**FINAL SUBMISSION**

**IBM-Naan Mudhalvan Data Analytics with Cognous**

**Student Name :**  Selvakumar.M

**Register Number :** 620821104104

**Branch :** B.E-CSE

**Year :** 3rd year

**Topic :** Data Analytics with Cognus

**Title :** Air Q Assessment TN

**College :** Gnanamani College of Technology

**Introduction:**

Air quality assessment is a critical endeavor in today's world, particularly in regions experiencing rapid urbanization and industrialization. Tamil Nadu, a diverse and populous state in southern India, is no exception to the challenges posed by air pollution. As the state continues to grow economically and infrastructurally, the need for a comprehensive understanding of air quality becomes increasingly apparent. This analysis seeks to delve into the intricacies of air quality assessment in Tamil Nadu, aiming to provide valuable insights into the state's environmental health, identify sources of pollution, and propose measures for mitigation and improvement.

**Abstract:**

Air quality assessment in Tamil Nadu is a multifaceted examination of the state's atmospheric conditions, pollution sources, and their impacts on human health and the environment. This analysis focuses on defining the objectives, methodologies, and potential outcomes of such an assessment.

The analysis objectives encompass a wide array of facets, including monitoring the levels of key air pollutants, assessing the spatial and temporal variation in air quality, and understanding the health implications for the population. Through data collection methods involving air quality monitoring stations, satellite data, and ground-level surveys, we aim to paint a comprehensive picture of air quality in Tamil Nadu.

The visualization strategy employs advanced tools and technologies to present the collected data in a comprehensible manner. Utilizing Geographic Information Systems (GIS) and data visualization software, we intend to create informative maps, graphs, and reports that reveal patterns and trends in air quality.

Furthermore, code integration will play a crucial role in data analysis and modeling. By employing coding languages and techniques for data processing, statistical analysis, and predictive modeling, we aim to derive actionable insights and suggest strategies for air quality improvement.

The ultimate goal of this air quality assessment is to inform policy decisions, raise public awareness, and catalyze efforts to reduce air pollution, ensuring a healthier and more sustainable future for the people of Tamil Nadu.

I. **Project Objectives**:

A. Define project objectives

* Analyzing air quality trends
* Identifying pollution hotspots
* Building a predictive model for RSPM/PM10 levels

II. **Analysis Approach**

A. Plan the analysis steps

* Data loading and collection
* Data preprocessing and cleaning
* Data analysis and modeling
* Data visualization

B. Define specific methodologies for each step

III. **Visualization Selection**

A. Determine visualization techniques

* Line charts for trend analysis
* Heat maps for pollution hotspot identification
* Scatter plots for predictive model representation

B. Consider tools and software for creating visualizations

This outline provides a structured approach to employing Design Thinking principles in the context of analyzing air quality data, with a focus on defining objectives, planning the analysis approach, and selecting effective visualization techniques.

**1. Data Collection:**

* Collect historical air quality data from various monitoring stations across Tennessee. This data should include information on pollutants such as PM2.5, PM10, ozone (O3), nitrogen dioxide (NO2), sulfur dioxide (SO2), and carbon monoxide (CO).
* Include meteorological data, traffic data, geographical information, and emissions data from industrial sources as additional features.

**2. Data Preprocessing:**

* Clean the data to handle missing values, outliers, and inconsistencies.
* Normalize or standardize the data to ensure that features are on a consistent scale.
* Create time-based features, such as hourly or daily averages, to capture temporal patterns.

**3. Feature Engineering:**

* Extract relevant features from the data, such as temperature, humidity, wind speed, wind direction, and time of day.
* Use geographical features like elevation and distance from pollution sources.
* Feature engineering can help the model capture the complex relationships between different variables.

**4. Model Selection:**

* Choose appropriate machine learning algorithms based on the nature of the problem:
  + For regression tasks (predicting pollutant concentrations), consider algorithms like Random Forest, Gradient Boosting, Support Vector Regression, or neural networks.
  + For classification tasks (air quality categories), algorithms like Random Forest, Decision Trees, or Support Vector Machines can be effective.

**5. Model Training and Validation:**

* Split the data into training, validation, and test sets to evaluate model performance.
* Employ cross-validation techniques to ensure the robustness of the model.
* Use relevant evaluation metrics, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or accuracy (for classification), to assess the model's accuracy.

**6. Hyperparameter Tuning:**

* Fine-tune model hyperparameters using techniques like grid search, random search, or Bayesian optimization to optimize model performance.

**7. Time Series Analysis:**

* If dealing with time-series data, consider using time series forecasting models like ARIMA, Prophet, or Long Short-Term Memory (LSTM) networks to capture temporal dependencies in air quality data.

**8. Ensemble Methods:**

* Implement ensemble methods like Random Forest or Gradient Boosting to improve model robustness and reduce overfitting.

**9. Incorporating External Data:**

* Integrate external data sources such as satellite imagery, weather forecasts, and traffic patterns to improve the model's predictive accuracy.
* These additional data sources can provide valuable context and information for air quality predictions.

**10. Real-time Monitoring and Feedback Loop:**

* Deploy the trained model for real-time air quality monitoring, incorporating data from live sensors and other sources.
* Implement a feedback loop to continuously update the model with new data and improve its accuracy over time.

**11. Model Interpretability:**

* Ensure that the machine learning model is interpretable to provide insights into how different factors affect air quality.
* Use techniques like SHAP (SHapley Additive exPlanations) or feature importance plots to explain model predictions.

**12. Public Engagement:**

* Develop user-friendly interfaces or mobile apps that provide real-time air quality information to the public, along with health recommendations based on air quality levels.
* Encourage public participation in air quality monitoring through citizen science initiatives and the use of mobile sensors.

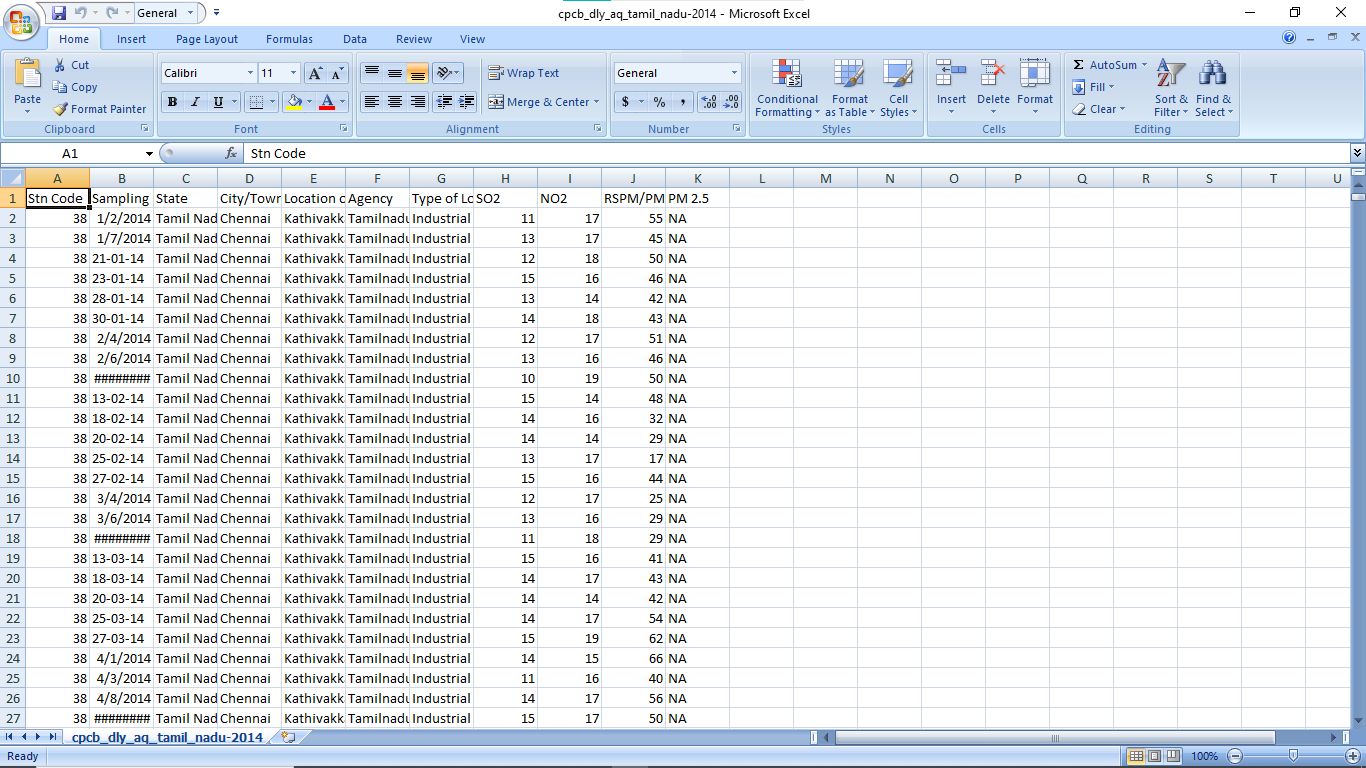
**13. Collaboration with Authorities:**

* Collaborate with environmental agencies, local authorities, and research institutions to ensure that the predictive model aligns with regulatory standards and policies.

**Data source**

A simple data source for an air quality assessment in Tennessee is the Tennessee Department of Environment and Conservation (TDEC) which collects and maintains air quality data through monitoring stations located throughout the state.

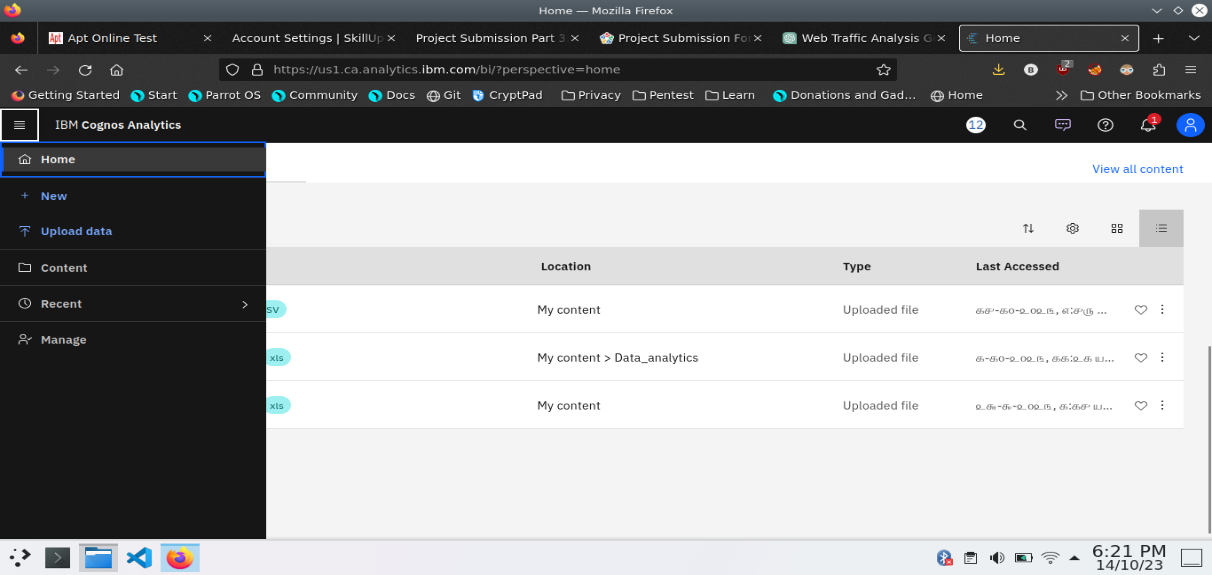
**Dataset link**: https://tn.data.gov.in/resource/location-wise-daily-ambient-air-quality-tamil-nadu-year-2014



