A

Course End Project Report on

Global Air Pollution Analysis

Is submitted in partial fulfillment of the Requirements for the Award of CIE of

DATA ANALYSIS AND VISUALIZATION-22ADE01

in

B.E, IV-SEM, INFORMATION TECHNOLOGY

Submitted by

A.Deepshika – 160122737004
G.Vyshnavi – 160122737005
M.Kavyasri - 160122737013

COURSE TAUGHT BY:

Dr Ramakrishna Kolikipogu Professor, Dept of IT.

DEPARTMENT OF INFORMATION TECHNOLOGY CHAITANYA

BHARATHI INSTITUTE OF TECHNOLOGY(A)

(Affiliated to Osmania University; Accredited by NBA, NAAC, ISO)

Kokapet (V), GANDIPET (M), HYDERABAD-500075

Website: www.cbit.ac.in

2023-2024

CERTIFICATE

This is to certify that the course end project work entitled **"Global Air Pollution Analysis"** is submitted by **A.Deepshika (160122737004),G.Vyshnavi (160122737005), and M.Kavyasri (160122737013)** in partial fulfillment of the requirements for the award of CIE Marks of **DATA ANALYSIS AND VISUALIZATION (22ADE01)** of **B.E, IV-SEM, INFORMATION TECHNOLOGY** to CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY(A) affiliated to OSMANIA UNIVERSITY, Hyderabad is a record of bonafide work carried out by them under my supervision and guidance. The results embodied in this report have not been submitted to any other University or Institute for the award of any other Degree or Diploma.

Signature of Course Faculty
Dr Ramakrishna Kolikipogu Professor of IT

Acknowledgement

The satisfaction that accompanies the successful completion of the task would be put incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success.

We wish to express our deep sense of gratitude to **Dr.Ramakrishna Kolikipogu, Professor of IT** for his able guidance and useful suggestions, which helped us in completing the Course End Project in time.

We are particularly thankful to **HoD, Principal and Management**, for their support and encouragement, which helped us to mould our project into a successful one.

We also thank all the staff members of IT Department for their valuable support and generous advice. Finally thanks to all our friends and family members for their continuous support and enthusiastic help.

A.Deepshika – 160122737004
G.Vyshnavi – 160122737005
M.Kavyasri - 160122737013

Abstract

Air pollution is a pervasive environmental issue with significant implications for human health, ecosystems, and the global climate. This paper explores the multifaceted nature of global air pollution, focusing on its sources, impacts, and mitigation strategies. Major pollutants, including particulate matter (PM), nitrogen oxides (NOx), sulphur dioxide (SO2), carbon monoxide (CO), and volatile organic compounds (VOCs), are examined in terms of their origins from industrial activities, transportation, agriculture, and natural phenomena. The adverse effects of these pollutants on respiratory and cardiovascular health, as well as their contribution to chronic diseases and mortality rates, are highlighted. Additionally, the ecological consequences, such as acid rain, eutrophication, and biodiversity loss, are discussed. The paper also addresses the role of air pollution in climate change, particularly through greenhouse gases like carbon dioxide (CO2) and methane (CH4).

The global distribution of air pollution reveals stark disparities, with developing regions often experiencing higher levels due to rapid industrialization and less stringent environmental regulations. International efforts and policies aimed at reducing air pollution, including the Paris Agreement, national regulations, and technological advancements in emission reduction, are critically analyzed. The effectiveness of these measures is assessed, along with the challenges of enforcement and the need for global cooperation.

Keywords: Pandas, NumPy, Matplotlib, Visualization, Analyzation, Illustrated Charts.

Table of Contents

List of Figures

CHAPTER 1

Introduction

Global air pollution analysis examines the presence and impact of harmful substances in the Earth's atmosphere. This field involves monitoring pollutants, such as particulate matter, nitrogen oxides, and ozone, to understand their sources, distribution, and effects on health and the environment. By analyzing air quality data and trends, researchers and policymakers can develop strategies to mitigate pollution, implement effective regulations, and promote sustainable practices. Addressing global air pollution is essential for protecting human health, preserving ecosystems, and combating climate change, requiring a collaborative effort across nations and disciplines.

