**IBM- Naan mudhalvan Data Analytics with Congnos**

**Phase-5**

**Documentation & Submission**

**Student Name.**  : HARISH A

**Register Number**. : 620821104043

**Branch**. : B.E CSE

**Year**. : 3rd year

**Topic**. : Data Analytics with Cognous

**Title**. : Air quality analysis in Tamilnadu

**College**. : Gnanamani College of Technology

**Introduction:**

Air quality analysis in Tamil Nadu is a multidisciplinary field that assesses the composition of the air and its impact on human health and the environment. Tamil Nadu ,a state in southern India, experiences various sources of air pollution, including industrial emissions, vehicular traffic, and natural factors. Understanding and monitoring air quality is essential to make informed decisions for public health, urban planning, and policy development.

**Abstract:**

Air pollution is a pressing concern in Tamil Nadu, a state in southern India, due to rapid urbanization, industrial growth, and their associated environmental impacts. This abstract outlines a comprehensive project proposal aimed at assessing and improving air quality in Tamil Nadu. The project seeks to establish an extensive air quality monitoring network, analyze data to identify pollution sources, raise public awareness, and advocate for evidence-based policies to mitigate air pollution.

**Understanding the problem:**

Air quality analysis is a crucial field that involves assessing the composition of the air we breathe to determine its safety and potential health risks. Understanding the problems and challenges in air quality analysis is essential for improving air quality management and safeguarding public health.

**Design and thinking:**

Design thinking is a creative problem-solving approach that can be applied to the field of air quality analysis to develop innovative and effective solutions.

**Project Definition:**

**Project Objectives:**

To assess and monitor air quality in [specific geographic area] to ensure public health and environmental safety.

To identify and quantify major air pollutants, including [list of pollutants], and their sources.

To develop actionable recommendations for air quality improvement based on data analysis and findings.

**2. Project Scope:**

Geographic Area: [Specify the geographic area or region covered by the project, e.g., a city, state, or industrial zone.]

Timeframe: [Specify the project’s start and end dates or duration.]

Data Sources: [List the sources of air quality data, such as monitoring stations, satellite data, or sensor networks.]

Key Parameters: [Specify the specific air quality parameters to be analyzed, e.g., PM2.5, PM10, NO2, SO2, O3, VOCs, etc.]

Stakeholders: [Identify the key stakeholders involved in or affected by the project, including government agencies, environmental organizations, researchers, and the public.]

**3. Project Goals:**

Improve public health by providing accurate and timely information on air quality.

Enhance environmental protection by identifying pollution sources and recommending mitigation strategies.

Ensure compliance with air quality standards and regulations.

Raise public awareness and promote informed decision-making regarding air quality.

4. Project Components:

a. Data Collection and Monitoring:

- Set up or utilize existing air quality monitoring stations and data sources.

- Ensure data accuracy, reliability, and real-time or near-real-time availability

**Project Deliverables**:

Regular air quality reports and updates for the geographic area.

Health impact assessments and risk communication materials.

Recommendations for air quality improvement.

Educational resources for the community.

## Common Machine Learning Models in Web Traffic Analysis

### 1. Regression Models

* Linear Regression
* Polynomial Regression
* Logistic Regression
* Purpose: Predict web traffic volume, conversion rates, and numerical metrics.

### 2. Classification Models

* Decision Trees
* Random Forests
* Support Vector Machines (SVMs)
* Purpose: Identify malicious traffic, categorize user behavior, and classify user interactions.

### 3. Clustering Models

* K-Means Clustering
* Hierarchical Clustering
* Purpose: Group users with similar behavior patterns for segmentation and personalization.

### 5. Ensemble Models

* Bagging (e.g., Random Forests)
* Boosting (e.g., AdaBoost, XGBoost)
* Purpose: Combine multiple models for more accurate predictions and robustness.

## Machine Learning Models with Cognos

Steps Involved

1. Data collection
2. Data preprocessing
3. Feature Engineering
4. Model Training
5. Deployment

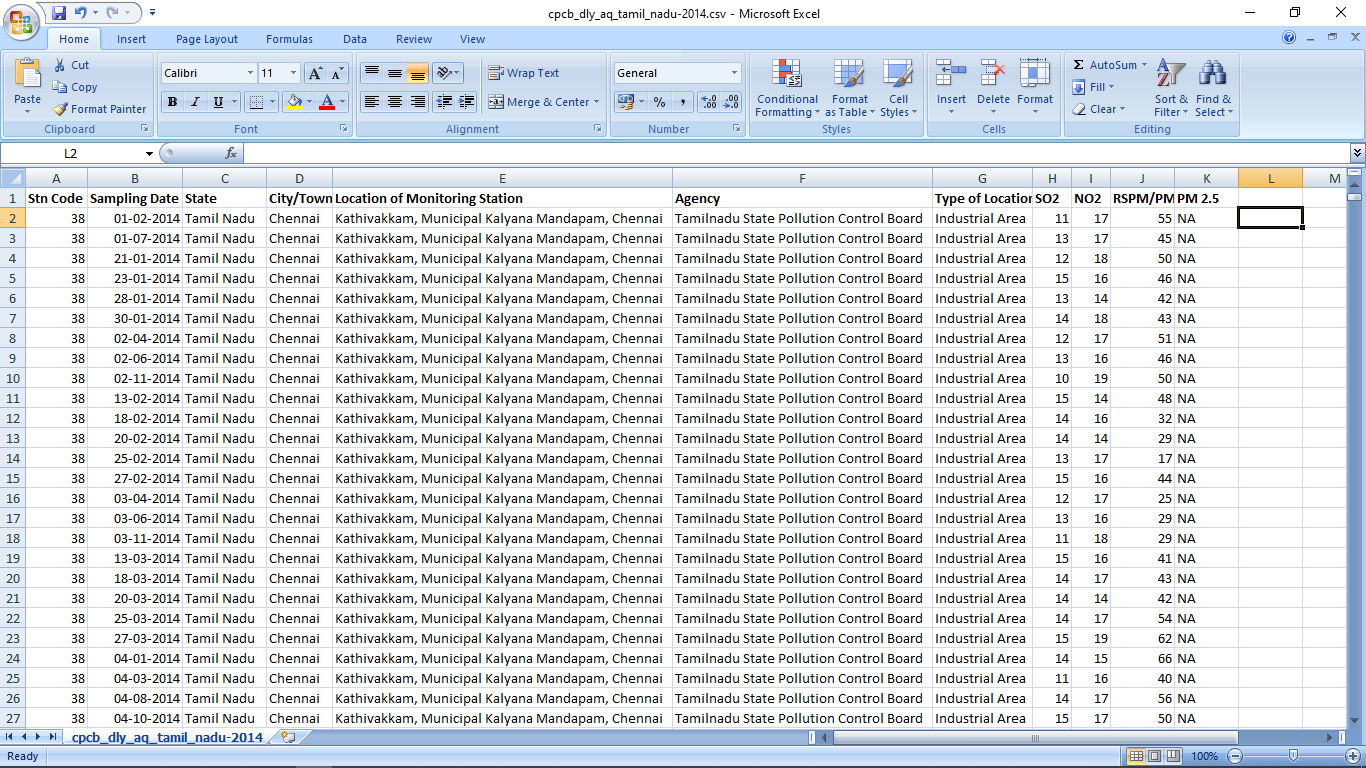
**Objectives**

In this phase defines start to building the Project by loading and preprocessing the dataset and perform different analysis and visualization using IBM Cognos.

**Data source**

Dataset is collected from the kaggle.com named “daily-website-visitors.csv” which has a data about the Days, Day of week, Date, page Loads, Unique visits, First-time visits, Returning Visits.

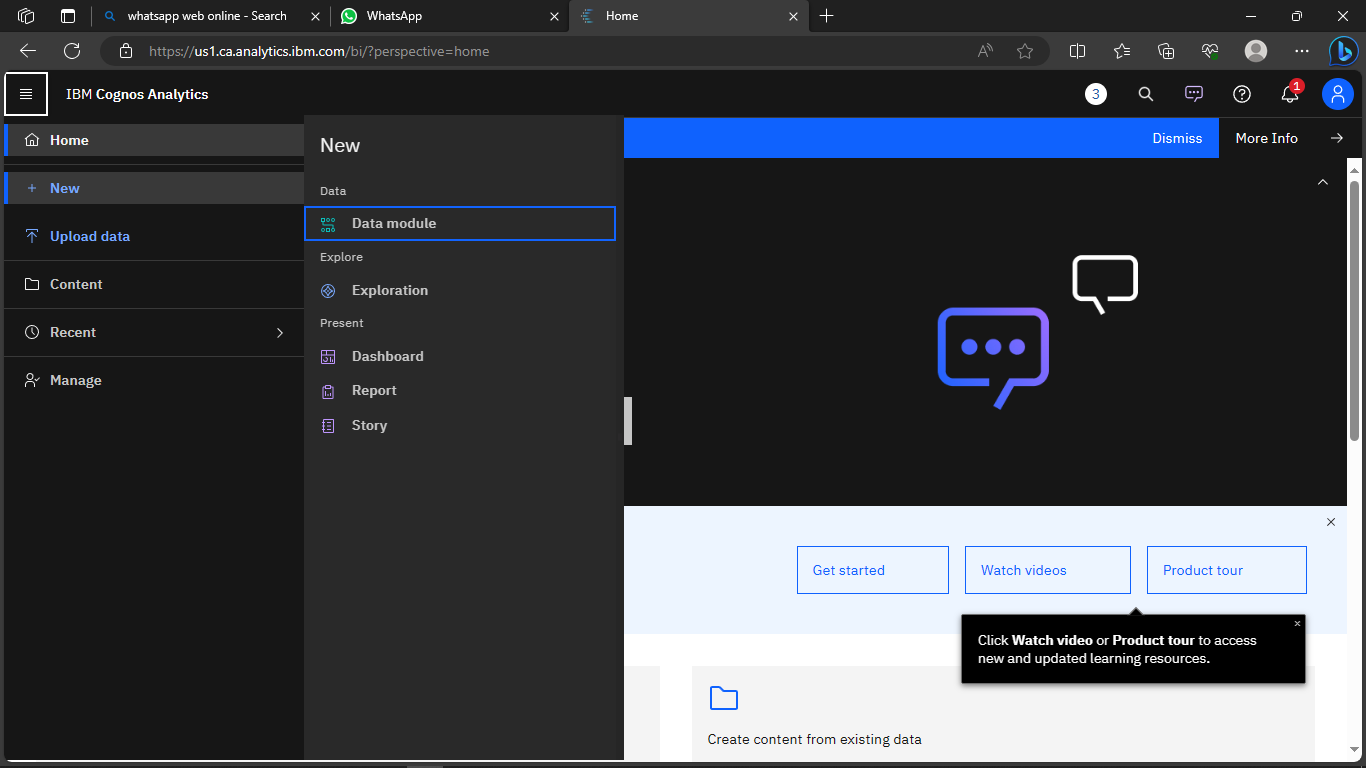
Dataset link: <https://www.kaggle.com/datasets/bobnau/daily-website-visitors>



