Air quality analysis in Tamilnadu

# Introduction:

Poor air quality has severe health implications, leading to respiratory diseases, cardiovascular issues, and other health problems. Air pollution not only affects human health but also harms the environment. It can lead to acid rain, damage to ecosystems, and the deterioration of historical monuments. Tamil Nadu has taken steps to monitor and control air pollution. Government agencies such as the Tamil Nadu Pollution Control Board (TNPCB) oversee air quality regulations and emission standards. Continuous monitoring stations are set up across the state to collect air quality data. Air quality analysis relies on advanced technologies, including air quality monitoring stations, satellite imagery, and data analytics. Innovations in sensor technology and data integration are helping to provide real-time air quality information.

# About phase 5:

in this part we create the document which contains steps involed in the project like the project's objectives, analysis approach, visualization techniques,code implementation, and the outputs of data analysis and visualization.

# About IBM cognos:

Cognos is a business intelligence performance management tools for IBM that allows technical and non-technical employees in any company to analyse, extract and create interactive dashboards that enable the company to take relevant key decisions. The Cognos is an intelligence-gathering platform for business that provides an analytical solution for business needs that is scalable and self-service. The highly interactive nature makes it a good way of creating user-friendly dashboards and reports for every company.

# Visualization using IBM cognos:

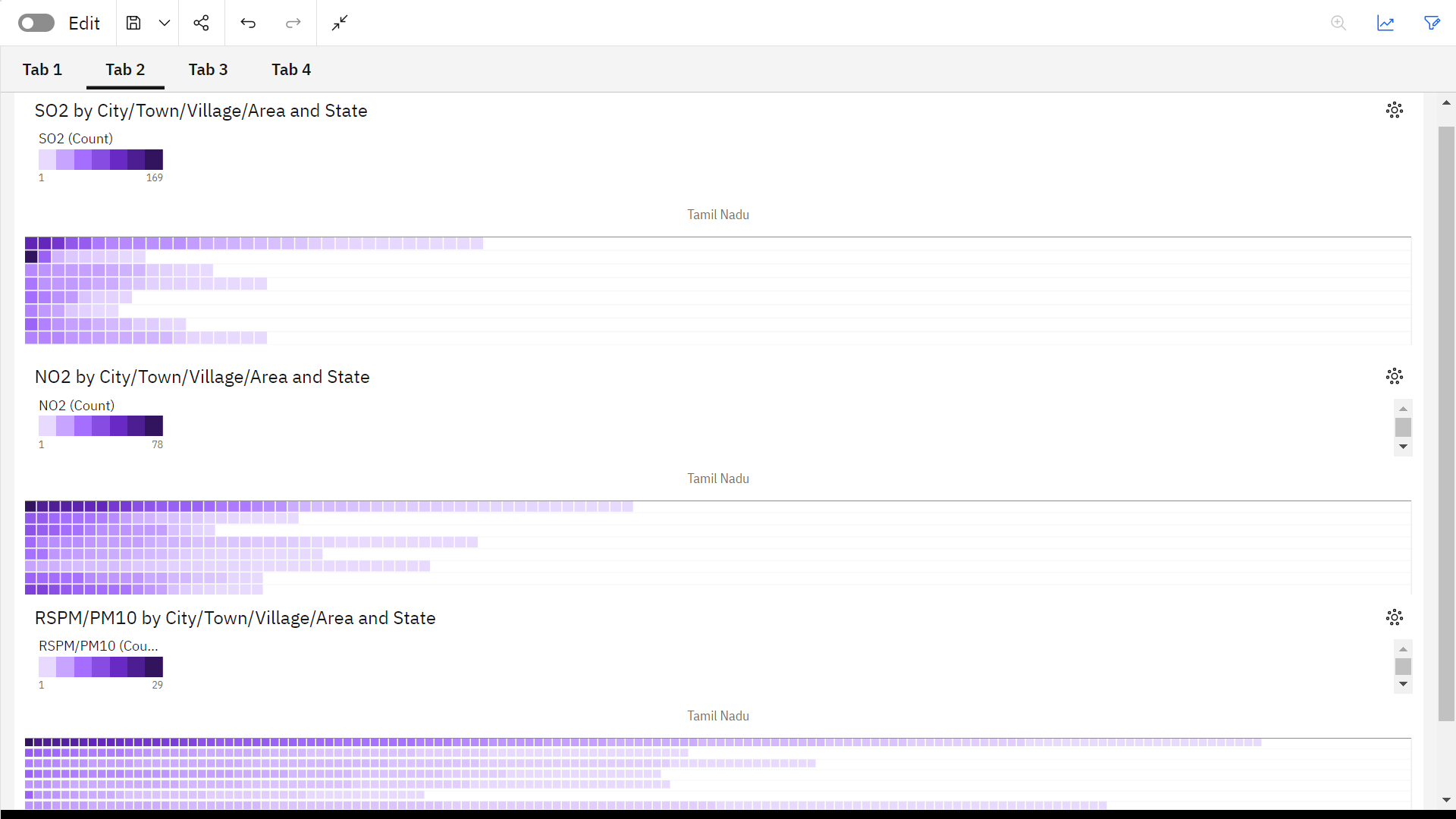
## Histogram:

A histogram is a graphical representation of the distribution of data. It's commonly used to visualize the frequency or count of data points within predefined intervals or "bins." The x-axis represents the data range or values, and the y-axis represents the frequency or count of data points falling into each bin.

## 

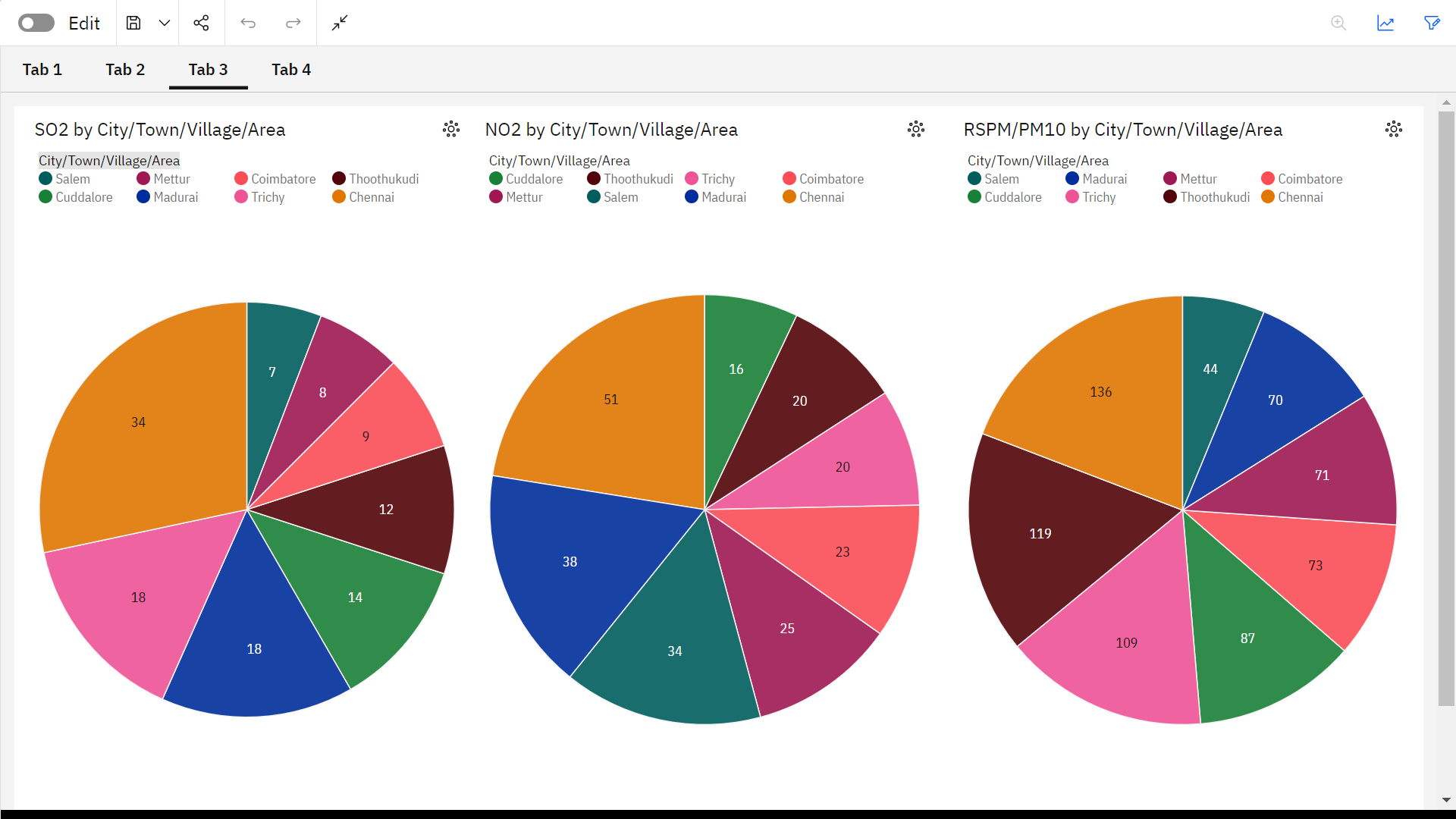
## Heatmap:

A heatmap is a graphical representation of data where individual values are represented as colors. It is a way to visualize data in a two-dimensional space, with each data point or value represented by a colored cell. Heatmaps are particularly useful for showing patterns, correlations, and variations in data, making them easier to understand and interpret.



## Pie chart:

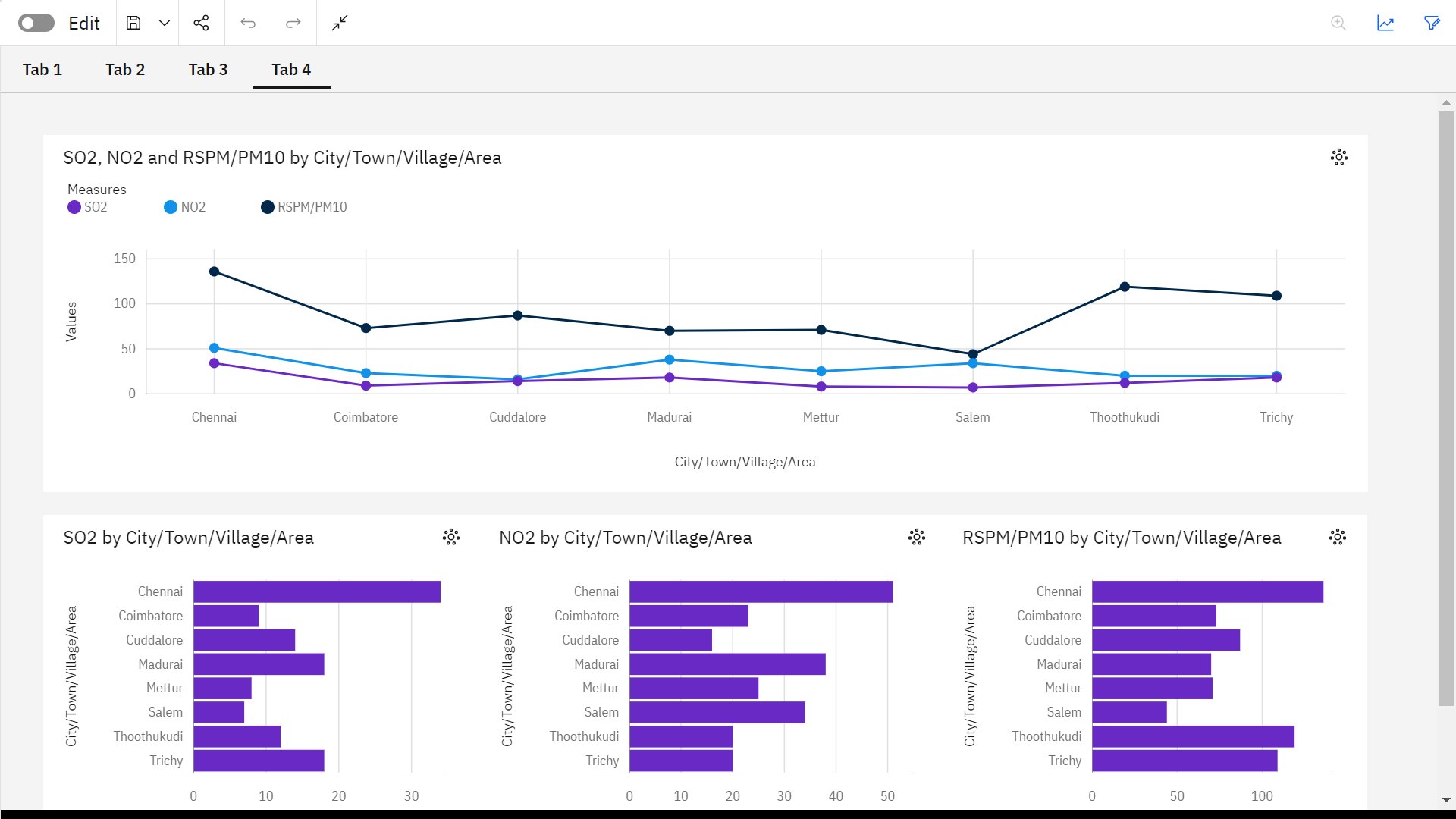
A pie chart is a circular statistical graphic that is used to represent data in a simple and visually appealing way. It is a type of chart that divides a circle into sectors or "slices," where each sector represents a proportion of the whole. The size of each sector is typically proportional to the quantity it represents. Pie charts are particularly useful for showing the composition or distribution of a data set.



