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:

IBM

कौशलं बलम्

Directorate General of Training

edunet foundation

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

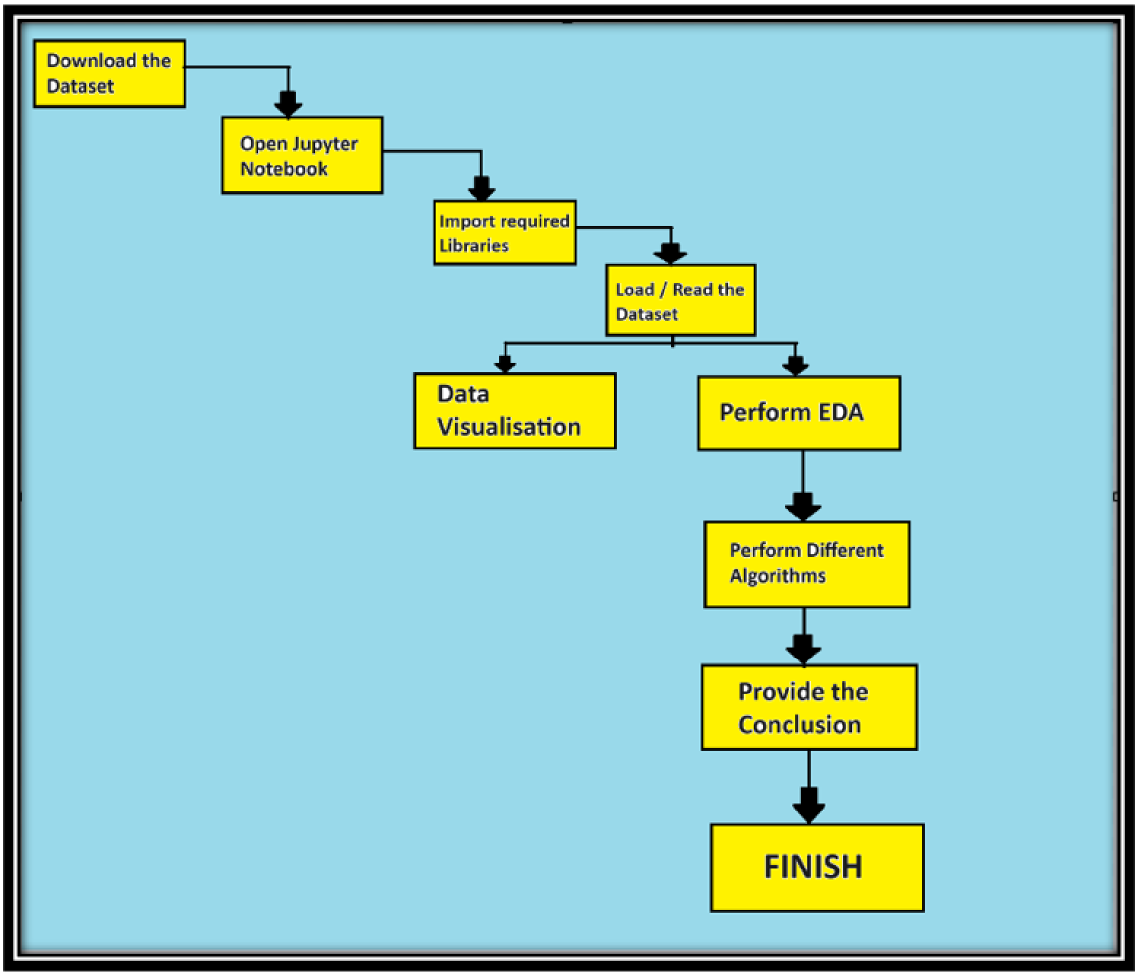
Table of Contents

1. 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.

