting, and neural networks.
- Consider specialized time series models like ARIMA, SARIMA, or LSTM for temporal data.

4. Cross-Validation:

   - Employ cross-validation techniques (e.g., k-fold cross-validation) to assess the generalization performance of your models.

5. Hyperparameter Tuning:

   - Optimize model hyperparameters using techniques like grid search or random search.
   - Adjust hyperparameters related to model complexity, learning rates, and regularization to find the best combination.

6. Ensemble Methods:

   - Combine predictions from multiple models using ensemble techniques like bagging (e.g., random forests) or boosting (e.g., gradient boosting).
   - Ensembles can often improve predictive accuracy.

7. Time Series Analysis:

   - Analyze the temporal patterns in your data and select appropriate time series forecasting models.
   - Test different seasonality and lag values to capture the underlying patterns in air quality data.

8. Validation Metrics:

   - Choose appropriate evaluation metrics for your air quality prediction task. Common metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2).
   - Evaluate your model's performance on both training and validation datasets.

9. Feature Importance Analysis:

    - Determine which features have the most significant impact on air quality predictions using feature importance techniques provided by some machine learning models.
    - Focus on the most influential features for model refinement.

10. Temporal Aggregation:

    - Adjust the temporal resolution of your predictions to match the application's needs (e.g., hourly, daily, weekly).
    - Higher temporal resolution may require more sophisticated modeling techniques.

11. Data Fusion:

    □ Combine air quality data with data from other sources, such as satellite imagery, traffic data, or emissions data, to improve predictions and gain a broader perspective on air quality.

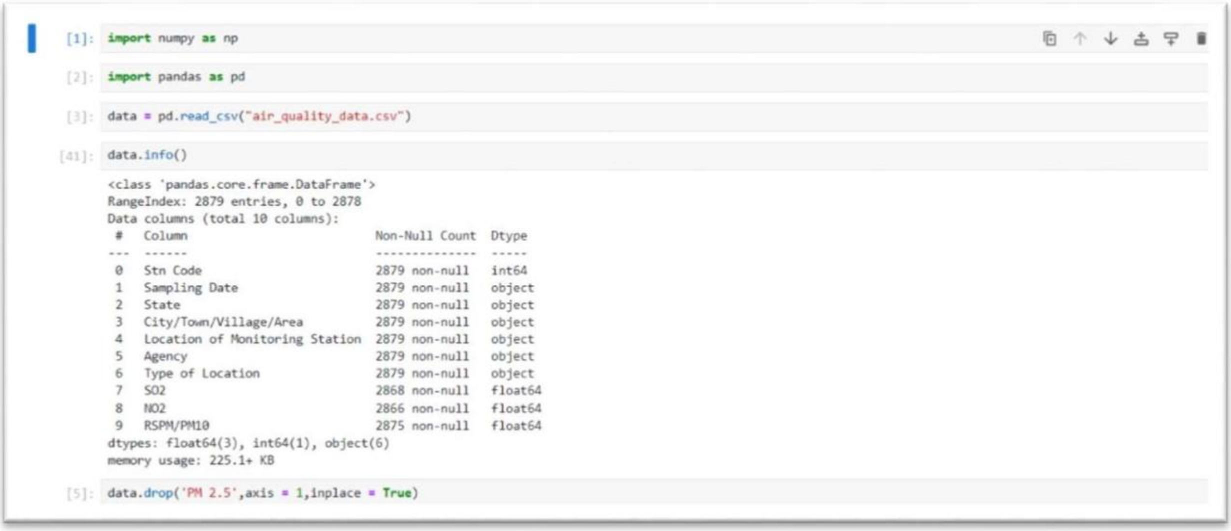
12. Continuous Monitoring and Model Updating:

    - Continuously monitor the performance of your air quality prediction models and update them as new data becomes available.
    - Adapt your models to changing environmental conditions or emerging pollution sources.

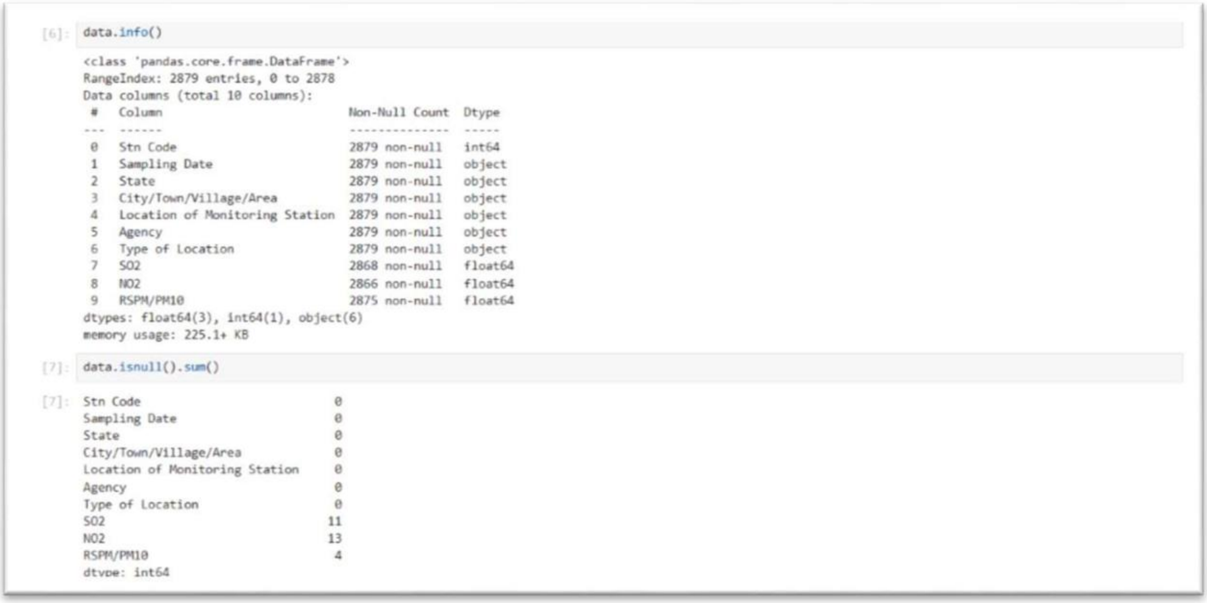
13. Domain Expertise:

    - Collaborate with domain experts who can provide valuable insights and guidance in selecting features and interpreting model results.

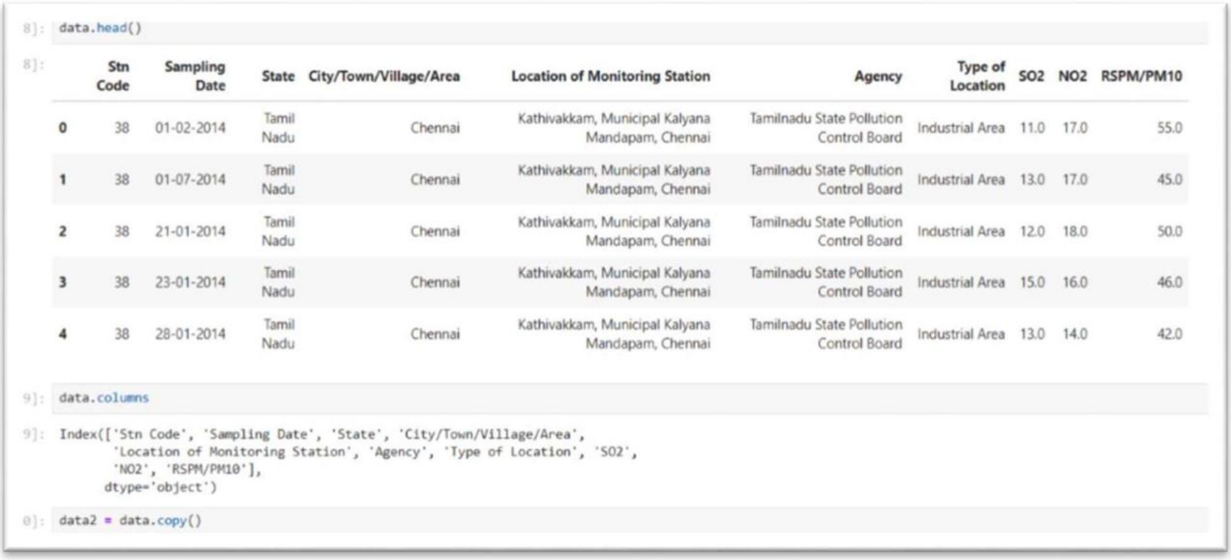
Execution:



```
[1]: import numpy as np

[2]: import pandas as pd

[3]: data = pd.read_csv("air_quality_data.csv")

[41]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2879 entries, 0 to 2878
Data columns (total 10 columns):
 #   Column                         Non-Null Count  Dtype  
---  ------                         --------------  -----  
 0   Stn Code                       2879 non-null   int64  
 1   Sampling Date                  2879 non-null   object 
 2   State                          2879 non-null   object 
 3   City/Town/Village/Area         2879 non-null   object 
 4   Location of Monitoring Station 2879 non-null   object 
 5   Agency                         2879 non-null   object 
 6   Type of Location               2879 non-null   object 
 7   SO2                            2868 non-null   float64
 8   NO2                            2866 non-null   float64
 9   RSPM/PM10                      2875 non-null   float64
dtypes: float64(3), int64(1), object(6)
memory usage: 225.1+ KB

[5]: data.drop('PM 2.5',axis = 1,inplace = True)
```



```
[6]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2879 entries, 0 to 2878
Data columns (total 10 columns):
 #   Column                         Non-Null Count  Dtype  
---  ------                         --------------  -----  
 0   Stn Code                       2879 non-null   int64  
 1   Sampling Date                  2879 non-null   object 
 2   State                          2879 non-null   object 
 3   City/Town/Village/Area         2879 non-null   object 
 4   Location of Monitoring Station 2879 non-null   object 
 5   Agency                         2879 non-null   object 
 6   Type of Location               2879 non-null   object 
 7   SO2                            2868 non-null   float64
 8   NO2                            2866 non-null   float64
 9   RSPM/PM10                      2875 non-null   float64
dtypes: float64(3), int64(1), object(6)
memory usage: 225.1+ KB

[7]: data.isnull().sum()

[7]: Stn Code                           0
     Sampling Date                      0
     State                              0
     City/Town/Village/Area             0
     Location of Monitoring Station     0
     Agency                             0
     Type of Location                   0
     SO2                               11
     NO2                               13
     RSPM/PM10                          4
     dtype: int64
```

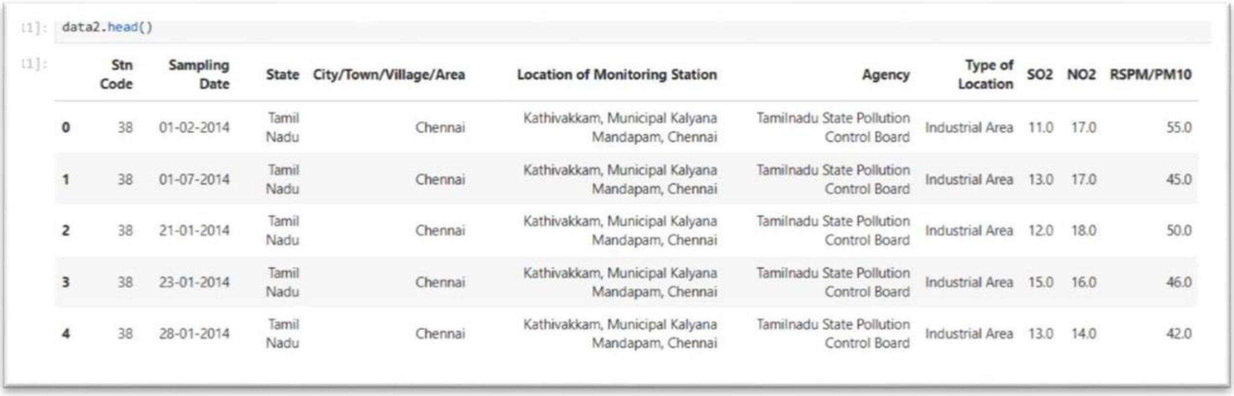


```
8]: data.head()
```

8]:

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10
0	38	01-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0
1	38	01-07-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0
2	38	21-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0
3	38	23-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0
4	38	28-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0

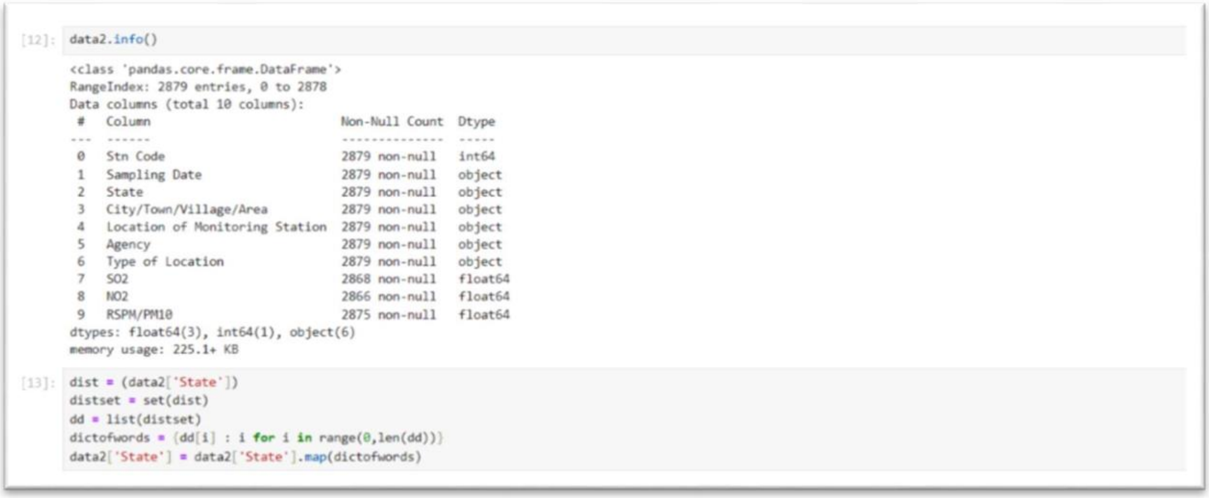
```
9]: data.columns

9]: Index(['Stn Code', 'Sampling Date', 'State', 'City/Town/Village/Area',
           'Location of Monitoring Station', 'Agency', 'Type of Location', 'SO2',
           'NO2', 'RSPM/PM10'],
          dtype='object')

0]: data2 = data.copy()
```



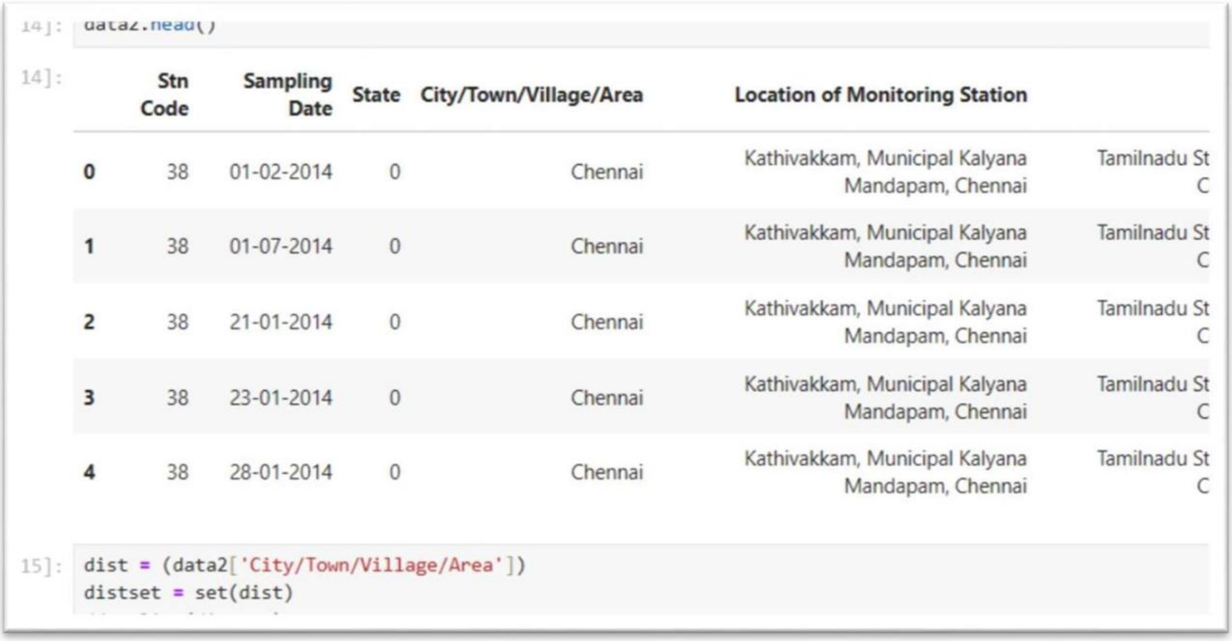
```
1]: data2.head()
```

1]:

