**Project Report on**

**India Air Quality Data Analysis**

Submitted by:

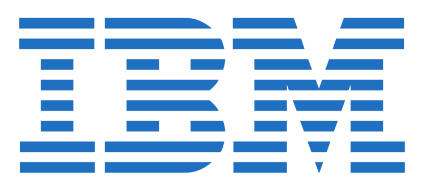
**Drishti**

**Abhishek Shukla**

**In partial fulfillment of completion of the course**

**Advanced Diploma in IT, Networking and Cloud Computing.**

**Under Guidance of:**



**Year 2022-2023**

# Abstract

Air pollution is a pressing environmental concern globally, and India faces significant challenges in maintaining air quality standards. This project presents a comprehensive analysis of air quality data in India, leveraging datasets obtained from authoritative sources. The primary objectives are to explore temporal and spatial patterns, identify key pollutants, and derive insights that can inform policy decisions.

The data preprocessing phase involves cleaning and organizing the dataset to ensure accuracy and reliability. Exploratory Data Analysis (EDA) techniques are applied to unveil trends in air quality over time, examining daily, monthly, and seasonal variations. Geospatial analysis further investigates regional disparities, offering a nuanced perspective on air quality distribution across different states and cities.

# Acknowledgement

At this juncture of our journey, we wish to express our heartfelt gratitude to all those who have contributed to the creation and success of **"India Air Quality Data Analysis".** This project has been a labor of passion and dedication, and it would not have been possible without the unwavering support and guidance we have received.

First and foremost, we offer our thanks to the boundless creativity and inspiration that flows from the universe. We are grateful for the opportunity to embark on this venture.

We extend our sincerest appreciation to our mentors, **Mrs. Mala Mishra & Ms. Ankita**

**Shukla**, whose wisdom and guidance have been instrumental in shaping the vision of **"India Air Quality Data Analysis".** Your support at every crucial turn has illuminated our path and fueled our determination to create a meaningful platform.

To our dedicated team of developers, designers, and content creators, we extend our deepest gratitude. Your tireless efforts, innovation, and creativity have breathed life into **"India Air Quality Data Analysis".** It is your collective dedication that has made this project a reality.

Our appreciation also goes to our colleagues and friends who provided invaluable insights and feedback during the development process. Your input has been instrumental in refining our ideas and enhancing the user experience.

We acknowledge the contributions of the broader IT community, whose open-source ethos has been a wellspring of knowledge and inspiration. The collaborative spirit of this community has been a guiding light.

Last but not least, we owe a debt of gratitude to our families and friends who have stood by us throughout this journey. Your unwavering support, encouragement, and belief in our vision have been our constant motivation.

**ADVANCE DIPLOMA IN IT NETWORKING**

**& CLOUD COMPUTING**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

The Advanced Diploma in IT Networking and Cloud Computing program offered by NSTI (W) Noida in collaboration with Edunet Foundation is a comprehensive course designed to equip students with advanced skills in information technology and cloud computing. This program covers a wide range of topics, including Computer Networking, Database Management, Virtualization, Cloud Technologies, and Cybersecurity. Students will gain hands-on experience through practical labs, workshops, and real-world projects, enabling them to excel in the rapidly evolving IT industry. Upon completion of the program, Graduates will have a strong foundation in both IT Fundamentals and Cloud Computing, making them highly sought-after professionals in the field.

**Project Requirements**

|  |  |
| --- | --- |
| **Project Name** | **India Air Quality Data Analysis** |
| **Languages Used** | **Python, Data wrangling and Data Visualisation tools** |
| **Editor** | **Jupyter Notebook, Google Colab** |
| **Web Browser** | **Google Chrome, Microsoft Edge** |

**Team Composition and Workload Division**

|  |  |
| --- | --- |
| Drishti | Data Analysis, Synopsis |
| Abhishek Shukla | Data Analysis |