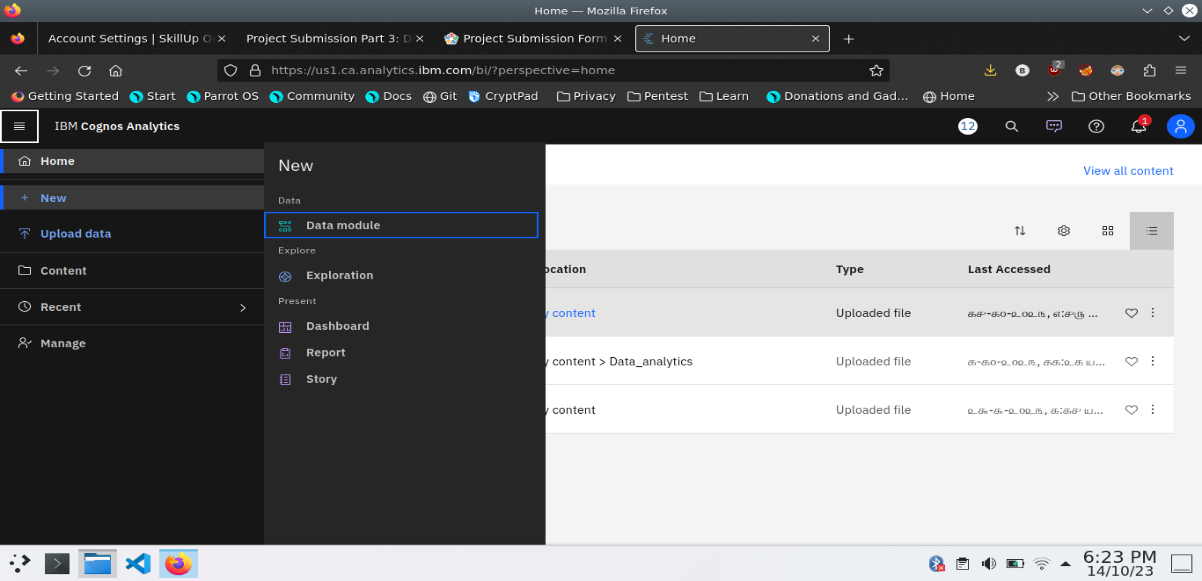
**Data Loading**

Steps Involved in data loading on IBM cognos.

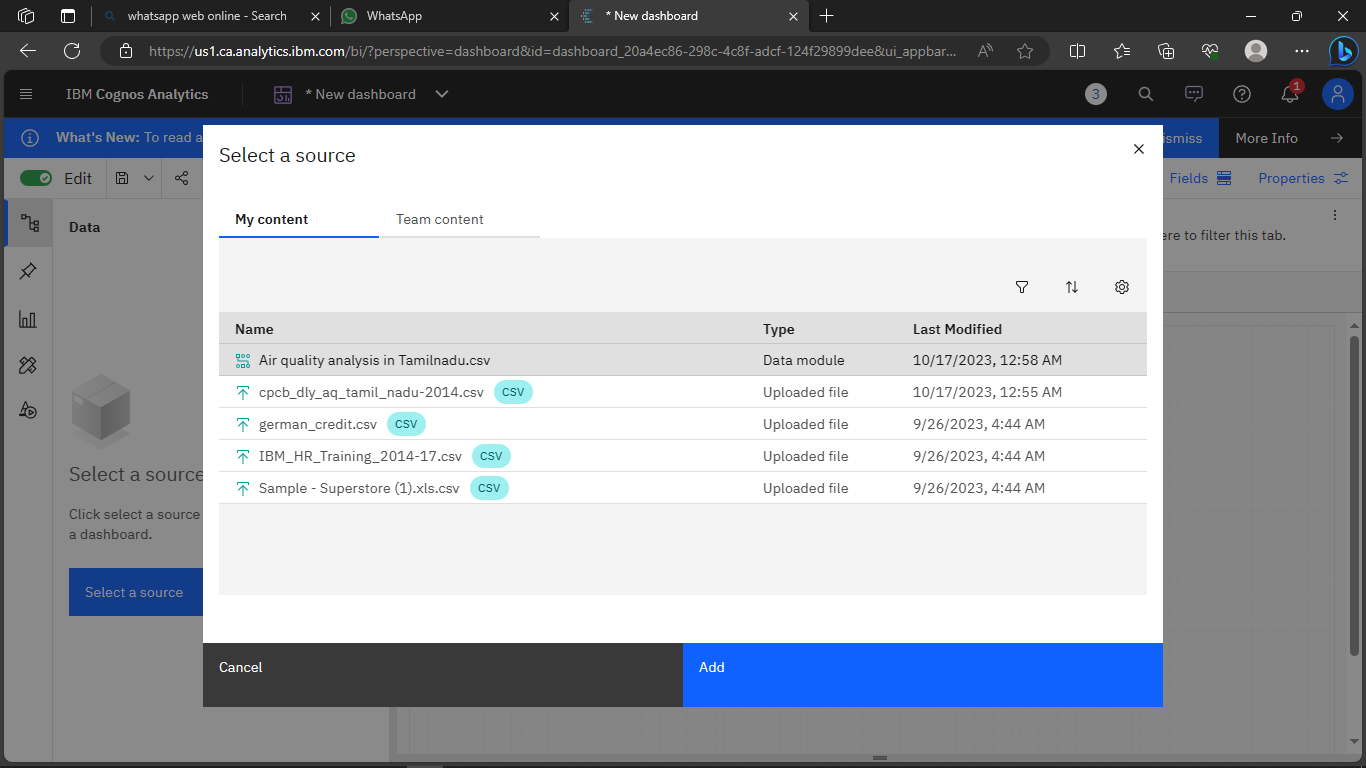
1. Login to your IBM cognos
2. Click more menu from the left side
3. Select new tab



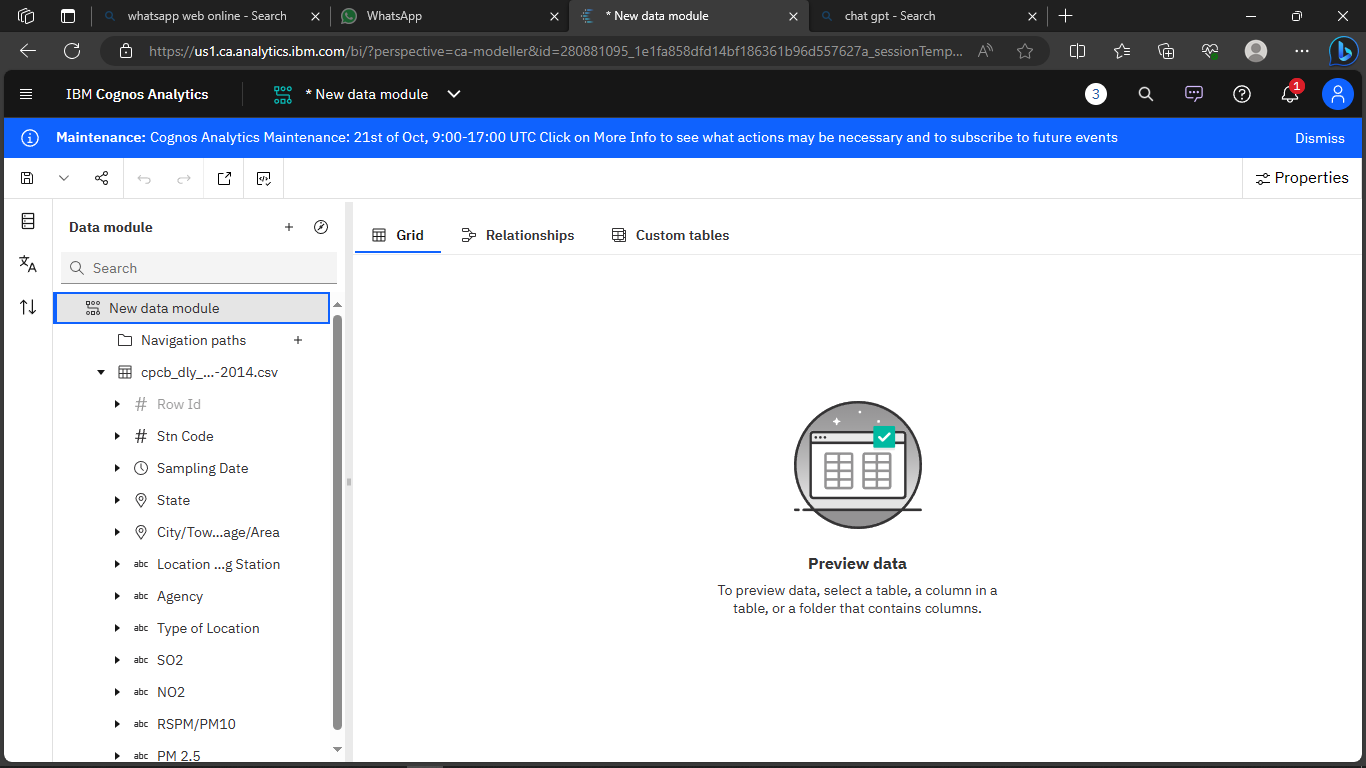
4. Click Data module tab

****

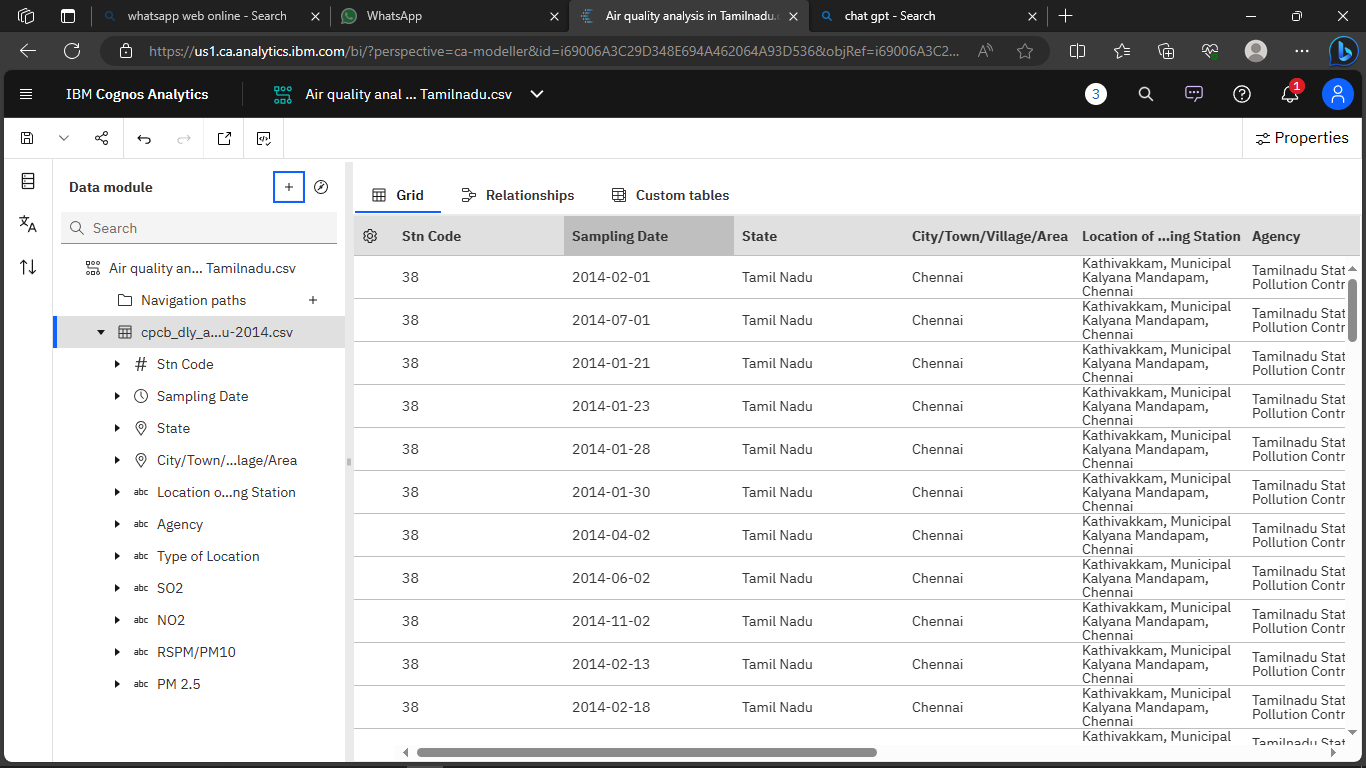
5.Upload the dataset for your project and select the Corresponding file