Air pollution refers to the presence of harmful substances in the atmosphere. These substances, known as pollutants, can be gases, particulates, or biological molecules that pose risks to human health, ecosystems, and the climate. The main types of air pollutants include:

Particulate Matter (PM10 and PM2.5): Tiny particles that can penetrate the respiratory system.

Nitrogen Oxides (NOx): Gases produced from burning fuels, which contribute to smog and acid rain.

Sulfur Dioxide (SO2): Produced from burning fossil fuels and industrial processes, contributing to acid rain.

Carbon Monoxide (CO): A colorless, odorless gas resulting from incomplete combustion.

Ozone (O3): A gas that forms in the atmosphere from reactions between sunlight and pollutants such as volatile organic compounds (VOCs) and NOx.

Volatile Organic Compounds (VOCs): Organic chemicals that easily vaporize and can cause health effects and contribute to ozone formation.

Sources of Air Pollution

Air pollution originates from various sources:

Natural Sources: Volcanic eruptions, wildfires, dust storms, and biogenic emissions (e.g., trees emitting VOCs).

Anthropogenic Sources: Human activities such as industrial processes, vehicle emissions, power plants, agriculture, and deforestation.

Impact of Air Pollution

The impacts of air pollution are vast and multifaceted:

Human Health: Causes respiratory diseases, cardiovascular diseases, and can lead to premature death.

Environmental Effects: Acid rain, reduced visibility (haze), and harm to wildlife and vegetation.

Climate Change: Some pollutants, like black carbon and methane, contribute to global warming.

Global Air Pollution Trends

Global air pollution trends show varying patterns depending on the region:

Developed Countries: Often have stricter regulations and better technology, leading to improvements in air quality.

Developing Countries: Face challenges with increasing industrialization, urbanization, and less stringent regulations, often resulting in worsening air quality.

1.1 Definition of Problem

Global Air Pollution refers to the contamination of the Earth's atmosphere by harmful substances, which poses serious risks to human health, ecosystems, and the climate. Analyzing global air pollution involves understanding the sources, distribution, impacts, and mitigation strategies related to these pollutants on a worldwide scale.

The analysis of global air pollution involves the comprehensive examination and evaluation of various pollutants present in the Earth's atmosphere on a worldwide scale. This includes the measurement, monitoring, and assessment of pollutants such as particulate matter, nitrogen oxides, sulfur dioxide, carbon monoxide, volatile organic compounds, and greenhouse gases emitted from natural and anthropogenic sources. The analysis aims to understand the spatial and temporal distribution of air pollutants, their sources, transport mechanisms, and the impacts on human health, ecosystems, climate, and the environment as a whole. It involves interdisciplinary approaches integrating atmospheric science, environmental monitoring, modeling, and policy analysis to develop strategies for mitigating air pollution and its adverse effects on global scales. Additionally, it encompasses the examination of international agreements, regulations, and initiatives aimed at addressing air quality issues collaboratively among nations to achieve sustainable development goals and ensure clean and healthy air for present and future generations.

1.2 Objectives

The objectives of global air pollution analysis are multifaceted, aiming to address various aspects of air quality management and environmental protection on a worldwide scale. Some key objectives include:

1. **Assessment of Air Quality:** To comprehensively evaluate the levels of various pollutants in the Earth's atmosphere across different regions, including urban, rural, and remote areas, to understand the extent of air pollution and its spatial distribution.

2. **Identification of Sources:** To identify and quantify the sources of air pollutants, including both natural sources such as wildfires and volcanic eruptions, and anthropogenic sources like industrial activities, transportation, agriculture, and energy production.

3. **Assessment of Health and Environmental Impacts:** To evaluate the impacts of air pollution on human health, ecosystems, biodiversity, climate change, and environmental quality, including the occurrence of respiratory diseases, cardiovascular problems, premature mortality, ecosystem degradation, and climate feedbacks.

4. **Development of Mitigation Strategies:** To develop and implement effective strategies and policies for mitigating air pollution and reducing emissions of harmful pollutants, including technological innovations, regulatory measures, emission controls, and sustainable development practices.