2. E-R Model



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

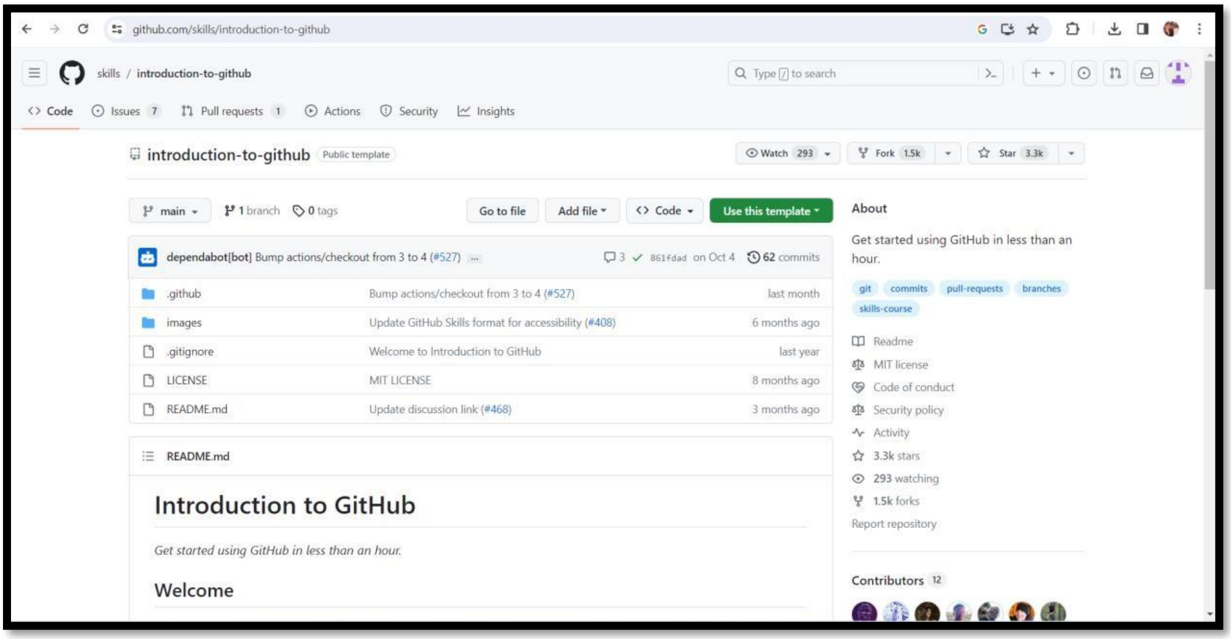
Operating System (OS)

Version Control System

Text Editors and Integrated Development Environments (IDEs)

3.4 Deployment Environment

Github



dependabot[bot] Bump actions/checkout from 3 to 4 (#527) ... 3 ✓ 861fdad on Oct 4 62 commits

.github	Bump actions/checkout from 3 to 4 (#527)	last month
images	Update GitHub Skills format for accessibility (#408)	6 months ago
.gitignore	Welcome to Introduction to GitHub	last year
LICENSE	MIT LICENSE	8 months ago
README.md	Update discussion link (#468)	3 months ago

README.md

Introduction to GitHub

Get started using GitHub in less than an hour.

Welcome

About

Get started using GitHub in less than an hour.

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.

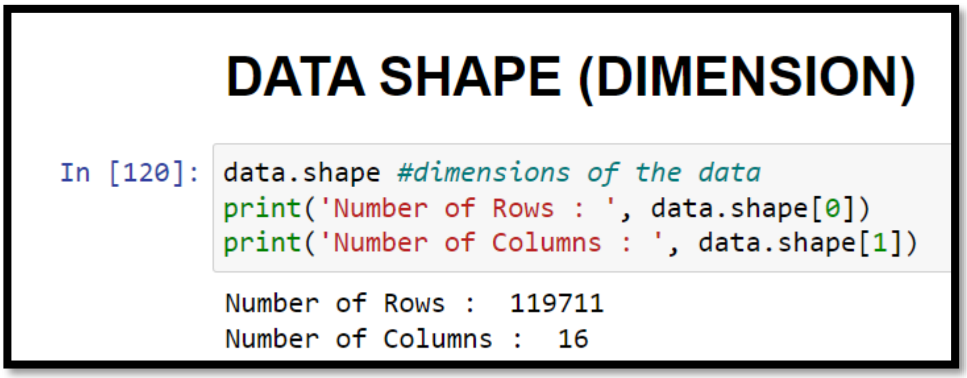
5. Data Collection

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:

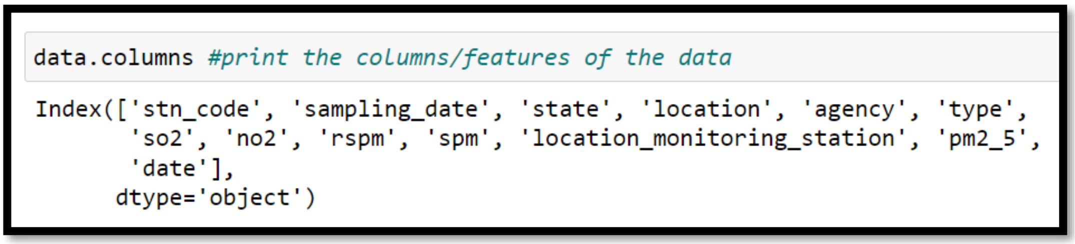
The dataset consists of [119711] rows and [16] columns.



DATA SHAPE (DIMENSION)

```
In [120]: data.shape #dimensions of the data
          print('Number of Rows : ', data.shape[0])
          print('Number of Columns : ', data.shape[1])

          Number of Rows :  119711
          Number of Columns :  16
```

Key variables include [list the essential variables, such as 'stn_code' , 'sampling_date, 'state,' 'location,' etc.].



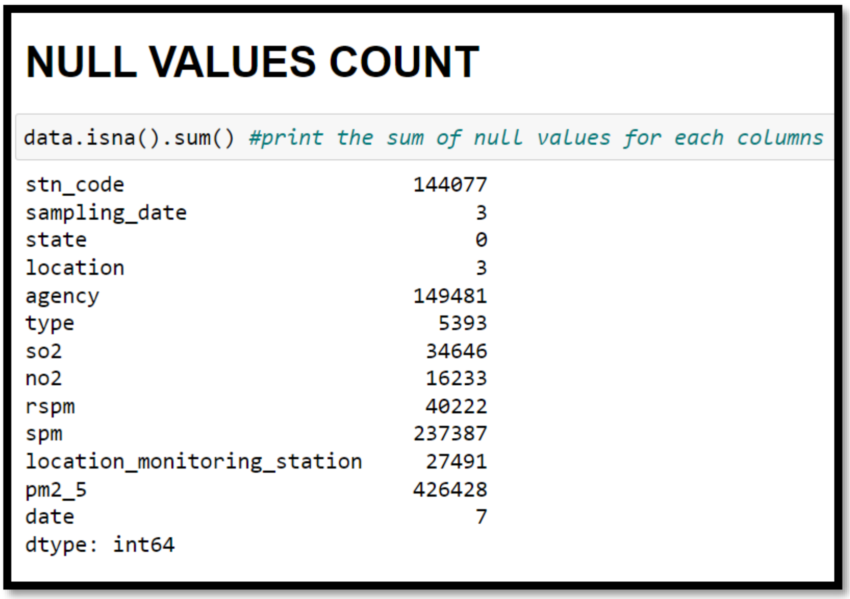
```
data.columns #print the columns/features of the data

Index(['stn_code', 'sampling_date', 'state', 'location', 'agency', 'type',
       'so2', 'no2', 'rspm', 'spm', 'location_monitoring_station', 'pm2_5',
       'date'],
      dtype='object')
```

5.2 Data Cleaning

Steps Taken:

Handling Missing Values:

Identified and assessed missing values across variables.

