DATA ANALYTICS WITH COGNOS

Name: Bharathi M

REG NO: 410121104010

DOMAIN: data analytics with cognos

PROJECT NAME : Air Quality Analaysis

PHASE 2 : Innovation

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



**Collect High-Quality Data:**

Ensure the dataset is clean, complete, and representative of the problem that we are trying to solve. We should address missing values and outliers appropriately.

**Data Preprocessing:**

Standardize or normalize our data to have consistent scales. Categorical variables may need to be encoded appropriately (e.g., one-hot encoding).

**Model Selection:**

Choose an appropriate machine learning algorithm for our problem. We should experiment with different algorithms to see which one performs best.

**Hyperparameter Tuning:**

Optimize the hyperparameters of our chosen algorithms. Techniques like grid search or random search can be helpful in finding the best combination of hyperparameters.

**Cross-Validation:**

Use techniques like k-fold cross-validation to assess our model's performance more reliably. This helps prevent overfitting and provides a better estimate of how well your model will generalize to new data.

**Ensemble Methods:**

Combine the predictions of multiple models (e.g., Random Forest, Gradient Boosting) to improve accuracy. Ensemble methods often outperform individual models.

**Regularization:**

Apply regularization techniques like L1 or L2 regularization to prevent overfitting.

**Feature Selection:**

Use feature selection techniques to identify and keep only the most important features. This can simplify our model and improve its generalization.

**CODE:**

import warnings

warnings.filterwarnings('ignore')