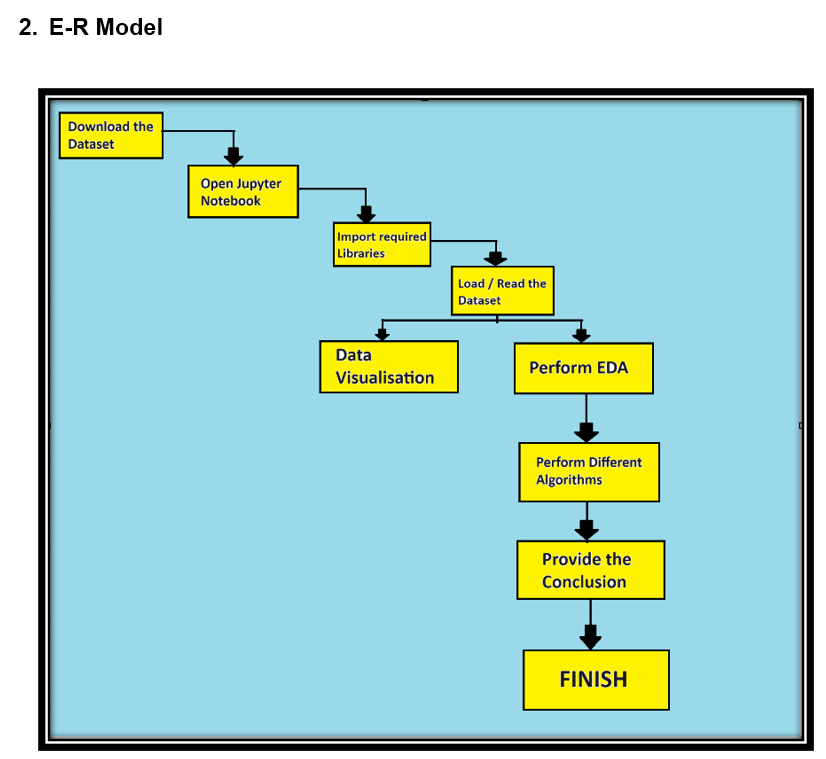
**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **SNO** | **TOPIC** | **Page No** |
| 1. | INTRODUCTION TO PROBLEM | 5 |
| 2. | ER MODEL | 5 |
| 3. | REQUIREMENTS | 5-6 |
| 4. | INTRODUCTION:  Background  Objective | 6 |
| 5. | DATA COLLECTION:  Data Source  Data Cleaning | 7-8 |
| 6. | EXPLORATORY DATA ANALYSIS [EDA] | 8-10 |
| 7. | OVERVIEW | 10 |
| 8. | PROJECT MODULE | 10-11 |
| 9. | SAMPLE SCREENSHOTS | 11-14 |
| 10. | SOURCE CODE | 14-27 |
| 11. | FUTURE SCOPE | 27 |
| 12 | CONCLUSION | 27 |
| 13 | REFERENCES | 28 |

# Introduction to Problem

Air pollution poses a critical threat to public health and the environment, and India, with its rapid industrialization and urbanization, faces substantial challenges in maintaining acceptable air quality standards. The problem at hand is the need for a comprehensive analysis of air quality data in India to understand the dynamics of pollution, identify key contributing factors, and inform evidence-based policies for mitigating the adverse effects of air pollution.

Rapid urbanization and industrialization have led to an alarming increase in pollution levels, adversely affecting air, water, and soil quality. The lack of comprehensive data analysis tools hinders our ability to understand the dynamic patterns and sources of pollution. This project aims to address this gap by conducting a thorough analysis of pollution data, identifying key contributors, and developing effective strategies for pollution control and mitigation. The goal is to provide actionable insights to policymakers and communities, fostering informed decision-making for a sustainable and healthier environment.



## 3.1 Technology Stack

**Python:** High-level programming language used for server-side scripting.

**Jupyter Notebook:** Jupyter Notebook is an open-source web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text, providing an interactive and collaborative environment for data science and analysis.

**3.2 Hardware**

Laptop/ Computer

## 3.3 Software

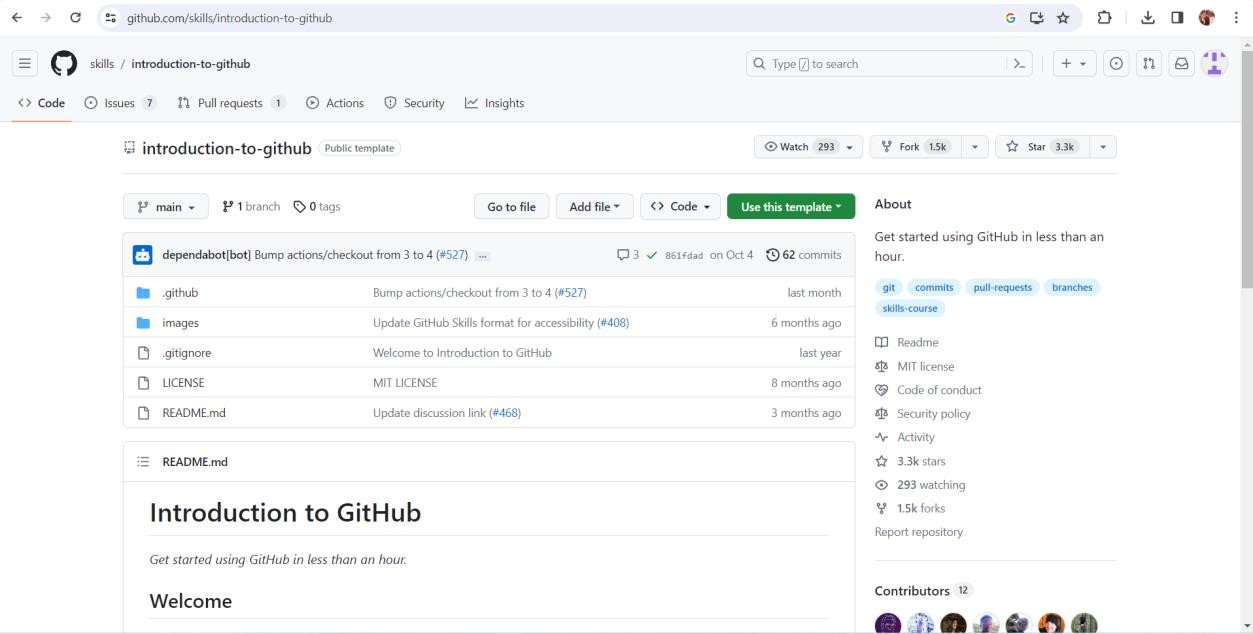
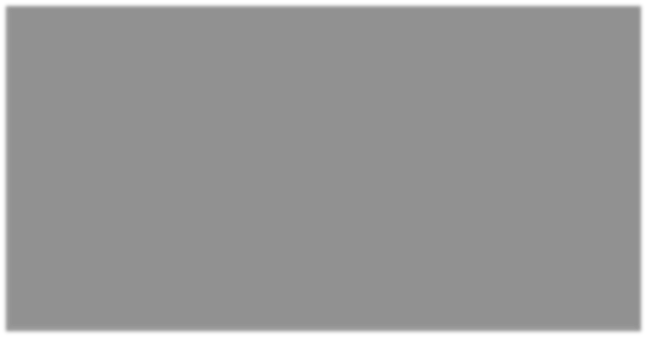
Operating System (OS)