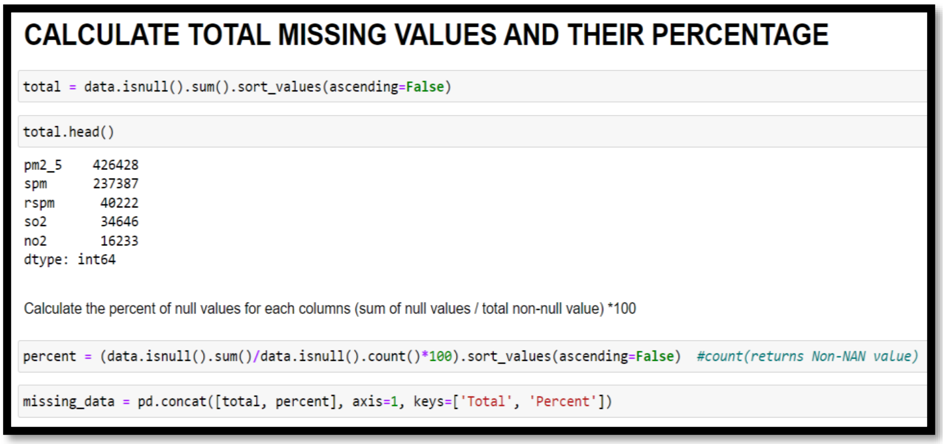


NULL VALUES COUNT

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

stn_code                       144077
sampling_date                       3
state                               0
location                            3
agency                         149481
type                             5393
so2                             34646
no2                             16233
rspm                            40222
spm                            237387
location_monitoring_station     27491
pm2_5                          426428
date                                7
dtype: int64
```

Duplicate Removal:

Checked for and removed duplicate entries to ensure data integrity.



CALCULATE TOTAL MISSING VALUES AND THEIR PERCENTAGE

```
total = data.isnull().sum().sort_values(ascending=False)
```

```
total.head()
```

```
pm2_5    426428
spm      237387
rspm      40222
so2       34646
no2       16233
dtype: int64
```

Calculate the percent of null values for each columns (sum of null values / total non-null value) *100

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

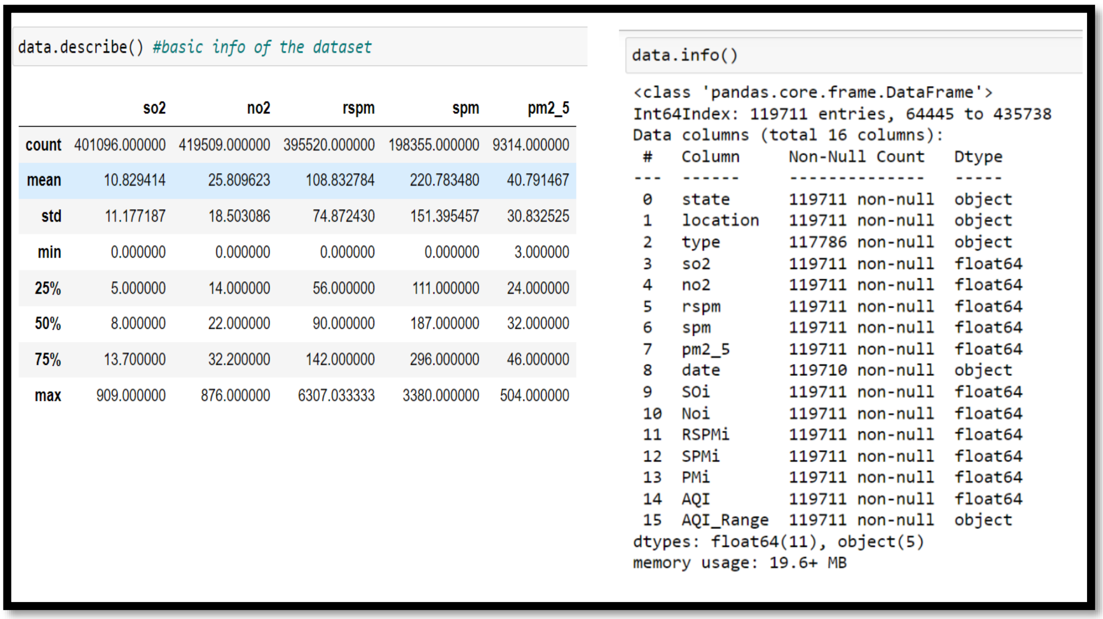
```
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
```

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.