import pandas as pd

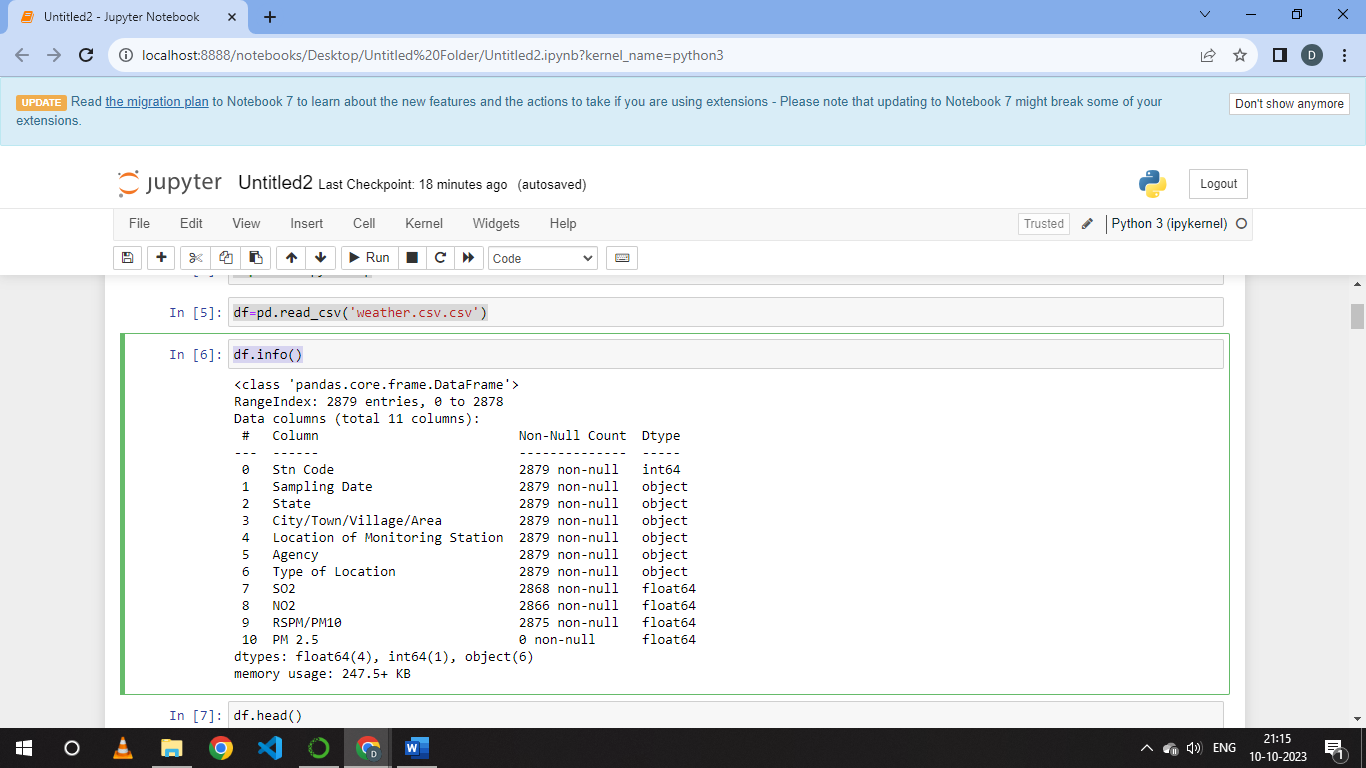
import matplotlib.pyplot as plt

import seaborn as sns

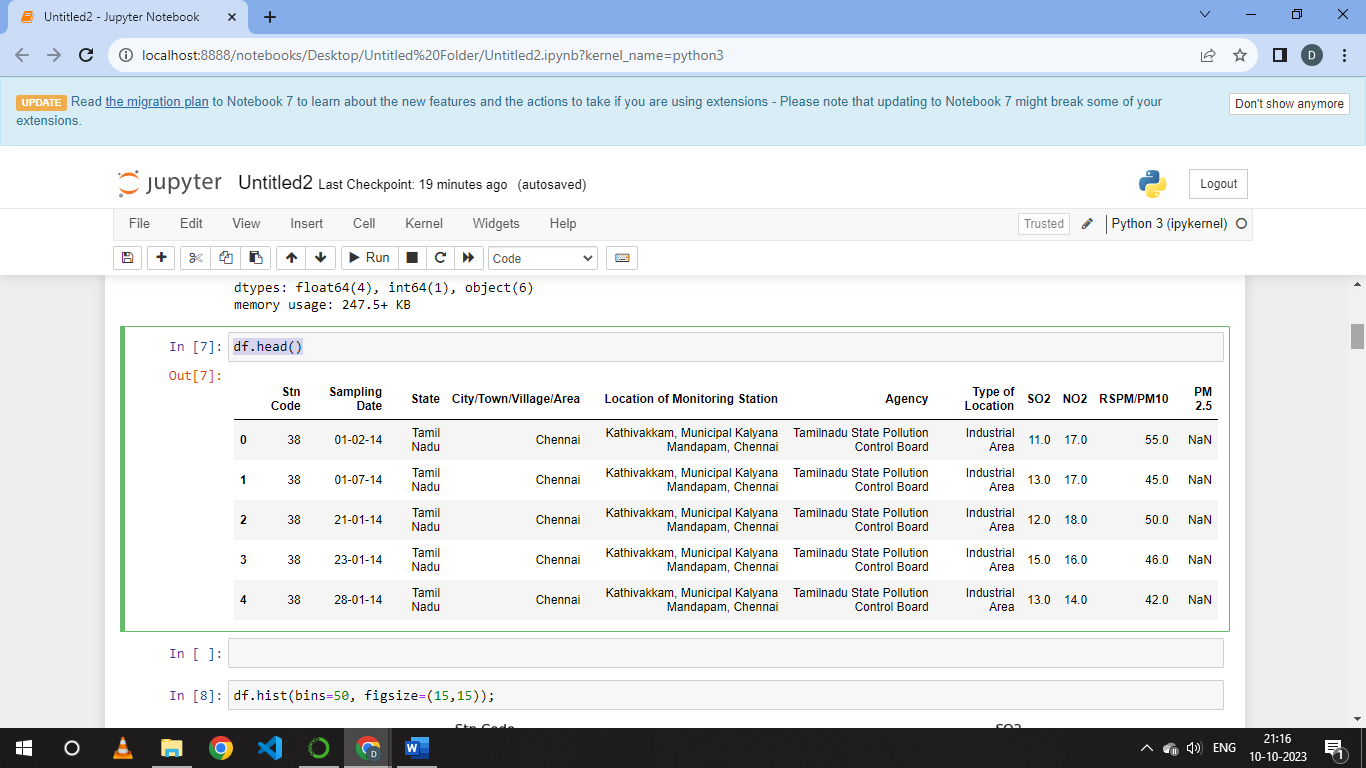
import numpy as np

df=pd.read\_csv('weather.csv')