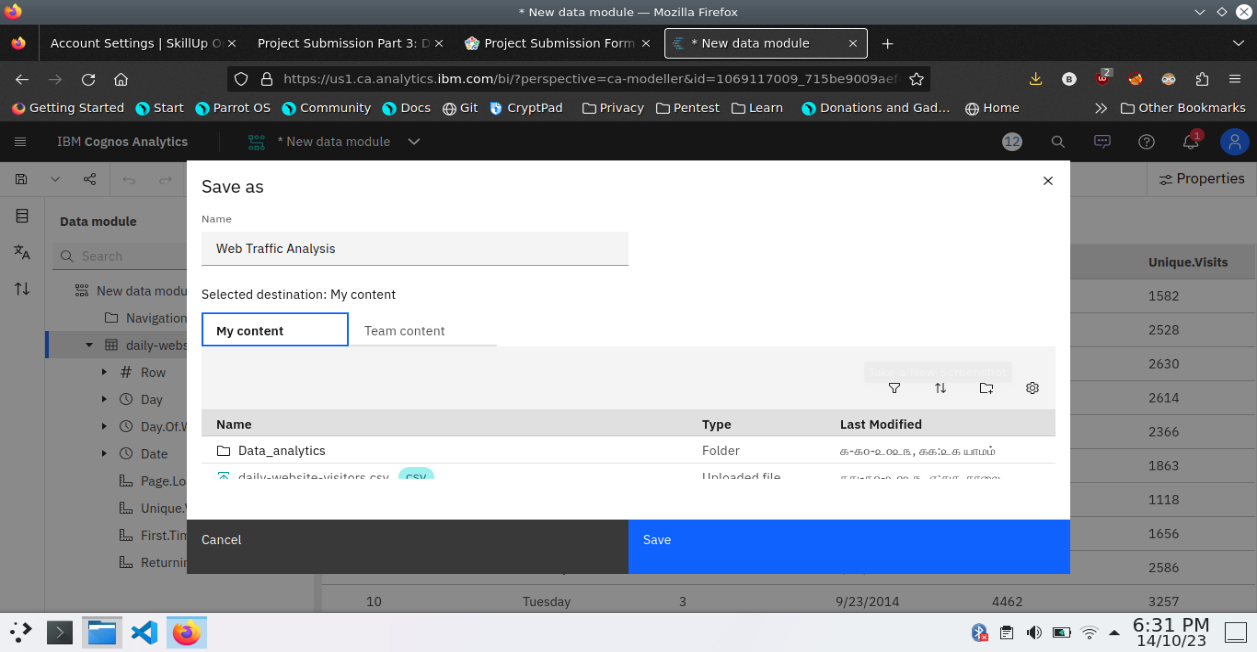
6. preview the data



7.Explore the data



8. save the data module



**Data Preprocessing and Cleaning**

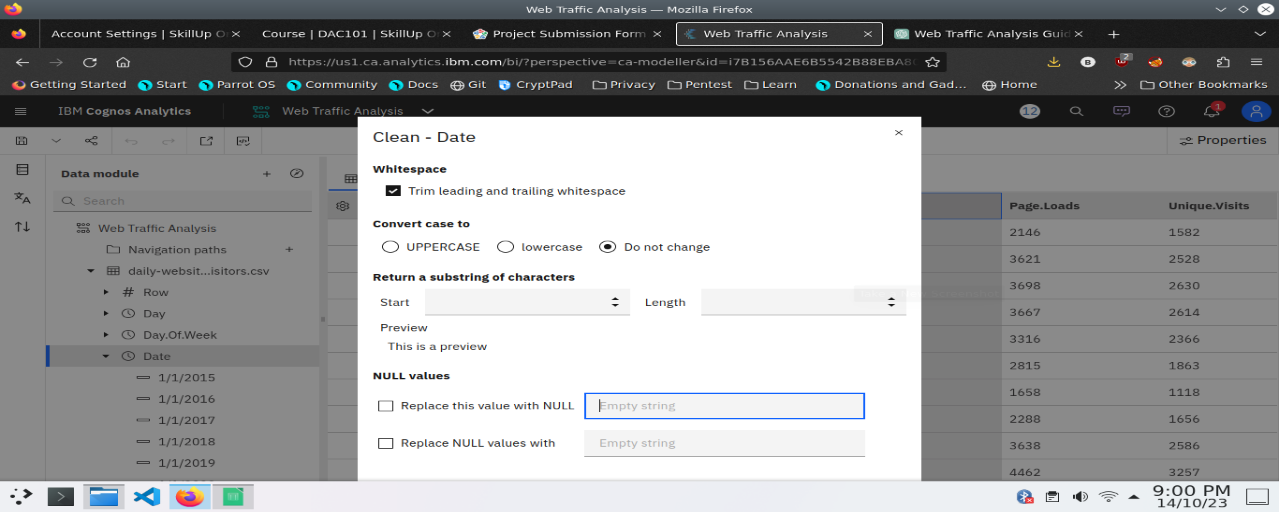
In this phase the following steps will taken

* 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



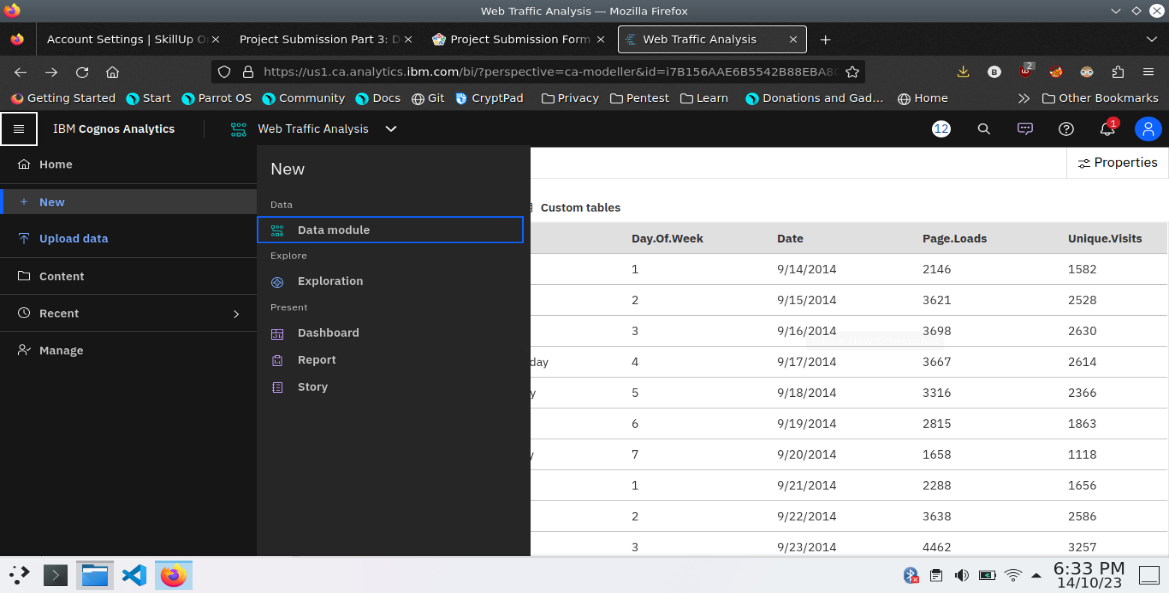
data module will be updated by doing the above process