```
data.describe() #basic info of the dataset
```

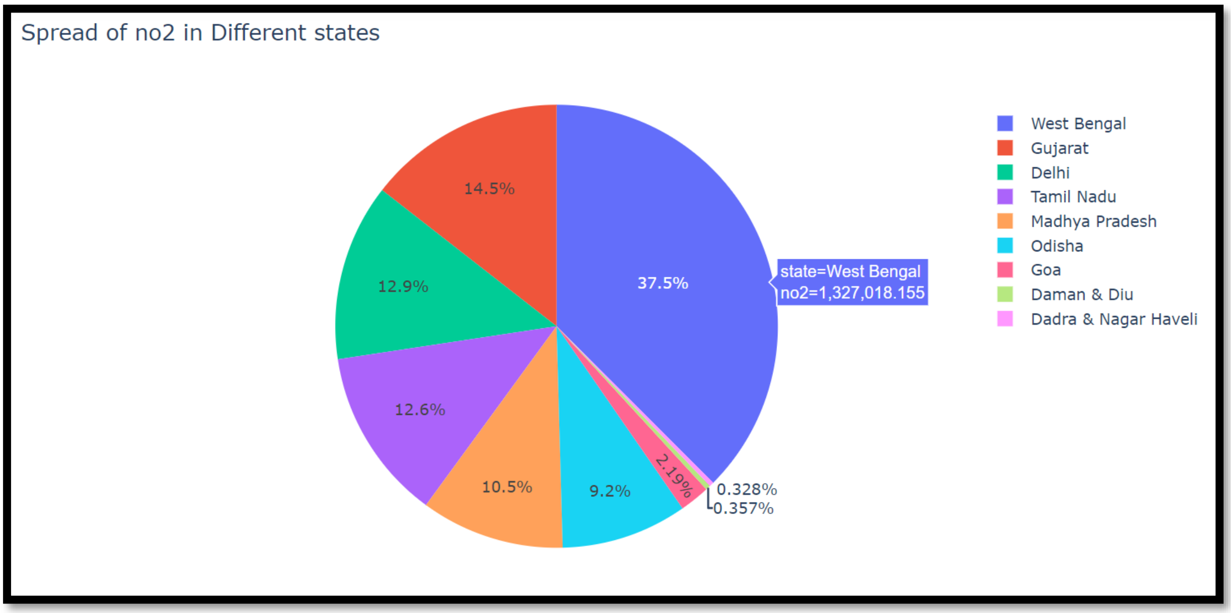
	so2	no2	rspm	spm	pm2_5
count	401096.000000	419509.000000	395520.000000	198355.000000	9314.000000
mean	10.829414	25.809623	108.832784	220.783480	40.791467
std	11.177187	18.503086	74.872430	151.395457	30.832525
min	0.000000	0.000000	0.000000	0.000000	3.000000
25%	5.000000	14.000000	56.000000	111.000000	24.000000
50%	8.000000	22.000000	90.000000	187.000000	32.000000
75%	13.700000	32.200000	142.000000	296.000000	46.000000
max	909.000000	876.000000	6307.033333	3380.000000	504.000000

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 119711 entries, 64445 to 435738
Data columns (total 16 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   state      119711 non-null  object
 1   location   119711 non-null  object
 2   type       117786 non-null  object
 3   so2        119711 non-null  float64
 4   no2        119711 non-null  float64
 5   rspm       119711 non-null  float64
 6   spm        119711 non-null  float64
 7   pm2_5      119711 non-null  float64
 8   date       119710 non-null  object
 9   SOi        119711 non-null  float64
 10  Noi        119711 non-null  float64
 11  RSPMi      119711 non-null  float64
 12  SPMi       119711 non-null  float64
 13  PMi        119711 non-null  float64
 14  AQI        119711 non-null  float64
 15  AQI_Range  119711 non-null  object
dtypes: float64(11), object(5)
memory usage: 19.6+ MB
```