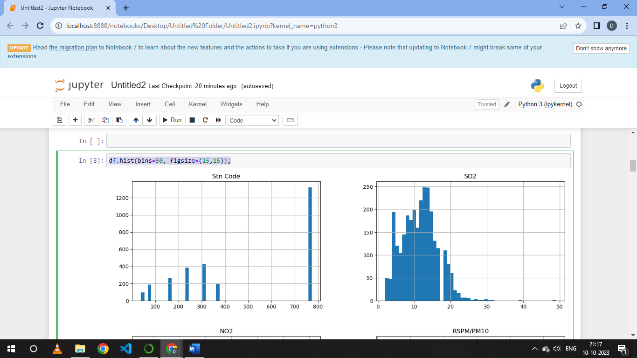
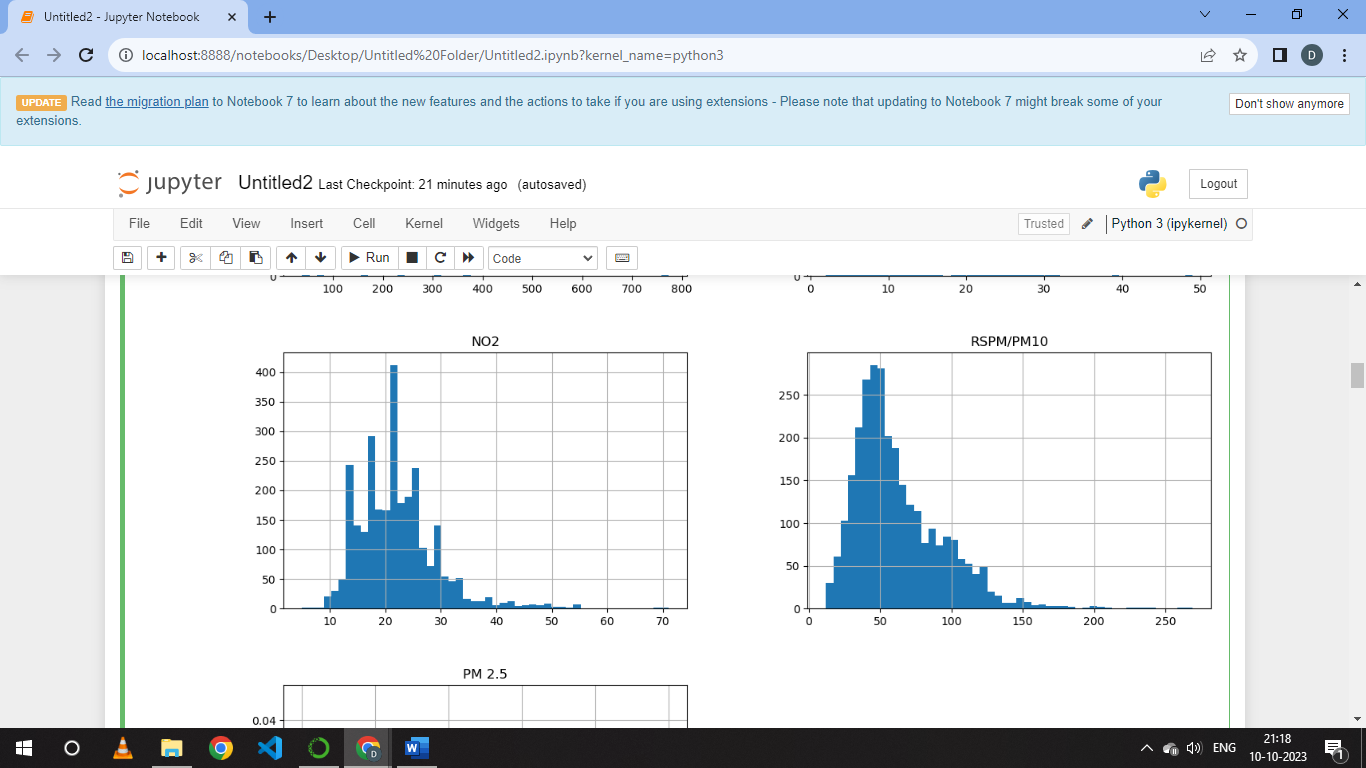
df.info()



df.head()



df.hist(bins=50, figsize=(15,15))

from sklearn.compose import ColumnTransformer

from sklearn.pipeline import Pipeline

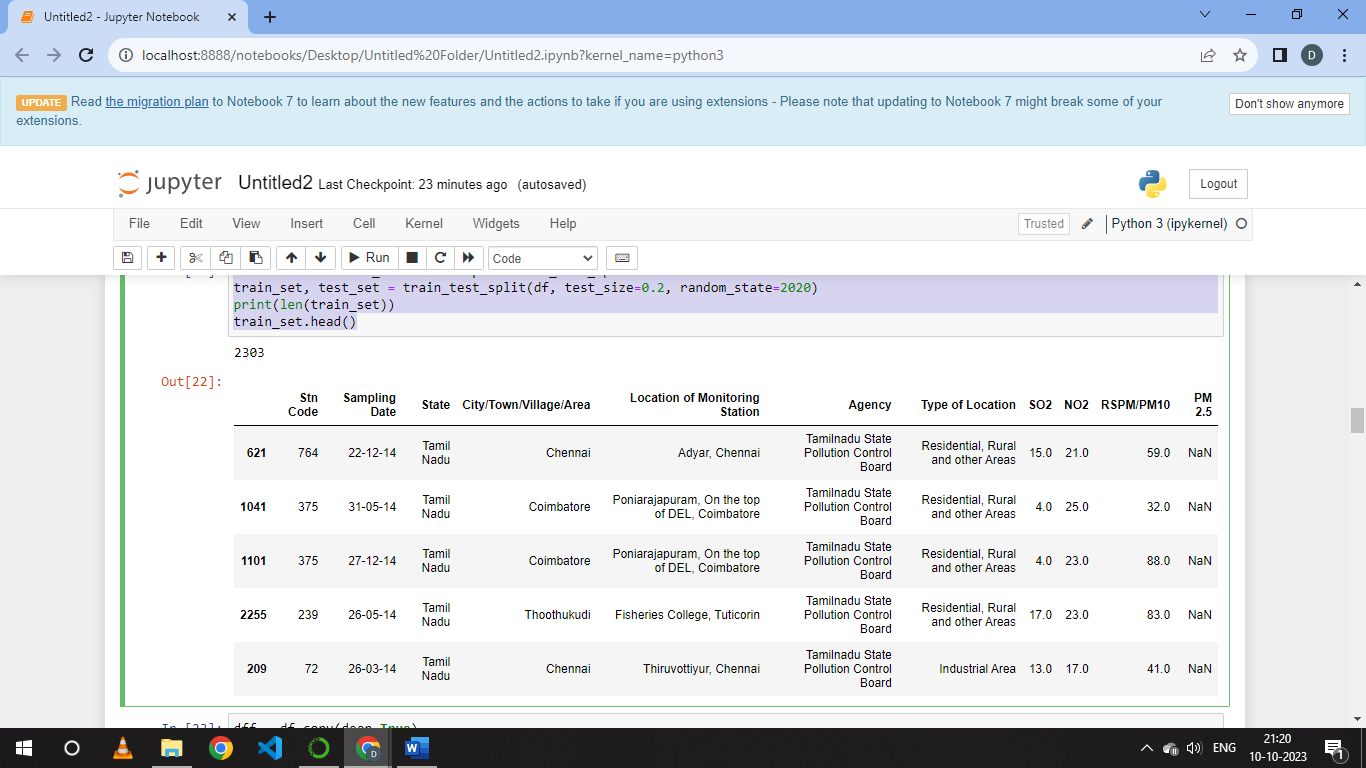
from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split