after the completion of process start creating the dashboard for Visualization

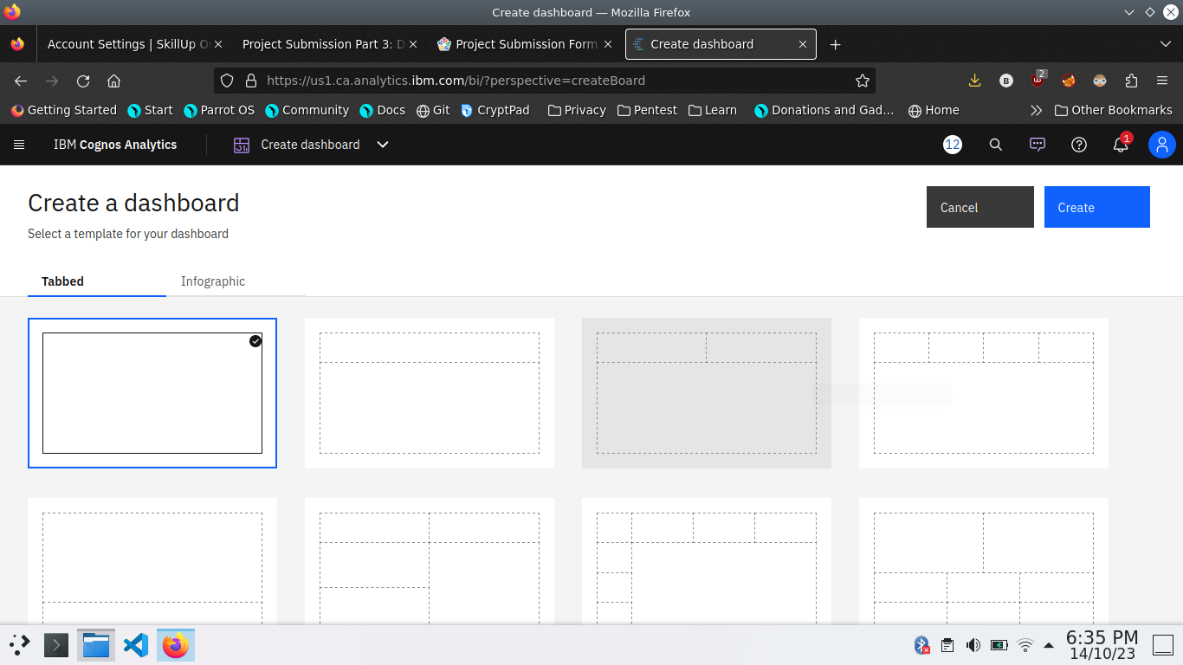
**Dashboard Creation**

Dashboard creation are helpful to visualizing the data

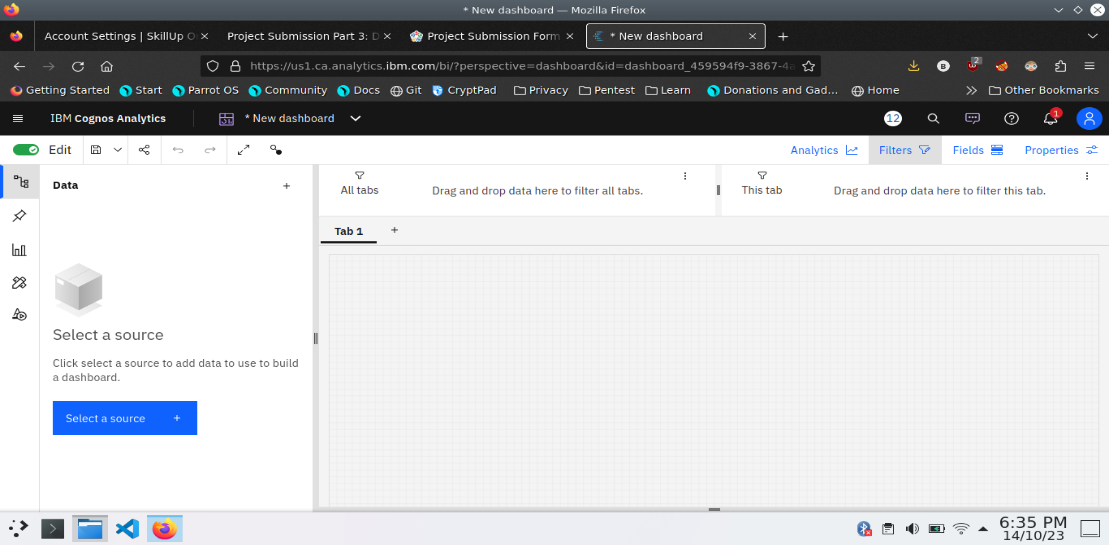
1. Goto Home menu
2. Select the new tab
3. Click dashboard



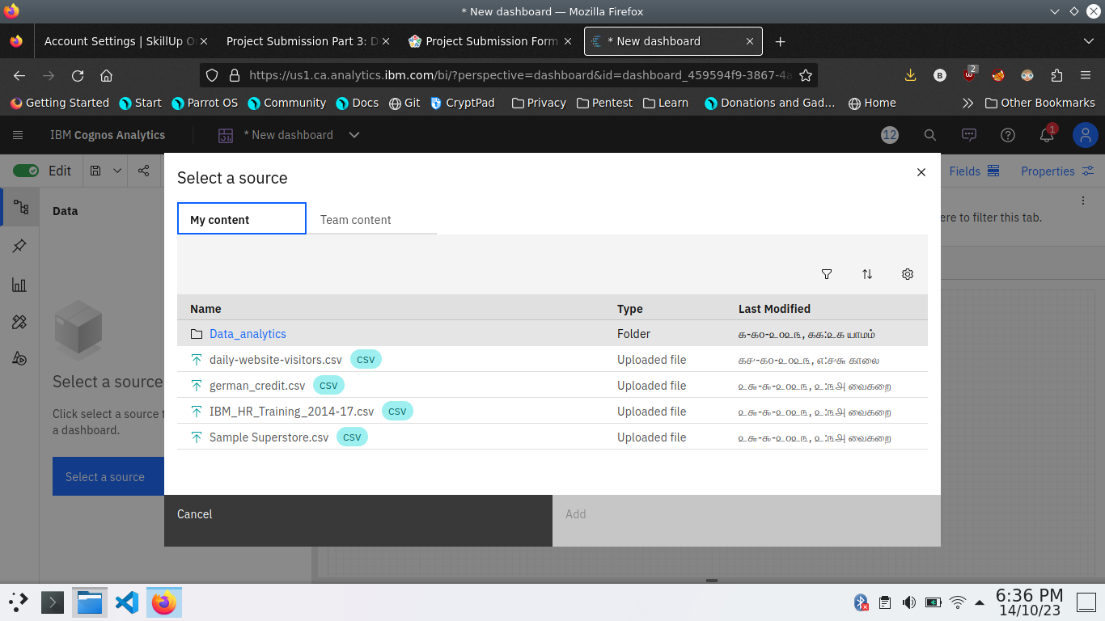
4. Choose the template for your project and click



5.Now Dashboard is created



6. Select the data source



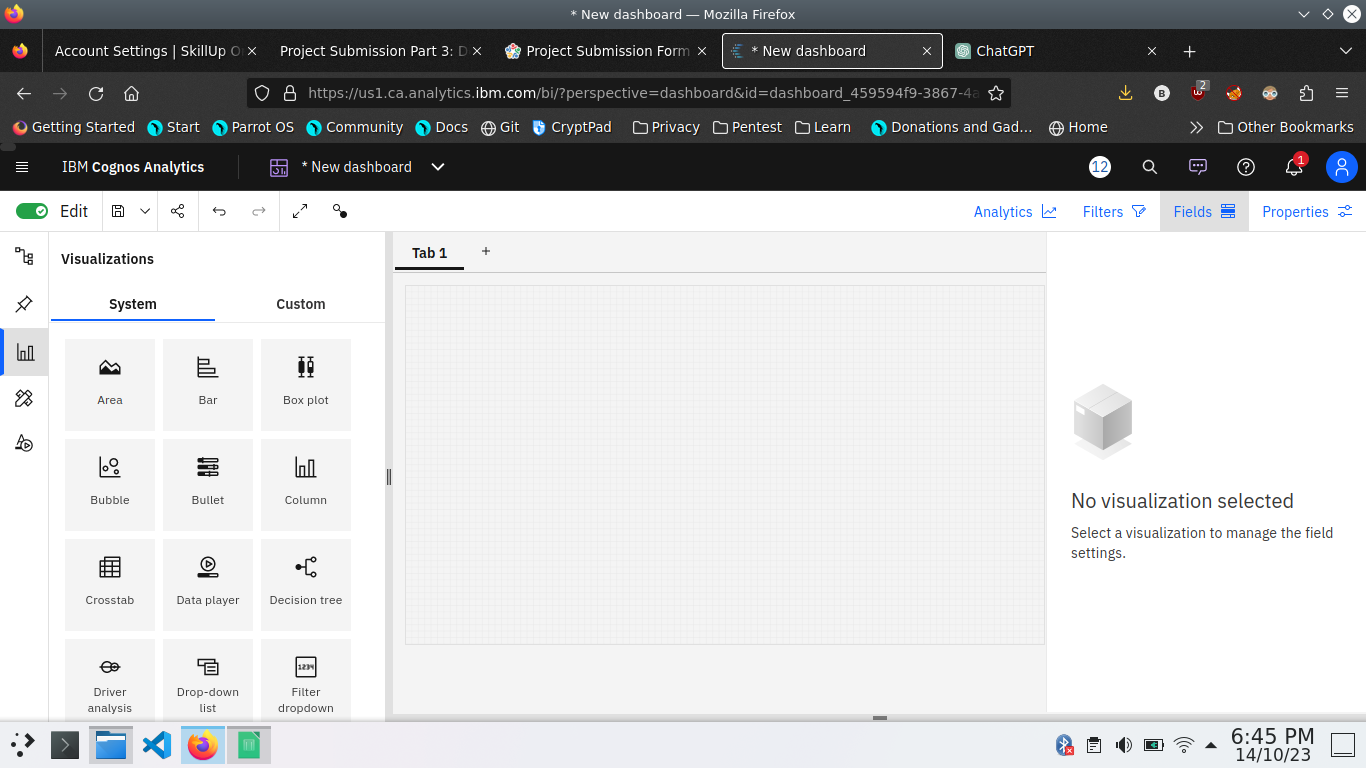
**Visualization**

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

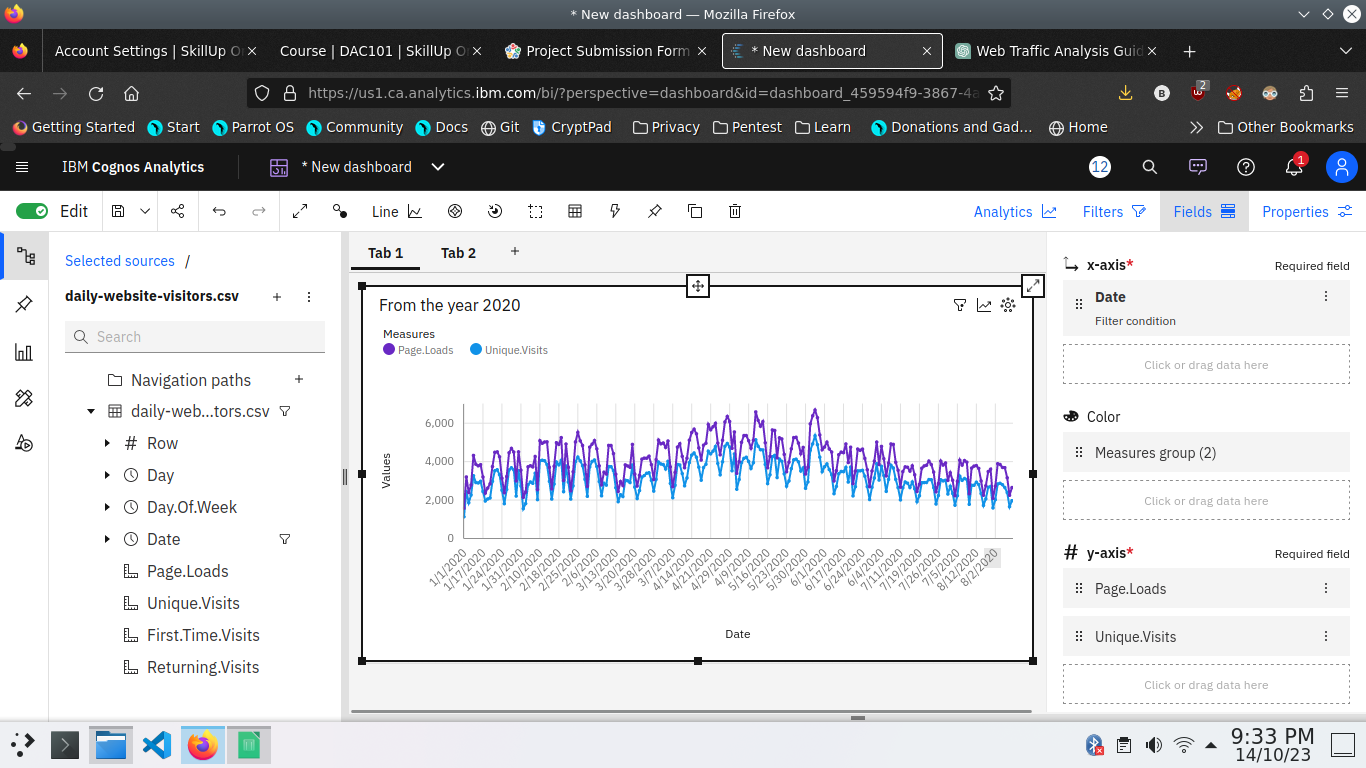
In IBM Cognos

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 “ Page.loads” and “Unique.visits” from the time period of 2020

X-axis =Dates

Y-axis = Page.Loads, Unique.visits.