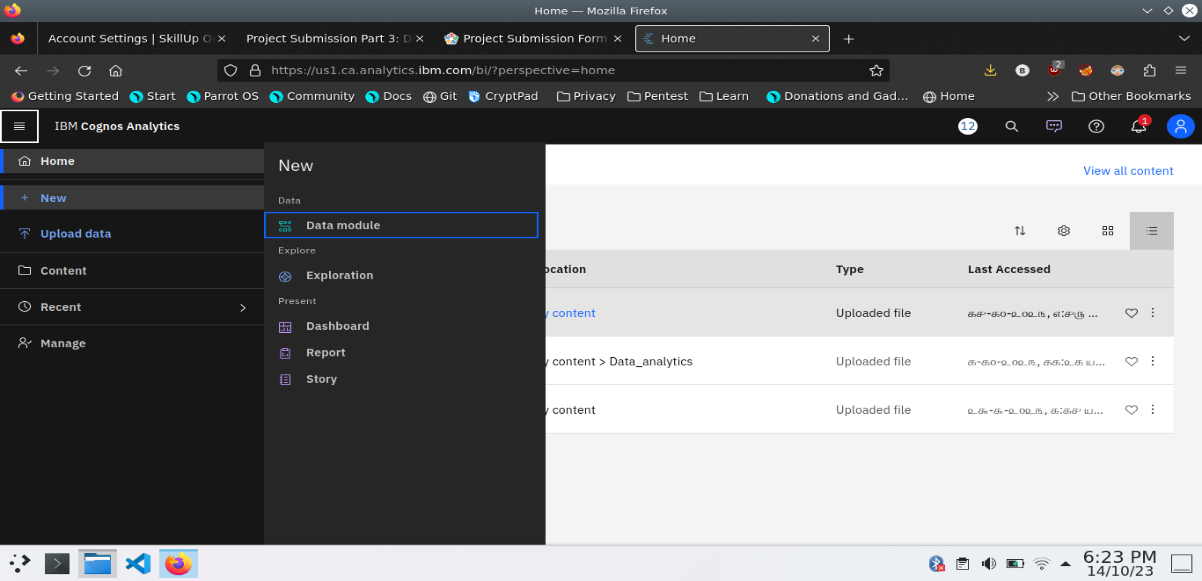
**Data Loading**

Steps Involved in data loading on IBM cognos.

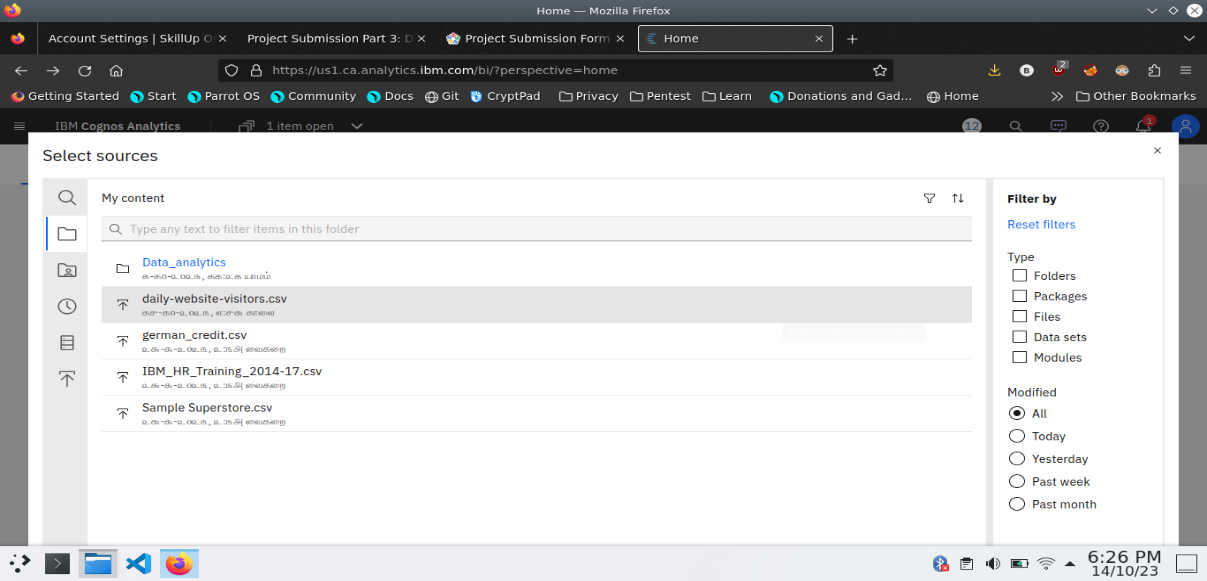
1. Login to your IBM cognos
2. Click more menu from the left side
3. Select new tab



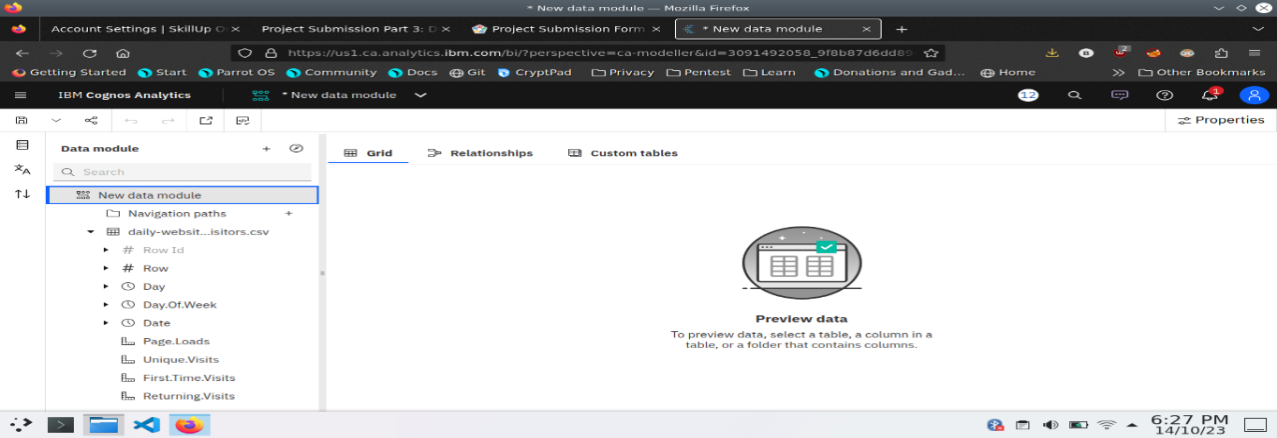
4. Click Data module tab

****

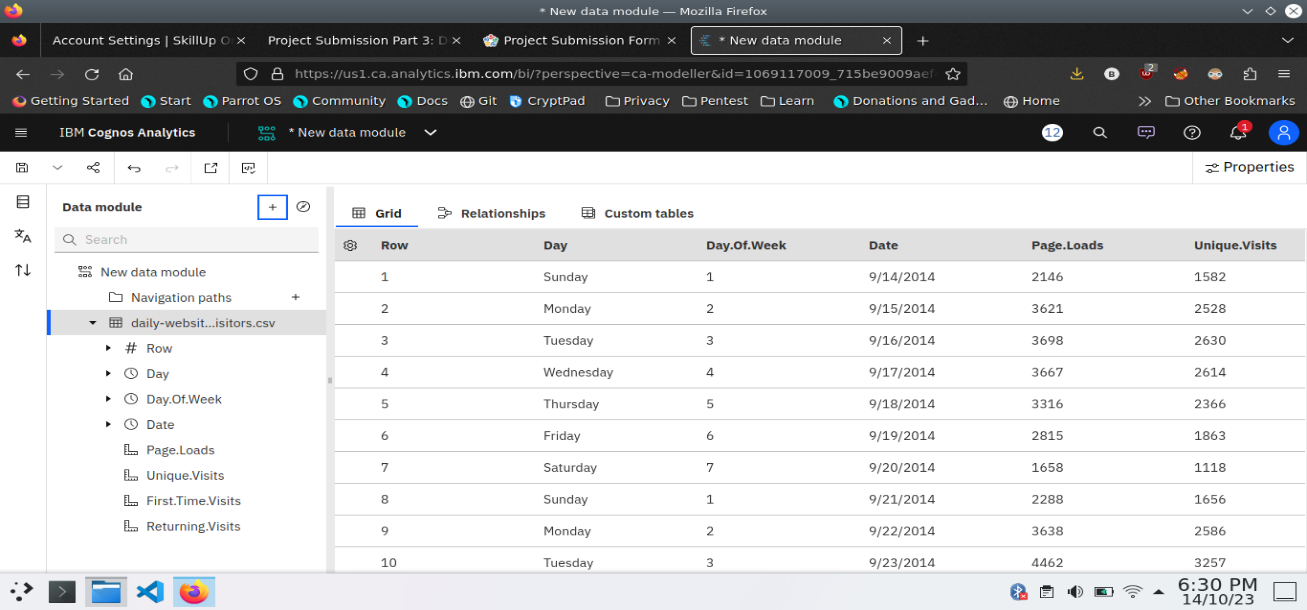
5.Upload the dataset for your project and select the Corresponding file