Version Control System

Text Editors and Integrated Development Environments (IDEs)

## 3.4 Deployment Environment

**Github**



# 4. Introduction

## Air quality is a crucial aspect of environmental well-being, directly impacting the health of individuals and the sustainability of ecosystems. With India's rapid industrial growth, urbanization, and expanding population, the issue of air pollution has become a pressing concern. The quality of the air we breathe is intricately linked to public health, climate change, and the overall quality of life.

## This project delves into a detailed analysis of air quality in India, aiming to unravel the complexities surrounding this multifaceted challenge. The introduction sets the stage by providing context and highlighting the significance of understanding and addressing air pollution in the Indian context.

## 4.1 Background

## Clean and breathable air is a fundamental requirement for sustaining life on Earth. The quality of the air we breathe has profound implications for human health, environmental well-being, and the overall balance of ecosystems. As an essential component of our daily existence, air quality plays a pivotal role in shaping the quality of life for individuals and communities.

## 4.2 Objective

The primary goals of our data analysis project are to:

* Understand the Distribution of Content.
* Identify Trends.
* Explore User Preferences.

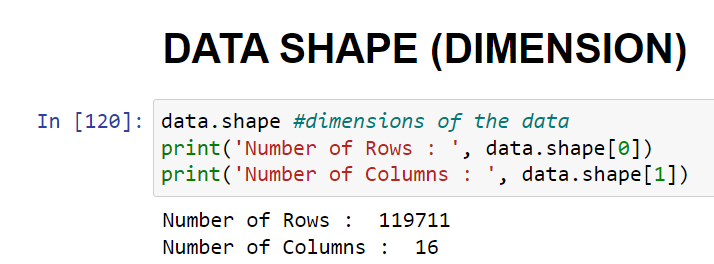
# 5. Data Collection

**5.1 Data Source**

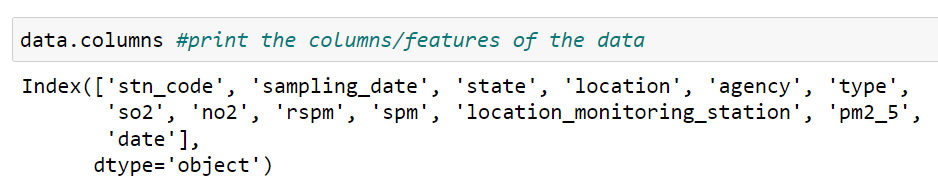
The India Air Quality dataset used in this analysis was sourced from **Kaggle**. When accessing air quality data, it's crucial to consider factors such as the type of pollutants measured, the frequency of data collection, the location of monitoring stations, and the reliability of the data.

**Dataset Structure:**

The dataset consists of [119711] rows and [16] columns.



Key variables include [list the essential variables, such as 'stn\_code' , 'sampling\_date, 'state,' 'location,' etc.].