6.2 Explore the effect of different gases on different states:

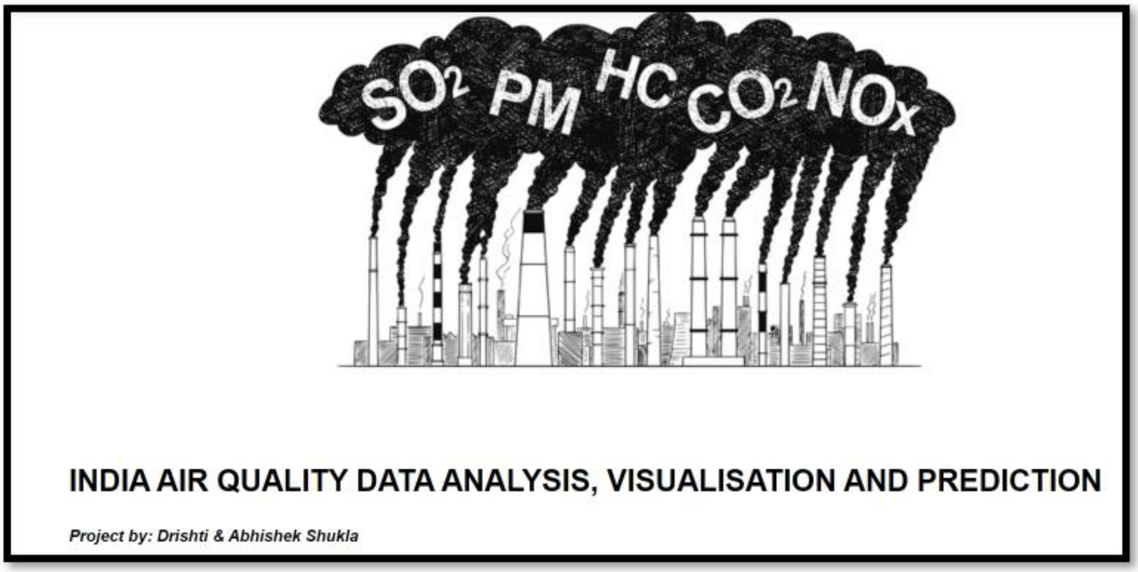
Visualizations, including pie charts, bar graphs, or heatmaps, highlight the harmful effect of pollution In different states.



Project Module

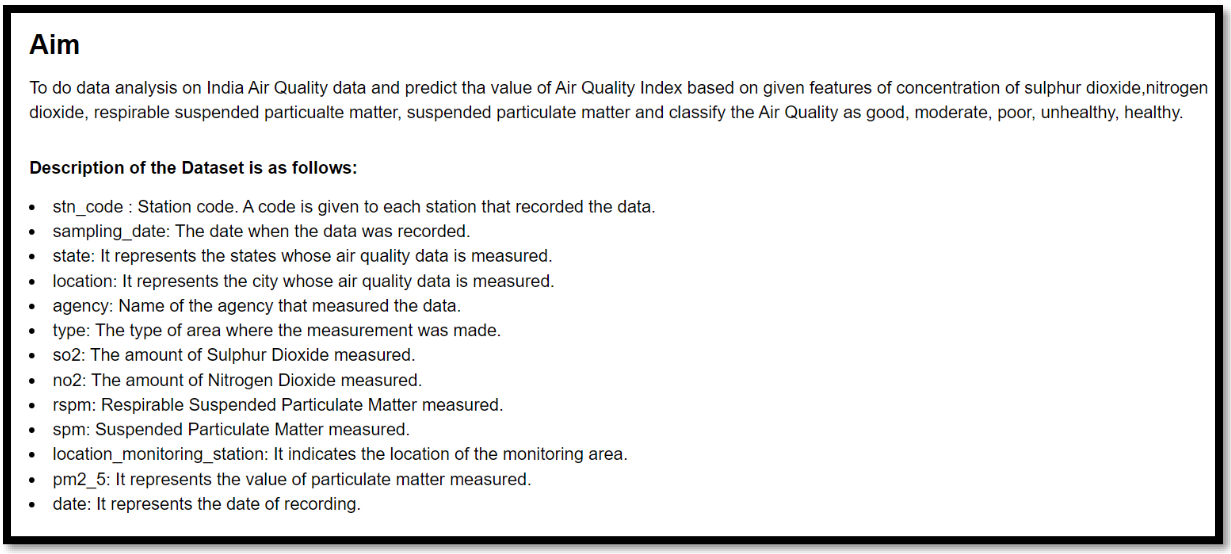
1. Import the required libraries.
2. Load/ Read the Dataset
3. Prepare EDA
4. Do Visualizations
5. Calculate Air Quality Index using the Formula
6. Prepare Heatmap
7. Make Predictions using different Models(Linear Regression, Logistic Regression, Random Forest, Decision Tree)
8. Prepare Profile Report