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

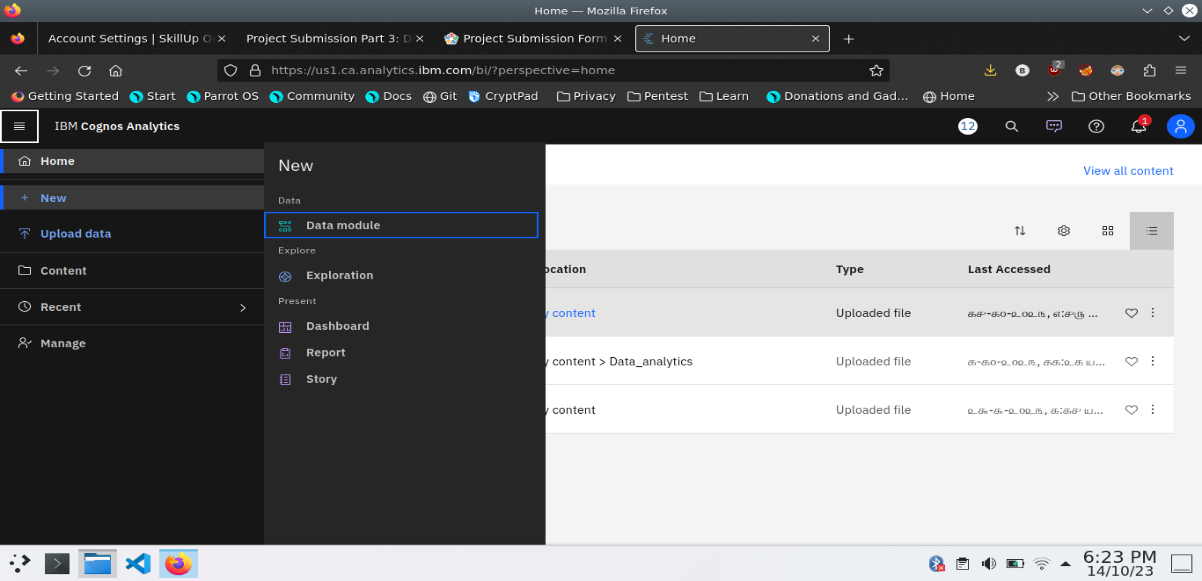
**Objectives**

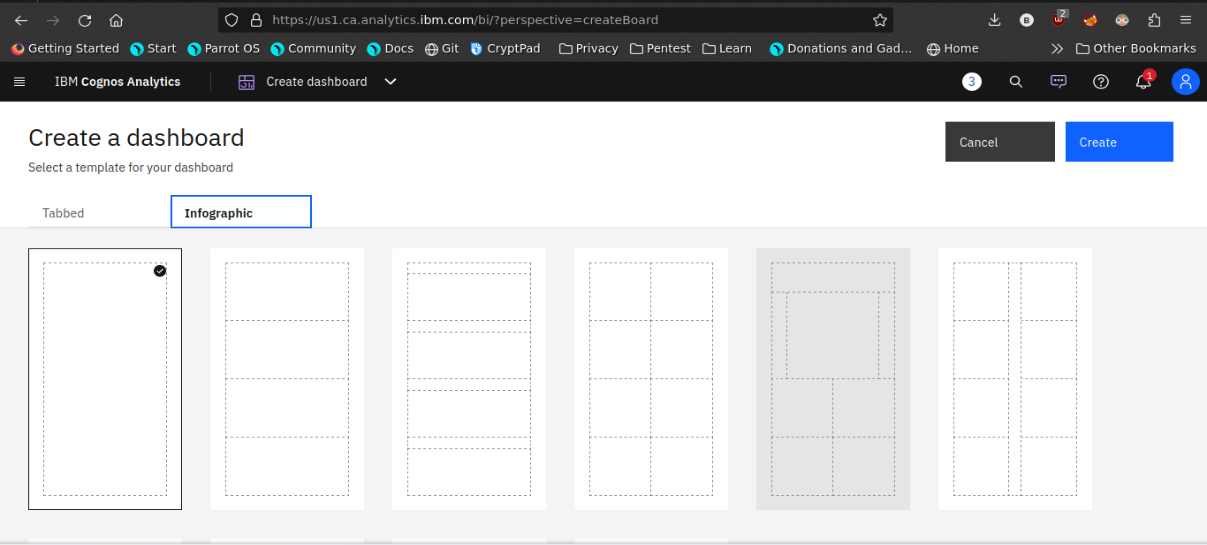
In this phase defines the visualization using IBM cognos, Integrating python code for advanced analysis and creating interactive dashboard. Use Python libraries and machine learning models for Complex analysis

**Visualization in IBM Cognos**

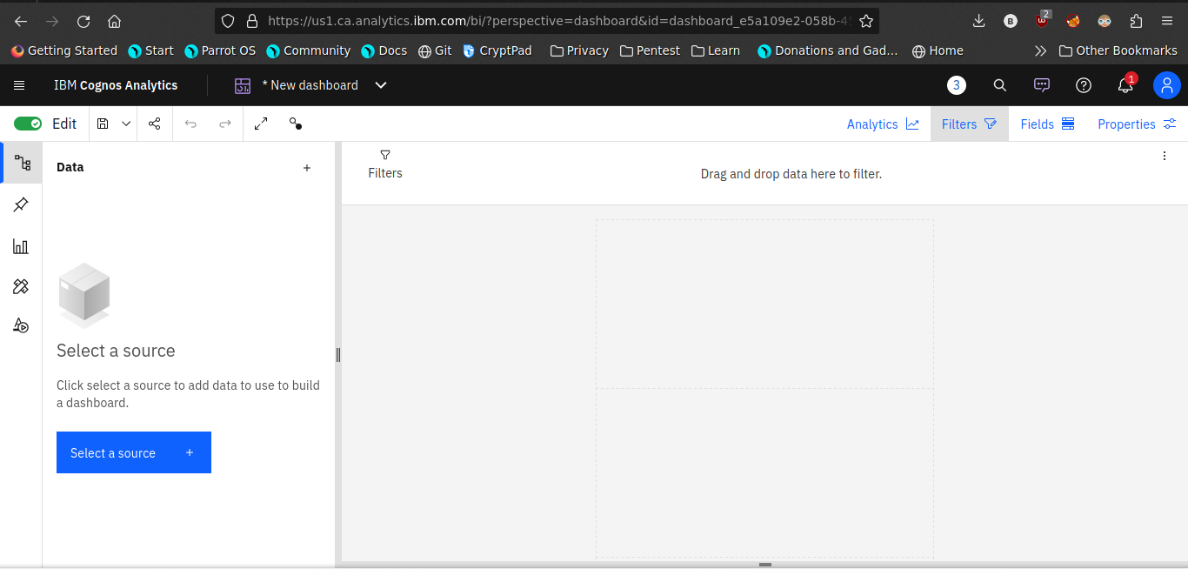
**Step 1:**

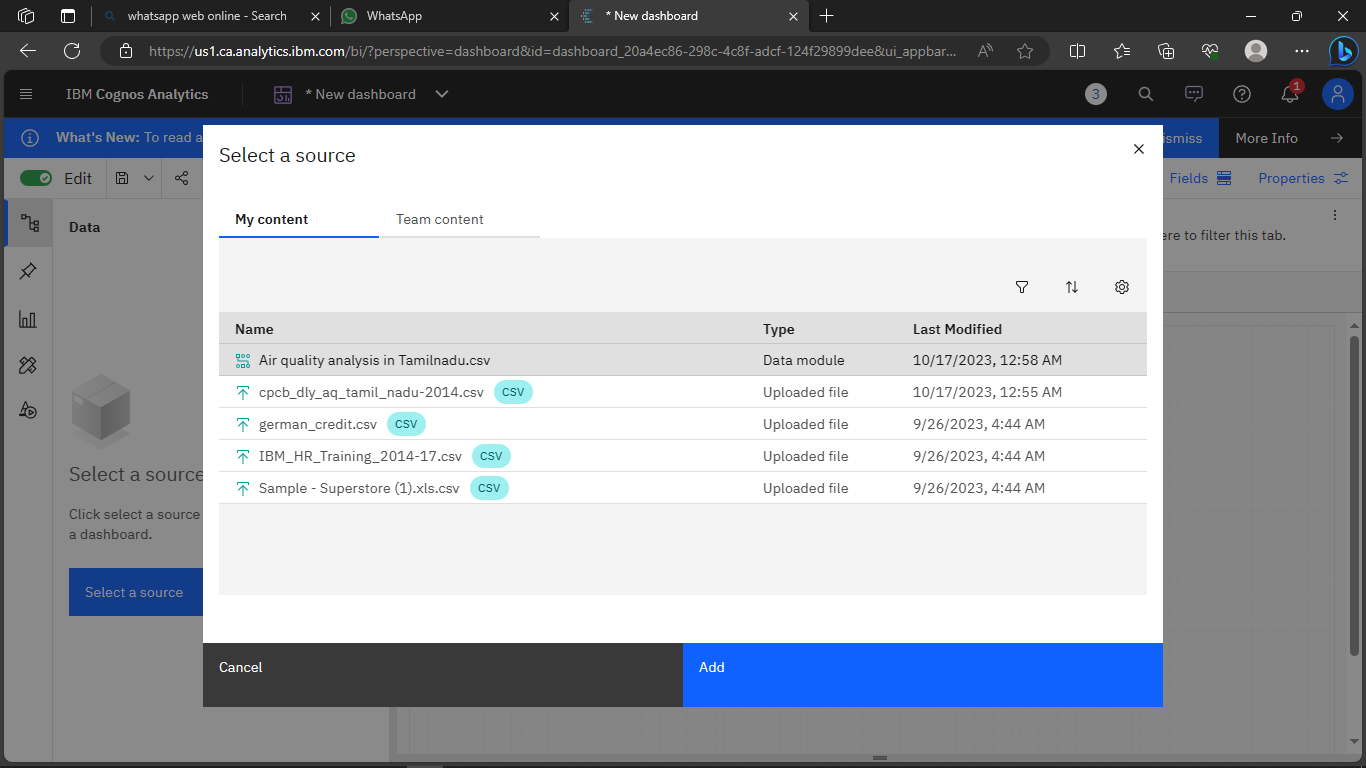
1. Login to your IBM cognos
2. Click more menu from the left side
3. Select new tab
4. Click Dashboard tab