## Line chart and bar chart:

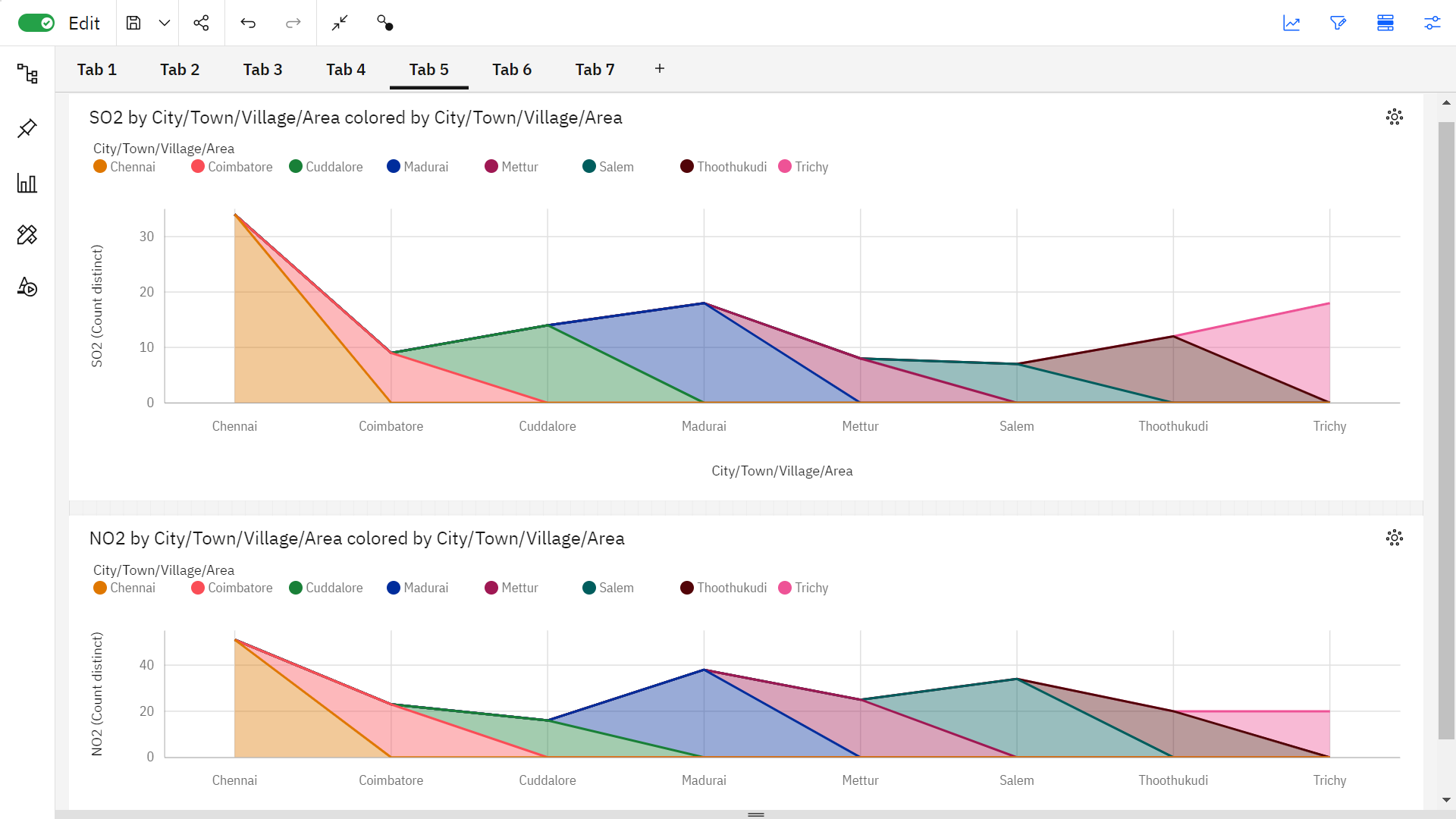
A bar chart is a graphical representation of data using rectangular bars or columns of varying heights. Each bar typically represents a category or group, and the height of the bar is proportional to the value it represents. Bar charts can be either vertical (column chart) or horizontal (bar chart), depending on the orientation of the bars.

A line chart is a type of data visualization that is used to represent data points over a continuous interval or time period, and it connects these data points with straight lines. It is particularly useful for showing trends, changes, or fluctuations in data over time. Each data point in a line chart is usually represented as a marker (e.g., a point or circle) at specific coordinates, and the markers are connected by lines to form a continuous representation of the data.