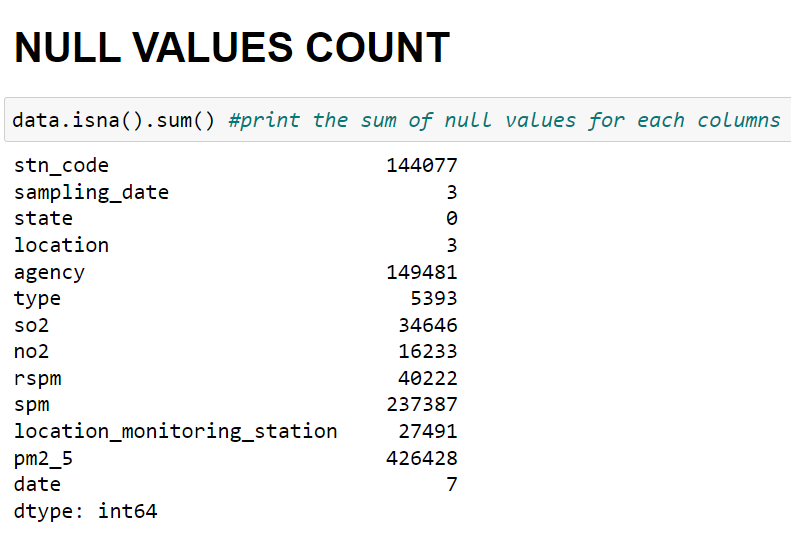


## 5.2 Data Cleaning

**Steps Taken:**

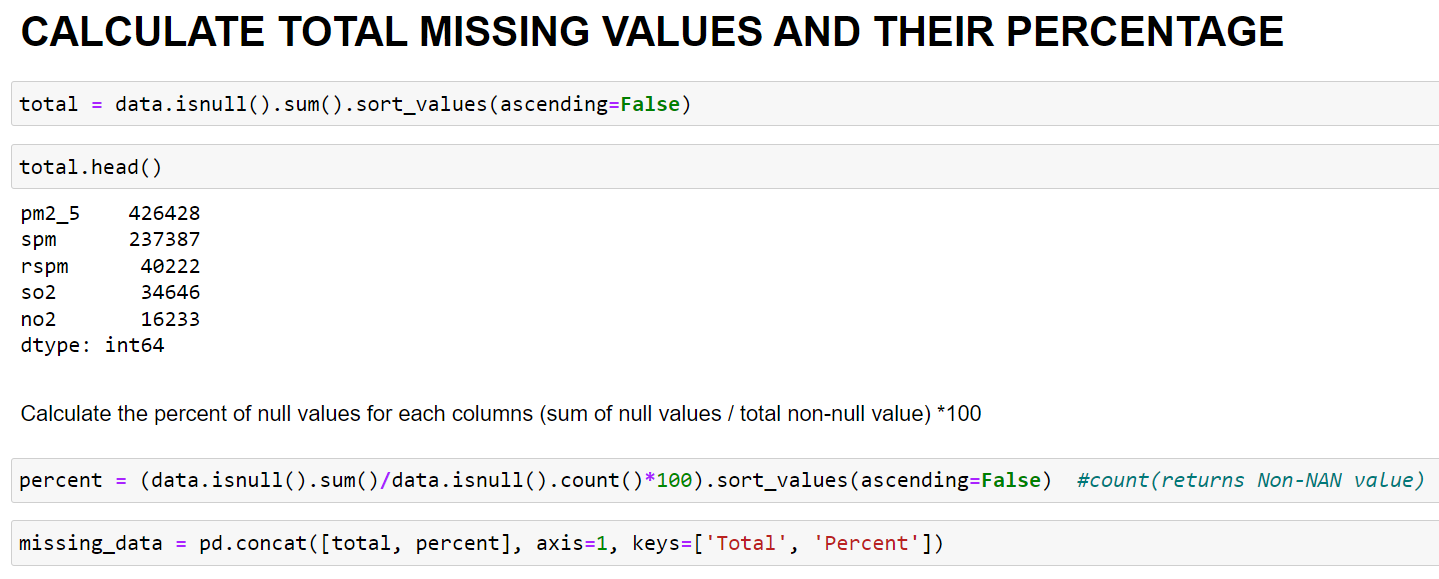
**Handling Missing Values:**

Identified and assessed missing values across variables.



**Duplicate Removal:**

Checked for and removed duplicate entries to ensure data integrity.