train\_set, test\_set = train\_test\_split(df, test\_size=0.2, random\_state=2020)

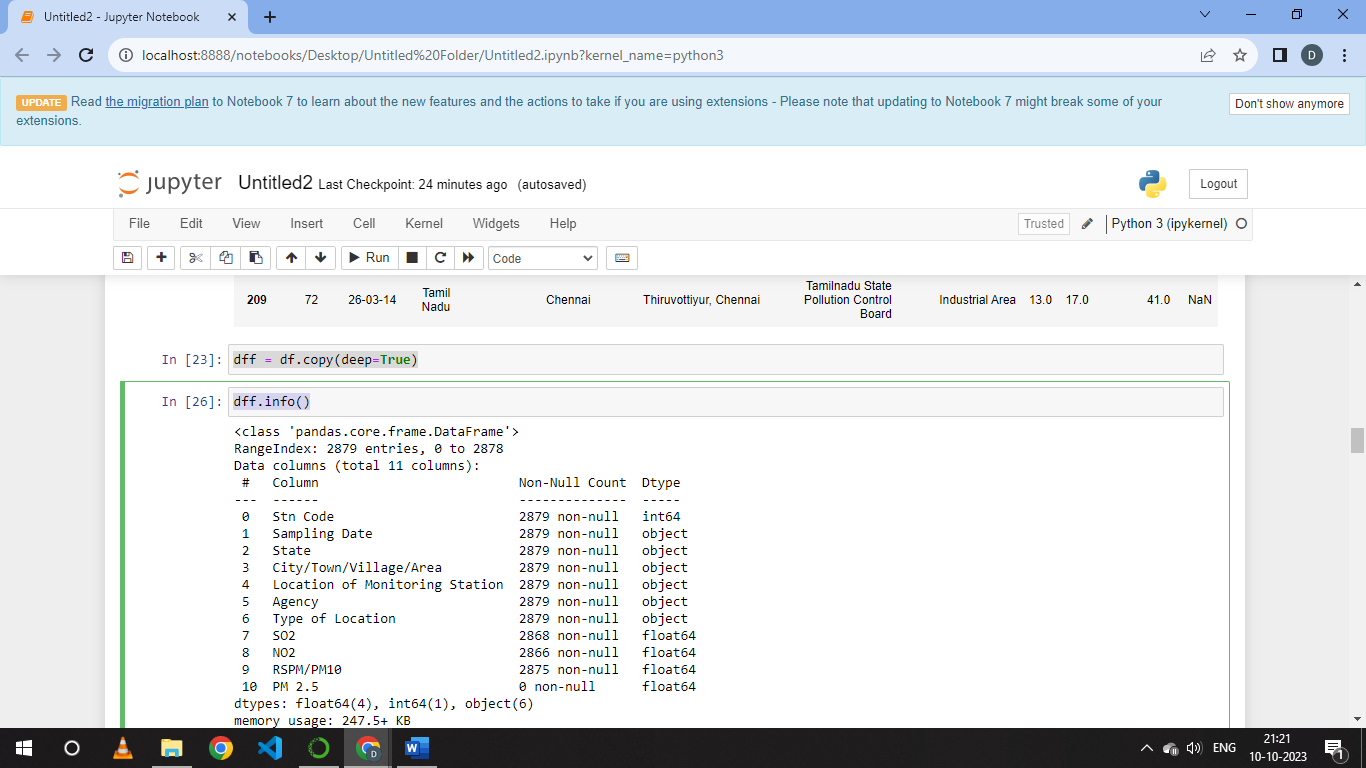
print(len(train\_set))

train\_set.head()



dff = df.copy(deep=True)

dff.info()