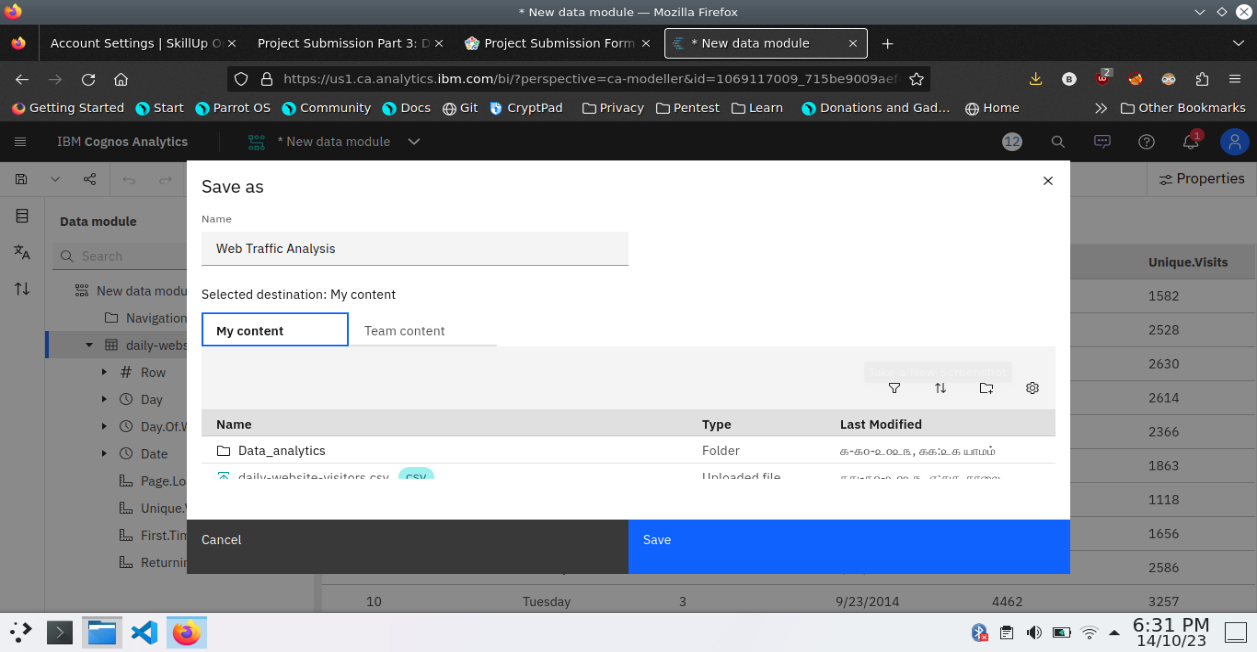
6. preview the data



7.Explore the data



8. save the data module



**Data Preprocessing and Cleaning**

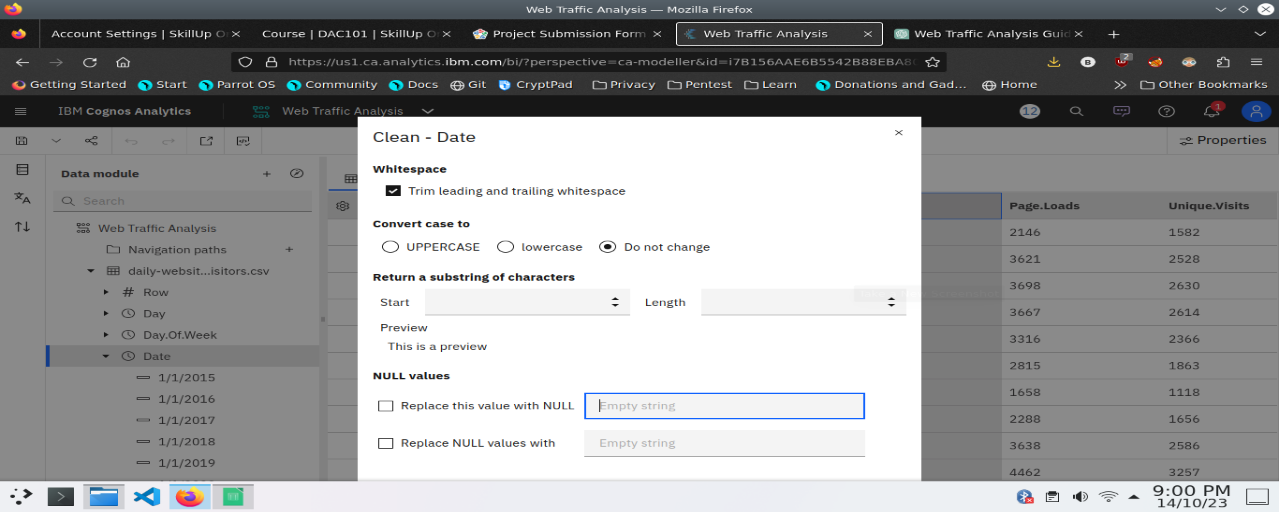
In this phase the following steps will taken

* Handling missing data
* Data Transformation
* Data Type Conversion
* Removing Duplicates
* Dealing Outliers

Once you saved the data module. Click the corresponding dataset on IBM cognos and Preview the mosule

Right Click the row where you want to clean the data

It provides the UI to Clean the data and makes the task easy one, Now Updating and Replacing the Null values are simple



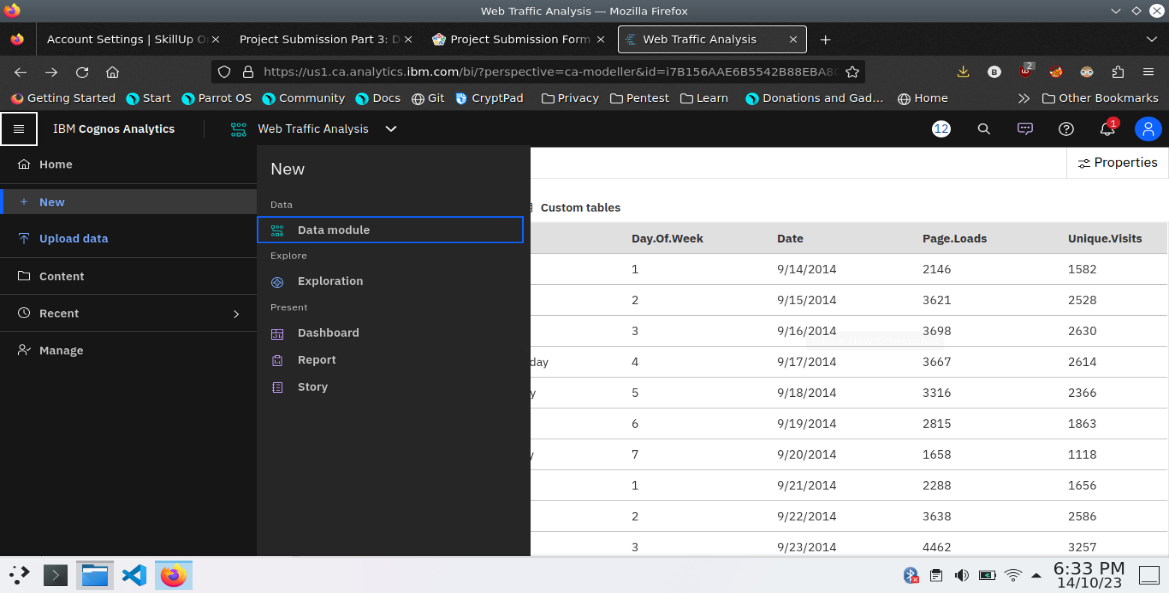
datamodule will be updated by doing the above process

after the completion ofprocess start creating the dashboard for Visualization

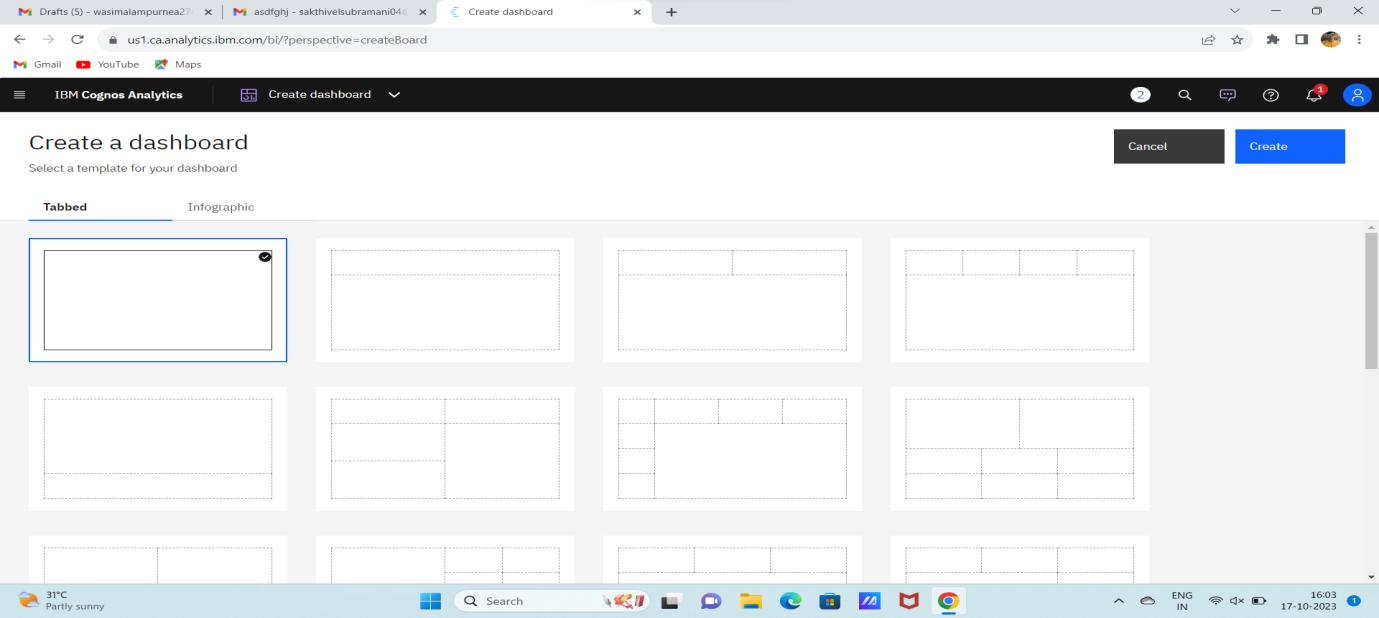
**Dashboard Creation**

Dashboard creation are helpful to visualizing the data

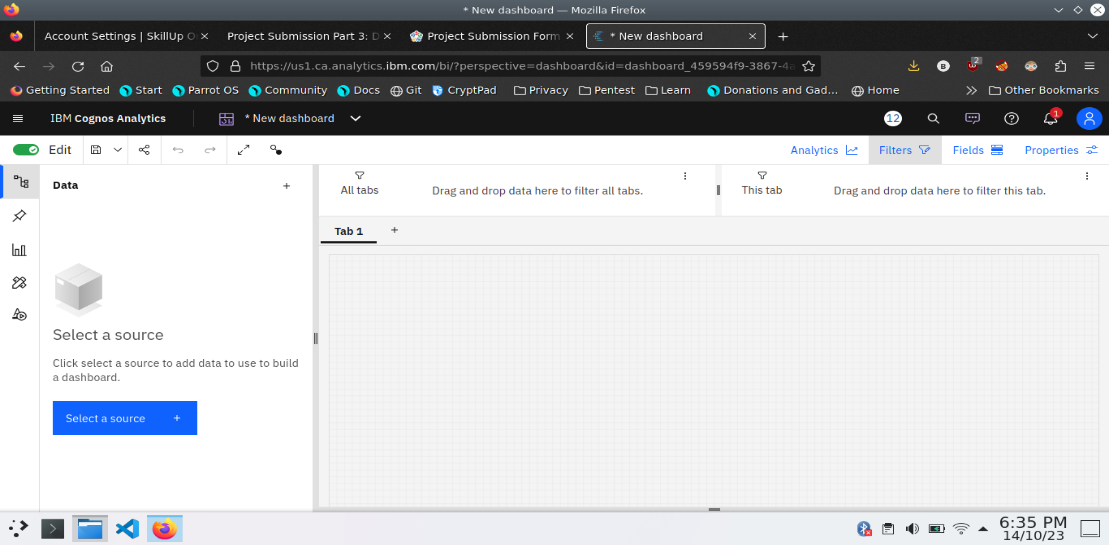
1. Goto Home menu
2. Select the new tab
3. Click dashboard



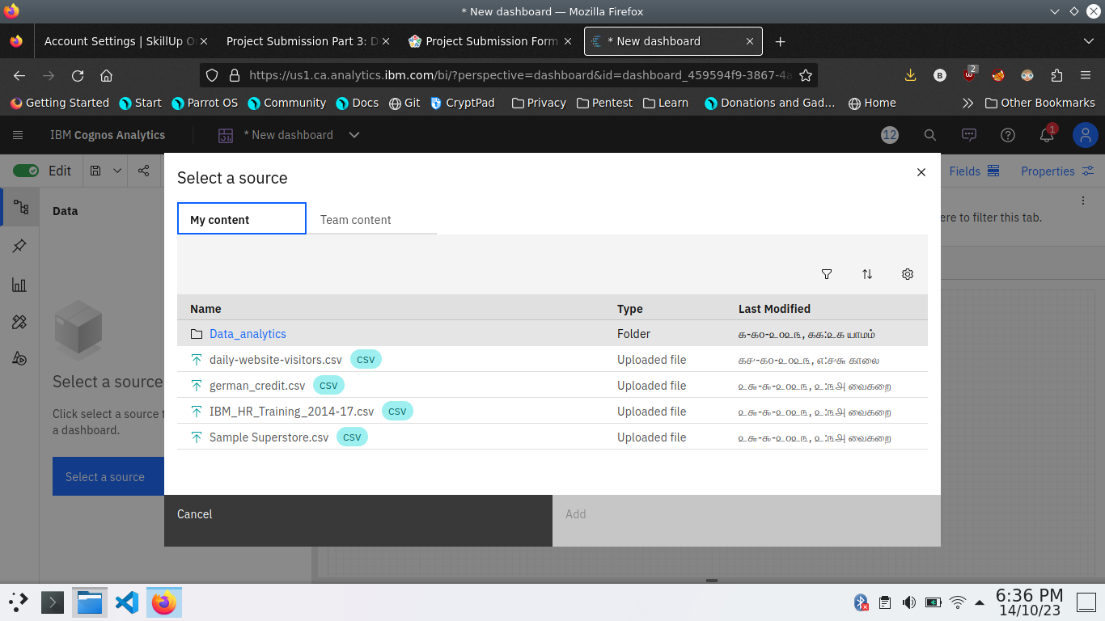
4. Choose the template for your project and click