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10
0	38	01-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0
1	38	01-07-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0
2	38	21-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0
3	38	23-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0
4	38	28-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0



```
[12]: data2.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2879 entries, 0 to 2878
Data columns (total 10 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Stn Code                        2879 non-null   int64
 1   Sampling Date                   2879 non-null   object
 2   State                           2879 non-null   object
 3   City/Town/Village/Area          2879 non-null   object
 4   Location of Monitoring Station  2879 non-null   object
 5   Agency                          2879 non-null   object
 6   Type of Location                2879 non-null   object
 7   SO2                             2868 non-null   float64
 8   NO2                             2866 non-null   float64
 9   RSPM/PM10                       2875 non-null   float64
dtypes: float64(3), int64(1), object(6)
memory usage: 225.1+ KB

[13]: dist = (data2['State'])
      distset = set(dist)
      dd = list(distset)
      dictofwords = {dd[i] : i for i in range(0,len(dd))}
      data2['State'] = data2['State'].map(dictofwords)
```

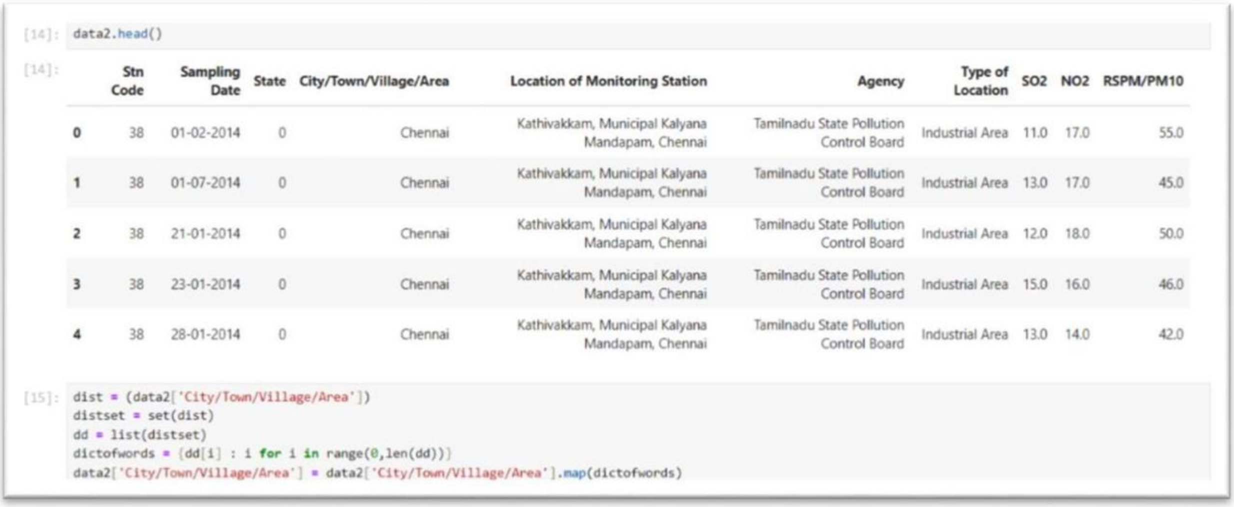


```
[14]: data2.head()
```

[14]:

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	
0	38	01-02-2014	0	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu St C
1	38	01-07-2014	0	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu St C
2	38	21-01-2014	0	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu St C
3	38	23-01-2014	0	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu St C
4	38	28-01-2014	0	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu St C

```
[15]: dist = (data2['City/Town/Village/Area'])
      distset = set(dist)
```

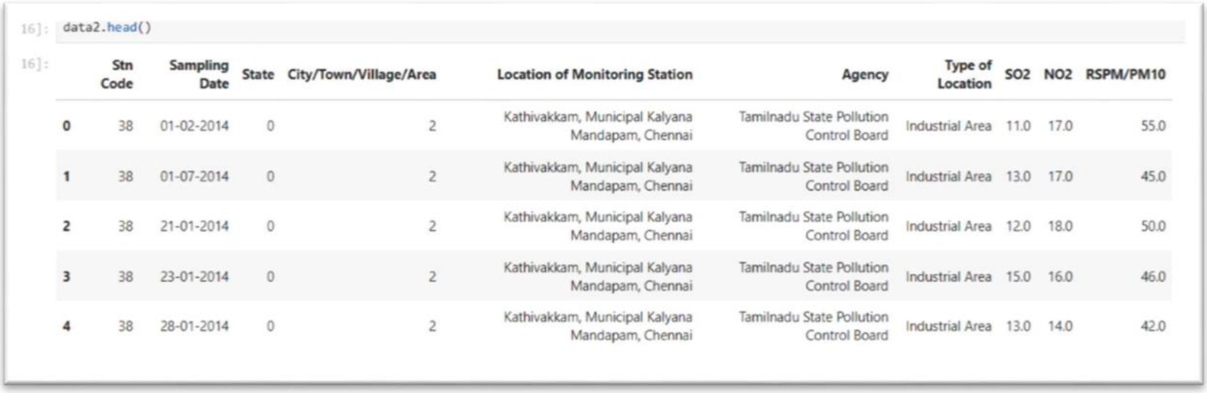


```
[14]: data2.head()
```

[14]:

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10
0	38	01-02-2014	0	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0
1	38	01-07-2014	0	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0
2	38	21-01-2014	0	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0
3	38	23-01-2014	0	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0
4	38	28-01-2014	0	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0

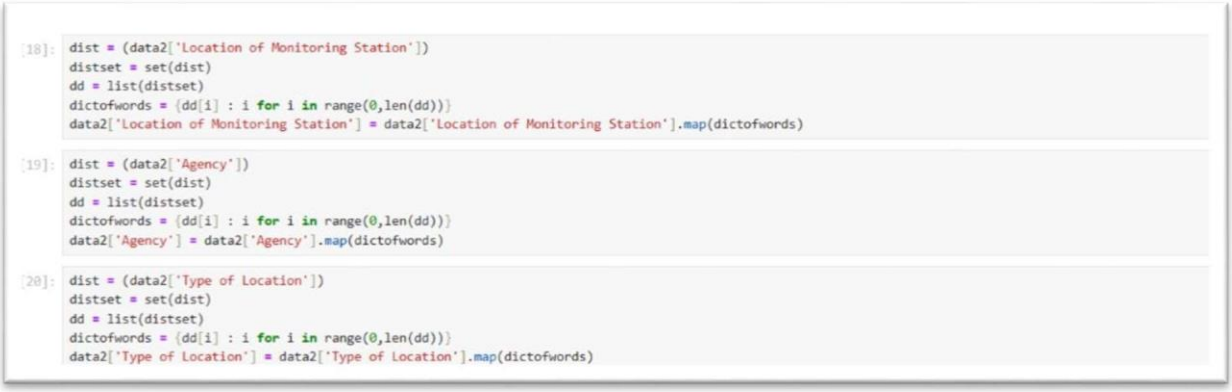
```
[15]: dist = (data2['City/Town/Village/Area'])
      distset = set(dist)
      dd = list(distset)
      dictofwords = {dd[i] : i for i in range(0,len(dd))}
      data2['City/Town/Village/Area'] = data2['City/Town/Village/Area'].map(dictofwords)
```



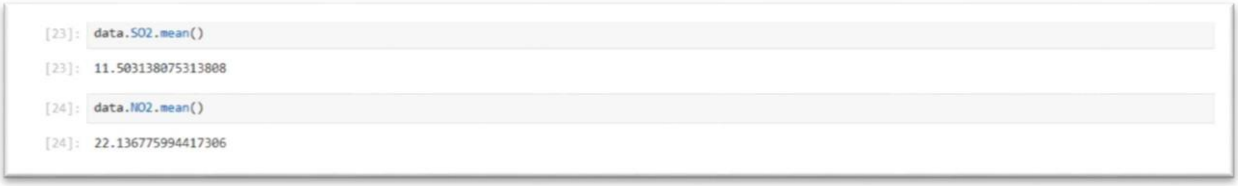
```
16]: data2.head()
```

16]:

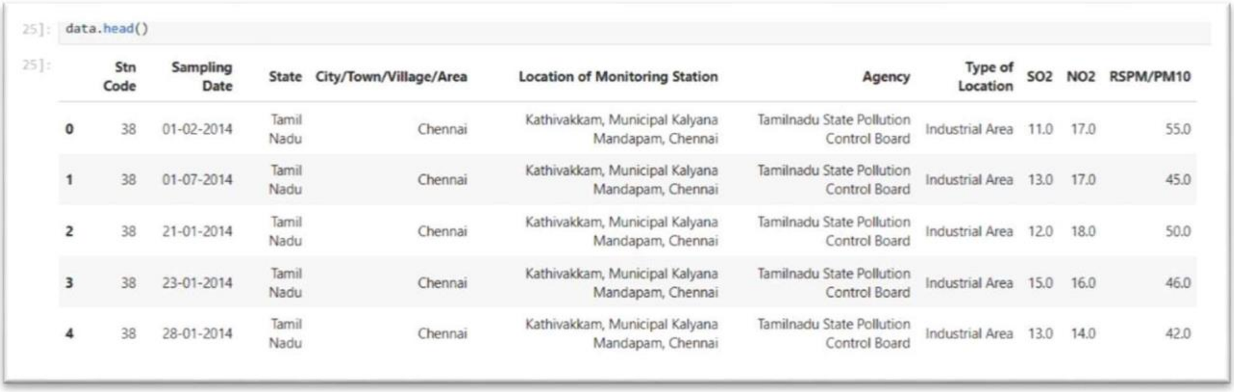
	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10
0	38	01-02-2014	0	2	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0
1	38	01-07-2014	0	2	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0
2	38	21-01-2014	0	2	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0
3	38	23-01-2014	0	2	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0
4	38	28-01-2014	0	2	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0



```
[18]: dist = (data2['Location of Monitoring Station'])
      distset = set(dist)
      dd = list(distset)
      dictofwords = {dd[i] : i for i in range(0,len(dd))}
      data2['Location of Monitoring Station'] = data2['Location of Monitoring Station'].map(dictofwords)

[19]: dist = (data2['Agency'])
      distset = set(dist)
      dd = list(distset)
      dictofwords = {dd[i] : i for i in range(0,len(dd))}
      data2['Agency'] = data2['Agency'].map(dictofwords)

[20]: dist = (data2['Type of Location'])
      distset = set(dist)
      dd = list(distset)
      dictofwords = {dd[i] : i for i in range(0,len(dd))}
      data2['Type of Location'] = data2['Type of Location'].map(dictofwords)
```



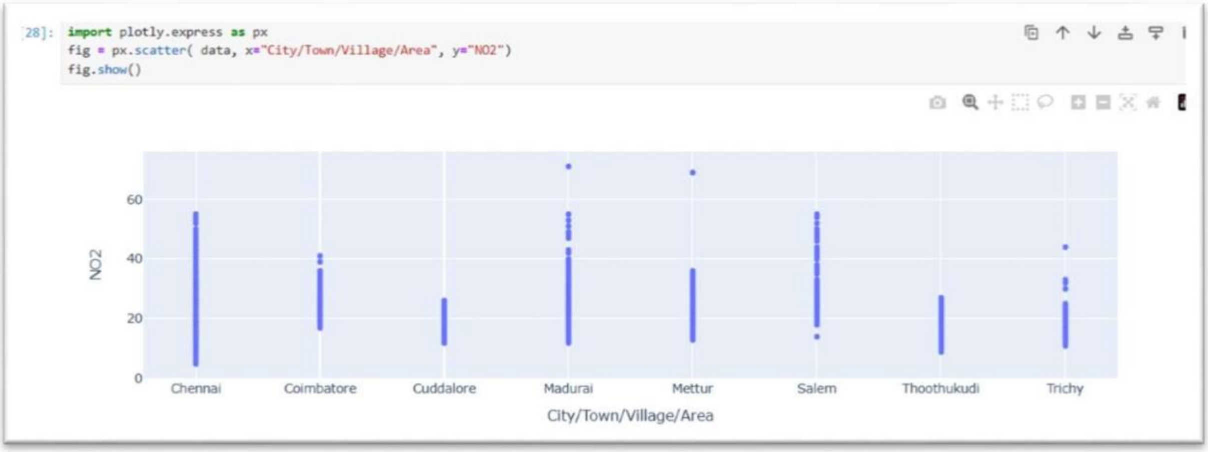
```
[23]: data.SO2.mean()

[23]: 11.503138075313808

[24]: data.NO2.mean()

[24]: 22.136775994417306
```



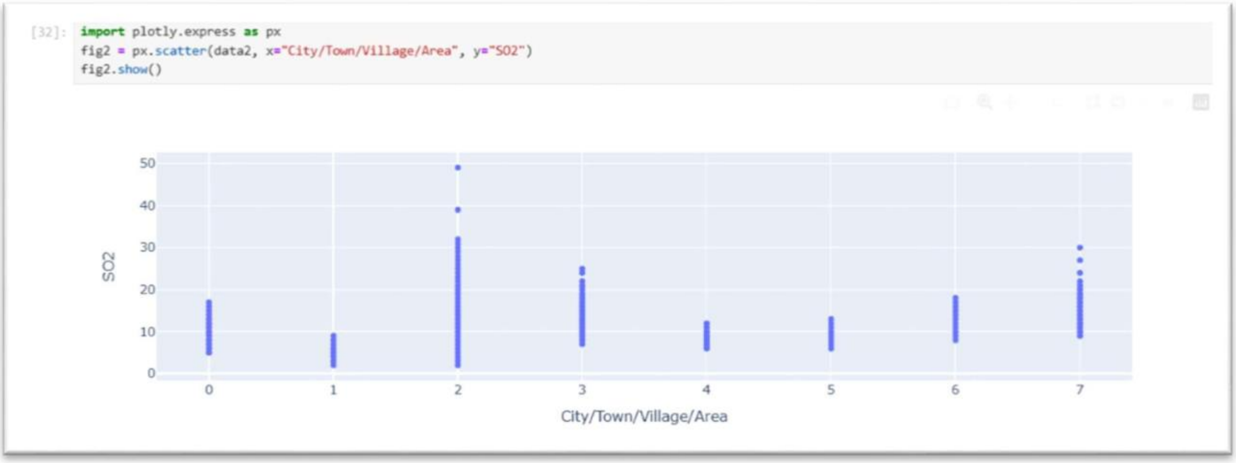
```
25]: data.head()
```

25]:

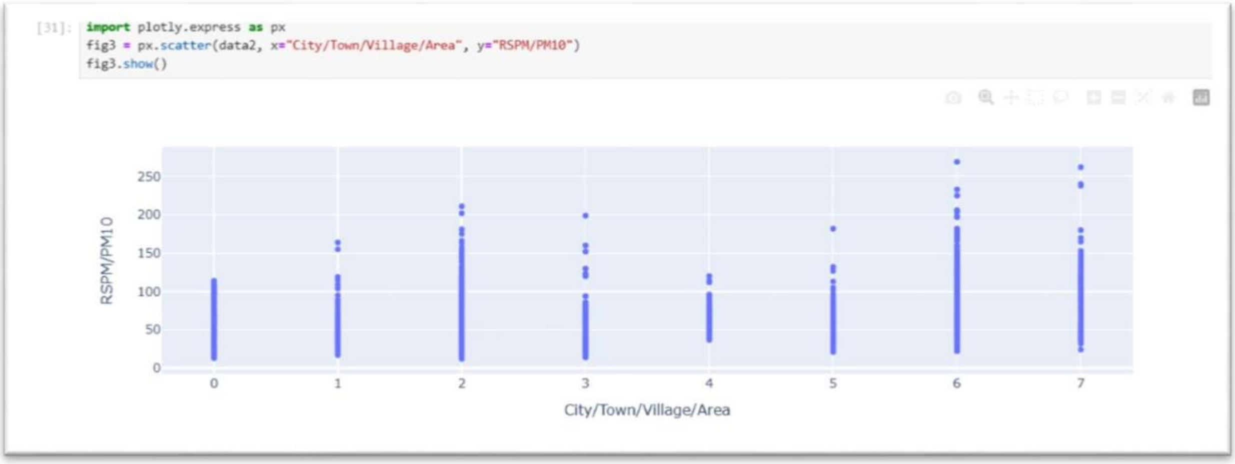
	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10
0	38	01-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0
1	38	01-07-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0
2	38	21-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0
3	38	23-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0
4	38	28-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0



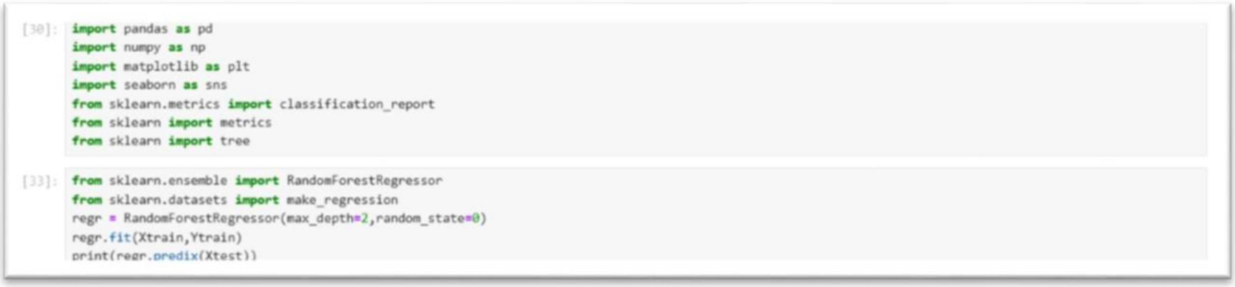
```
[28]: import plotly.express as px
      fig = px.scatter( data, x="City/Town/Village/Area", y="NO2")
      fig.show()
```



```
[32]: import plotly.express as px
      fig2 = px.scatter(data2, x="City/Town/Village/Area", y="SO2")
      fig2.show()
```



```
[31]: import plotly.express as px
      fig3 = px.scatter(data2, x="City/Town/Village/Area", y="RSPM/PM10")
      fig3.show()
```



```
[30]: import pandas as pd
      import numpy as np
      import matplotlib as plt
      import seaborn as sns
      from sklearn.metrics import classification_report
      from sklearn import metrics
      from sklearn import tree
```

```
[33]: from sklearn.ensemble import RandomForestRegressor
      from sklearn.datasets import make_regression
      regr = RandomForestRegressor(max_depth=2,random_state=0)
      regr.fit(Xtrain,Ytrain)
      print(regr.predix(Xtest))
```

Phase 3 : Development Part 1

Begin the analysis by loading and preprocessing the Air quality dataset

Data Preprocessing in Air quality analytics

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

Some common steps in data preprocessing include:

Data preprocessing is an important step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis. Some common steps in data preprocessing include:

1. Data Cleaning:

   This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.

2. Data Integration:

   This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.

3. Data Transformation:

   This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.

4. Data Reduction:

   This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.

5. Data Discretization:

   This involves dividing continuous data into discrete categories or intervals.

Discretization is often used in data mining and machine learning algorithms that require categorical data. Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering.

6. Data Normalization:

This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.

Data preprocessing plays a crucial role in ensuring the quality of data and the accuracy of the analysis results. The specific steps involved in data preprocessing may vary depending on the nature of the data and the analysis goals.

Preprocessing in Data Mining

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Steps Involved in Data Preprocessing:

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

(a). Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are.

□ Ignore the tuples:□

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

□ Fill the Missing values:□

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

(b). Noisy Data:

Noisy data is a meaningless data that can’t be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

□ Binning Method:□

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

□ Regression:□

Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

□ Clustering:□

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation:

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

□ Normalization:□

It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

□ Attribute Selection:□

In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

□ Discretization:□

This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

□ Concept Hierarchy Generation:□

Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. Data Reduction:

Data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information. This is done to improve the efficiency of data analysis and to avoid overfitting of the model. Some common steps involved in data reduction are:

□ Feature Selection:□

This involves selecting a subset of relevant features from the dataset. Feature selection is often performed to remove irrelevant or redundant features from the dataset. It can be done using various techniques such as correlation analysis, mutual information, and principal component analysis (PCA).

□ Feature Extraction:□

This involves transforming the data into a lower-dimensional space while preserving the important information. Feature extraction is often used when the original features are high-dimensional and complex. It can be done using techniques such as PCA, linear discriminant analysis (LDA), and non-negative matrix factorization (NMF).

□ Sampling:□

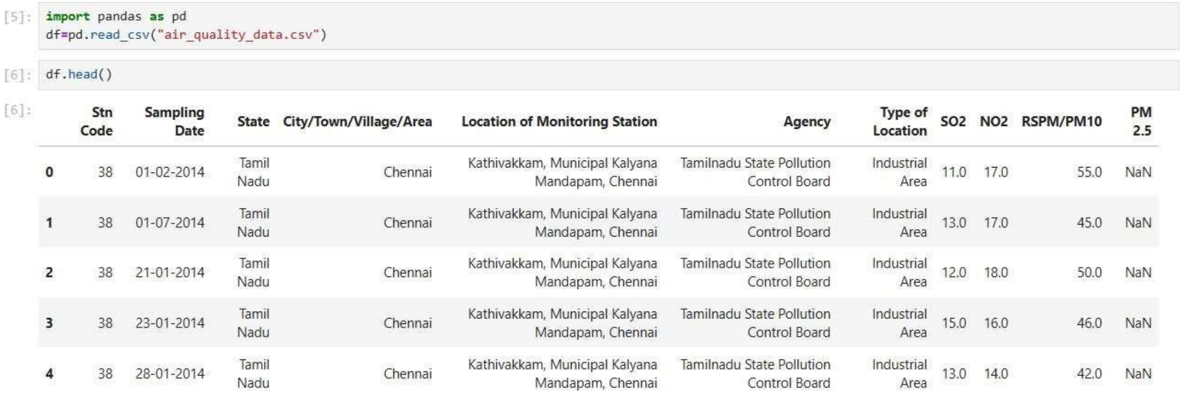
This involves selecting a subset of data points from the dataset. Sampling is often used to reduce the size of the dataset while preserving the important information. It can be done using techniques such as random sampling, stratified sampling, and systematic sampling.

□ Clustering:□

This involves grouping similar data points together into clusters. Clustering is often used to reduce the size of the dataset by replacing similar data points with a representative centroid. It can be done using techniques such as k-means, hierarchical clustering, and density-based clustering.

□ Compression:□

This involves compressing the dataset while preserving the important information. Compression is often used to reduce the size of the dataset for storage and transmission purposes. It can be done using techniques such as wavelet compression, JPEG compression, and gzip compression.

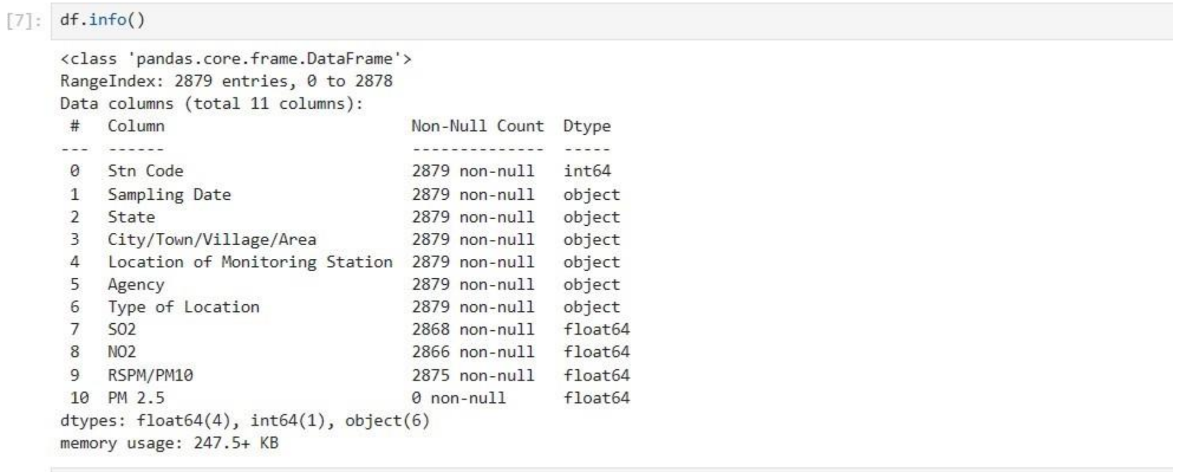


[5]:
```
import pandas as pd
df=pd.read_csv("air_quality_data.csv")
```

[6]:
```
df.head()
```

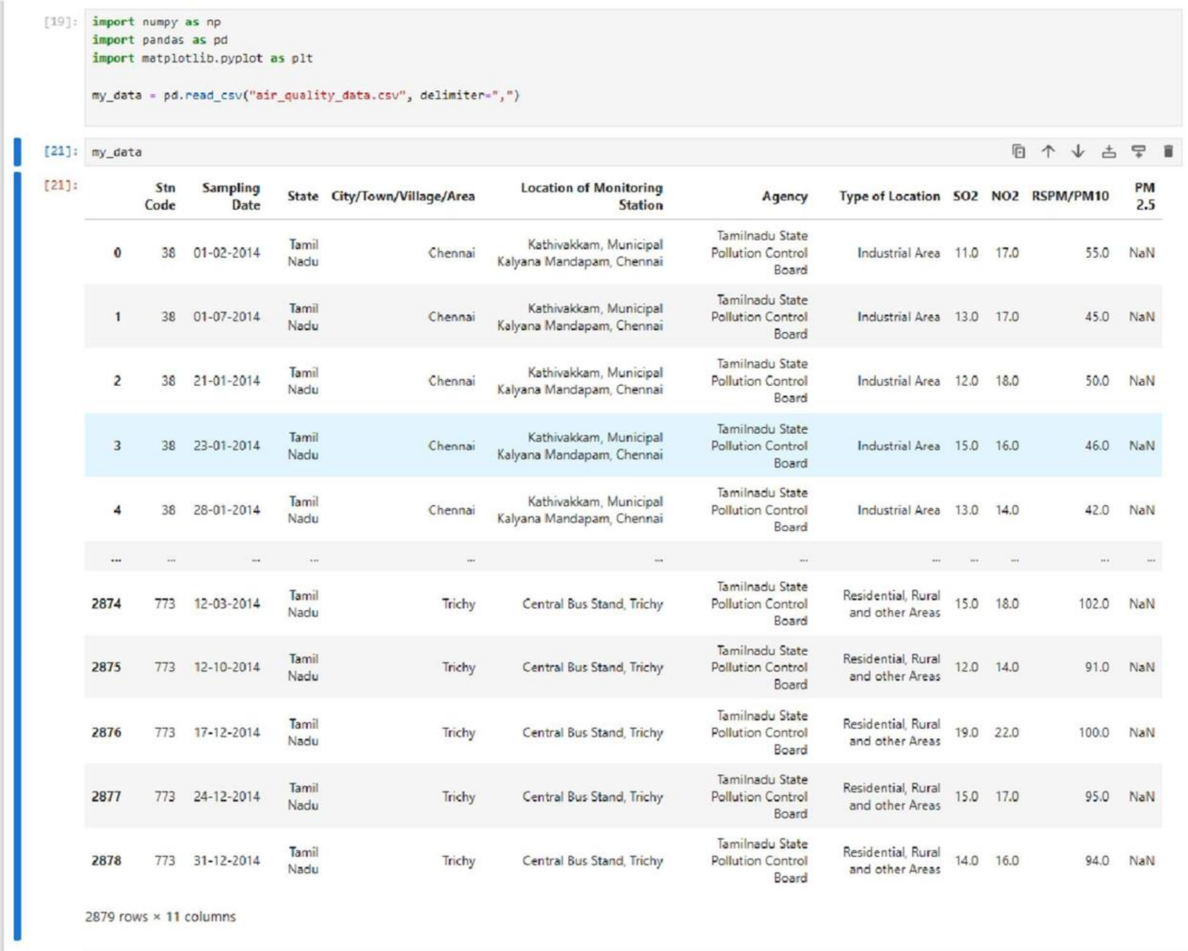
[6]:

	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10	PM 2.5
0	38	01-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0	NaN
1	38	01-07-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0	NaN
2	38	21-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0	NaN
3	38	23-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0	NaN
4	38	28-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0	NaN



[7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2879 entries, 0 to 2878
Data columns (total 11 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   Stn Code                        2879 non-null   int64  
 1   Sampling Date                   2879 non-null   object 
 2   State                           2879 non-null   object 
 3   City/Town/Village/Area          2879 non-null   object 
 4   Location of Monitoring Station  2879 non-null   object 
 5   Agency                          2879 non-null   object 
 6   Type of Location                2879 non-null   object 
 7   SO2                             2868 non-null   float64
 8   NO2                             2866 non-null   float64
 9   RSPM/PM10                       2875 non-null   float64
 10  PM 2.5                          0 non-null      float64
dtypes: float64(4), int64(1), object(6)
memory usage: 247.5+ KB
```