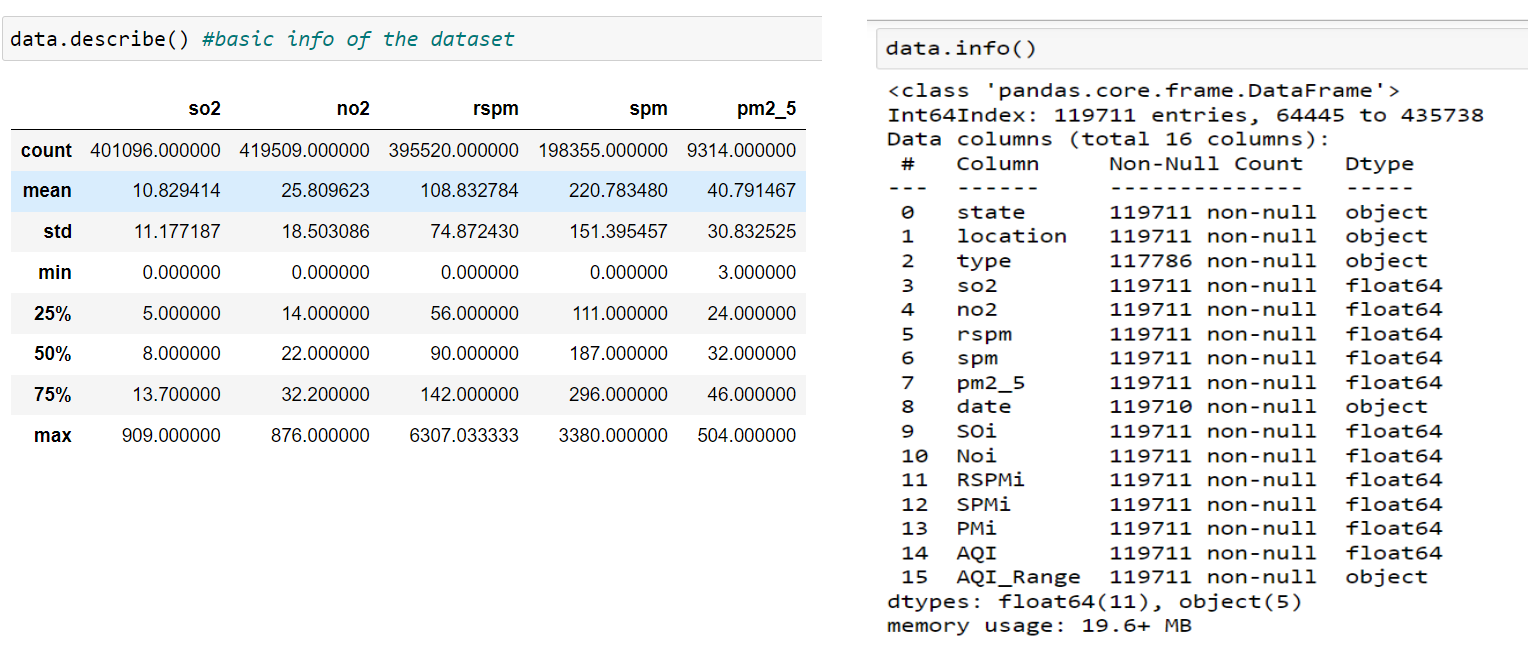


# 6. Exploratory Data Analysis (EDA)

The exploratory data analysis phase has provided foundational insights into the India Air Quality dataset.