5.Now Dashboard is created



6. Select the data source



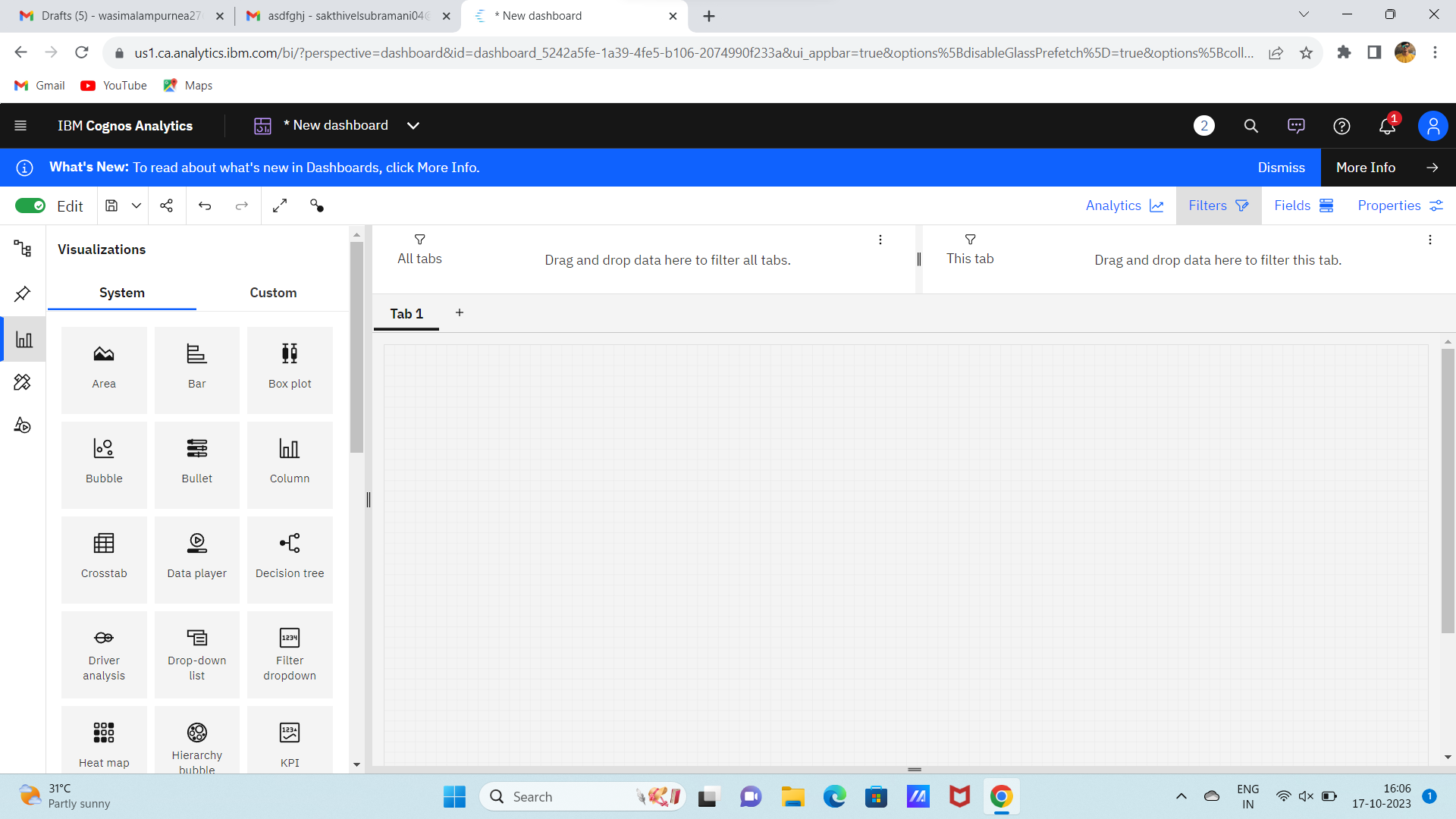
**Visualization**

After creating the dashboard, the next step is to visualize the data

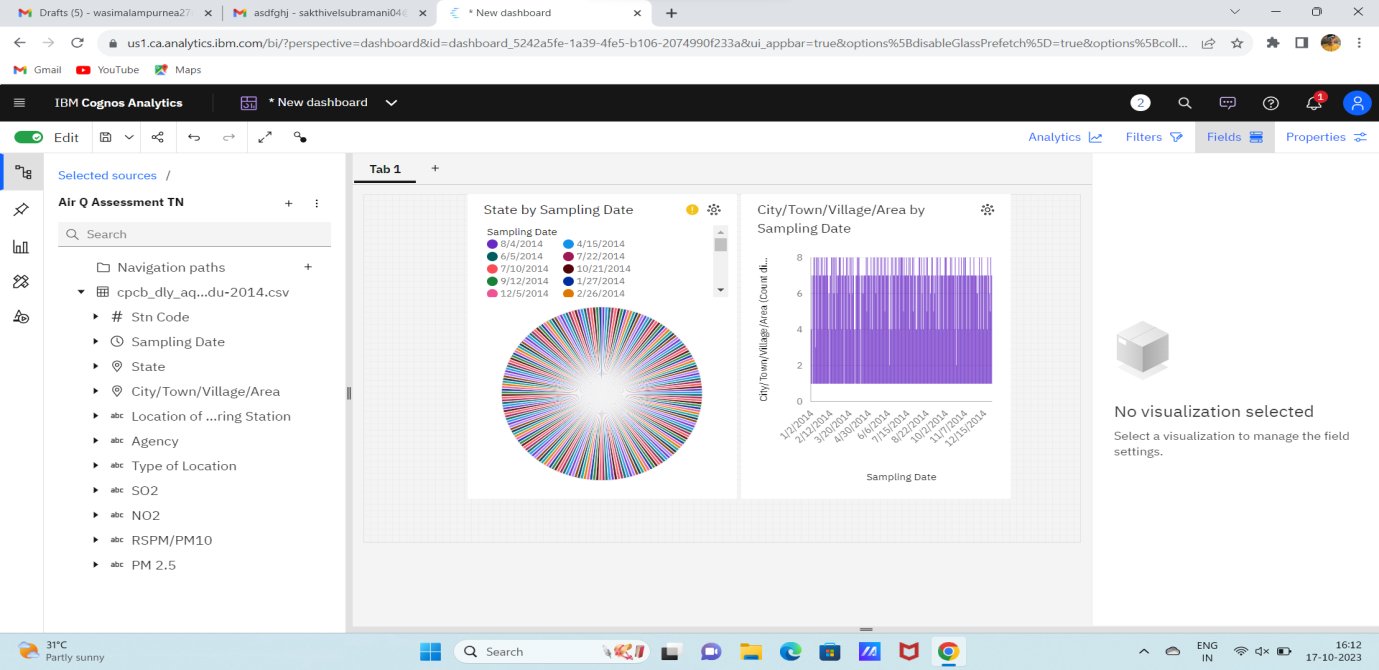
In IBM Cognos

1. Goes to the Corresponding Dashboard

2. select the visualizations tab in the left side of title bar



3.Choose the system as you want and put the data source for the required columns



In the above screen shot displays the Line graph and model compares the “sampling date” and “city/town/village/area” from the time period of 2020

X-axis =sampling Dates

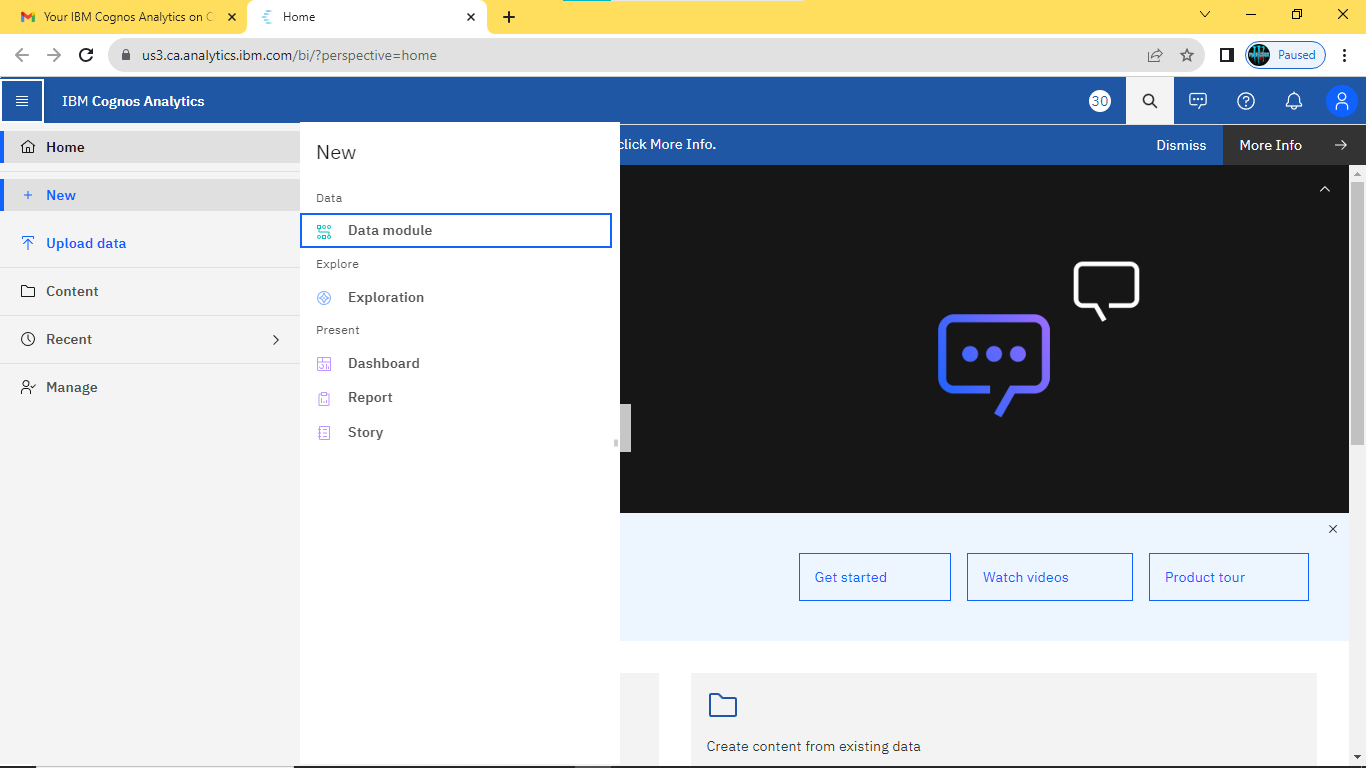
Y-axis = city/town/village/area

After performing these activities a comprehensive document will be created to demonstrate the ability to Communicate and share finding.