8 Sample Screenshots



INDIA AIR QUALITY DATA ANALYSIS, VISUALISATION AND PREDICTION

Project by: Drishti & Abhishek Shukla

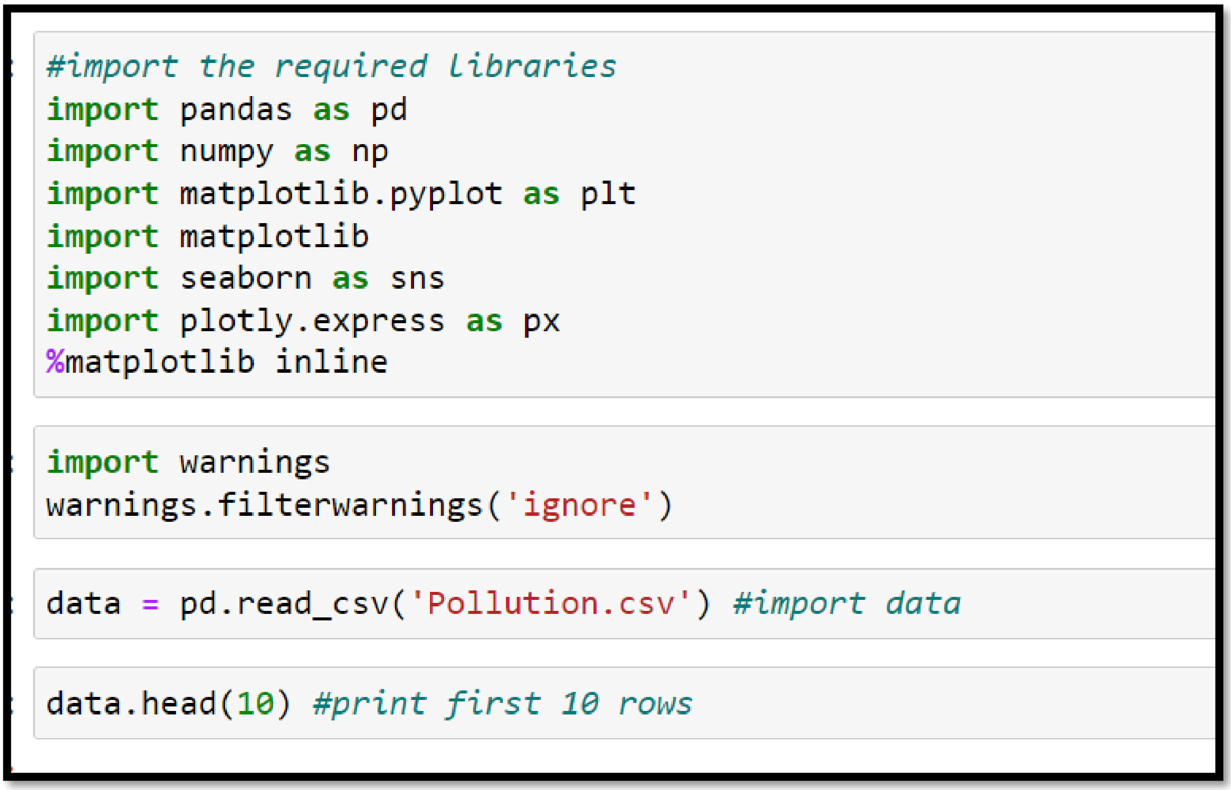


Aim

To do data analysis on India Air Quality data and predict tha value of Air Quality Index based on given features of concentration of sulphur dioxide,nitrogen dioxide, respirable suspended particualte matter, suspended particulate matter and classify the Air Quality as good, moderate, poor, unhealthy, healthy.