****

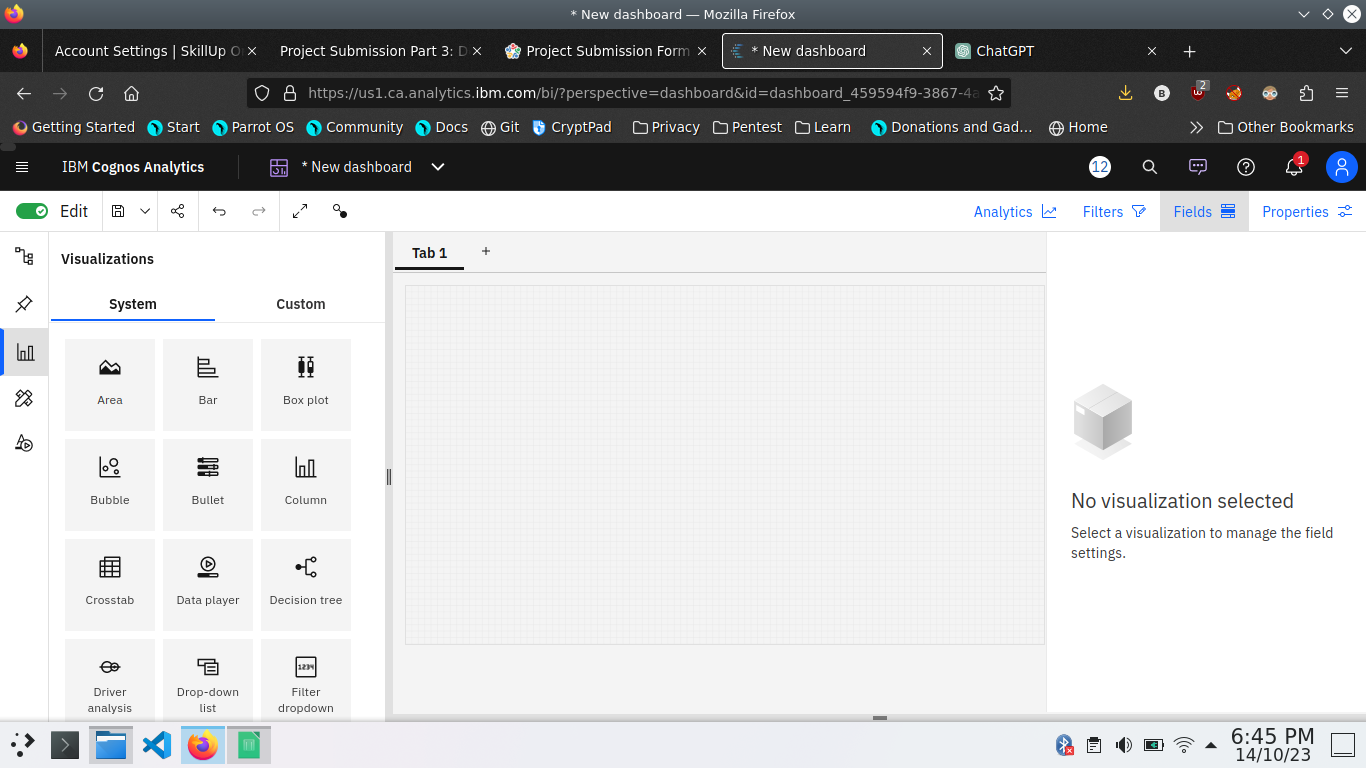
1. Select Template for your dashboard

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



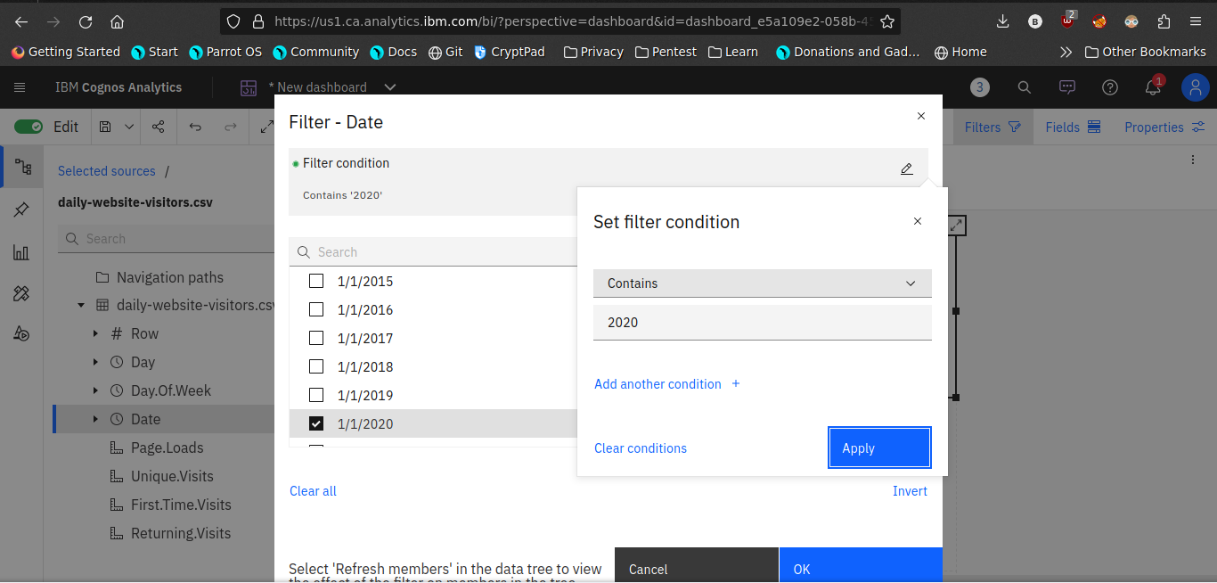
Select your Corresponding dataset

1. Select favourable visualization system



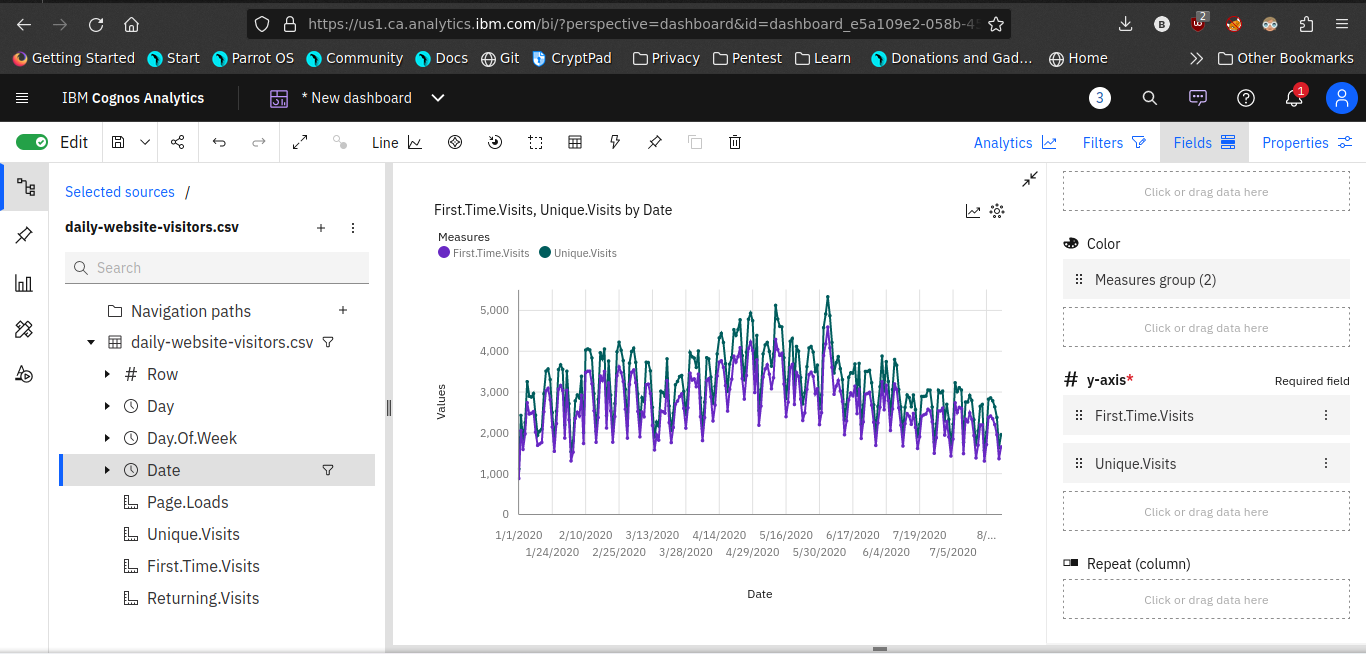
1. If needed Filter the data

Here the datasets are filtered by the date Which contains 2020



1. Line plot

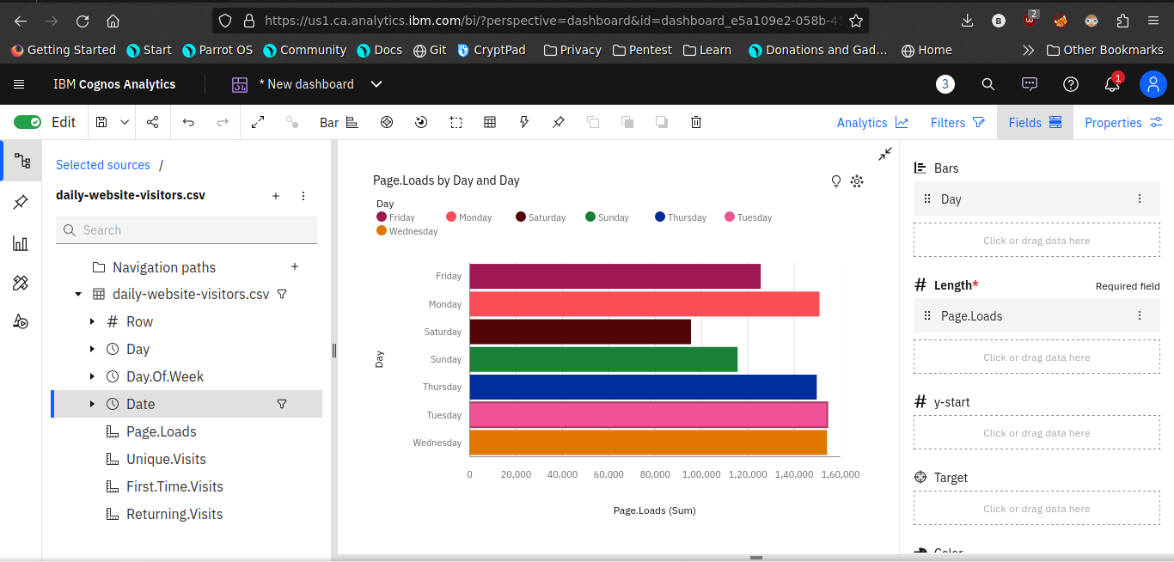
In this line plot X\_axis are dates and Y\_axis are First time visits and Unique visits