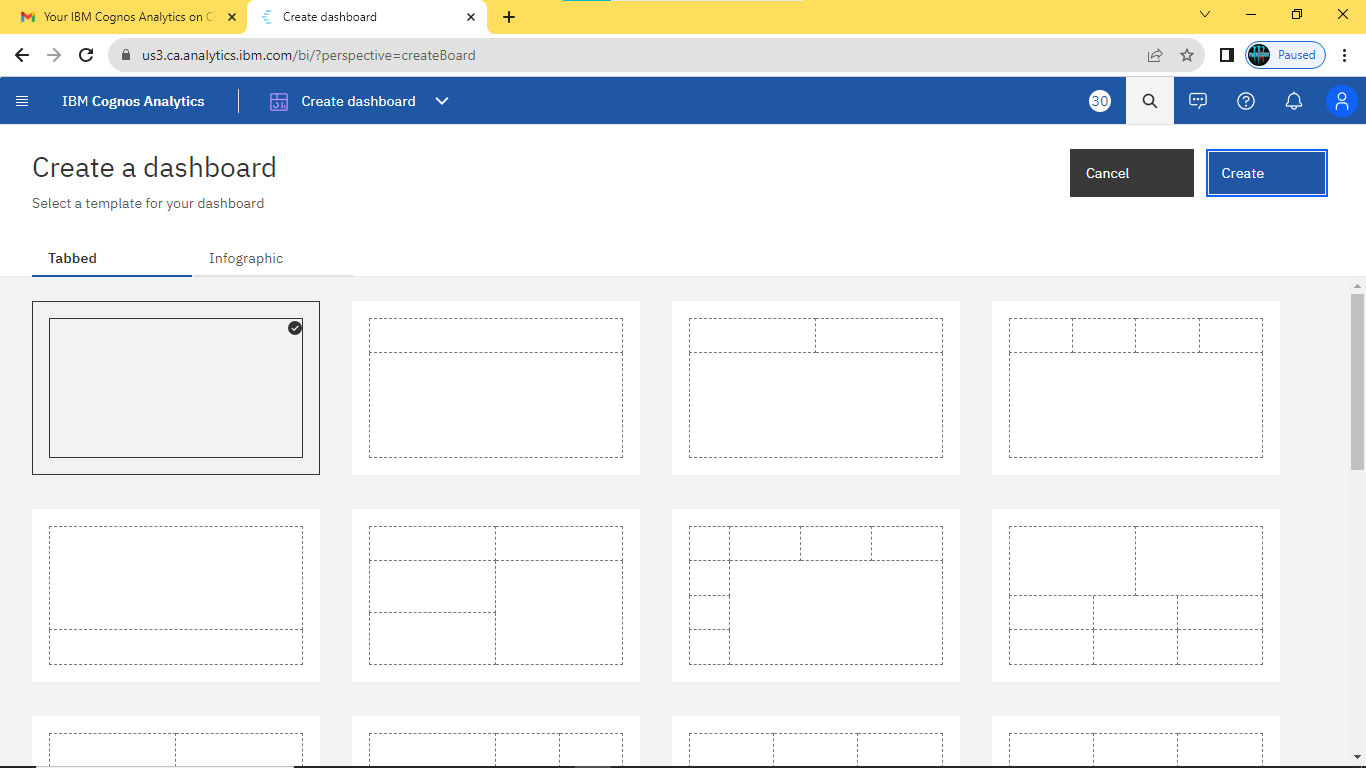
**Visualization in IBM Cognos**

**Step 1:**

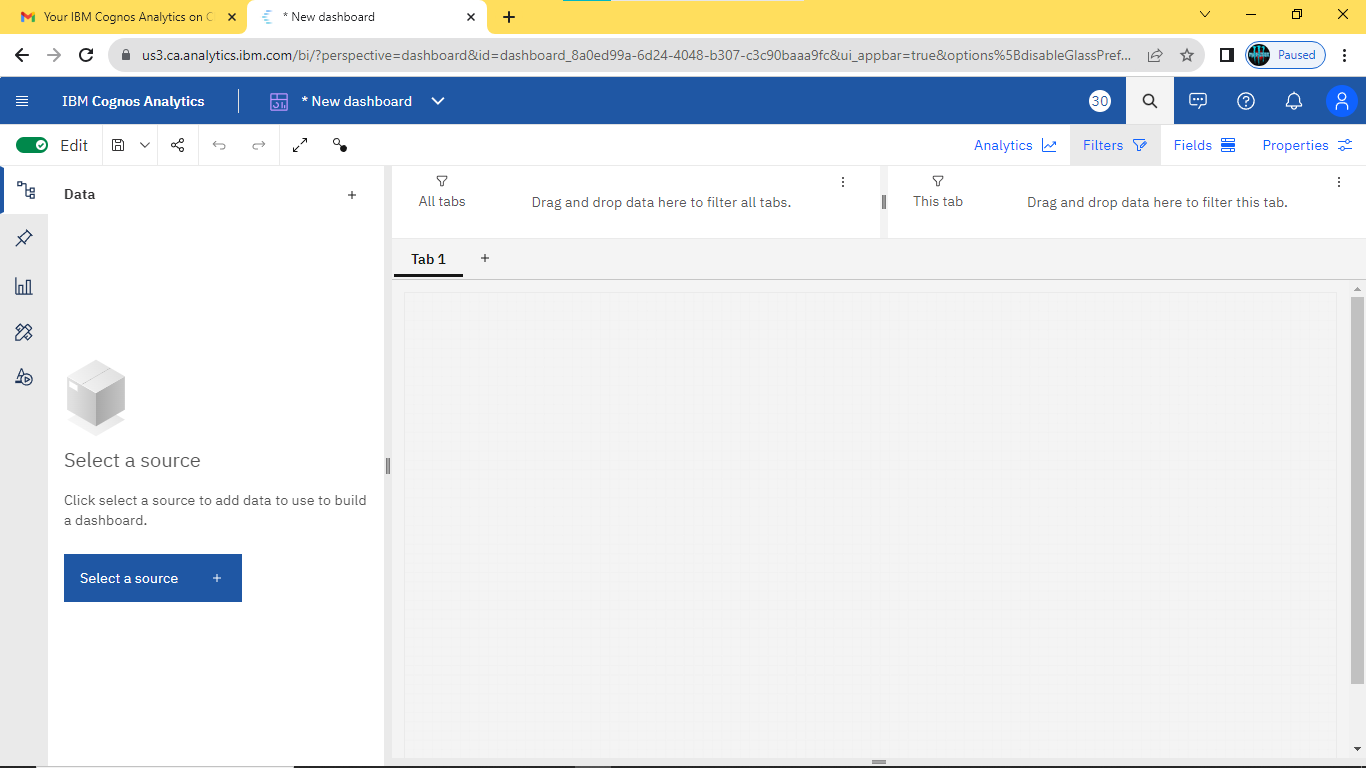
1. Login to your IBM cognos
2. Click more menu from the left side
3. Select new tab
4. Click Dashboard tab