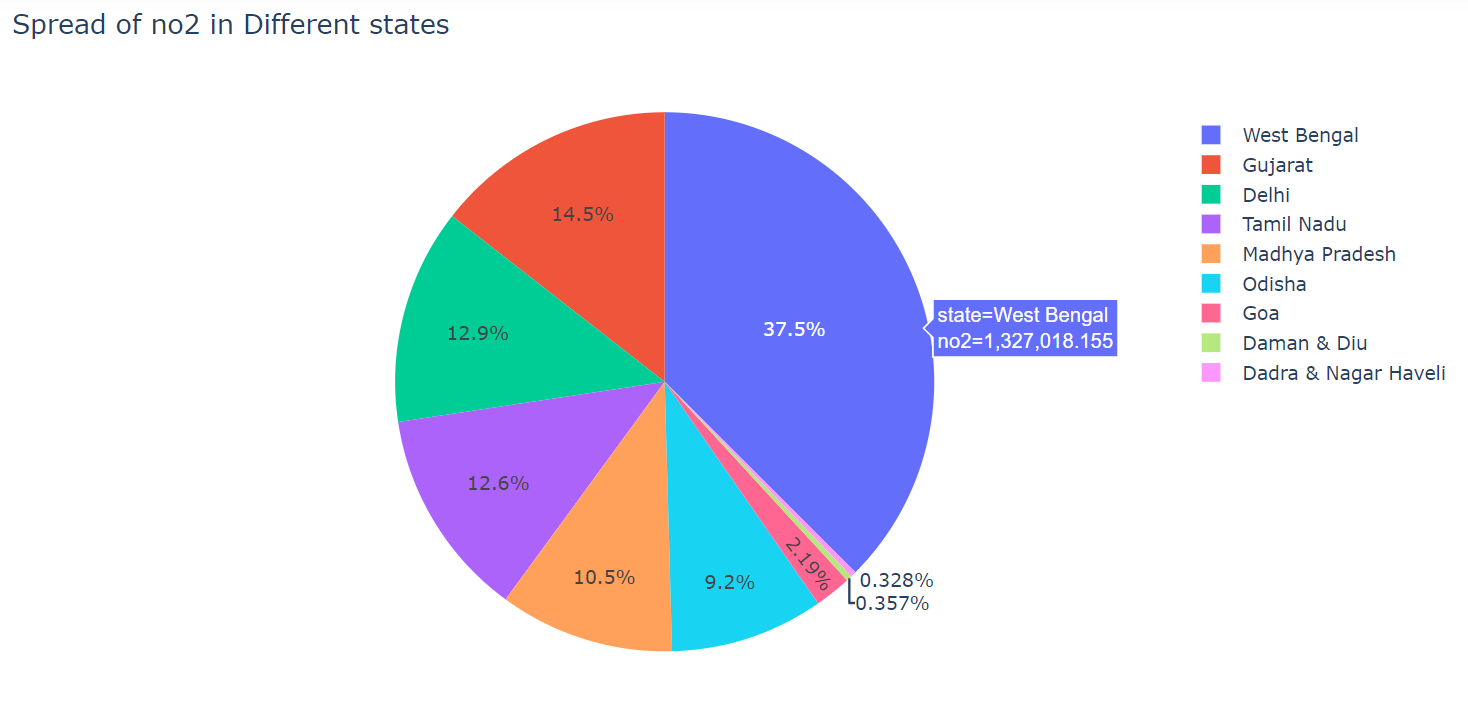
**6.1 Overview**

The dataset under consideration comprises [119711] records and [16] features, offering a comprehensive view of India Air Quality. Initial statistical analysis reveals [brief summary of key statistics, such as mean, median, and standard deviation], providing a foundation for further exploration.



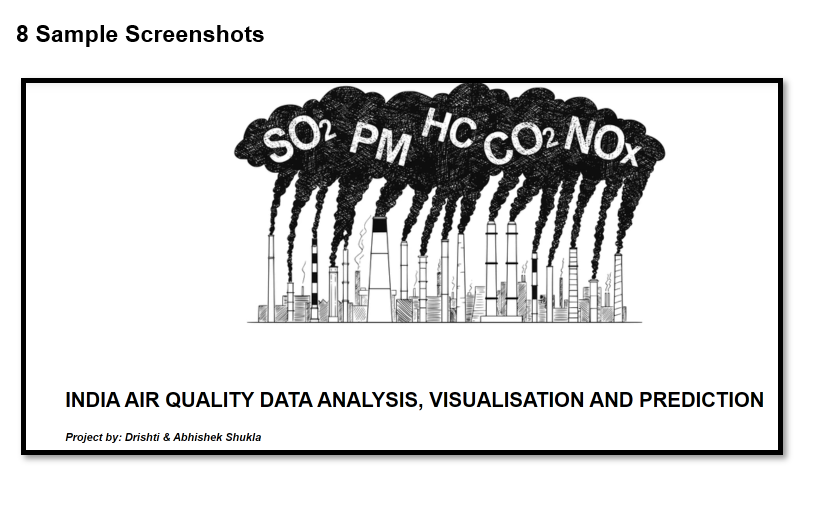
**6.2 Explore the effect of different gases on different states:**

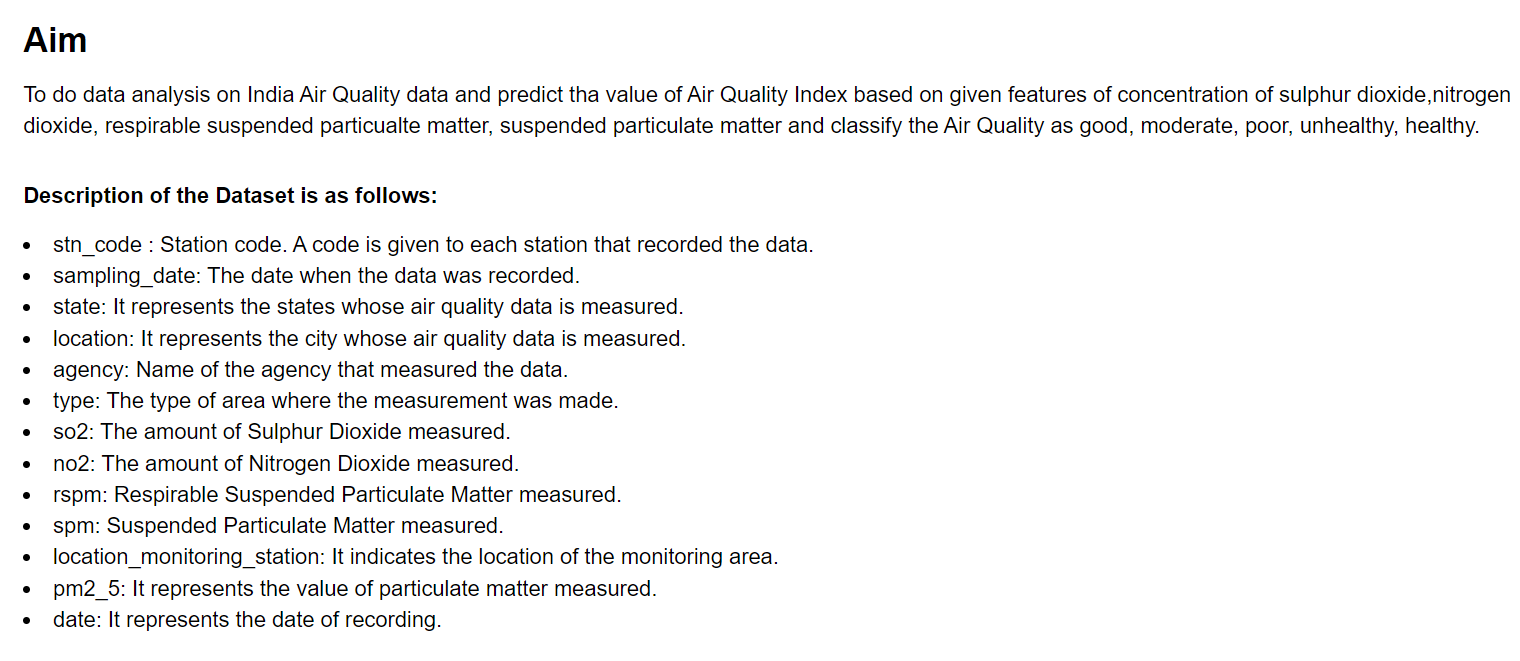
Visualizations, including pie charts, bar graphs, or heatmaps, highlight the harmful effect of pollution In different states.