```
[19]: import numpy as np
      import pandas as pd
      import matplotlib.pyplot as plt

      my_data = pd.read_csv("air_quality_data.csv", delimiter=",")
```

```
[21]: my_data
```

[21]:

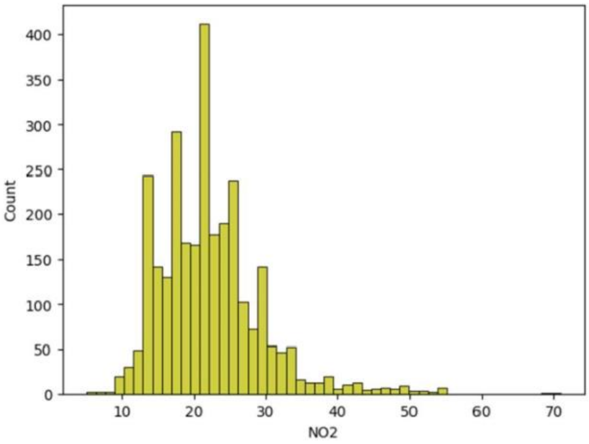
	Stn Code	Sampling Date	State	City/Town/Village/Area	Location of Monitoring Station	Agency	Type of Location	SO2	NO2	RSPM/PM10	PM 2.5
0	38	01-02-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	11.0	17.0	55.0	NaN
1	38	01-07-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	17.0	45.0	NaN
2	38	21-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	12.0	18.0	50.0	NaN
3	38	23-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	15.0	16.0	46.0	NaN
4	38	28-01-2014	Tamil Nadu	Chennai	Kathivakkam, Municipal Kalyana Mandapam, Chennai	Tamilnadu State Pollution Control Board	Industrial Area	13.0	14.0	42.0	NaN
...	...	...	...	...	...	...	...	...	...	...	...
2874	773	12-03-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	18.0	102.0	NaN
2875	773	12-10-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	12.0	14.0	91.0	NaN
2876	773	17-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	19.0	22.0	100.0	NaN
2877	773	24-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	15.0	17.0	95.0	NaN
2878	773	31-12-2014	Tamil Nadu	Trichy	Central Bus Stand, Trichy	Tamilnadu State Pollution Control Board	Residential, Rural and other Areas	14.0	16.0	94.0	NaN

2879 rows × 11 columns

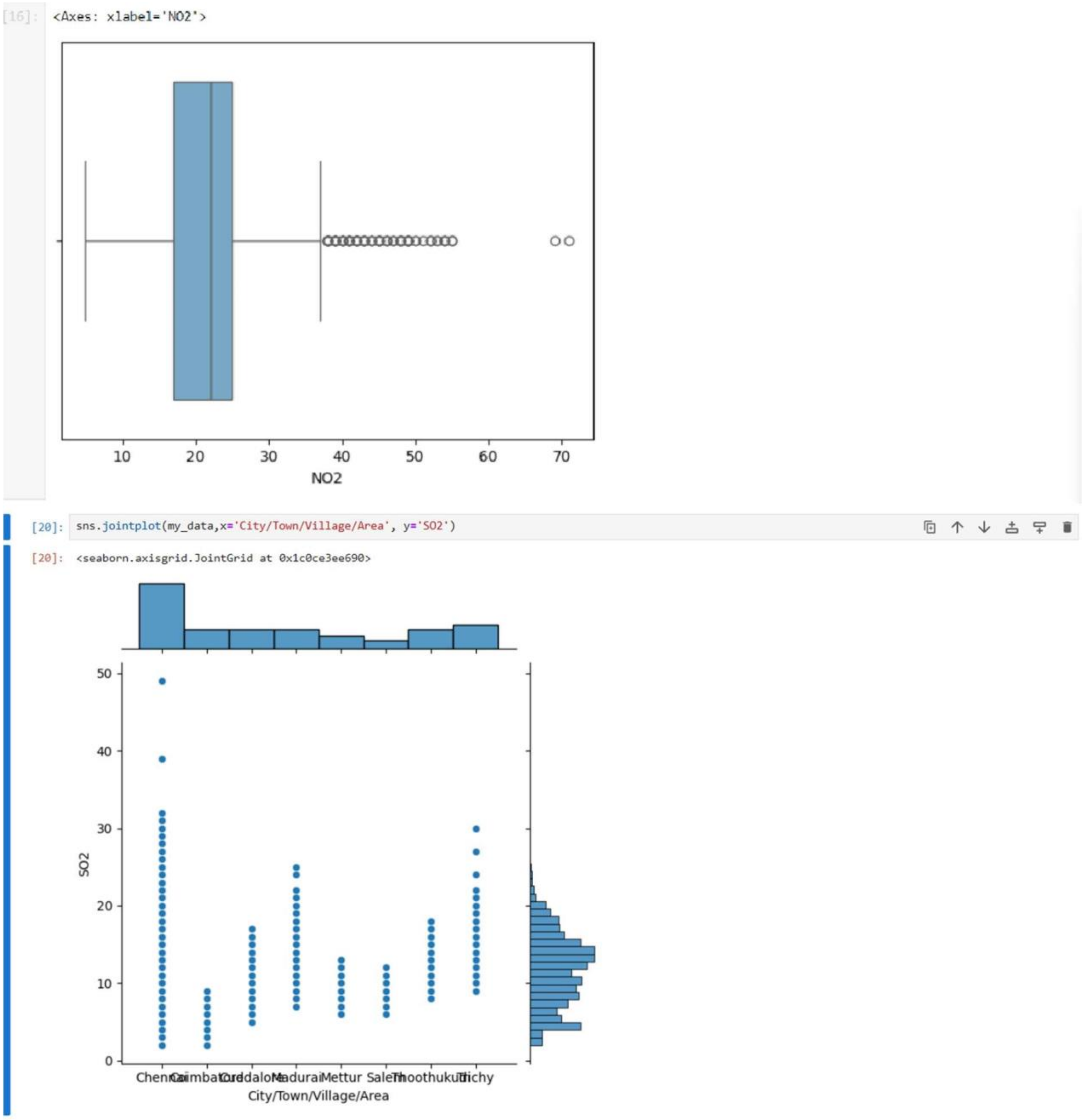
```
[13]: import pandas as pd
      import numpy as np
      import seaborn as sns
      import matplotlib.pyplot as plt
```

```
[15]: sns.histplot(my_data, x='NO2', bins=50, color='y')
```

```
[15]: <Axes: xlabel='NO2', ylabel='Count'>
```



```
[16]: sns.boxplot(my_data, x='NO2', palette='Blues')
```



```
[16]: <Axes: xlabel='NO2'>
```

```
[20]: sns.jointplot(my_data,x='City/Town/Village/Area', y='SO2')
```

```
[20]: <seaborn.axisgrid.JointGrid at 0x1c0ce3ee690>
```

Visualization In Air quality

Visualizing air quality data can provide valuable insights and make it easier to interpret the information. There are various visualization tools and libraries available in Python, with Matplotlib and Seaborn being some of the most commonly used ones. Here's a general approach to visualizing an air quality dataset using Python:

Import Necessary Libraries:

import pandas as pd import
matplotlib.pyplot as plt import
seaborn as sns Load Your
Dataset:

python
Copy code

```
# Load your dataset into a DataFrame
df = pd.read_csv('your_air_quality_data.csv')
```

Data Preprocessing:

Before visualizing, you might need to clean and prepare your data. You've already shown some code for data preprocessing in your previous questions, which includes converting columns to datetime and removing commas.

Exploratory Data Analysis (EDA):

1. Univariate Analysis:
   Start by exploring individual variables. You can create histograms, box plots, or violin plots to understand the distribution of air quality parameters like SO2, NO2, and RSPM/PM10.

2. Time Series Analysis:
   Since you converted 'Sampling Date' to a datetime object, you can create time series plots to see how air quality parameters change over time. This can be particularly insightful for long-term trends and seasonality.

3. Multivariate Analysis:
   You can create scatter plots and pair plots to analyze relationships between different variables in your dataset. For example, you might want to explore how SO2 and NO2 levels correlate.

4. Geospatial Visualization (if applicable):

If your dataset contains geographical information, you can create maps to visualize air quality variations across different locations.

5. Matplotlib:

   This is a powerful and flexible library for creating various types of plots.

6. Seaborn:

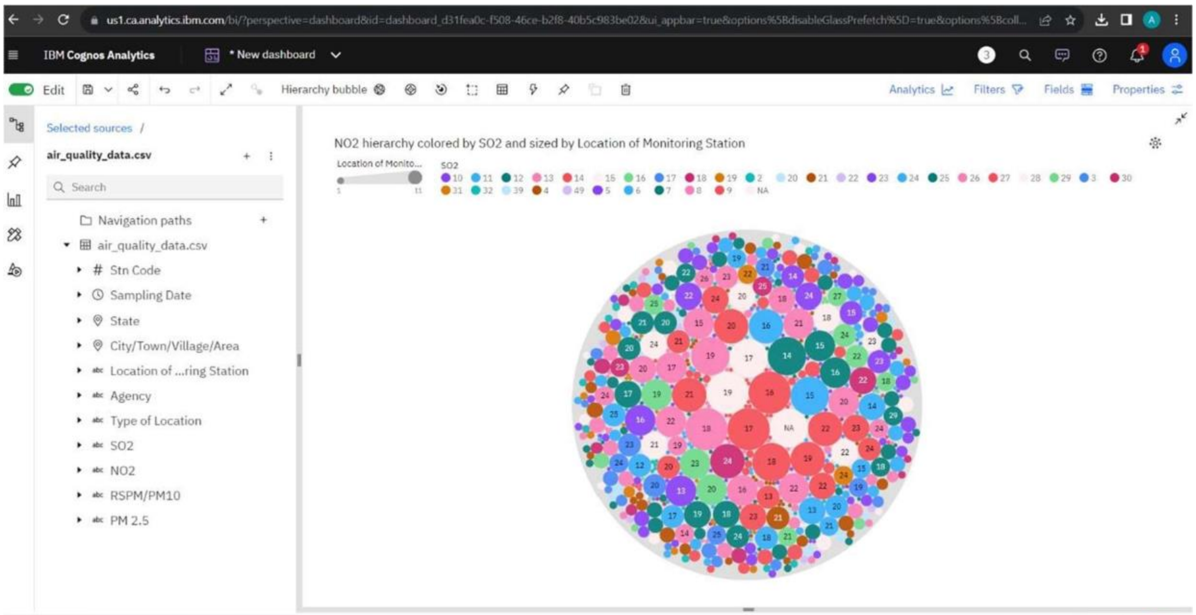
   Seaborn is built on top of Matplotlib and provides a high-level interface for creating informative and attractive statistical graphics.

7. Plotly:

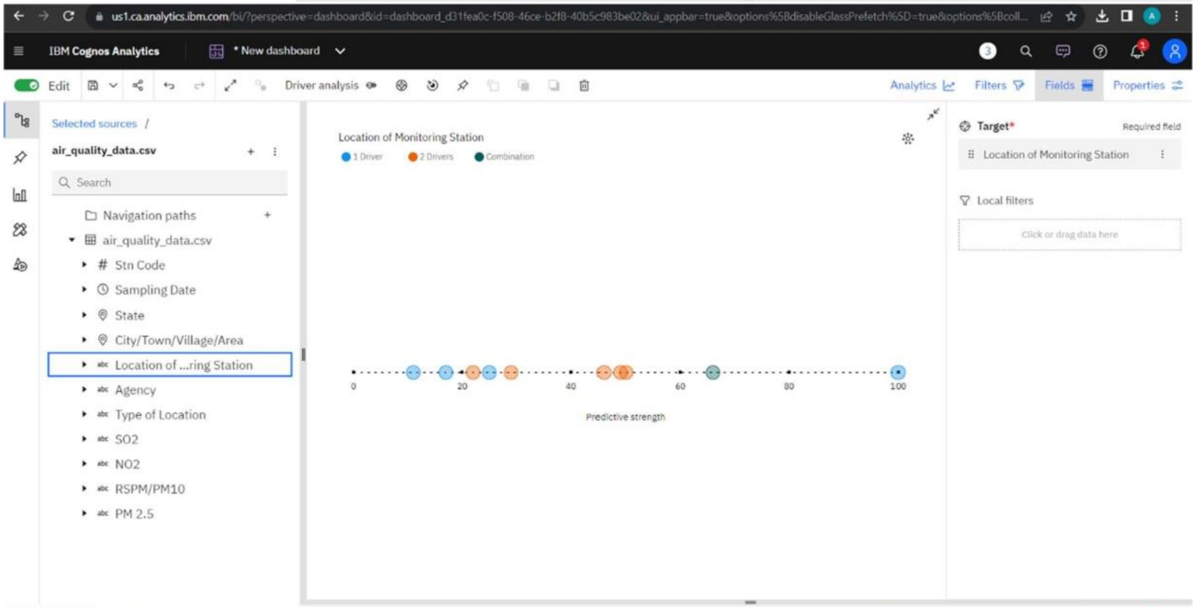
   If you want to create interactive visualizations, Plotly is a great choice.

Example Code:

```
# Scatter plot between SO2 and NO2
plt.figure(figsize=(8, 6))
sns.scatterplot(x='SO2', y='NO2', data=df)
plt.title('SO2 vs. NO2')
plt.xlabel('SO2')
plt.ylabel('NO2')
plt.show()
```

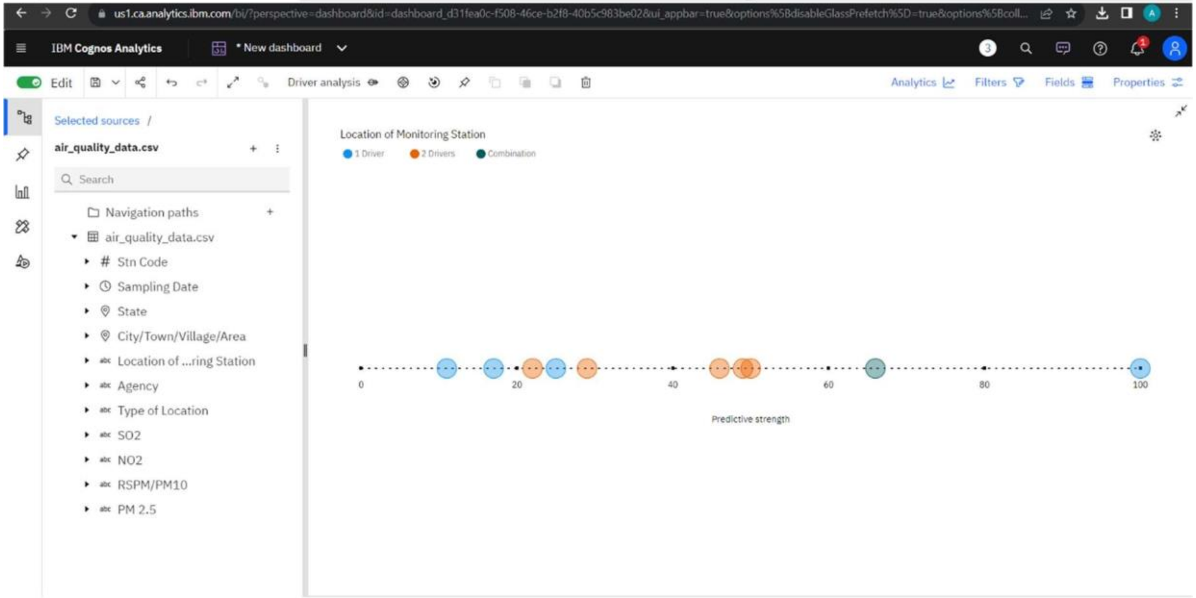


NO2 hierarchy colored by SO2 and sized by Location of Monitoring Station

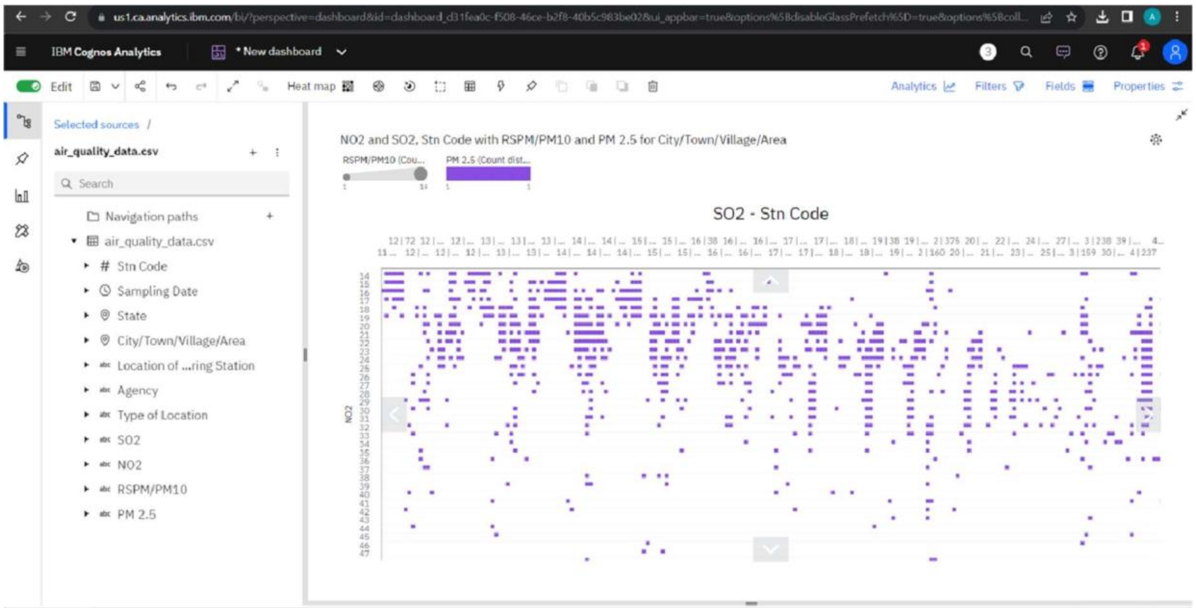


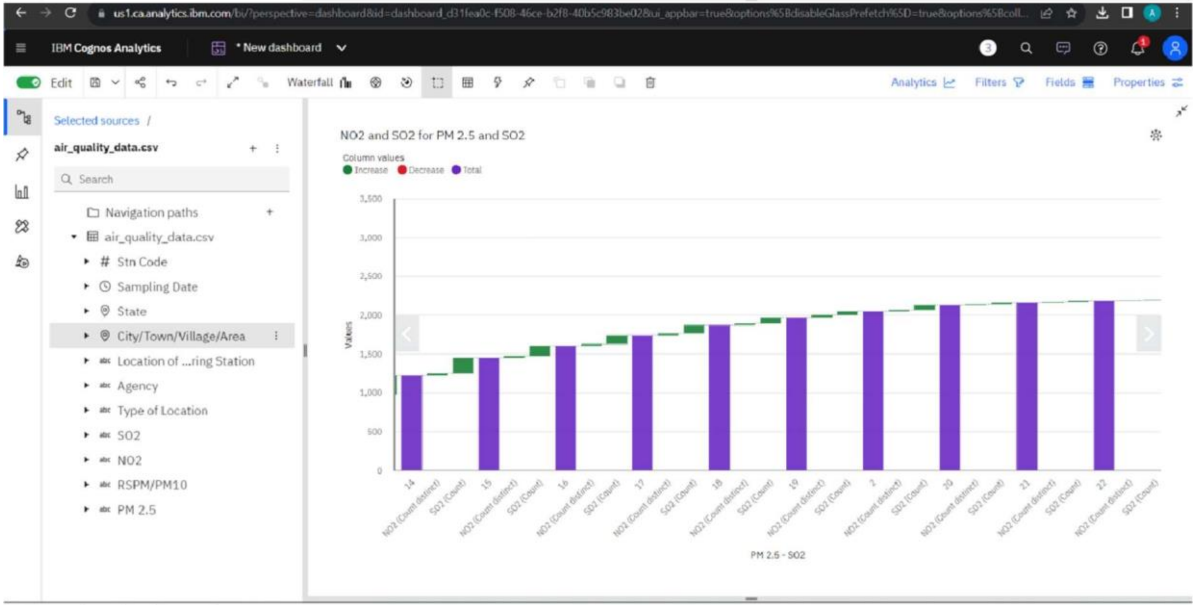
Location of Monitoring Station

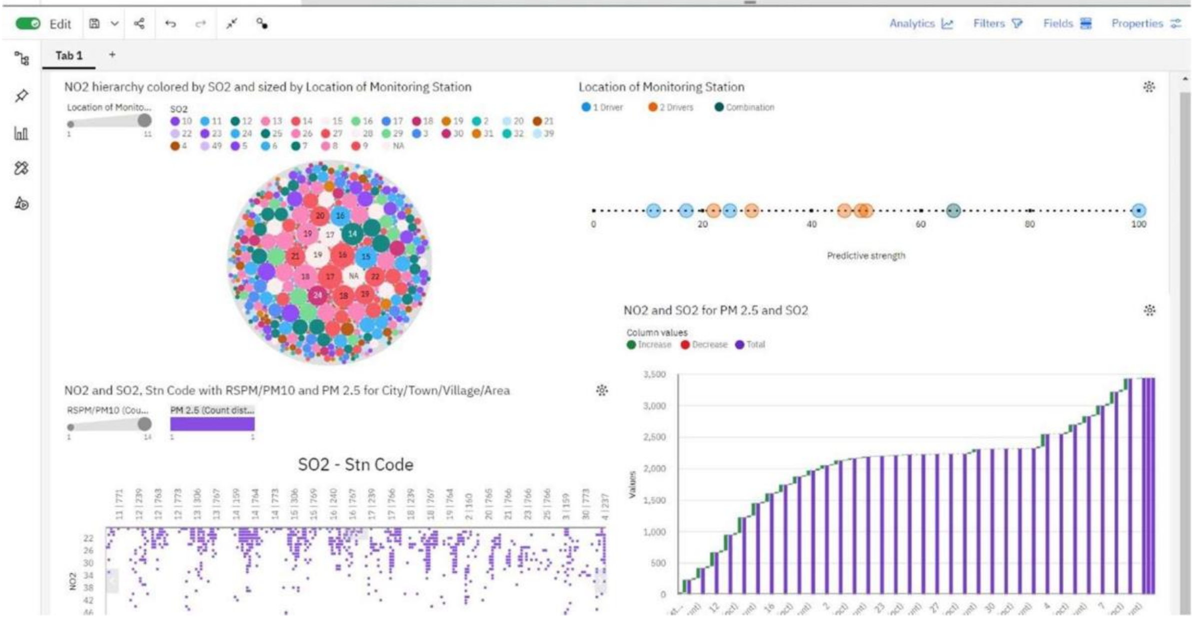
Target* Location of Monitoring Station

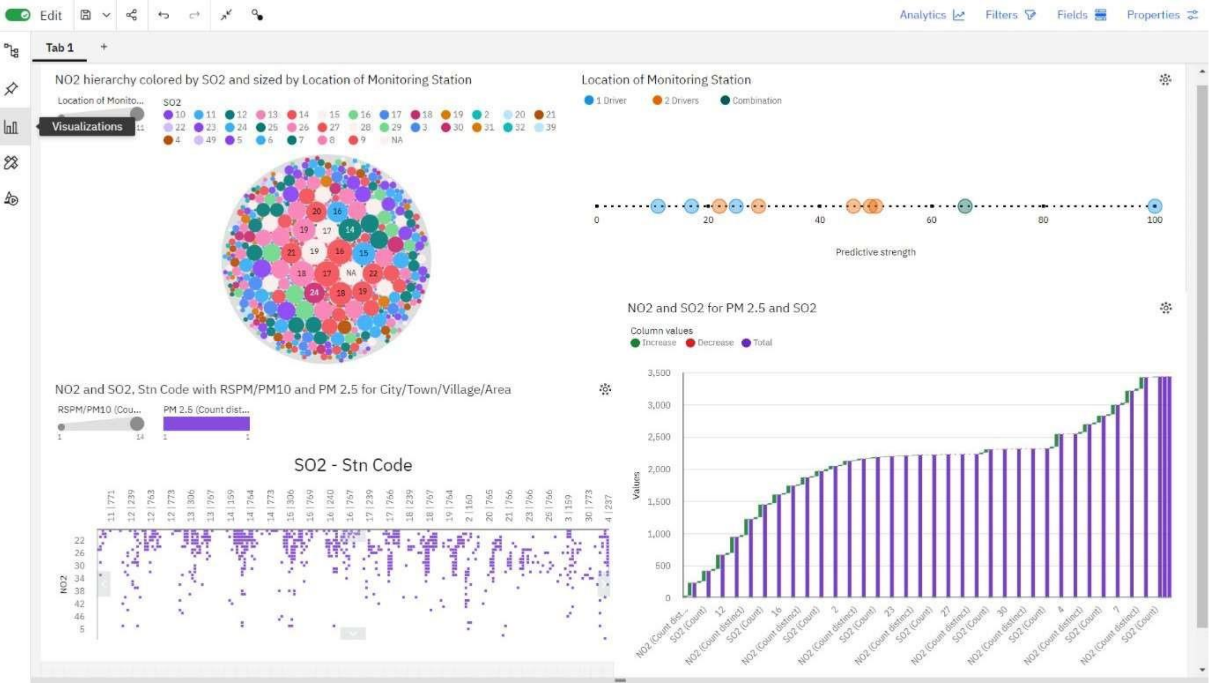


Location of Monitoring Station


us1.ca.analytics.ibm.com/bi/?perspective=dashboard&id=dashboard_d31fea0c-f508-46ce-b2f8-40b5c983be02&ui_appbar=true&options%5BdisableGlassPrefetch%5D=true&options%5Bcoll...
IBM Cognos Analytics
* New dashboard
Edit
Heat map
Analytics
Filters
Fields
Properties
Selected sources /
air_quality_data.csv
Search
Navigation paths
air_quality_data.csv
Stn Code
Sampling Date
State
City/Town/Village/Area
Location of ...ring Station
Agency
Type of Location
SO2
NO2
RSPM/PM10
PM 2.5
NO2 and SO2, Stn Code with RSPM/PM10 and PM 2.5 for City/Town/Village/Area
RSPM/PM10 (Cou...
PM 2.5 (Count dist...
SO2 - Stn Code
NO2


us1.ca.analytics.ibm.com/bi/?perspective=dashboard&id=dashboard_d31fea0c-f508-46ce-b2f8-40b5c983be02&ui_appbar=true&options%5BdisableGlassPrefetch%5D=true&options%5Bcoll...
IBM Cognos Analytics
* New dashboard
Edit
Waterfall
Analytics
Filters
Fields
Properties
Selected sources /
air_quality_data.csv
Search
Navigation paths
air_quality_data.csv
Stn Code
Sampling Date
State
City/Town/Village/Area
Location of ...ring Station
Agency
Type of Location
SO2
NO2
RSPM/PM10
PM 2.5
NO2 and SO2 for PM 2.5 and SO2
Column values
Increase
Decrease
Total
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Values
14
NO2 (Count distinct)
SO2 (Count)
15
16
17
18
19
2
20
21
22
PM 2.5 - SO2


Edit
Analytics
Filters
Fields
Properties
Tab 1
NO2 hierarchy colored by SO2 and sized by Location of Monitoring Station
Location of Monito...
SO2
Location of Monitoring Station
1 Driver
2 Drivers
Combination
0
20
40
60
80
100
Predictive strength
NO2 and SO2 for PM 2.5 and SO2
Column values
Increase
Decrease
Total
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Values
NO2 and SO2, Stn Code with RSPM/PM10 and PM 2.5 for City/Town/Village/Area
RSPM/PM10 (Cou...
PM 2.5 (Count dist...
SO2 - Stn Code
NO2


Edit
Analytics
Filters
Fields
Properties
Tab 1
+
Visualizations
NO2 hierarchy colored by SO2 and sized by Location of Monitoring Station
Location of Monito...
SO2
10 11 12 13 14 15 16 17 18 19 2 20 21
22 23 24 25 26 27 28 29 3 30 31 32 39
4 49 5 6 7 8 9 NA
Location of Monitoring Station
1 Driver
2 Drivers
Combination
0
20
40
60
80
100
Predictive strength
NO2 and SO2 for PM 2.5 and SO2
Column values
Increase
Decrease
Total
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Values
NO2 (Count distinct)
SO2 (Count)
NO2 and SO2, Stn Code with RSPM/PM10 and PM 2.5 for City/Town/Village/Area
RSPM/PM10 (Cou...
PM 2.5 (Count dist...
SO2 - Stn Code
NO2
22
26
30
34
38
42
46
5

Phase 4 : Development Part 2

Perform the air quality analysis and create visualizations.

Perform the Air quality in Data analytics

Performing air quality analysis in data analytics involves collecting, processing, and analyzing data to assess the quality of the air in a specific region or location. This process is crucial for environmental monitoring, public health assessment, and policy-making. Here is an overview of the key steps and considerations in air quality analysis using data analytics:

1. Data Collection:

   Gather relevant data sources, including air quality monitoring stations, meteorological data, satellite imagery, and more. These sources may provide data on pollutants such as PM2.5, PM10, ozone (O3), sulfur dioxide (SO2), nitrogen dioxide (NO2), and carbon monoxide (CO).

2. Data Preprocessing:

   Clean and preprocess the data to address missing values, outliers, and inconsistencies. This step ensures that the data is reliable and suitable for analysis.

3. Data Integration:

   Combine data from various sources to create a comprehensive dataset that includes air quality measurements and relevant contextual information, such as weather conditions and geographic coordinates.

4. Exploratory Data Analysis (EDA):

   Conduct EDA to gain insights into the data. Visualizations, statistical summaries, and correlation analyses can help identify trends, patterns, and potential relationships between variables.

5. Feature Engineering:

   Create new features or variables that might enhance the analysis, such as daily averages, pollution indices, or spatial aggregation of data.

6. Time-Series Analysis:

   Given that air quality data is often collected over time, perform time-series analysis to identify long-term trends, seasonality, and patterns in air quality.

7. Spatial Analysis:

Utilize geographic information systems (GIS) to analyze the spatial distribution of air quality data. This can help identify areas with higher or lower pollution levels.

8. Machine Learning Models:

Develop predictive models using machine learning techniques to forecast air quality conditions based on historical data and relevant features. Models may include regression, classification, or time-series forecasting algorithms.

9. Anomaly Detection:

Implement anomaly detection algorithms to identify unusual spikes or dips in air quality, which could be indicative of pollution events or equipment malfunctions.

10. Data Visualization:

Create interactive dashboards and visualizations to make the results more accessible to non-technical stakeholders. Tools like Tableau, Power BI, or custom web-based dashboards can be useful for this purpose.

11. Interpretation and Insights:

Interpret the results of the analysis and provide actionable insights. This information can be used for policy recommendations, public awareness campaigns, or pollution control strategies.

12. Model Validation:

Evaluate the performance of predictive models through validation techniques like cross-validation, root mean square error (RMSE), or coefficient of determination (R-squared) to ensure their accuracy and reliability.

13. Continuous Monitoring:

Establish a system for continuous monitoring of air quality and update the analysis as new data becomes available. This allows for the detection of emerging trends or changes in air quality over time.

14. Reporting and Communication:

Communicate the findings to relevant stakeholders, including government agencies, environmental organizations, and the public, using clear and accessible reports and presentations.

Visualization for Air quality analysis

Creating visualizations in air quality analysis is crucial for interpreting data and communicating findings effectively. Here are some common types of visualizations and the tools you can use to create them:

1. Time Series Plots:
   - Display the variation of air quality parameters over time, such as daily, monthly, or yearly trends.
   - Tools: Python libraries like Matplotlib, Seaborn, or R's ggplot2, or data visualization tools like Tableau and Power BI.

2. Heatmaps:
   - Visualize the spatial distribution of air quality parameters using color-coded grids.
   - Tools: Python libraries (e.g., Matplotlib, Seaborn), GIS software (e.g., QGIS), or mapping tools like Google Maps API.

3. Box Plots:
   - Show the distribution of air quality data, including median, quartiles, and potential outliers. Tools: Matplotlib, Seaborn, R's ggplot2, or data visualization tools.

4. Bar Charts and Histograms:
   - Display the frequency distribution of air quality values or comparisons between different locations or time periods.
   - Tools: Matplotlib, Seaborn, ggplot2, or data visualization tools.
5. Scatter Plots:
   - Illustrate relationships between air quality parameters and other variables, such as weather conditions. Tools: Matplotlib, Seaborn, ggplot2, or data visualization tools.

6. Contour Maps:
   - Represent air quality data on geographic maps using contour lines to show spatial patterns.
   - Tools: GIS software (e.g., ArcGIS, QGIS), Python libraries like Basemap or Cartopy, or data visualization tools with mapping capabilities.

7. Pie Charts:
   - Show the composition of air quality data by pollutant type or source.
   - Tools: Matplotlib, Seaborn, ggplot2, or data visualization tools.

8. Radar Charts:
   - Compare air quality parameters across multiple categories or locations.
   - Tools: Libraries like Plotly, Matplotlib, or specialized radar chart tools.

9. Dashboard Visualizations:
   - Create interactive dashboards with multiple visualizations to provide a comprehensive view of air quality data.
   - Tools: Data visualization platforms like Tableau, Power BI, Plotly Dash, or custom web development with JavaScript libraries.

10. Animation:
    - Animate time-series data to show how air quality parameters change over time.
    - Tools: Python libraries like Matplotlib (for creating animations) or specialized animation tools.