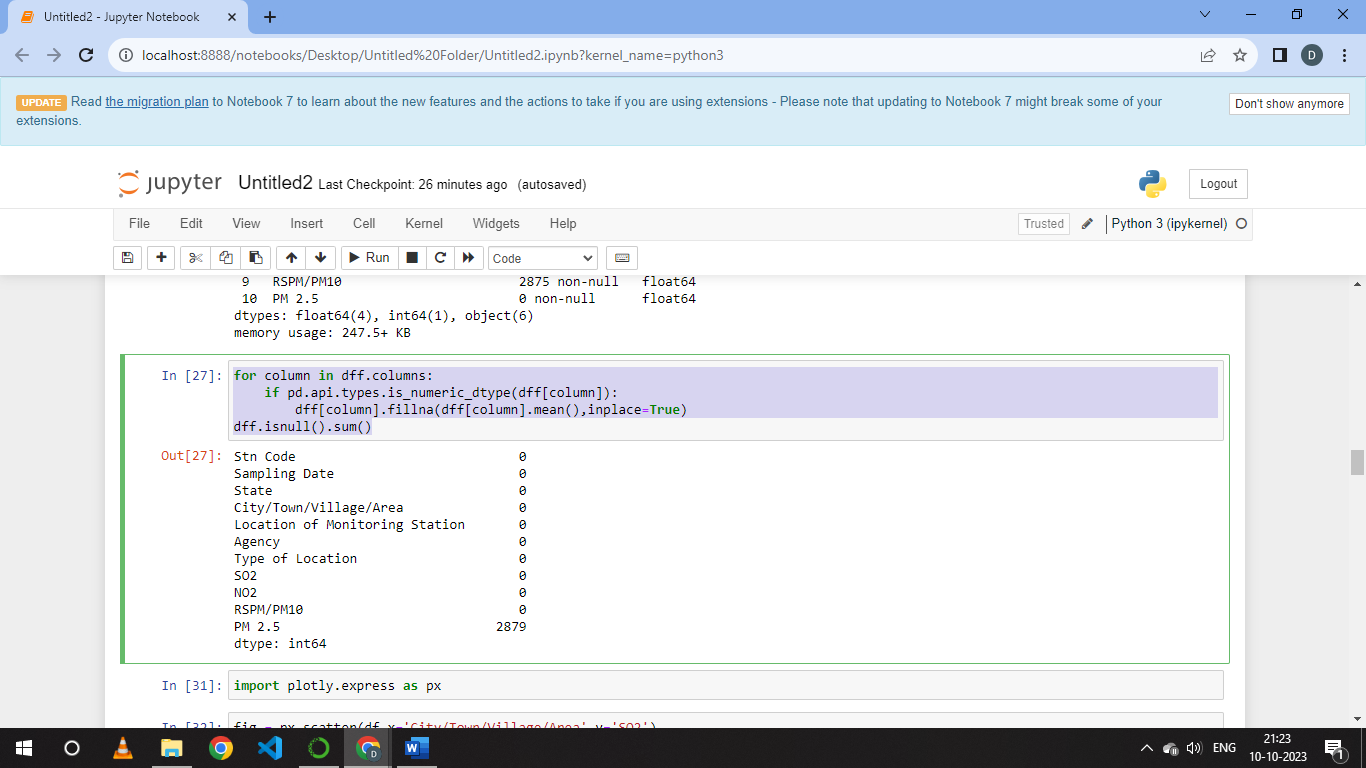


for column in dff.columns:

if pd.api.types.is\_numeric\_dtype(dff[column]):

dff[column].fillna(dff[column].mean(),inplace=True)

dff.isnull().sum()



import plotly.express as px

fig = px.scatter(df,x='City/Town/Village/Area',y='SO2')