# Project Module

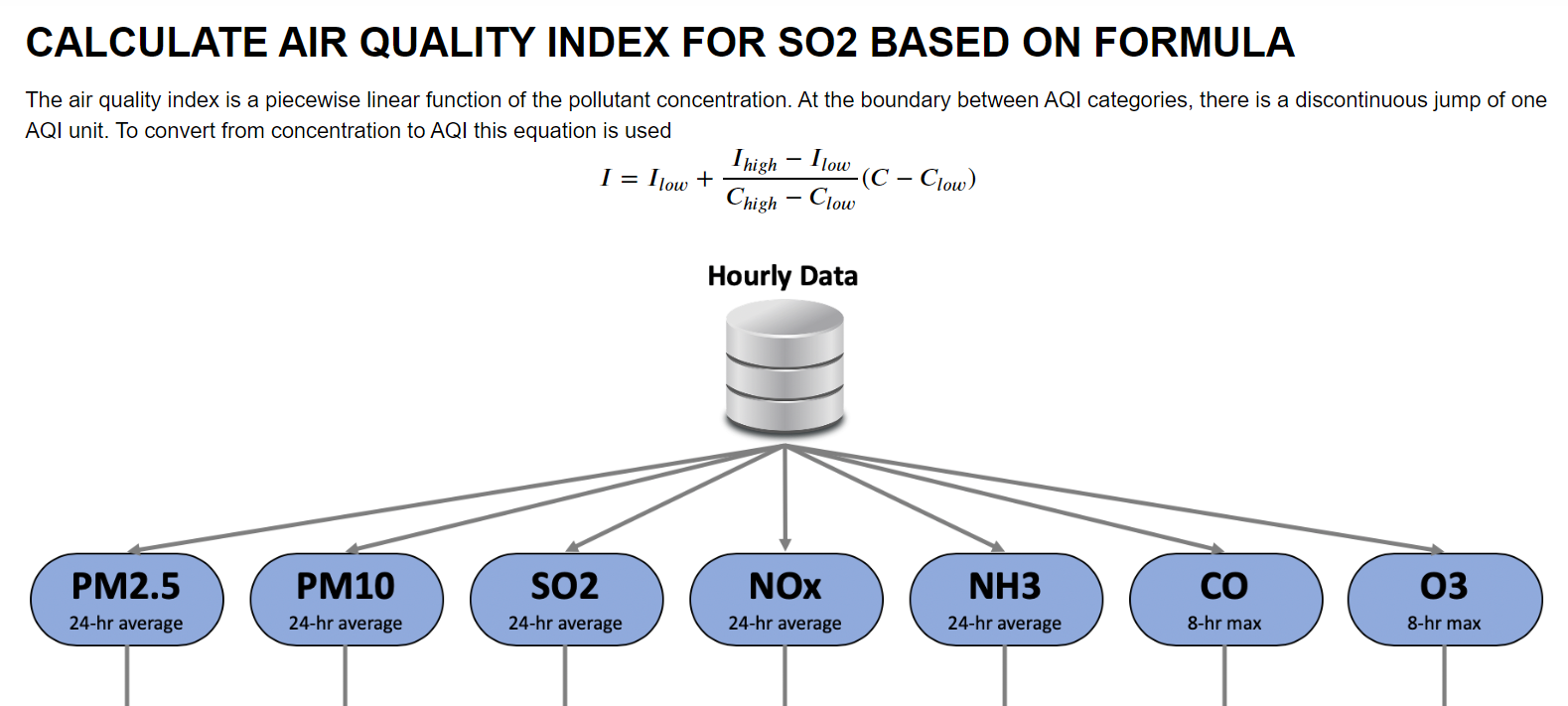
1. Import the required libraries.
2. Load/ Read the Dataset
3. Prepare EDA
4. Do Visualizations
5. Calculate Air Quality Index using the Formula
6. Prepare Heatmap
7. Make Predictions using different Models(Linear Regression, Logistic Regression, Random Forest, Decision Tree)
8. Prepare Profile Report