11. Geospatial Heatmaps:
    - Visualize geographic patterns using heatmap layers on maps.
    - Tools: Libraries like Folium (for Python), Leaflet, Google Maps API, or GIS software.

12. 3D Plots:
    - Represent air quality data in a three-dimensional space to highlight complex relationships or patterns.
    - Tools: Libraries like Plotly, Matplotlib (3D toolkit), or specialized 3D visualization software.

Execution:

Diagram 1:

Bar graph using Location of monitoring station by City/Town/Village/Area coloured by SO2

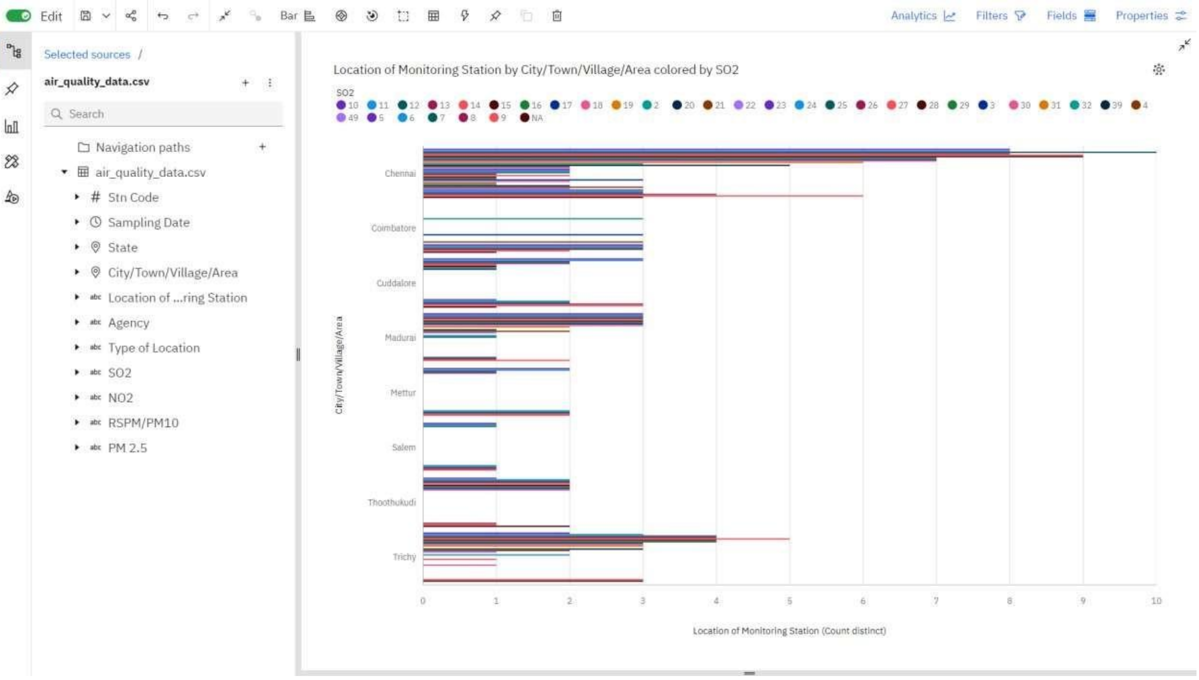


Diagram 2 :

Bar diagram using SO2 by location of Monitoring station coloured by City/Town/Village/Area

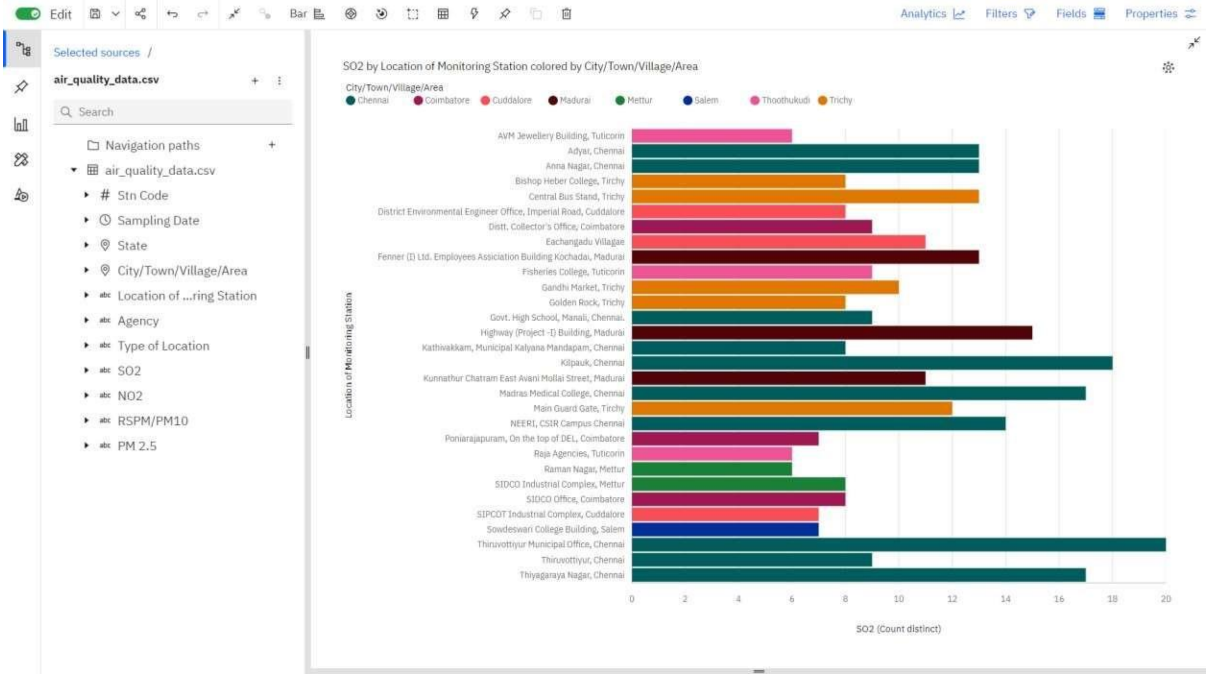

Edit
Bar
Analytics
Filters
Fields
Properties
Selected sources /
air_quality_data.csv
Search
Navigation paths
air_quality_data.csv
Stn Code
Sampling Date
State
City/Town/Village/Area
Location of ...ring Station
Agency
Type of Location
SO2
NO2
RSPM/PM10
PM 2.5
SO2 by Location of Monitoring Station colored by City/Town/Village/Area
City/Town/Village/Area
Chennai
Coimbatore
Cuddalore
Madurai
Mettur
Salem
Thoothukudi
Trichy
Location of Monitoring Station
AVM Jewellery Building, Tuticorin
Adyar, Chennai
Anna Nagar, Chennai
Bishop Heber College, Tirchy
Central Bus Stand, Trichy
District Environmental Engineer Office, Imperial Road, Cuddalore
Distt. Collector's Office, Coimbatore
Eachangadu Villagae
Fenner (I) Ltd. Employees Asiciation Building Kochadai, Madurai
Fisheries College, Tuticorin
Gandhi Market, Trichy
Golden Rock, Trichy
Govt. High School, Manali, Chennai.
Highway (Project -I) Building, Madurai
Kathivakkam, Municipal Kalyana Mandapam, Chennai
Kilpauk, Chennai
Kunnathur Chatram East Avani Mollai Street, Madurai
Madras Medical College, Chennai
Main Guard Gate, Tirchy
NEERI, CSIR Campus Chennai
Poniarajapuram, On the top of DEL, Coimbatore
Raja Agencies, Tuticorin
Raman Nagar, Mettur
SIDCO Industrial Complex, Mettur
SIDCO Office, Coimbatore
SIPCOT Industrial Complex, Cuddalore
Sowdeswari College Building, Salem
Thiruvottiyur Municipal Office, Chennai
Thiruvottiyur, Chennai
Thiyagaraya Nagar, Chennai
0 2 4 6 8 10 12 14 16 18 20
SO2 (Count distinct)

Diagram 3:

Rubble graph using NO2 by type of location coloured by City/Town/Village And sized by Location of Monitoring Station

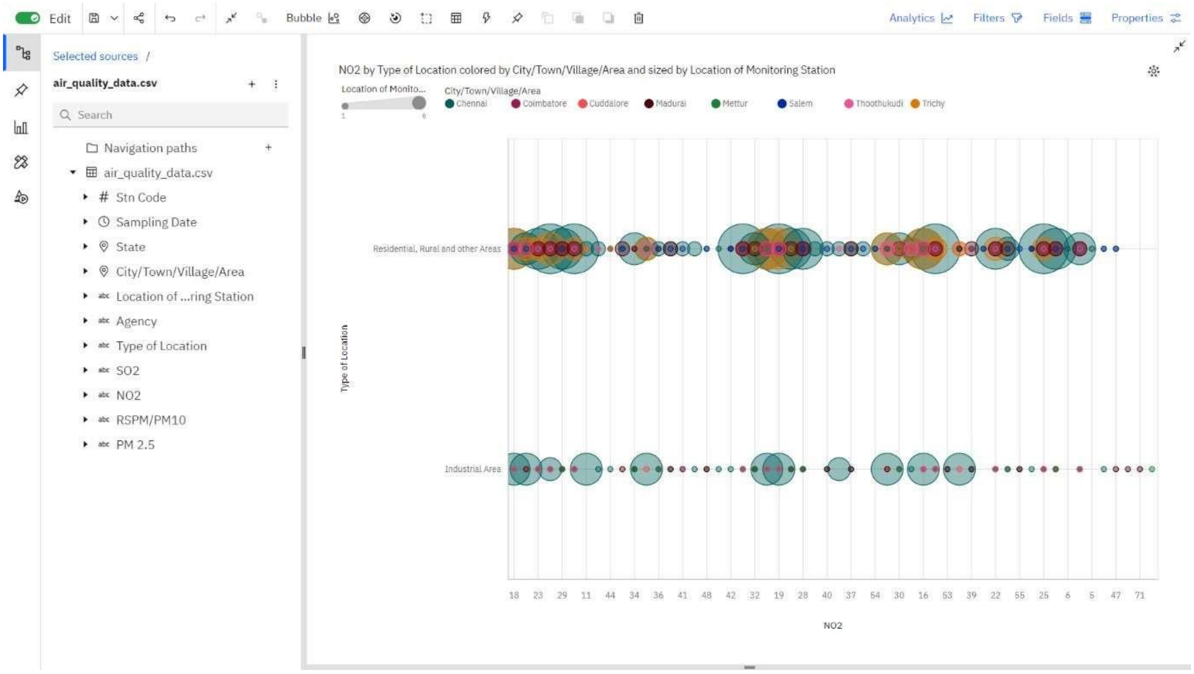


Diagram 4:

Column graph using SO2 by using Location of Monitoring Station by City/Town/Village/Area

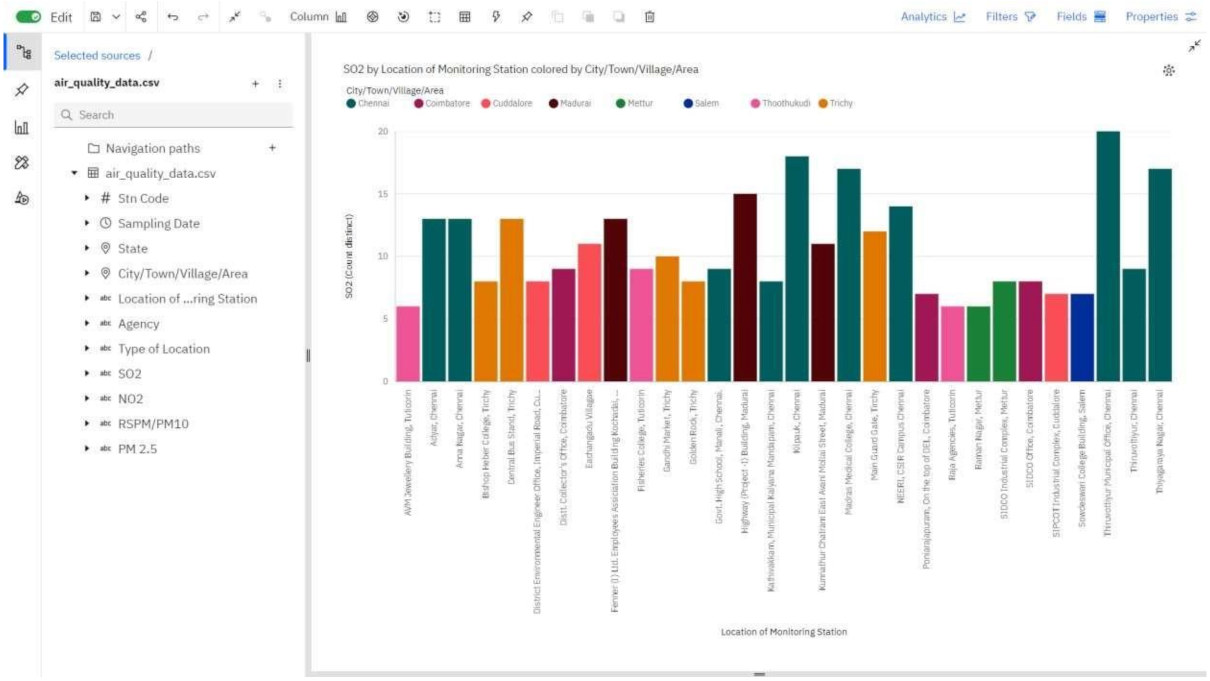

Edit
Column
Analytics
Filters
Fields
Properties
Selected sources /
air_quality_data.csv
Search
Navigation paths
air_quality_data.csv
Stn Code
Sampling Date
State
City/Town/Village/Area
Location of ...ring Station
Agency
Type of Location
SO2
NO2
RSPM/PM10
PM 2.5
SO2 by Location of Monitoring Station colored by City/Town/Village/Area
City/Town/Village/Area
Chennai
Coimbatore
Cuddalore
Madurai
Mettur
Salem
Thoothukudi
Trichy
SO2 (Count distinct)
20
15
10
5
0
AVM Jewellery Building, Tuticorin
Adyar, Chennai
Anna Nagar, Chennai
Bishop Heber College, Trichy
Central Bus Stand, Trichy
District Environmental Engineer Office, Imperial Road, Cu...
Distt. Collector's Office, Coimbatore
Eachangadu Villagae
Fenner (I) Ltd. Employees Assiciation Building Kochadai, ...
Fisheries College, Tuticorin
Gandhi Market, Trichy
Golden Rock, Trichy
Govt. High School, Manali, Chennai.
Highway (Project -I) Building, Madurai
Kathivakkam, Municipal Kalyana Mandapam, Chennai
Kilpauk, Chennai
Kunnathur Chatram East Avani Moolai Street, Madurai
Madras Medical College, Chennai
Main Guard Gate, Trichy
NEERI, CSIR Campus Chennai
Ponlarajapuram, On the top of DEL, Coimbatore
Raja Agencies, Tuticorin
Raman Nagar, Mettur
SIDCO Industrial Complex, Mettur
SIDCO Office, Coimbatore
SIPCOT Industrial Complex, Cuddalore
Sowdeswari College Building, Salem
Thiruvottiyur Municipal Office, Chennai
Thiruvottiyur, Chennai
Thiyagaraya Nagar, Chennai
Location of Monitoring Station

Diagram 5:

Heat map graph using PM 2.5 by Location of Monitoring station and City/Town/Village/Area

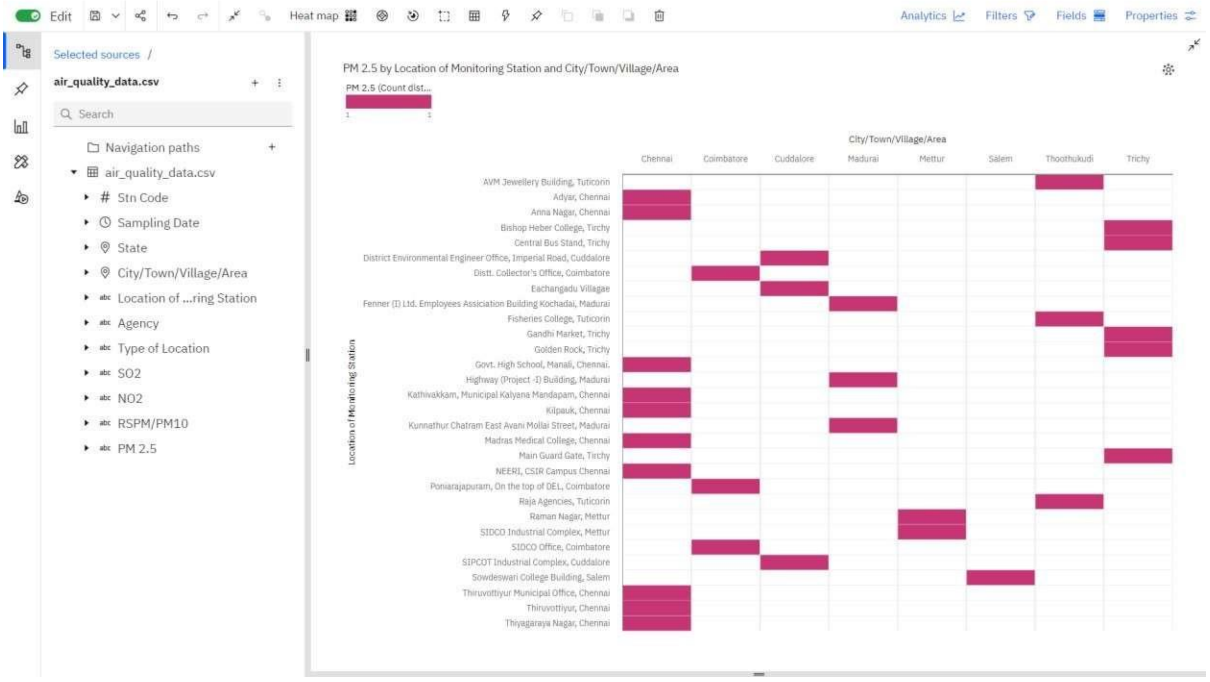

Edit
Heat map
Analytics
Filters
Fields
Properties
Selected sources /
air_quality_data.csv
Search
Navigation paths
air_quality_data.csv
Stn Code
Sampling Date
State
City/Town/Village/Area
Location of ...ring Station
Agency
Type of Location
SO2
NO2
RSPM/PM10
PM 2.5
PM 2.5 by Location of Monitoring Station and City/Town/Village/Area
PM 2.5 (Count dist...
1
1
City/Town/Village/Area
Chennai
Coimbatore
Cuddalore