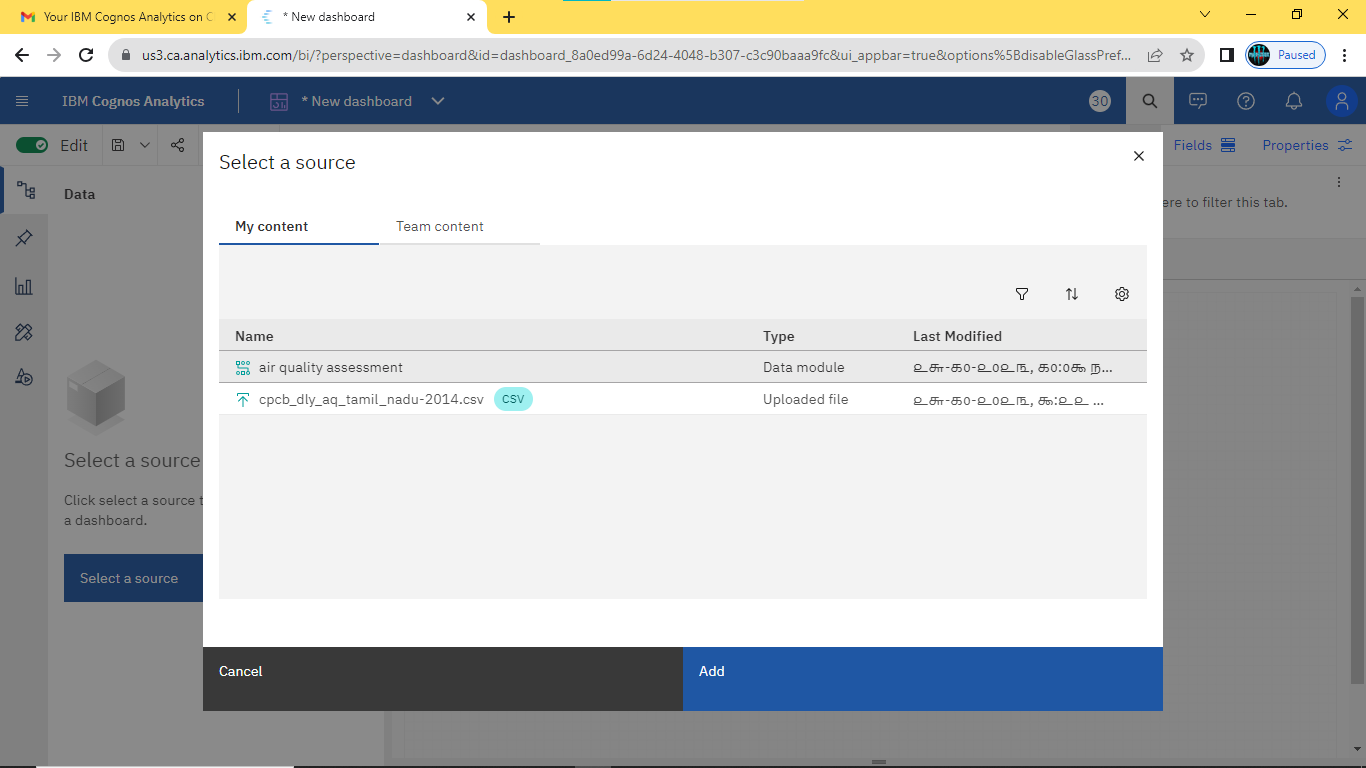
1. Select Template for your dashboard



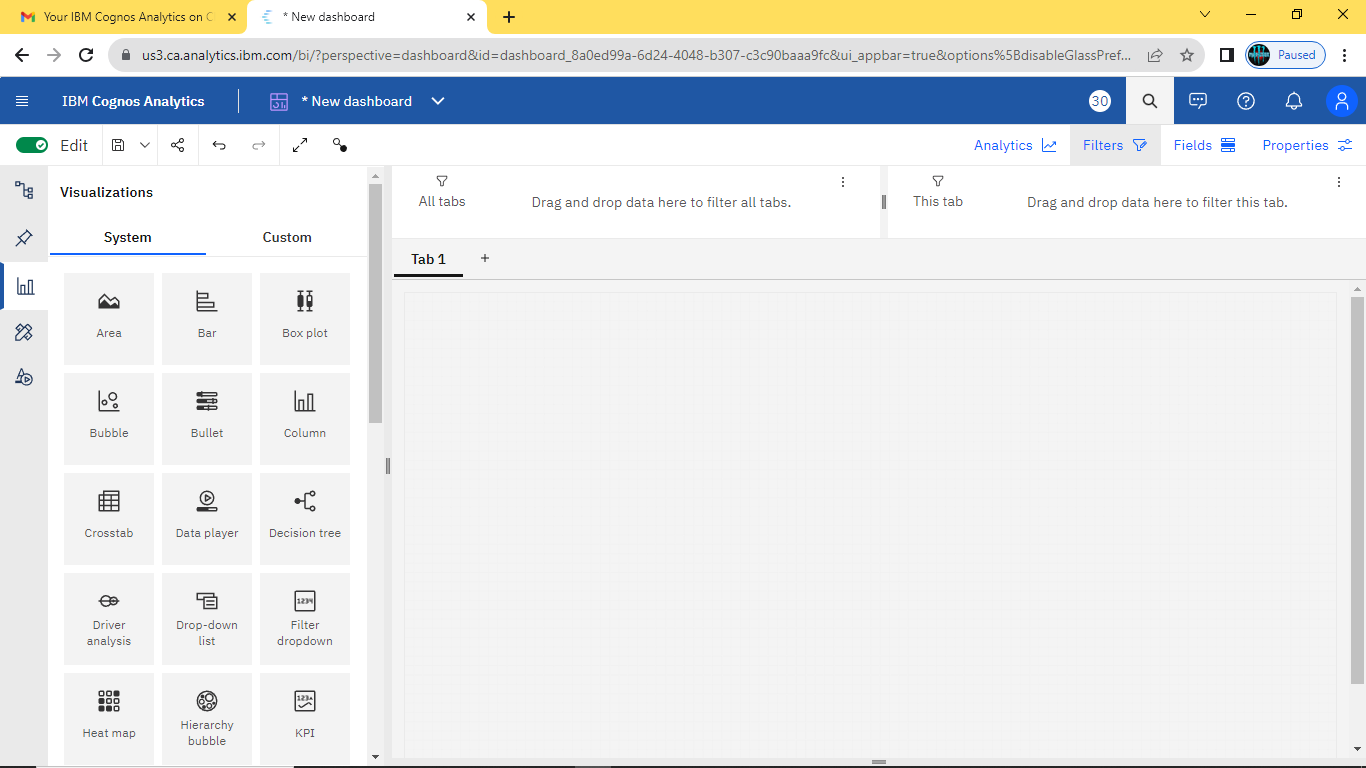
1. Now the dashboard is created and select your data-source.



1. Select your Corresponding dataset

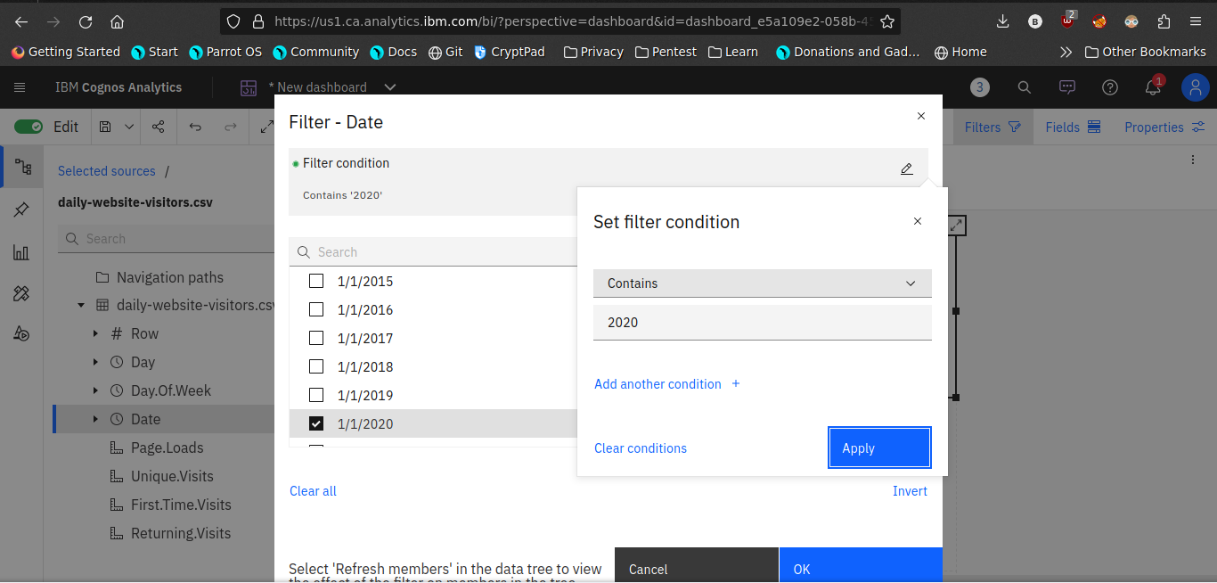


1. Select favourable visualization system



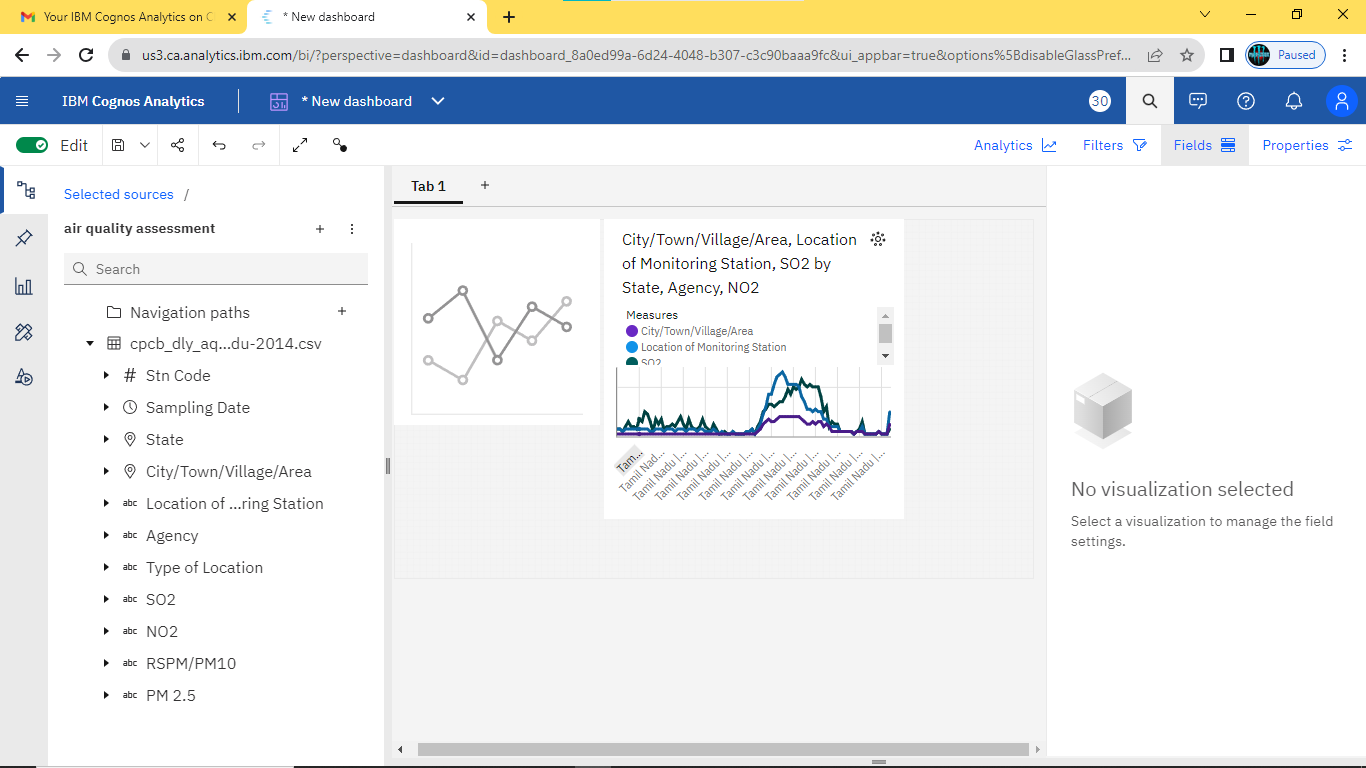
1. If needed Filter the data

Here the datasets are filtered by the date Which contains 2020



1. Line plot

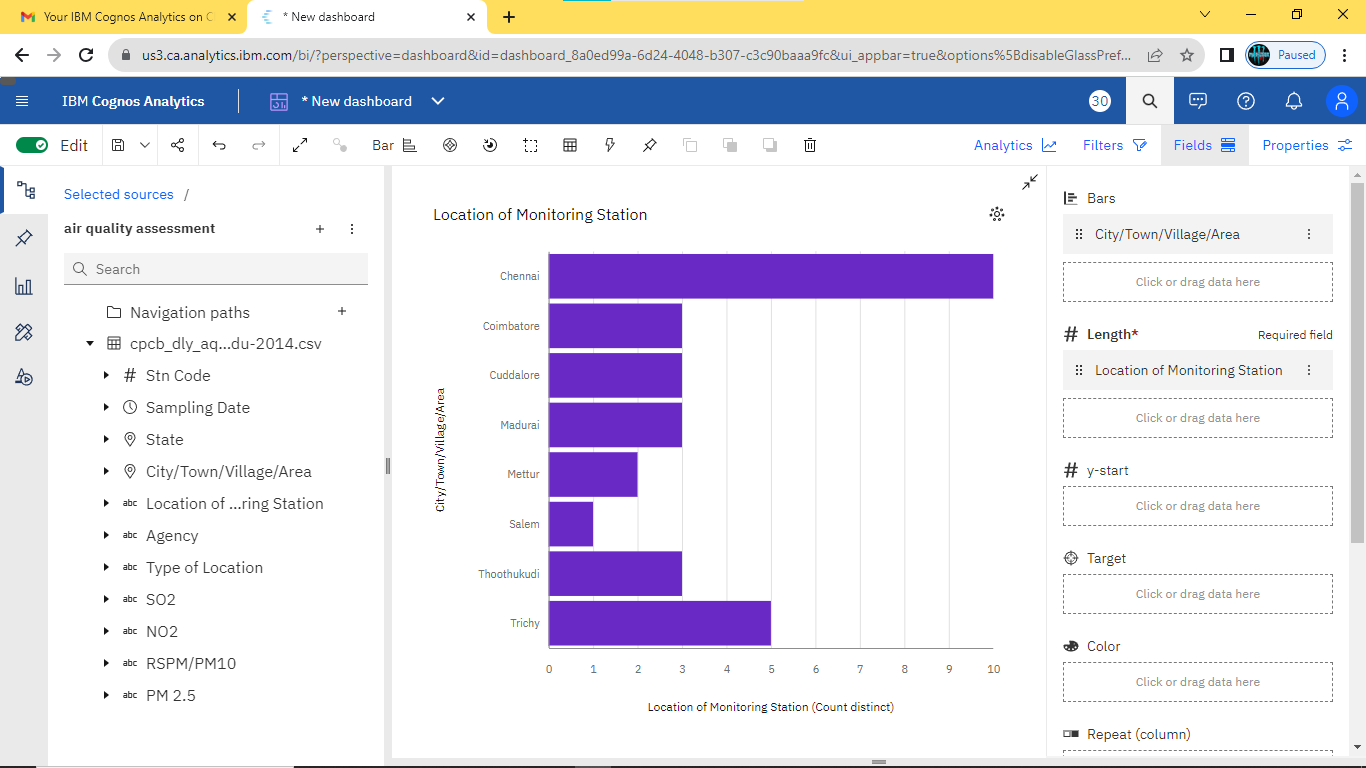
In this line plot X\_axis are dates and Y\_axis are First time visits and Unique visits