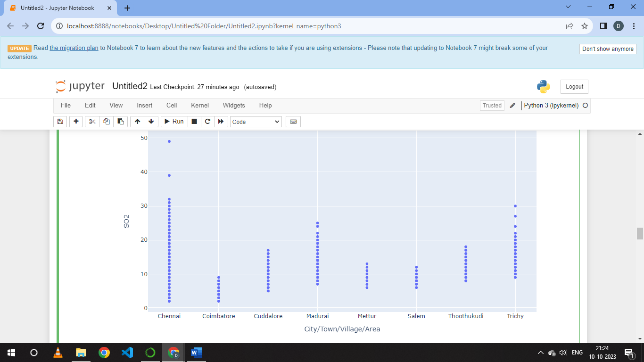
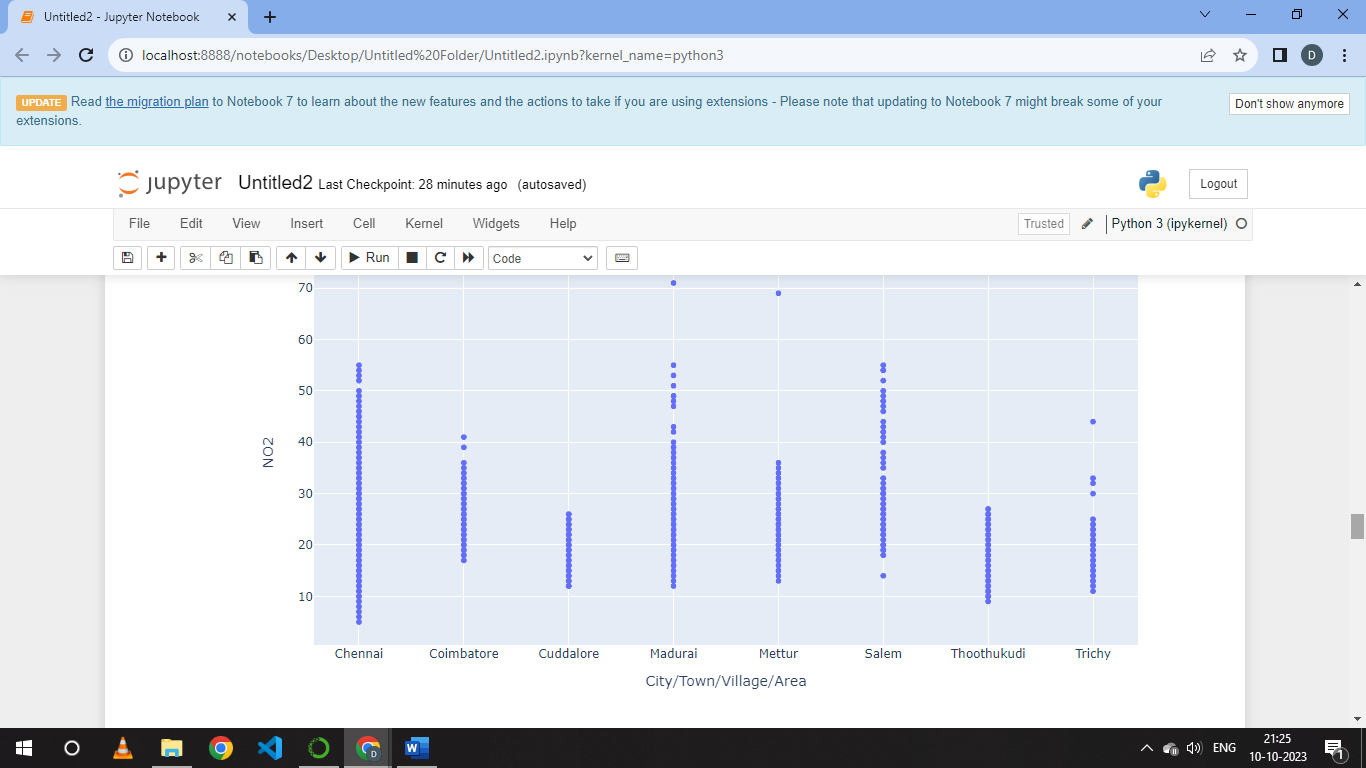
fig1 = px.scatter(df,x='City/Town/Village/Area',y='NO2')

fig2 = px.scatter(df,x='City/Town/Village/Area',y='RSPM/PM10')

fig.show()

fig1.show()

fig2.show()

features=df[['SO2','NO2']]

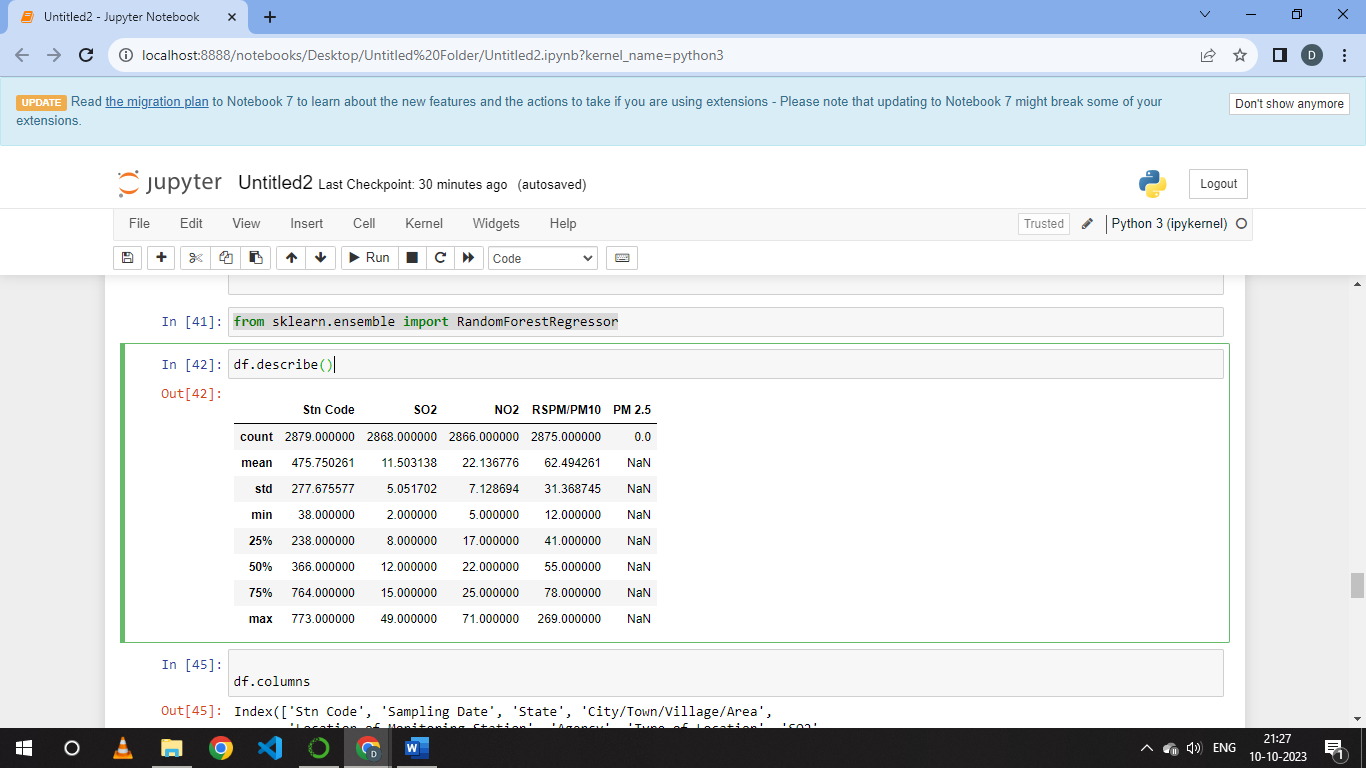
labels=df['RSPM/PM10']