## Area chart:

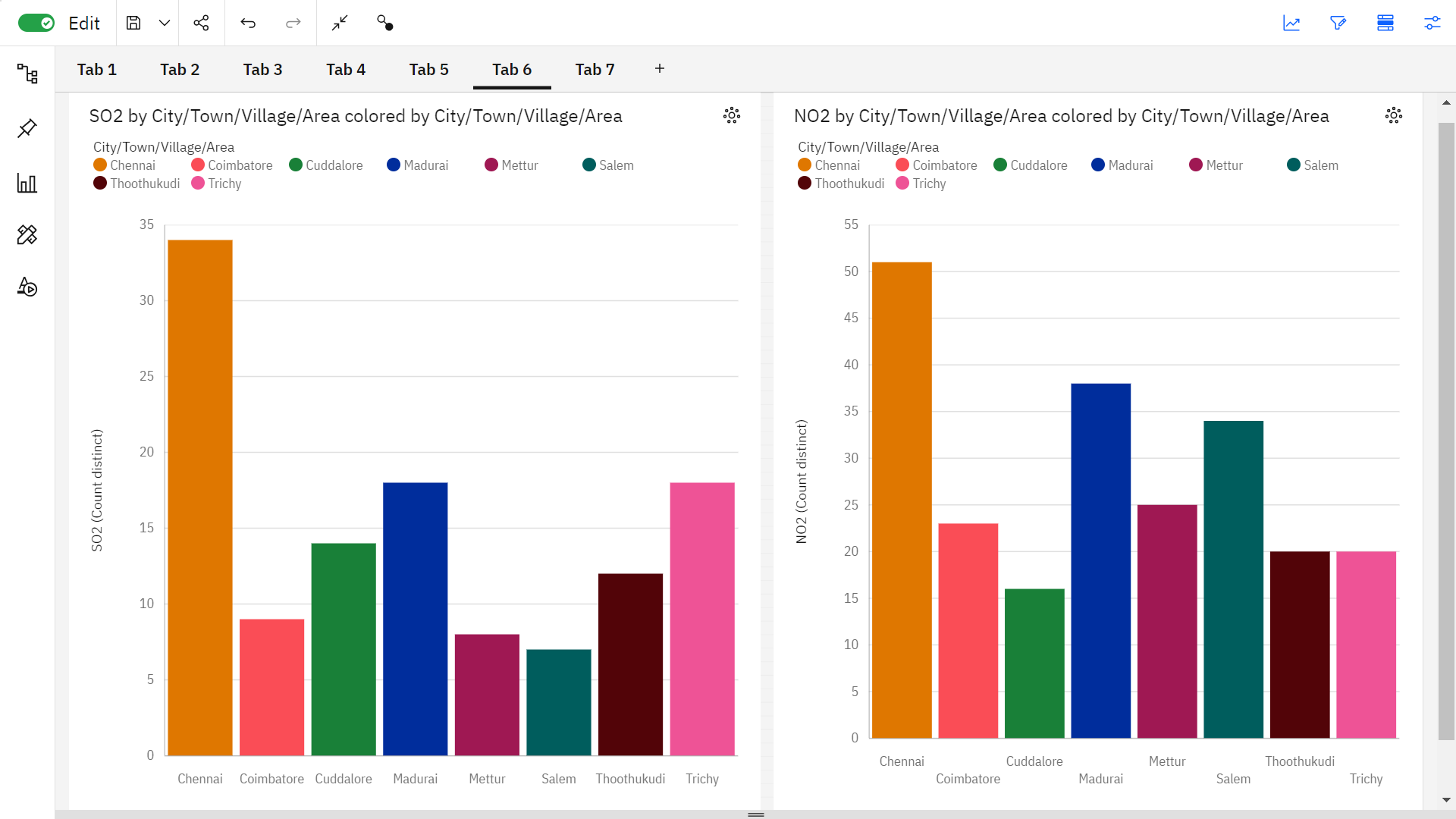
## An area chart or area graph is based on the line chart but is used to primarily communicate the summation of data rather than represent individual data values, as in line charts. The area between axis and line is usually emphasized with colors, textures or hatchings. The area underneath the line can help one to depict how data progressed with time and can be an excellent way to compare values without going too deep.