1. **Barchart**

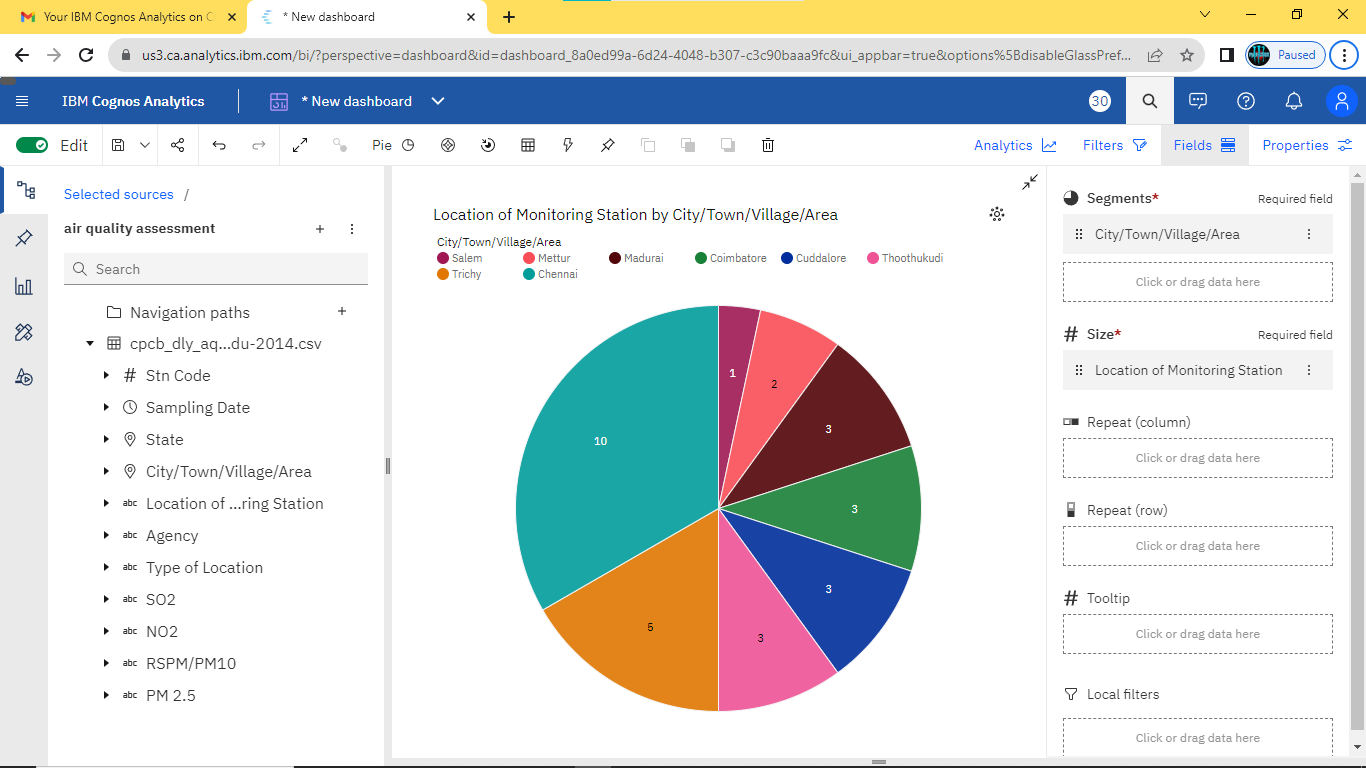
In this Bar chart the bars represent the ‘days in week’ and length defines ‘Page.Loads’

It helpful to visualize the maximum pageloads occurs on a day



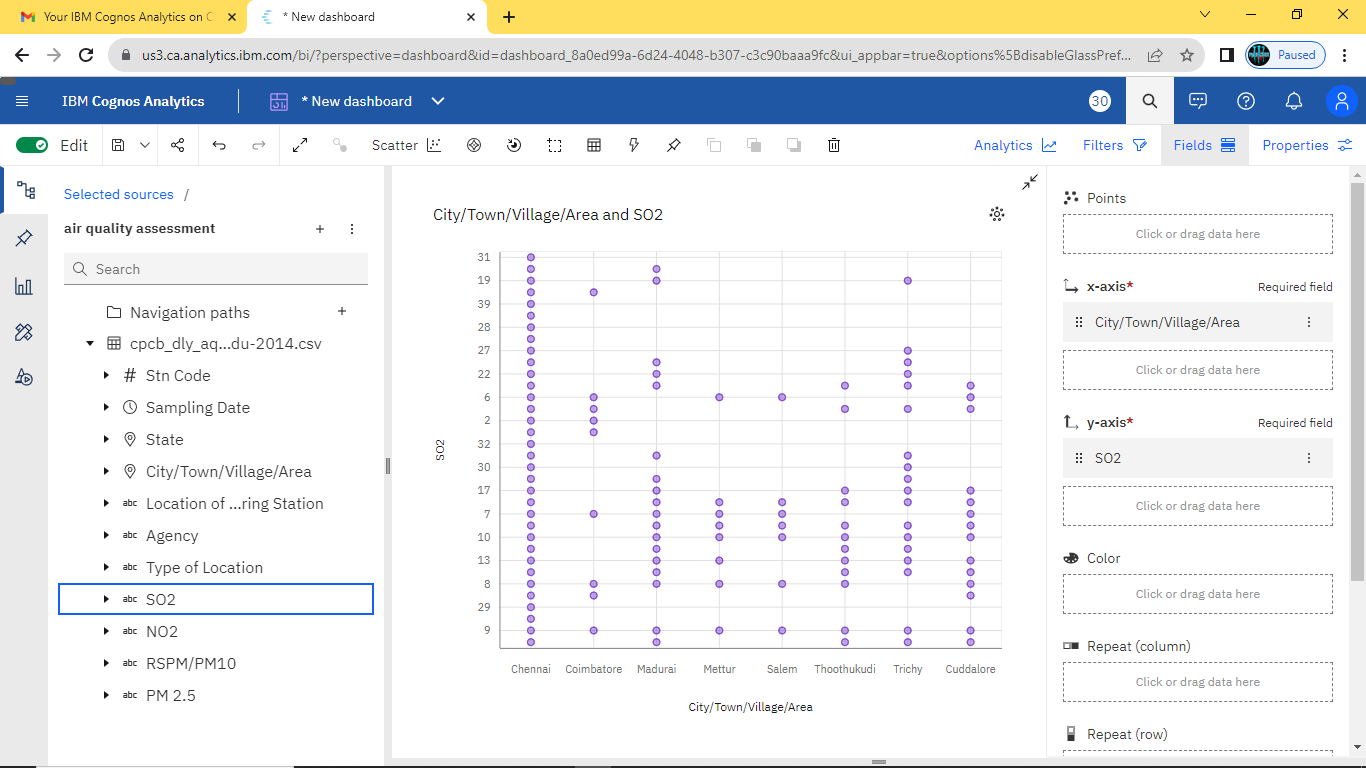
1. **Piechart**

This is same as a bar chart. it helpful to analyze the Returing visits occurs on a particular day



1. **Scatter plot**

It is used to display the relationship between two variables and observe the nature of the relationship. The relationships observed can either be positive or negative, non-linear or linear



Now the visualization phase where over. lets start analyze the dataset using Python libraries use machine learning models for predictive analysis.