Description of the Dataset is as follows:

- stn_code : Station code. A code is given to each station that recorded the data.
- sampling_date: The date when the data was recorded.
- state: It represents the states whose air quality data is measured.
- location: It represents the city whose air quality data is measured.
- agency: Name of the agency that measured the data.
- type: The type of area where the measurement was made.
- so2: The amount of Sulphur Dioxide measured.
- no2: The amount of Nitrogen Dioxide measured.
- rspm: Respirable Suspended Particulate Matter measured.
- spm: Suspended Particulate Matter measured.
- location_monitoring_station: It indicates the location of the monitoring area.
- pm2_5: It represents the value of particulate matter measured.
- date: It represents the date of recording.



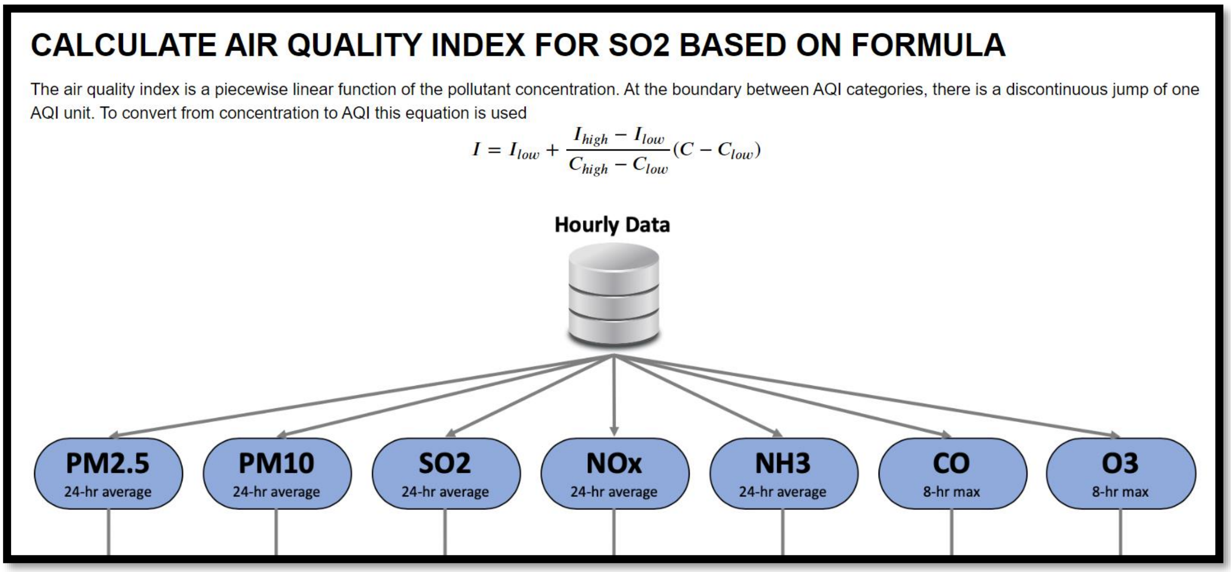
```
#import the required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib
import seaborn as sns
import plotly.express as px
%matplotlib inline
```

```
import warnings
warnings.filterwarnings('ignore')
```

```
data = pd.read_csv('Pollution.csv') #import data
```

```
data.head(10) #print first 10 rows
```

Visualization for states with highest pollutants



CALCULATE AIR QUALITY INDEX FOR SO2 BASED ON FORMULA

The air quality index is a piecewise linear function of the pollutant concentration. At the boundary between AQI categories, there is a discontinuous jump of one AQI unit. To convert from concentration to AQI this equation is used

$$I = I_{low} + \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - C_{low})$$

Hourly Data

PM2.5
24-hr average

PM10
24-hr average

SO2
24-hr average

NOx
24-hr average

NH3
24-hr average

CO
8-hr max

O3
8-hr max

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