# 9. Source Code

#import the required libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import matplotlib

import seaborn as sns

import plotly.express as px

%matplotlib inline

import warnings

warnings.filterwarnings('ignore')

data = pd.read\_csv('Pollution.csv') #import data

data.columns #print the columns/features of the data

data.describe()

data.info()

data.shape #dimensions of the data

print('Number of Rows : ', data.shape[0])

print('Number of Columns : ', data.shape[1])

data.isna().sum() #print the sum of null values for each columns

data.drop(['stn\_code','agency','sampling\_date','location\_monitoring\_station'],axis=1,inplace=True)

total = data.isnull().sum().sort\_values(ascending=False)

percent = (data.isnull().sum()/data.isnull().count()\*100).sort\_values(ascending=False) #count(returns Non-NAN value)

missing\_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])

sns.barplot(x=missing\_data.index, y=missing\_data['Percent'])

plt.xlabel('Features', fontsize=20)

plt.ylabel('Percent of missing values', fontsize=20)

plt.title('Percent missing data by feature', fontsize=20)

plt.show()

plt.hist(data.spm,range=(0.0,4000)) #spm

plt.show()

data['rspm']=grp\_state['rspm'].transform(mean) #fill value with mean value group by state

data['so2']=grp\_state['so2'].transform(mean)

data['no2']=grp\_state['no2'].transform(mean)

data['spm']=grp\_state['spm'].transform(mean)

data['pm2\_5']=grp\_state['pm2\_5'].transform(mean)

#Creating a pie chart

piechart = px.pie(data, values='no2', names='state', title='Spread of no2 in Different states')

piechart.show()

#Scatter plots of all columns

sns.set()

cols = ['so2', 'no2', 'rspm', 'spm', 'pm2\_5']

sns.pairplot(data[cols], size = 2.5)

plt.show()

# #Correlation matrix

# corrmat = data.corr()

# f, ax = plt.subplots(figsize = (15, 10))

# sns.heatmap(corrmat, vmax = 1, square = True, annot = True)

# 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

data['SOi']=data['so2'].apply(cal\_SOi)

df= data[['so2','SOi']]

df.head()

def cal\_aqi(si,ni,rspmi,spmi):

aqi=0

if(si>ni and si>rspmi and si>spmi):

aqi=si

if(ni>si and ni>rspmi and ni>spmi ):

aqi=ni

if(rspmi>si and rspmi>ni and rspmi>spmi ):

aqi=rspmi

if(spmi>si and spmi>ni and spmi>rspmi):

aqi=spmi

return aqi

data['AQI']=data.apply(lambda x:cal\_aqi(x['SOi'],x['Noi'],x['RSPMi'],x['SPMi']),axis=1)

df= data[['state','SOi','Noi','RSPMi','SPMi','AQI']]

df.head()

def AQI\_Range(x):

if x<=50:

return "Good"

elif x>50 and x<=100:

return "Moderate"

elif x>100 and x<=200:

return "Poor"

elif x>200 and x<=300:

return "Unhealthy"

elif x>300 and x<=400:

return "Very unhealthy"

elif x>400:

return "Hazardous"

data['AQI\_Range'] = data['AQI'] .apply(AQI\_Range)

data.head()

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

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_log\_error

from sklearn.metrics import mean\_squared\_error

from sklearn.metrics import r2\_score, mean\_squared\_error

from sklearn.model\_selection import KFold

from sklearn.model\_selection import cross\_val\_score

prof = ProfileReport(data)

prof.to\_file(output\_file = 'output.html')

# 10. Future Scope

With the increasing volume and variety of data generated, the future will likely see a greater emphasis on big data analytics, exploring large datasets to extract meaningful patterns and insights.

As the need for instant insights grows, real-time data analysis will become more prominent, especially in industries like finance, healthcare, and IoT (Internet of Things).

# 11. Conclusion

In conclusion, this data analysis project has successfully unveiled valuable insights, revealing patterns and trends within the dataset. The systematic exploration of relationships between variables has provided a deeper understanding of the underlying dynamics. The findings offer a foundation for informed decision-making, guiding future strategies and actions. The project's use of advanced analytical tools and methodologies showcases the evolving landscape of data science. Moving forward, continuous advancements in machine learning, artificial intelligence, and big data analytics will shape the future of data analysis. Ethical considerations must remain at the forefront to ensure responsible data usage and unbiased results. Collaboration between data scientists and domain experts will further refine analyses for specific industries. The project highlights the importance of transparency and reproducibility in analytical workflows for fostering trust in results. As we embrace emerging technologies, the scope for data analysis remains expansive, promising innovative solutions to complex challenges across diverse domains.

## 12. References <https://www.kaggle.com/datasets>