**Data Analysis using python**

In this steps are used to analyse the given dataset using python libraries

**Steps:**

1. **Import Necessarypackages**

Pandas

seaborn

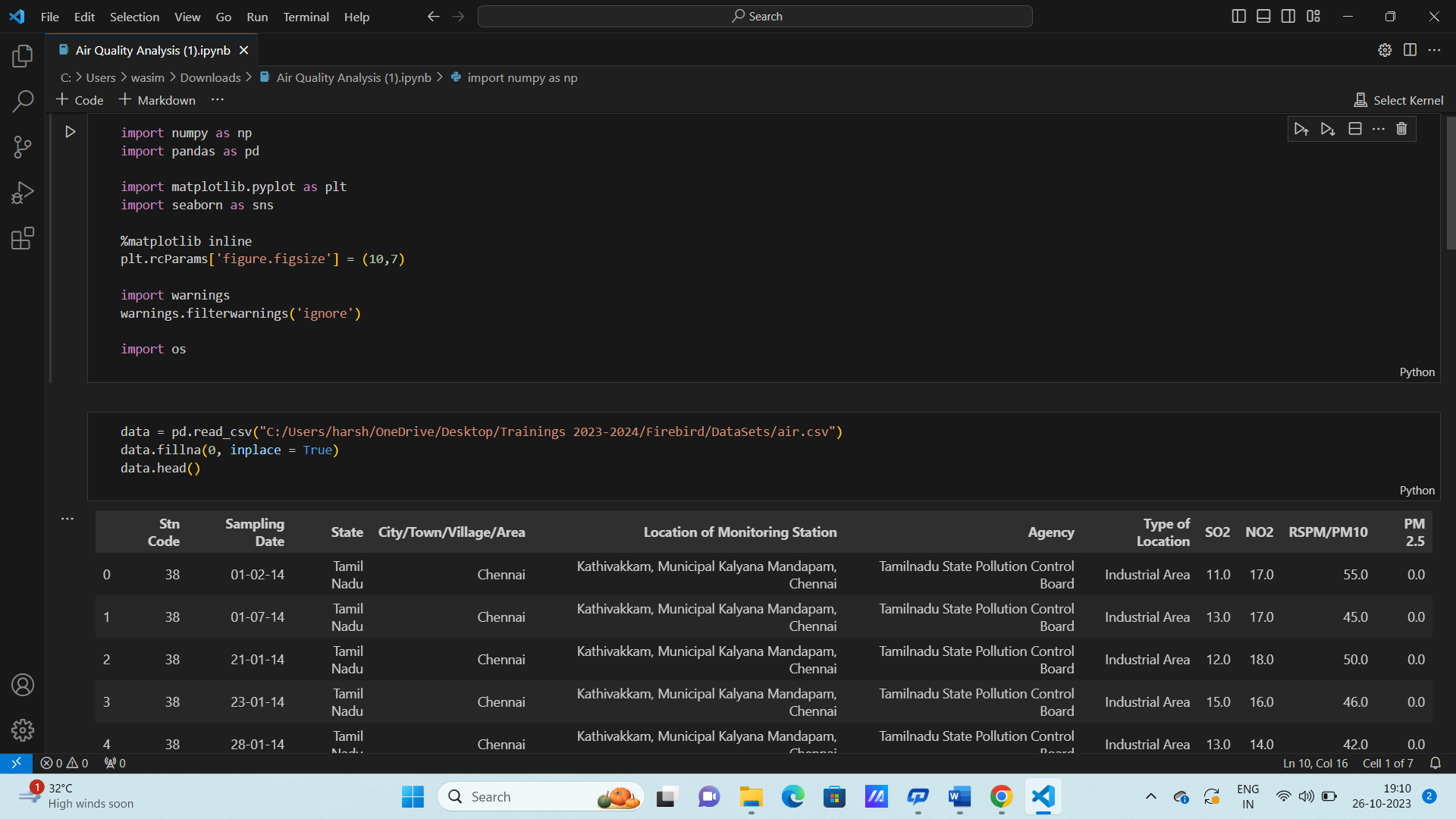
Machine learning models

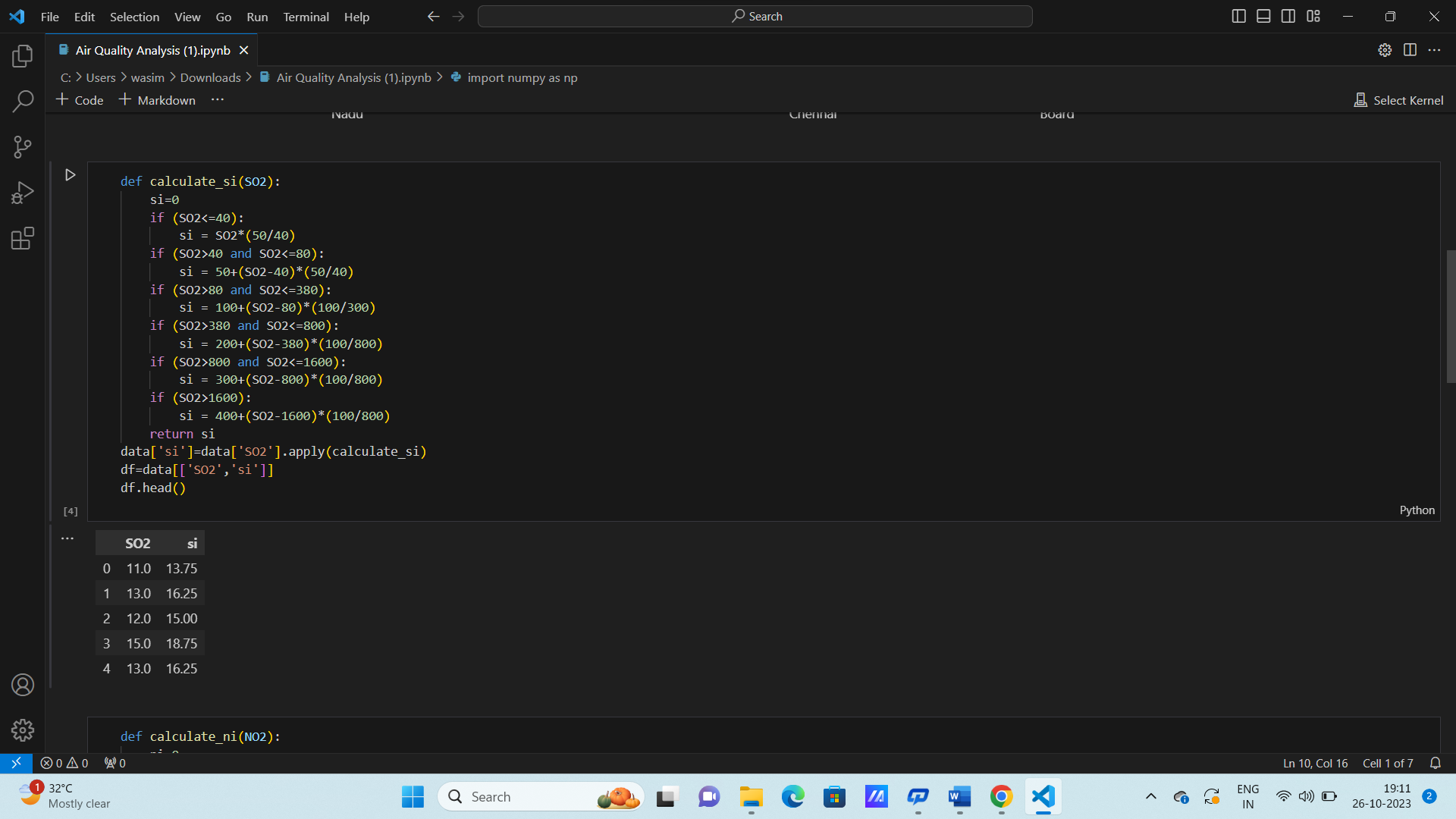
Linear regression

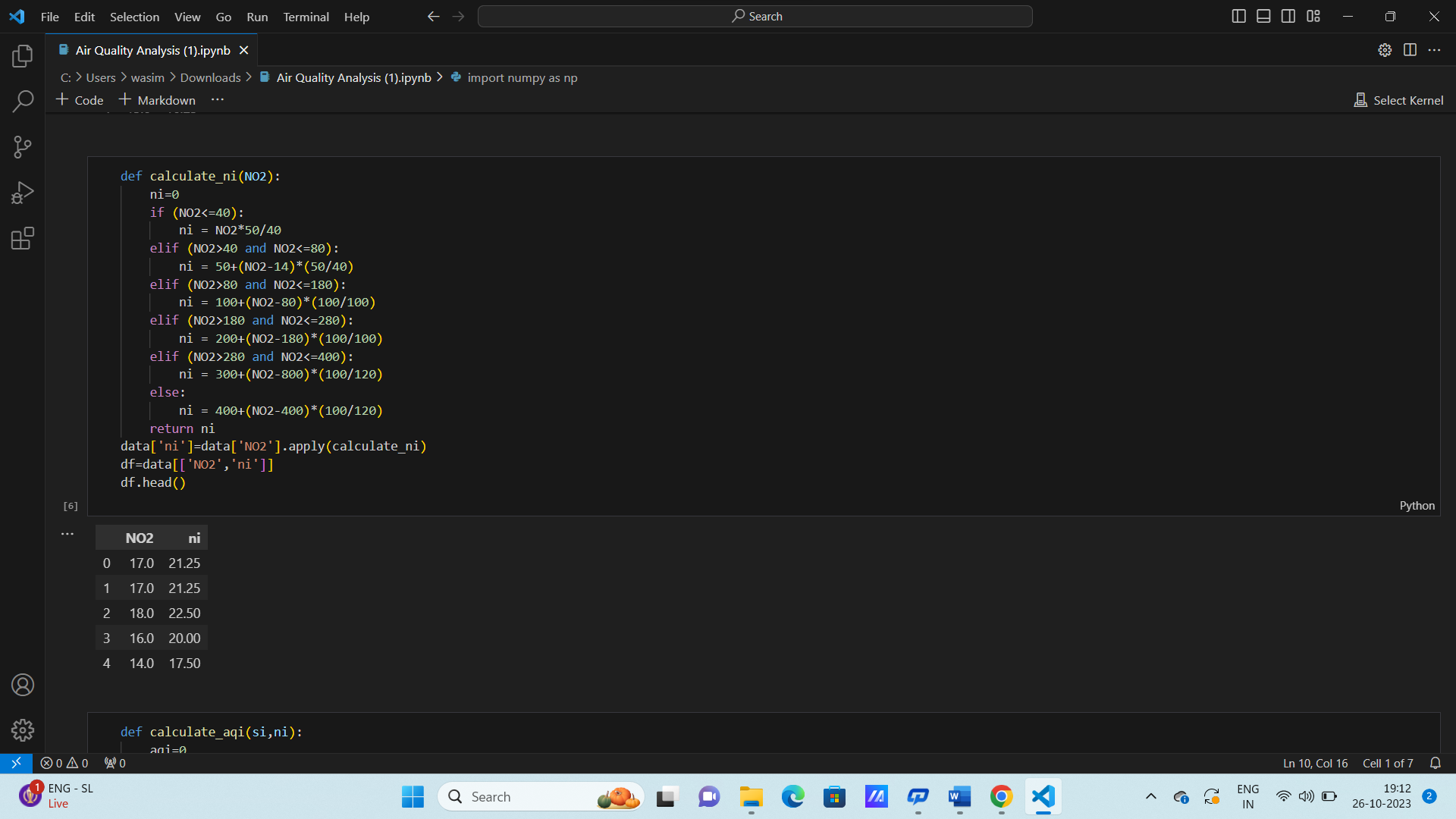
1. **Make a training and test data**

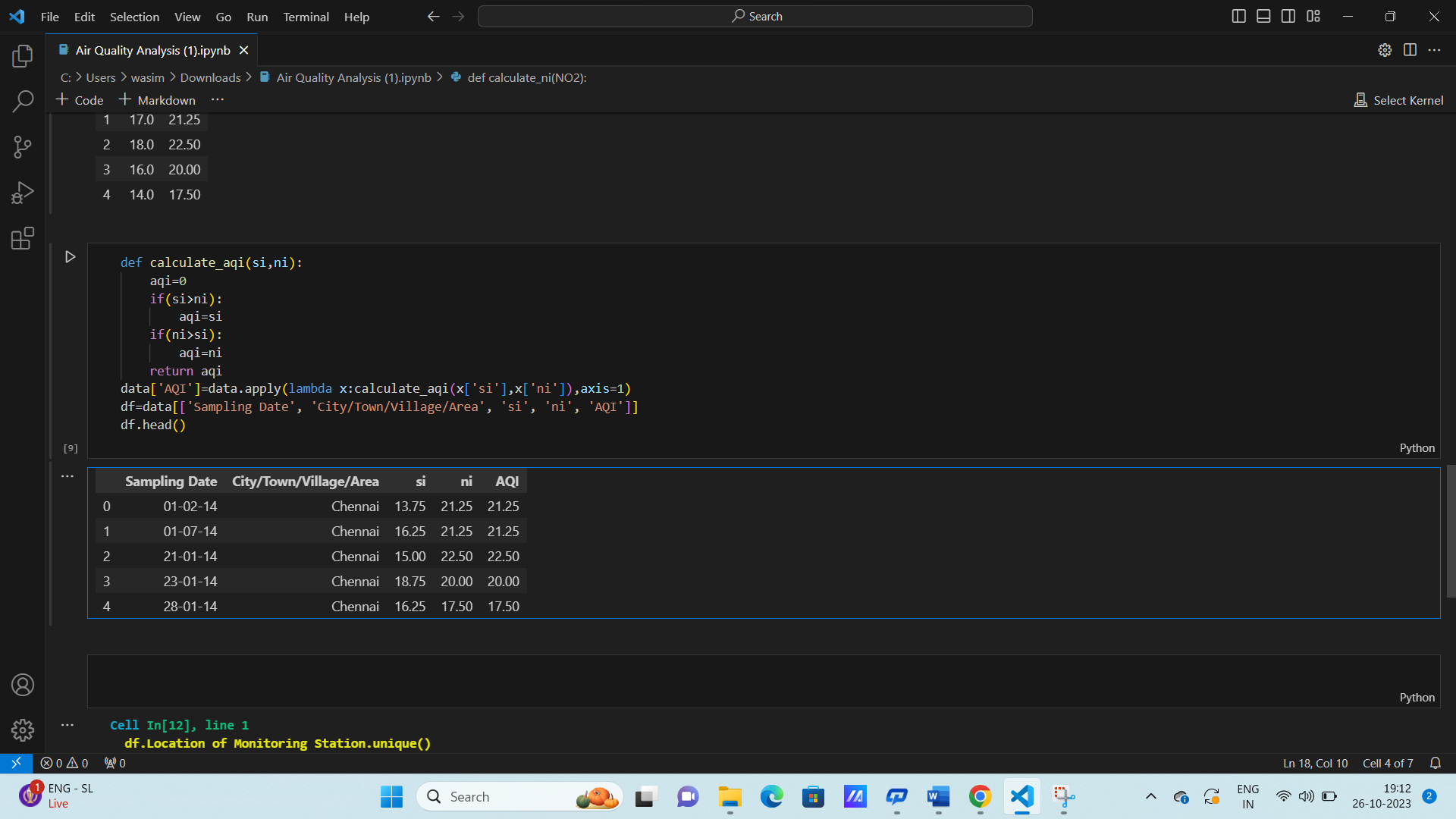
Use the train test split model

Compare the testing and training data set by visualization library

****

****

****

****

**Conclusion:**

In closing, our air quality assessment of Tamil Nadu highlights the critical need for action to combat air pollution in the state. Through our analysis, we have uncovered trends, identified pollution hotspots, and developed predictive models, providing valuable insights for policymakers and the public. It is now imperative to translate these findings into concrete measures that improve air quality, protect public health, and preserve the environment for future generations.