1. **Barchart**

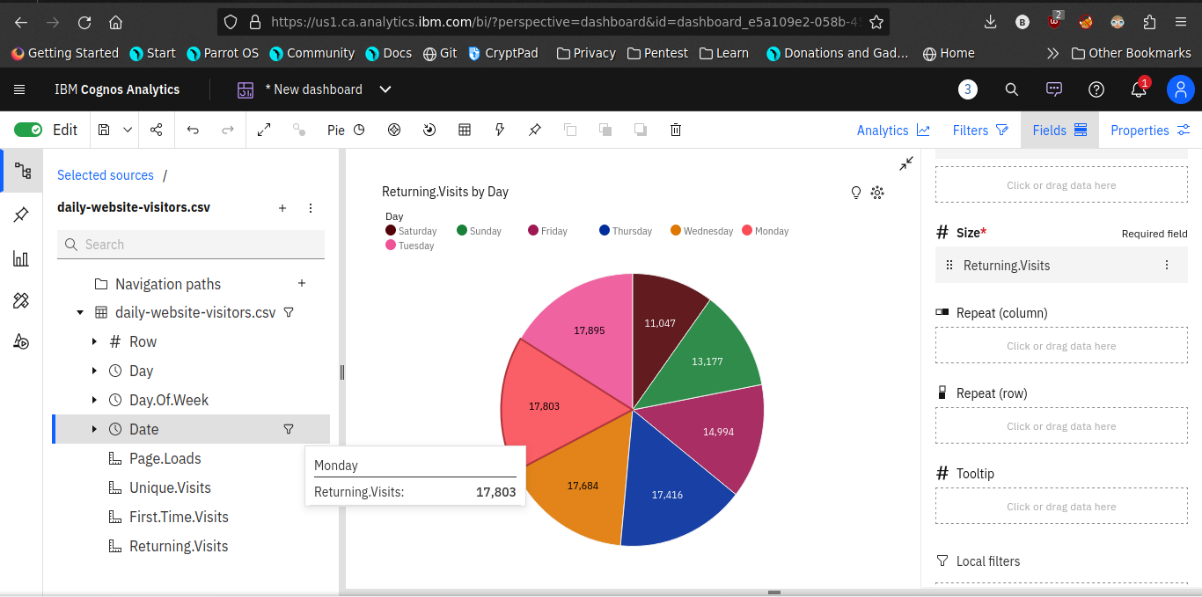
In this Bar chart the bars represent the ‘days in week’ and length defines ‘Page.Loads’

It helpful to visualize the maximum pageloads occurs on a day



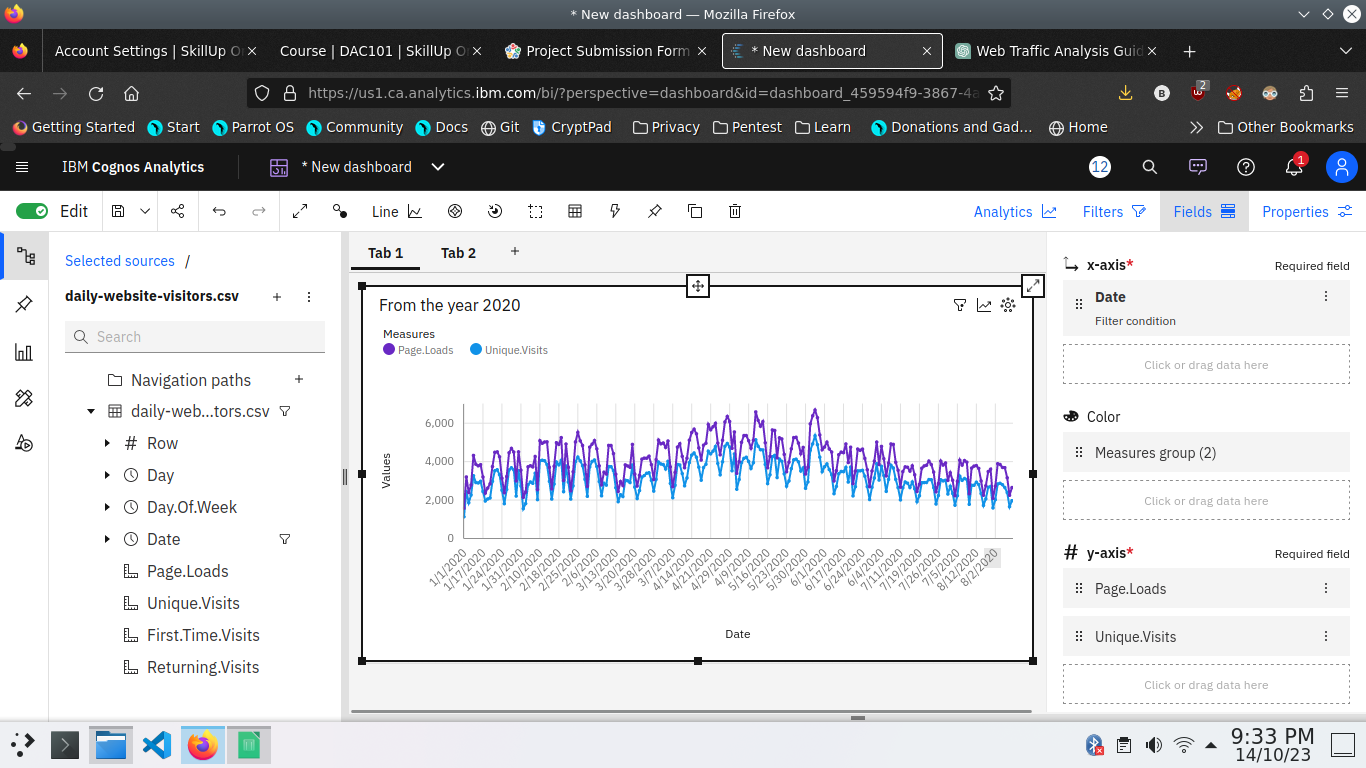
1. **Pie chart**

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



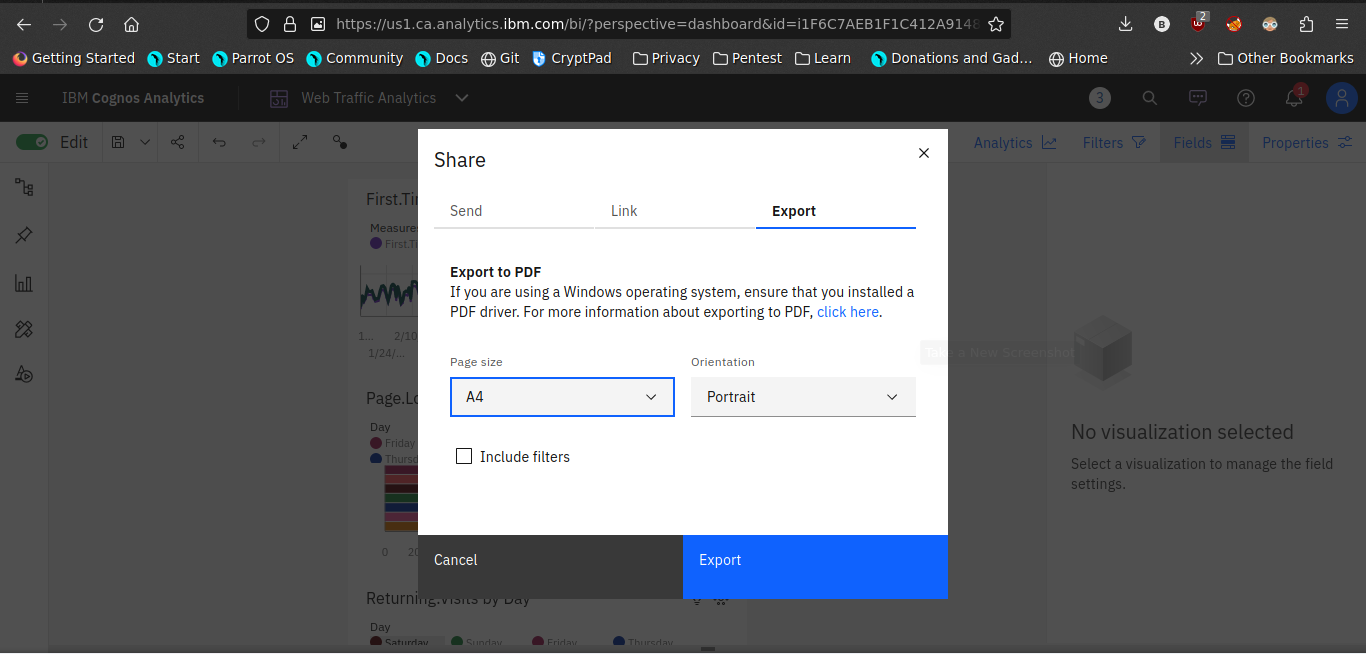
1. **Scatter plot**

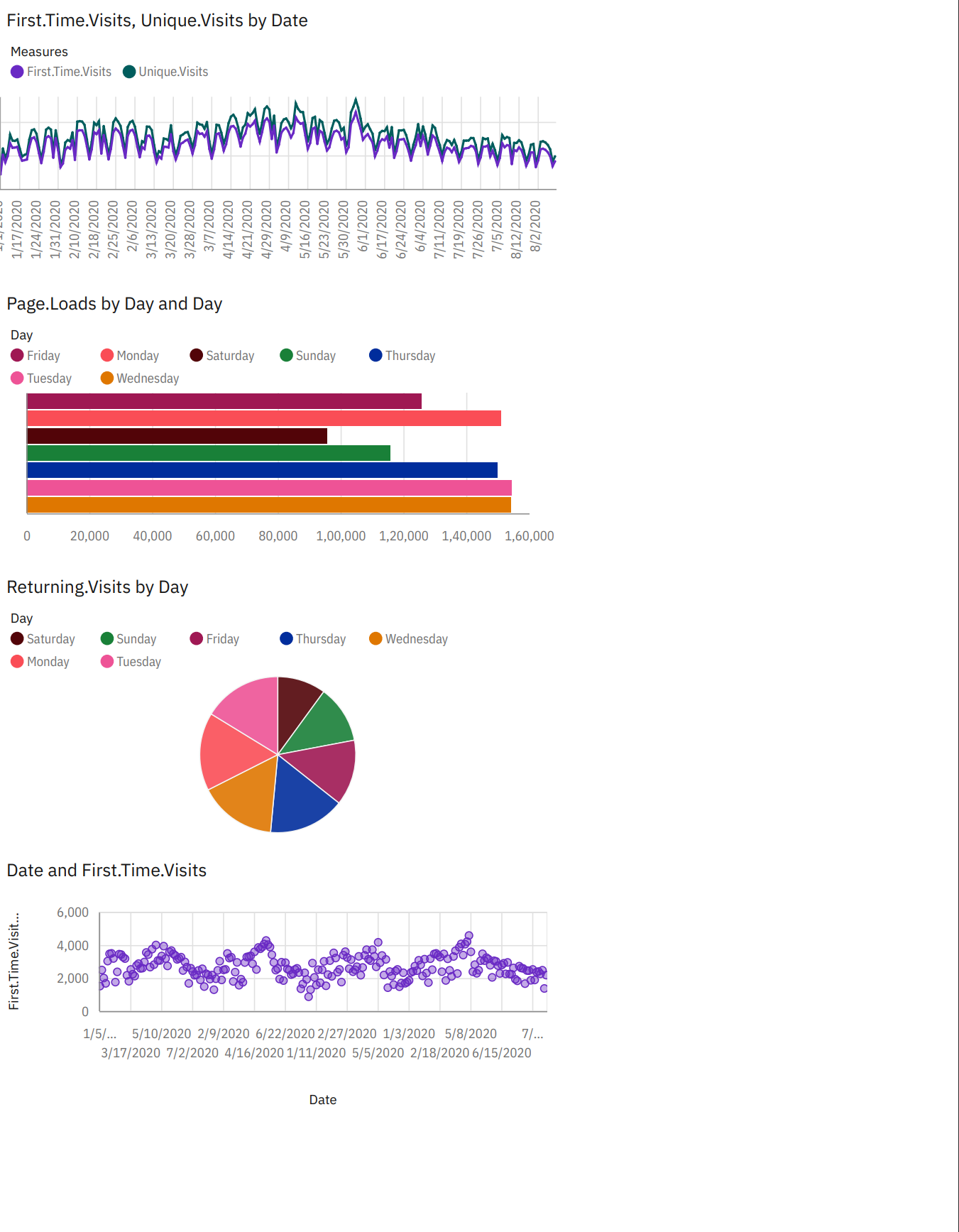
It is used to display the relationship between two variables and observe the nature of the relationship. The relationships observed can either be positive or negative, non-linear or linear