from sklearn.model\_selection import train\_test\_split

xtrain,xtest,ytrain,ytest = train\_test\_split(features,labels,test\_size=0.2,random\_state=2)

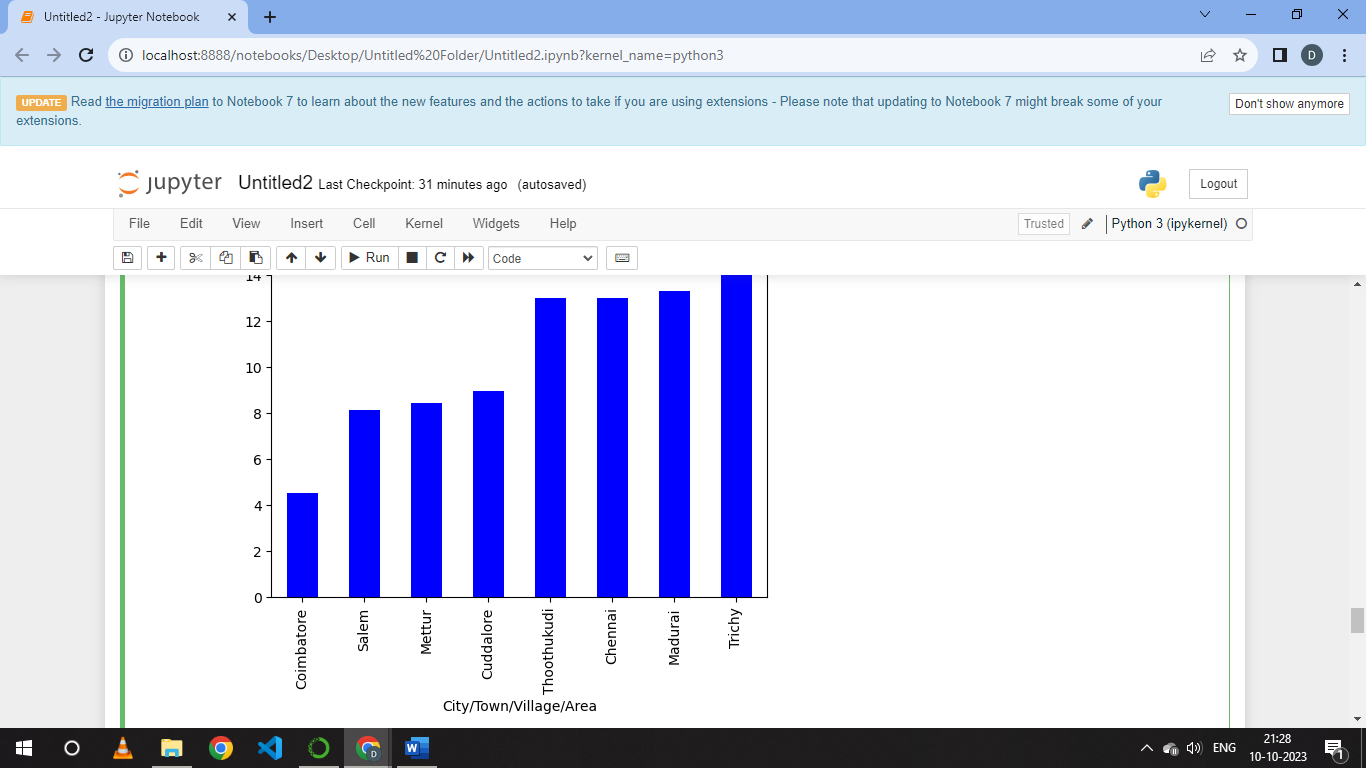
from sklearn.ensemble import RandomForestRegressor