## Stacked column chart:

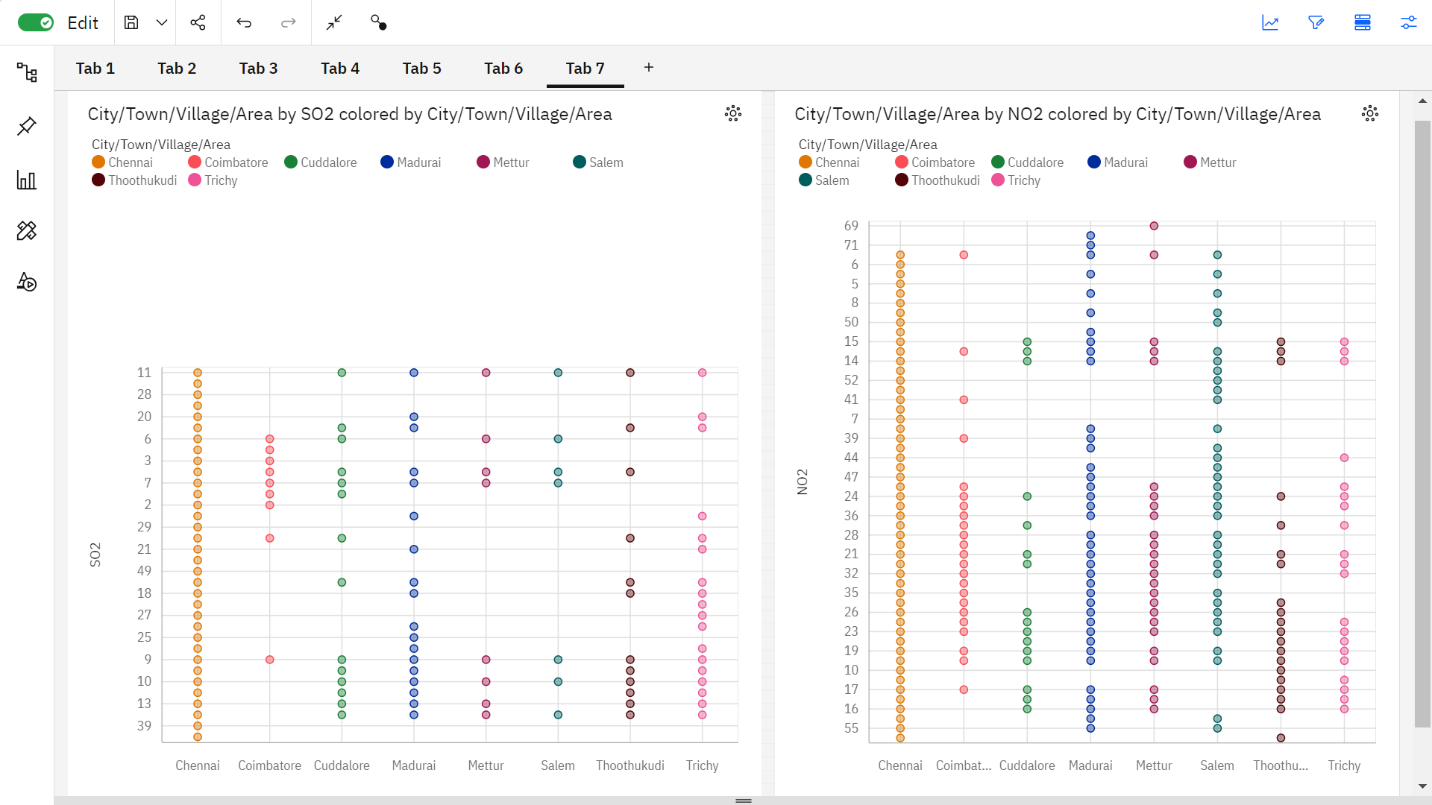
A stacked column chart is a basic Excel chart type to allow part-to-whole comparisons over time, or across categories. In a stacked column chart, data series are stacked one on top of the other in vertical columns.

Stacked column charts can show change over time because it's easy to compare total column lengths.