Madurai
Mettur
Salem
Thoothukudi
Trichy
Location of Monitoring Station
AVM Jewellery Building, Tuticorin
Adyar, Chennai
Anna Nagar, Chennai
Bishop Heber College, Tirchy
Central Bus Stand, Trichy
District Environmental Engineer Office, Imperial Road, Cuddalore
Distt. Collector's Office, Coimbatore
Eachangadu Villagae
Fenner (I) Ltd. Employees Assiciation Building Kochadai, Madurai
Fisheries College, Tuticorin
Gandhi Market, Trichy
Golden Rock, Trichy
Govt. High School, Manali, Chennai.
Highway (Project -I) Building, Madurai
Kathivakkam, Municipal Kalyana Mandapam, Chennai
Kilpauk, Chennai
Kunnathur Chatram East Avani Mollai Street, Madurai
Madras Medical College, Chennai
Main Guard Gate, Tirchy
NEERI, CSIR Campus Chennai
Poniarajapuram, On the top of DEL, Coimbatore
Raja Agencies, Tuticorin
Raman Nagar, Mettur
SIDCO Industrial Complex, Mettur
SIDCO Office, Coimbatore
SIPCOT Industrial Complex, Cuddalore
Sowdeswari College Building, Salem
Thiruvottiyur Municipal Office, Chennai
Thiruvottiyur, Chennai
Thiyagaraya Nagar, Chennai

Diagram 6:

Area graph using Location of monitoring station by SO2 colored by City/Town/Village/Area

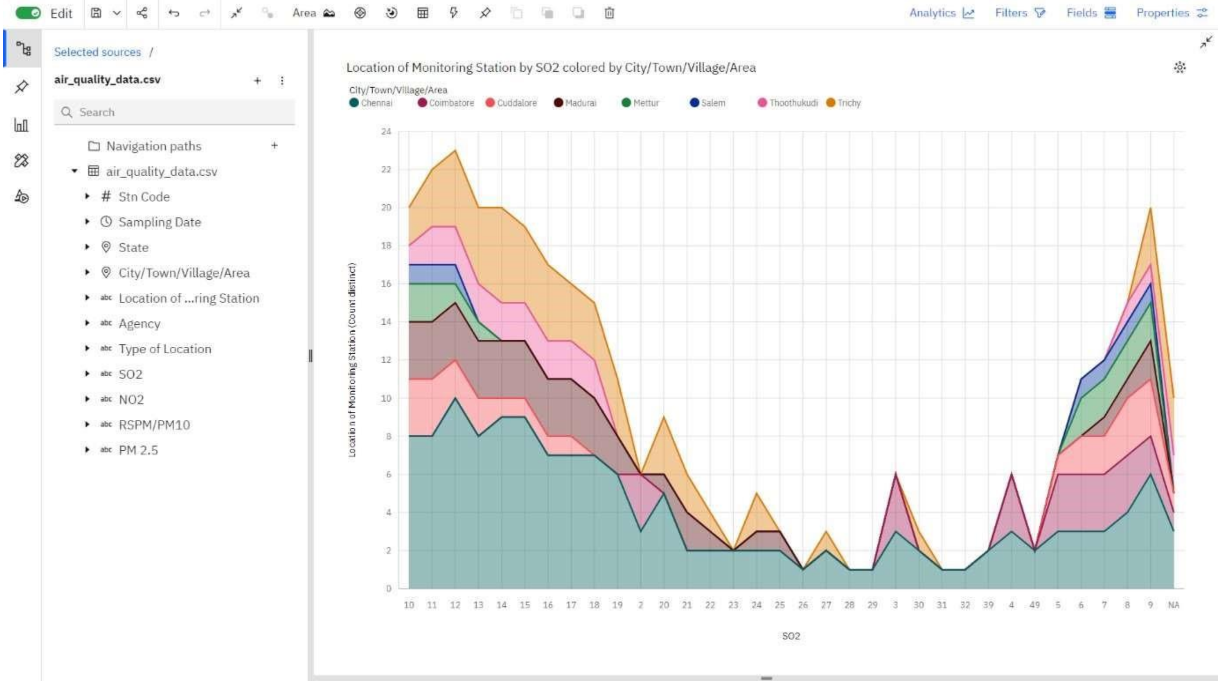


Diagram 7:

Bar graph using Type of location by SO2 colored by City/Town/Village /Area

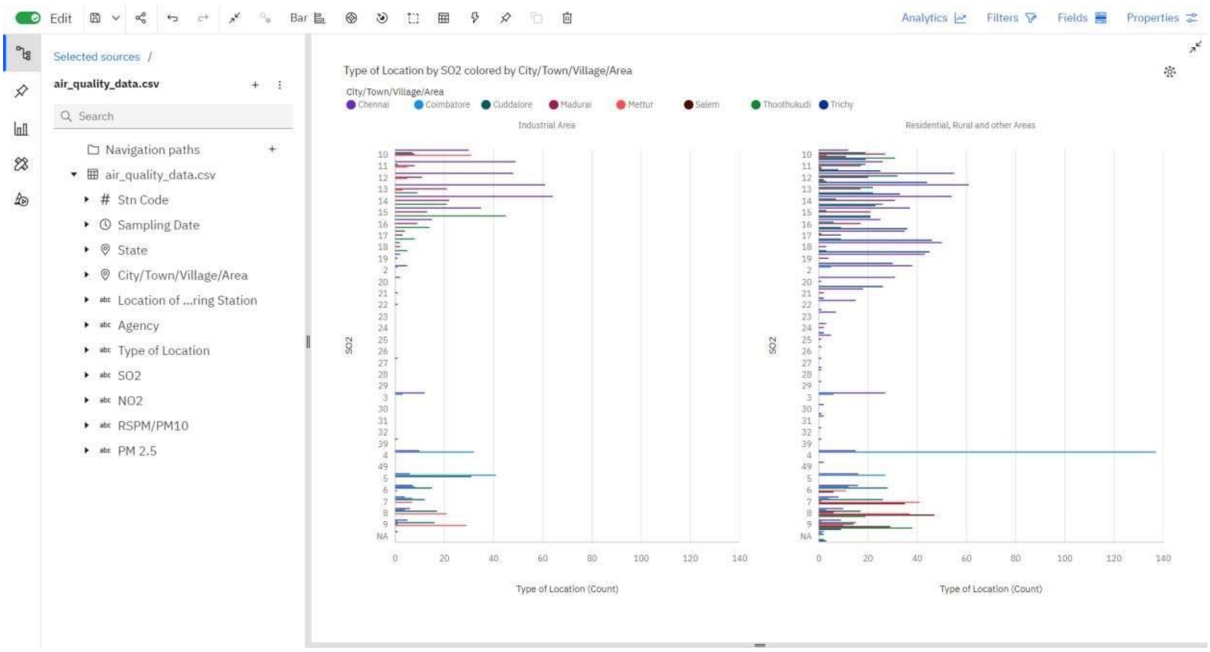


Diagram 8:

Bubble graph using SO2 and NO2 with City/Town/Village/Area and Location station for Type of location

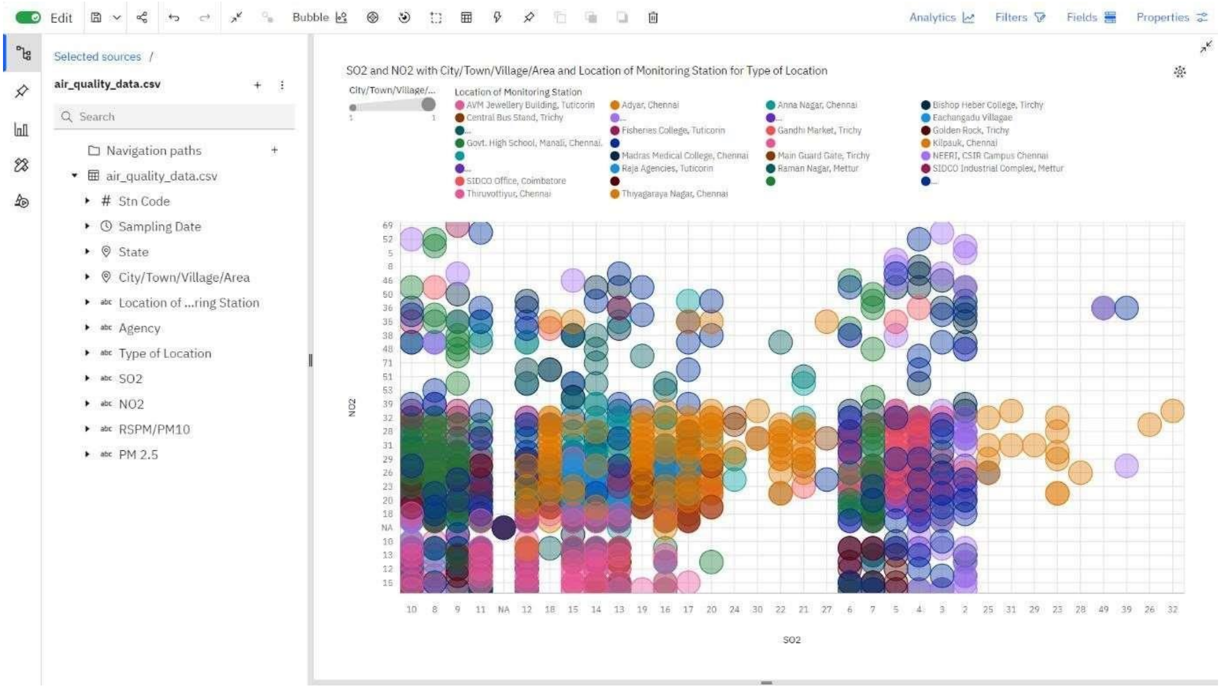


Diagram 9:

Pie graph using City/Town/Village/Area by SO2

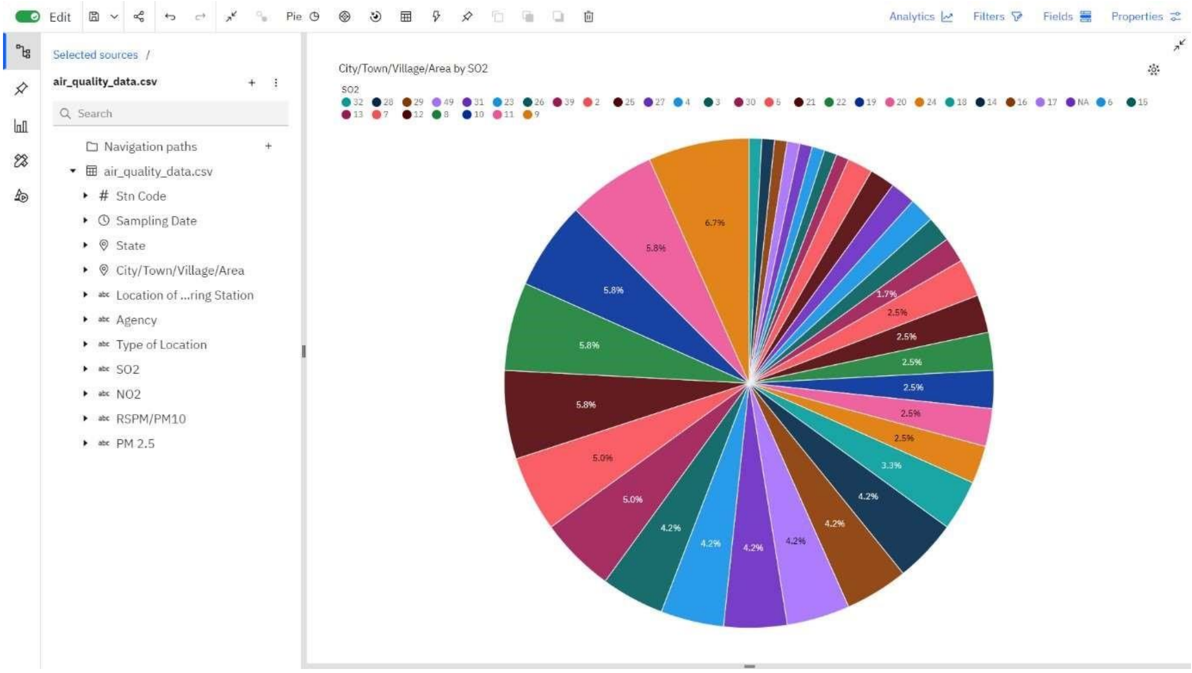


Diagram 10:

Stacked column graph using NO2,SO2,PM 2.5,RSP</PM10, City/Town/Vilage/Area

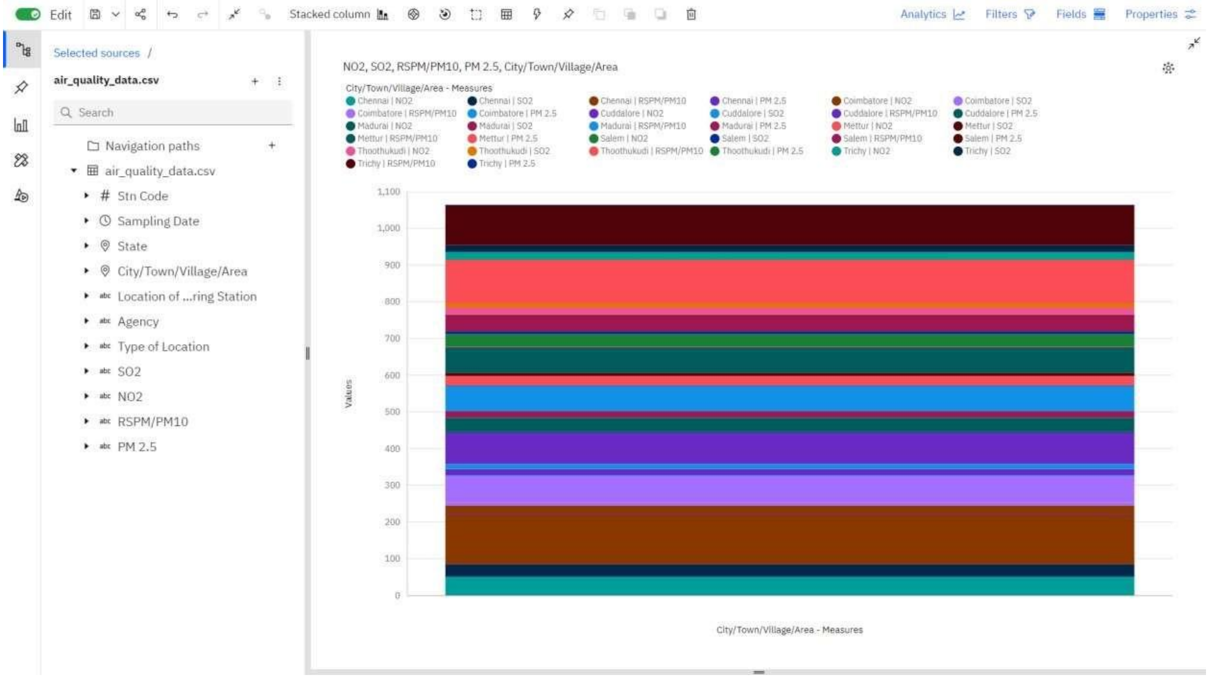


Diagram 11:

Scatter graph using NO2 by SO2 with points for Type of Location

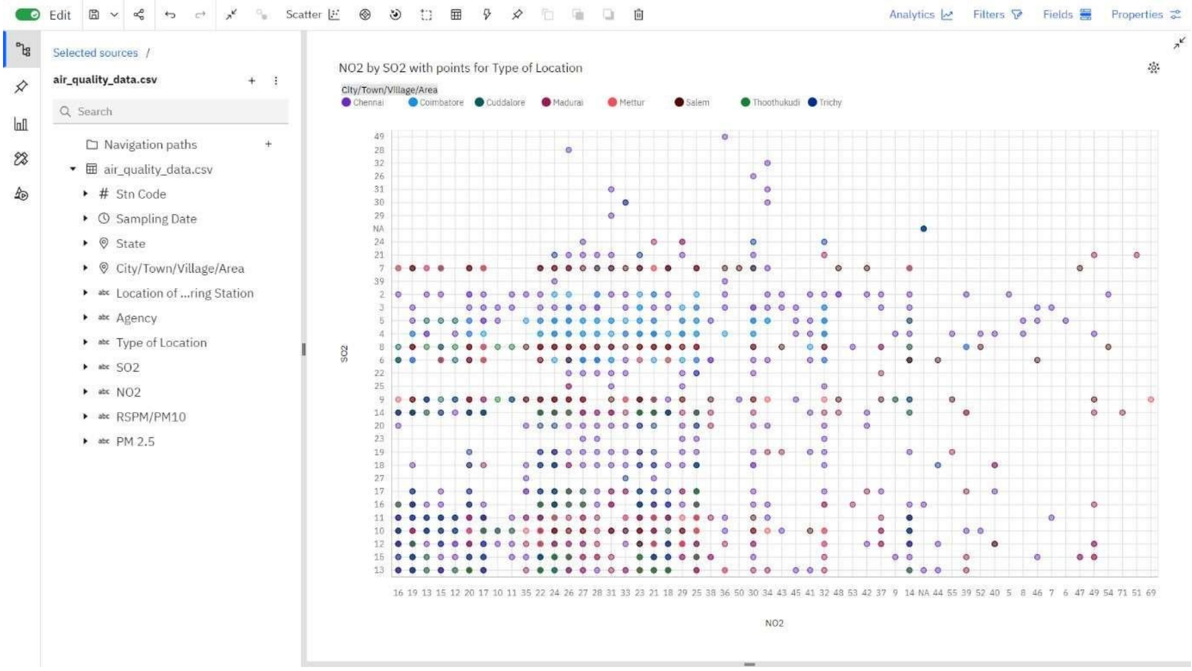
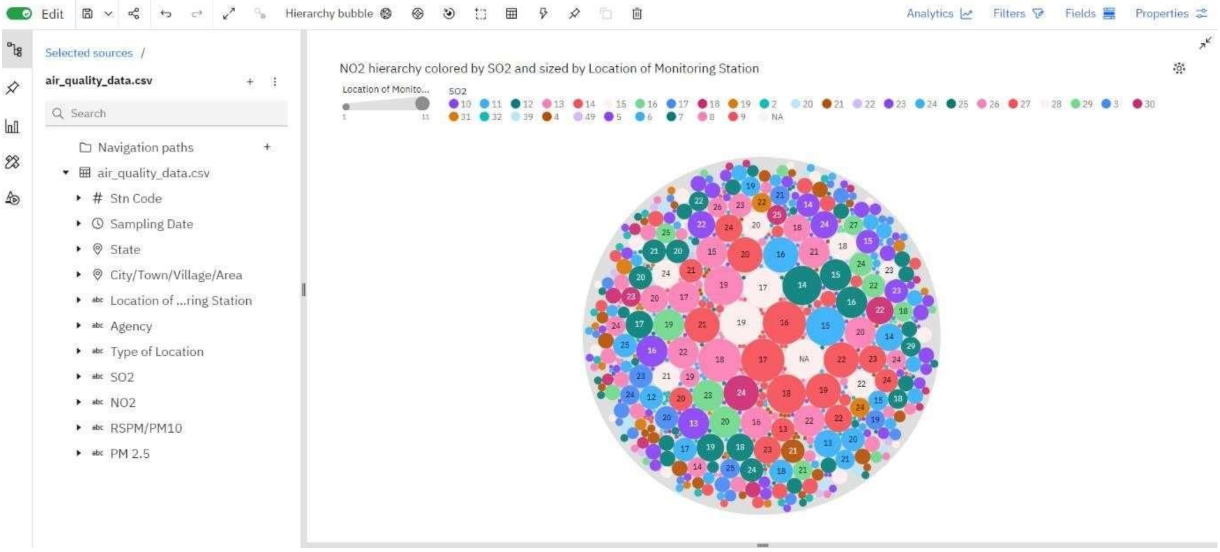


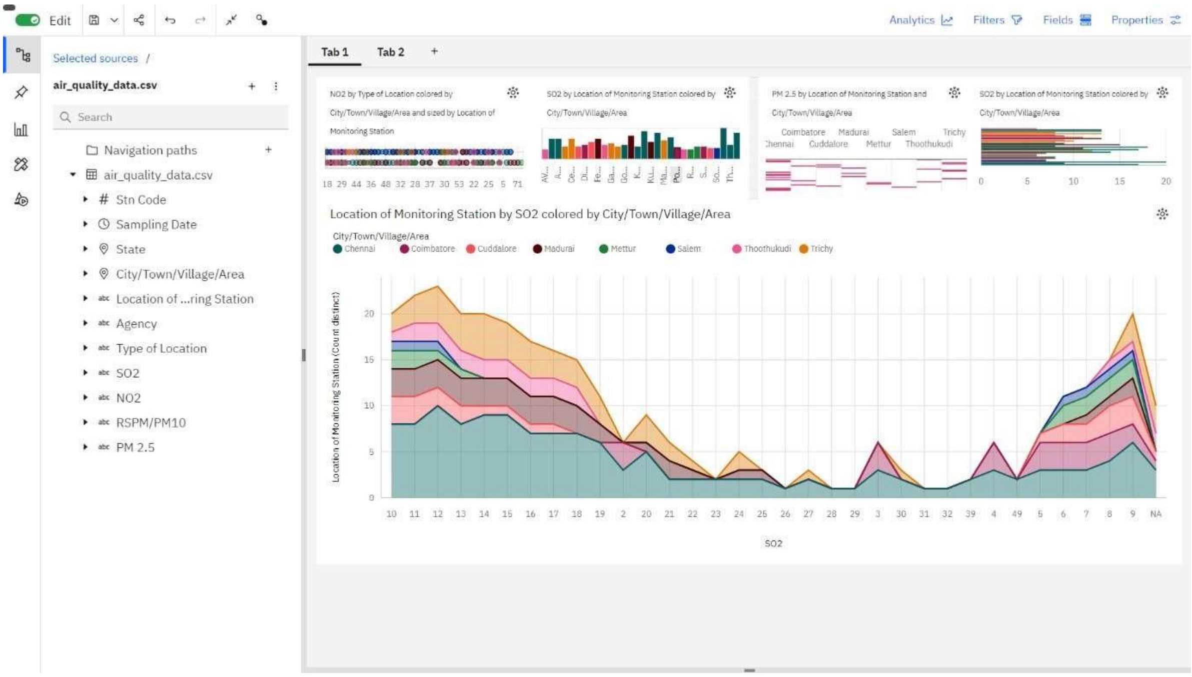
Diagram 12:

Hierarchy bubble graph using NO2 hierarchy colored by SO2 and sized by Location of Monitoring station

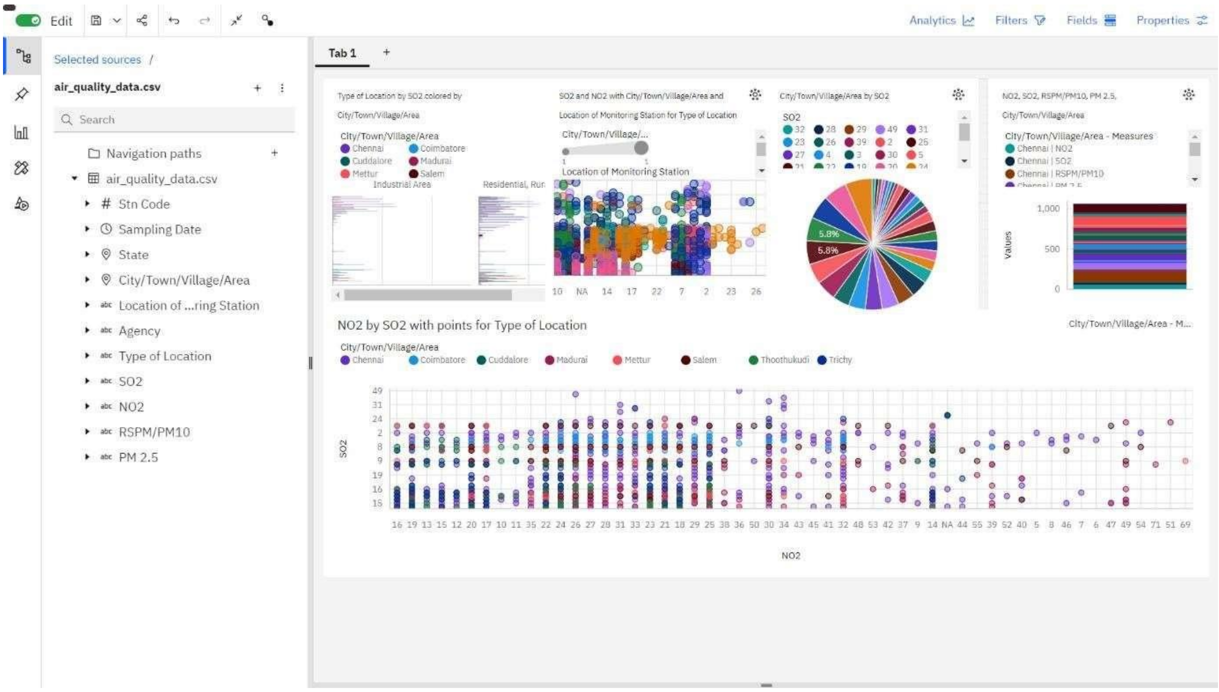


Overview :

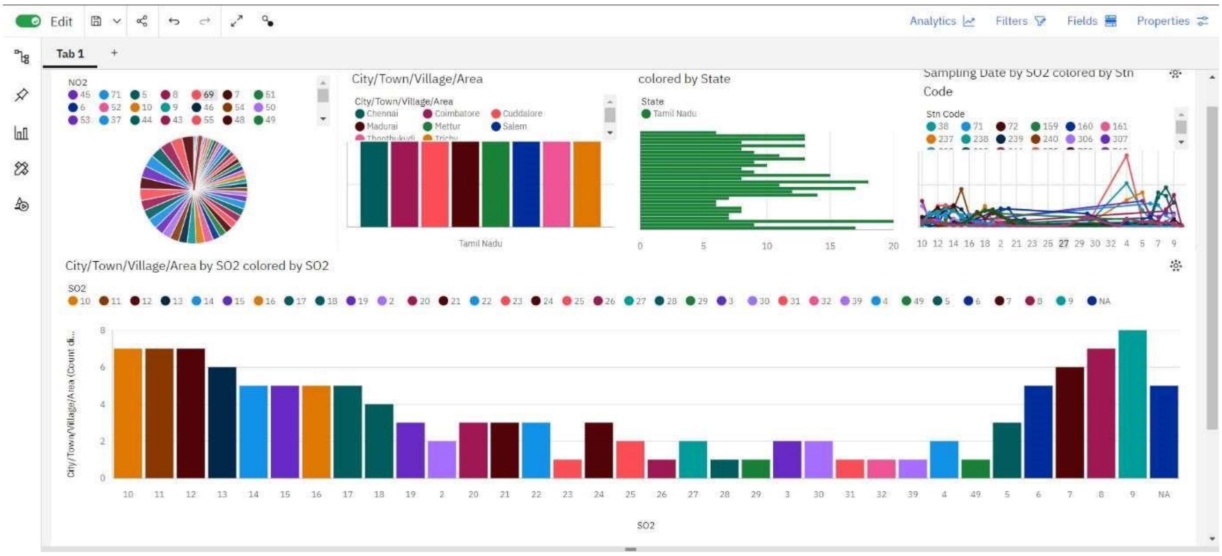
Graph 1:


Edit
Analytics
Filters
Fields
Properties
Selected sources /
air_quality_data.csv
Search
Navigation paths
air_quality_data.csv
Stn Code
Sampling Date
State
City/Town/Village/Area
Location of ...ring Station
Agency
Type of Location
SO2
NO2
RSPM/PM10
PM 2.5
Tab 1
Tab 2
NO2 by Type of Location colored by City/Town/Village/Area and sized by Location of Monitoring Station
SO2 by Location of Monitoring Station colored by City/Town/Village/Area
PM 2.5 by Location of Monitoring Station and City/Town/Village/Area
SO2 by Location of Monitoring Station colored by City/Town/Village/Area
Location of Monitoring Station by SO2 colored by City/Town/Village/Area
City/Town/Village/Area
Chennai
Coimbatore
Cuddalore
Madurai
Mettur
Salem
Thoothukudi
Trichy
Location of Monitoring Station (Count distinct)
SO2
0
5
10
15
20
10 11 12 13 14 15 16 17 18 19 2 20 21 22 23 24 25 26 27 28 29 3 30 31 32 39 4 49 5 6 7 8 9 NA

Graph 2:


Edit
Analytics
Filters
Fields
Properties
Selected sources /
air_quality_data.csv
Search
Navigation paths
air_quality_data.csv
Stn Code
Sampling Date
State
City/Town/Village/Area
Location of ...ring Station
Agency
Type of Location
SO2
NO2
RSPM/PM10
PM 2.5
Tab 1
Type of Location by SO2 colored by City/Town/Village/Area
Industrial Area
Residential, Rur
SO2 and NO2 with City/Town/Village/Area and Location of Monitoring Station for Type of Location
City/Town/Village/Area by SO2
NO2, SO2, RSPM/PM10, PM 2.5, City/Town/Village/Area
City/Town/Village/Area - Measures
Chennai | NO2
Chennai | SO2
Chennai | RSPM/PM10
Values
0
500
1,000
5.8%
5.8%
NO2 by SO2 with points for Type of Location
City/Town/Village/Area
Chennai
Coimbatore
Cuddalore
Madurai
Mettur
Salem
Thoothukudi
Trichy
SO2
NO2
16 19 13 15 12 20 17 10 11 35 22 24 26 27 28 31 33 23 21 18 29 25 38 36 50 30 34 43 45 41 32 48 53 42 37 9 14 NA 44 55 39 52 40 5 8 46 7 6 47 49 54 71 51 69

Graph 3:



Conclusion:

In conclusion, the integration of data analytics with IBM Cognos or similar platforms revolutionizes air quality analysis. It equips decision-makers, researchers, and environmentalists with powerful tools to comprehensively assess air quality, respond to immediate concerns, and plan for a cleaner, healthier, and more sustainable future. As air quality continues to be a global concern, data analytics offers a vital means to better understand, address, and ultimately improve the quality of the air we breath