df.describe()



df[['SO2','City/Town/Village/Area']].groupby(["City/Town/Village/Area"]).mean().sort\_values(by='SO2').head(20).plot.bar(color='b')

plt.show()



def cal\_SOi(SO2):

si=0

if (SO2<=40):

si= SO2\*(50/40)

elif (SO2>40 and SO2<=80):

si= 50+(SO2-40)\*(50/40)

elif (SO2>80 and SO2<=380):

si= 100+(SO2-80)\*(100/300)

elif (SO2>380 and SO2<=800):

si= 200+(SO2-380)\*(100/420)

elif (SO2>800 and SO2<=1600):

si= 300+(SO2-800)\*(100/800)

elif (SO2>1600):

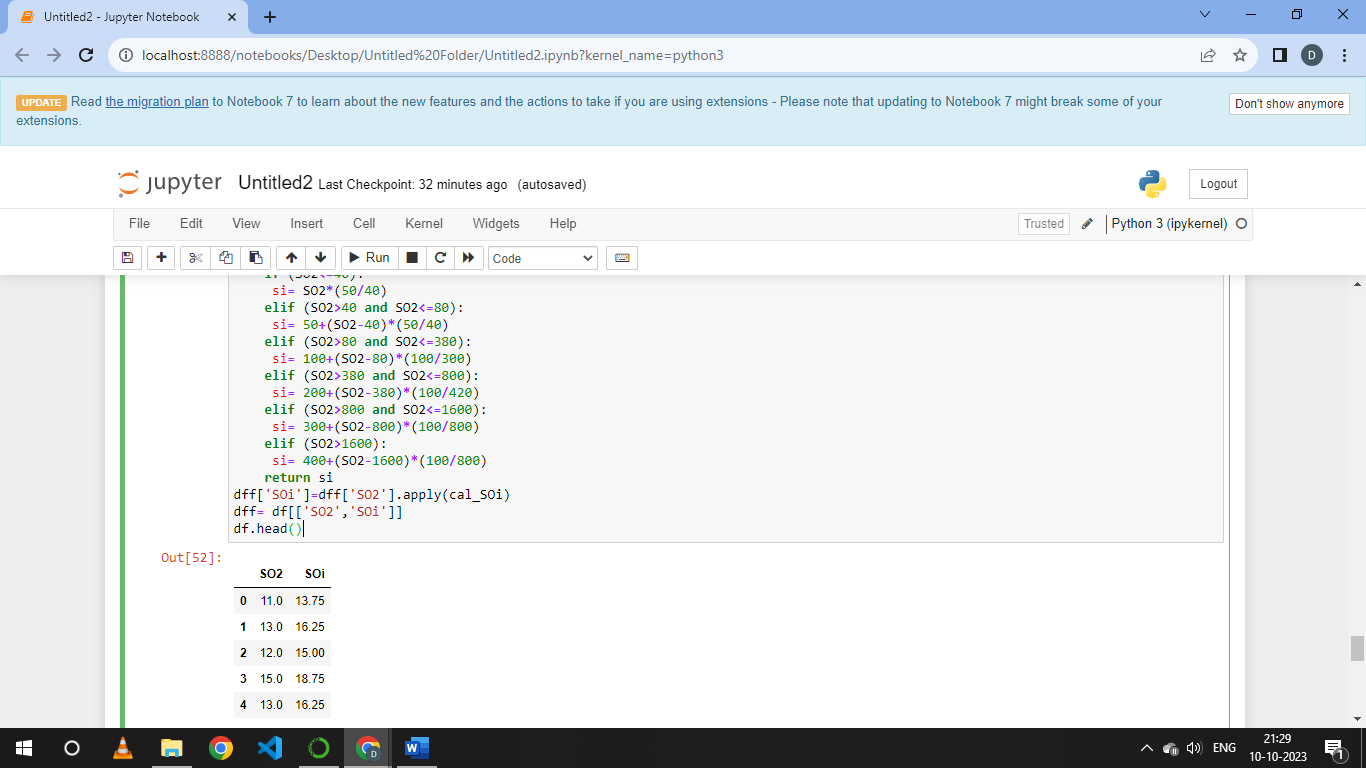
si= 400+(SO2-1600)\*(100/800)

return si

dff['SOi']=dff['SO2'].apply(cal\_SOi)

dff= df[['SO2','SOi']]

df.head()



**CONCLUSION :**  
 The air quality analysis module is essential for understanding the state of air pollution, its impact on public health, and the effectiveness of air quality management measures. Advancements in sensor technology, data analytics, and modelling techniques continue to enhance the accuracy and timeliness of air quality assessments, contributing to better-informed decisions for environmental and public health management.