## Bubble chart:

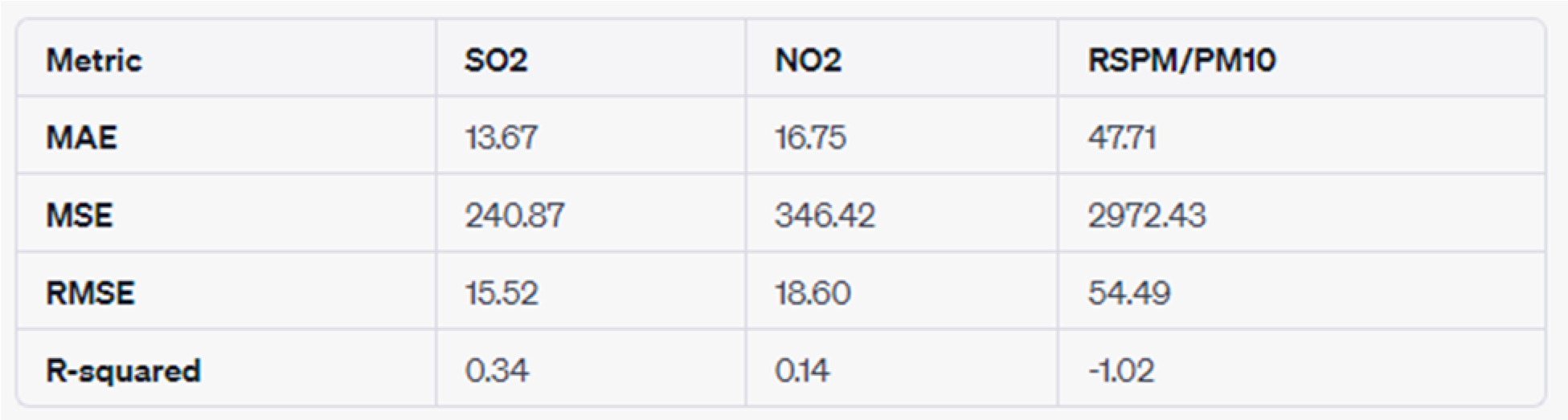
A bubble chart (aka bubble plot) is an extension of the scatter plot used to look at relationships between three numeric variables. Each dot in a bubble chart corresponds with a single data point, and the variables’ values for each point are indicated by horizontal position, vertical position, and dot size.



# Performance metrics:

Performance metrics are measurable data used to track processes within a business using activities, employee behavior and productivity as key metrics. These metrics track and measure the achievement of overall business goals.

Accuracy is a metric that generally describes how the model performs across all classes. It is useful when all classes are of equal importance. It is calculated as the ratio between the number of correct predictions to the total number of predictions.



## Mean Absolute Error (MAE):

SO2: The average absolute difference between the predicted and actual SO2 values is approximately 13.67. This represents the typical magnitude of error in the model's predictions for SO2.

NO2: For NO2, the MAE is around 16.75, indicating a slightly larger average absolute error compared to SO2.

RSPM/PM10: The model's predictions for RSPM/PM10 have an average absolute error of about 47.71.

## Mean Squared Error (MSE):

SO2: The MSE for SO2 is 240.87, reflecting the average squared difference between predicted and actual SO2 values. A lower MSE indicates better model performance.

NO2: MSE is 346.42 for NO2, suggesting a larger spread of squared errors compared to SO2.

RSPM/PM10: The model has a higher MSE of 2972.43 for RSPM/PM10, signifying more variability in the squared errors.

## Root Mean Squared Error (RMSE):

SO2: The RMSE of 15.52 is the square root of the MSE for SO2, providing an interpretable scale similar to the original target variable.

NO2: RMSE is 18.60 for NO2, indicating the average magnitude of error in the model's predictions for NO2.

RSPM/PM10:The RMSE of 54.49 for RSPM/PM10 is relatively high, suggesting larger errors in predicting this variable.

# R-squared:

SO2: The R-squared value of 0.34 indicates that the model explains about 34% of the variability in SO2 values. A higher R-squared is desirable.

NO2: With an R-squared of 0.14, the model's explanatory power is limited for NO2.

RSPM/PM10: The negative R-squared (-1.02) suggests that the model doesn't fit well to the RSPM/PM10 data, and its predictive power is worse than a simple mean.

## Conclusion:

The analysis of air quality data involves several important steps, including data loading, preprocessing, visualization, statistical analysis, and machine learning. The specific steps and libraries used may vary depending on the dataset and analysis goals. The findings from the analysis can provide valuable insights for decision-making and improving air quality.