1. After completing the dashboard lets export to another format

Like pdf





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 Necessary packages**

Pandas

seaborn

Machine learning models

Linear regression

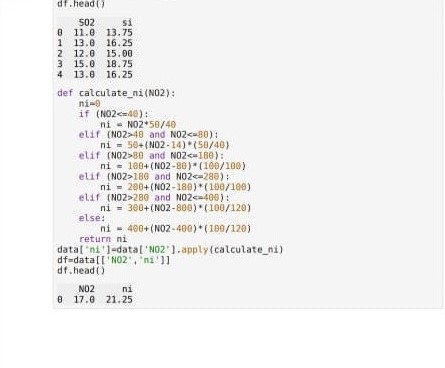
1. **Make a training and test data**

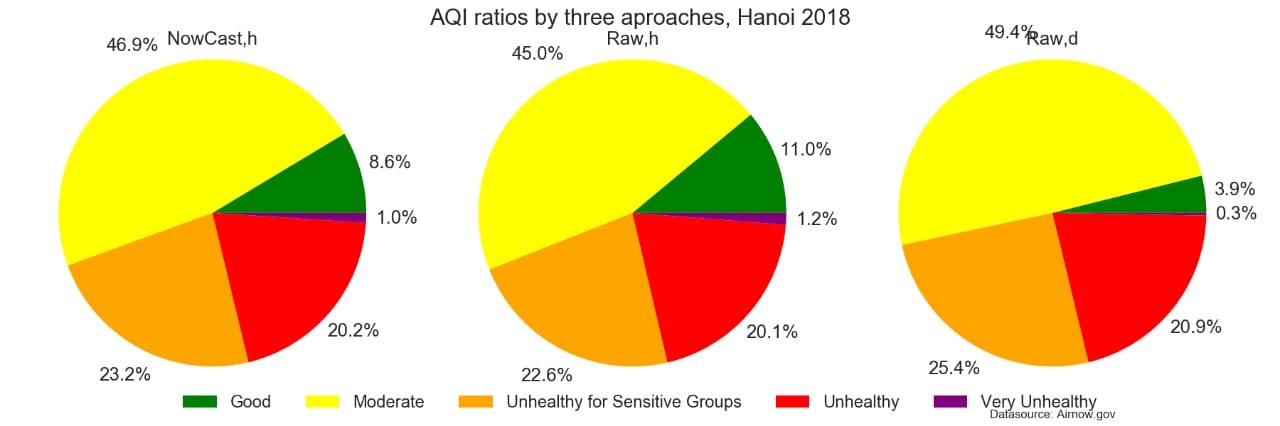
Use the train test split model

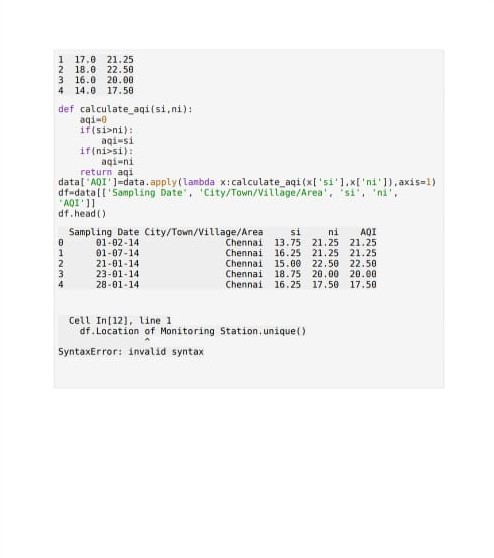
Compare the testing and training data set by visualization library

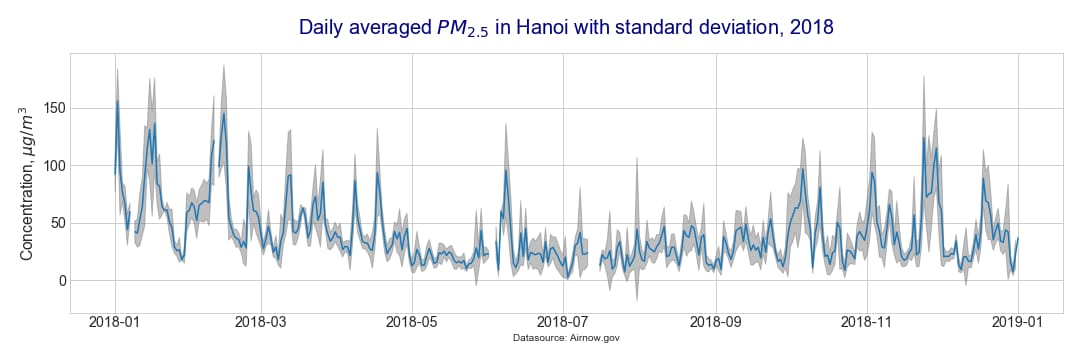
1. **Calculate the accuracy of the model**

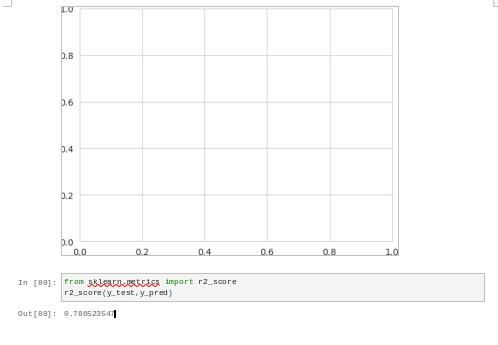
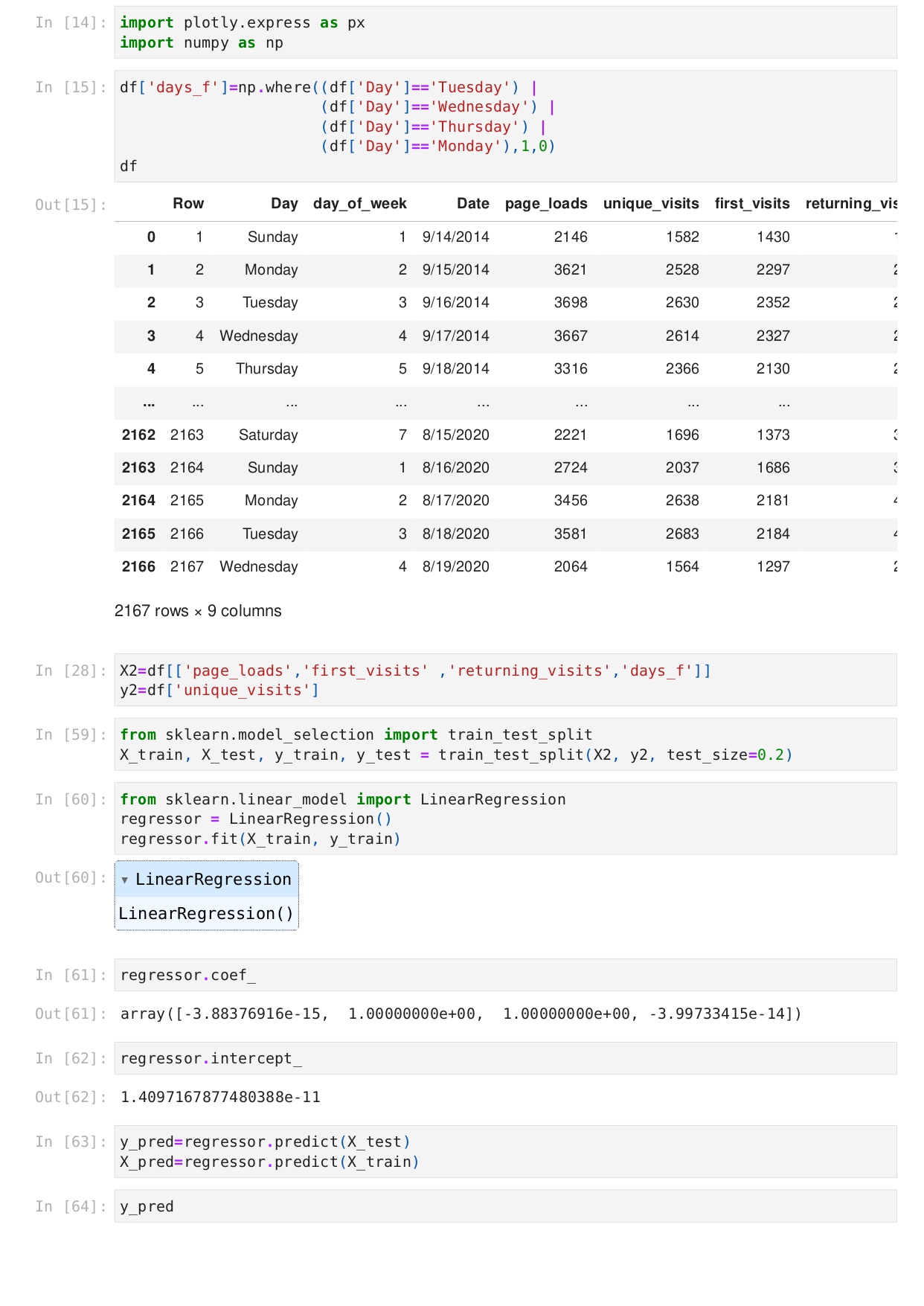
Use r2\_score to measure the accuracy of the model



****

****

****

****

The accuracy for the above machine learning model is

0.780523547

**Conclusion:**

A comprehensive analysis of air quality in Tamil Nadu reveals several critical insights. While the state has made progress in managing air pollution, there are ongoing challenges that requires immediate attention to safeguard public health and the environment.