5. **Public Awareness and Education:** To raise public awareness about the health and environmental impacts of air pollution, promote understanding of the causes and consequences of poor air quality, and encourage individual and collective actions to reduce emissions and improve air quality.

CHAPTER 2
Methodology

In this section, we outline the methodologies employed in our research to achieve the objectives outlined in the preceding sections. Our approach encompasses a combination of quantitative and qualitative methods, each tailored to address specific aspects of the research inquiry. We adopt a multi- faceted methodology to comprehensively explore the research problem, gather relevant data, and derive meaningful insights. The methodologies utilized in this study include:

1. **Data collection and DATASET description.**
2. **Data Analysis Methodology Overview.**

2.1 Data collection and Dataset description

The dataset used in this project was sourced from the popular data science platform Kaggle, under the title " Global air pollution analysis". Kaggle provides a huge range of datasets for research and analysis purposes, and the "Global air pollution analysis, serves as valuable resource for exploring trends in the pollution of the countries.

Dataset Description:

The dataset contains structured information gathered from various sources, representing a diverse range of attributes and phenomena. It offers insights into [specific domain or topic], featuring variables such as [list some key variables or attributes]. With entries spanning [timeframe or geographic area], it provides a comprehensive view of [subject area], suitable for analysis, modeling, and research purposes. This dataset serves as a valuable resource for [potential users or stakeholders], enabling informed decision-making and deeper understanding of [relevant domain]."

	Country	City	AQI Value	AQI Categ	CO AQI Va	CO AQI Ca	Ozone AQI	Ozone AQI	NO2 AQI V	NO2 AQI C	PM2.5 AQI	PM2.5 AQI Category
2	Russian Fe	Praskovey	51	Moderate	1	Good	36	Good	0	Good	51	Moderate
3	Brazil	Presidente	41	Good	1	Good	5	Good	1	Good	41	Good
4	Italy	Priolo Gar	66	Moderate	1	Good	39	Good	2	Good	66	Moderate
5	Poland	Przasnysz	34	Good	1	Good	34	Good	0	Good	20	Good
6	France	Punaauia	22	Good	0	Good	22	Good	0	Good	6	Good
7	United Sta	Punta Gor	54	Moderate	1	Good	14	Good	11	Good	54	Moderate
8	Germany	Puttlingen	62	Moderate	1	Good	35	Good	3	Good	62	Moderate
9	Belgium	Puurs	64	Moderate	1	Good	29	Good	7	Good	64	Moderate
10	Russian Fe	Pyatigorsk	54	Moderate	1	Good	41	Good	1	Good	54	Moderate
11	Egypt	Qalyub	142	Unhealthy	3	Good	89	Moderate	9	Good	142	Unhealthy for Sensitive Groups
12	China	Qinzhou	68	Moderate	2	Good	68	Moderate	1	Good	58	Moderate
13	Netherlan	Raalte	41	Good	1	Good	24	Good	6	Good	41	Good
14	India	Radaur	158	Unhealthy	3	Good	139	Unhealthy	1	Good	158	Unhealthy
15	Pakistan	Radhan	158	Unhealthy	1	Good	50	Good	1	Good	158	Unhealthy
16	Republic o	Radovis	83	Moderate	1	Good	46	Good	0	Good	83	Moderate
17	France	Raismes	59	Moderate	1	Good	30	Good	4	Good	59	Moderate
18	India	Rajgir	154	Unhealthy	3	Good	100	Unhealthy	2	Good	154	Unhealthy
19	Italy	Ramacca	55	Moderate	1	Good	47	Good	0	Good	55	Moderate
20	United Sta	Phoenix	72	Moderate	1	Good	4	Good	23	Good	72	Moderate
21	India	Phulabani	161	Unhealthy	2	Good	71	Moderate	0	Good	161	Unhealthy
22	Poland	Piaseczno	28	Good	1	Good	28	Good	2	Good	28	Good
23	India	Pimpri	118	Unhealthy	2	Good	30	Good	2	Good	118	Unhealthy for Sensitive Groups
24	Brazil	Pindobacu	33	Good	0	Good	10	Good	1	Good	33	Good
25	China	Pingyin	150	Unhealthy	3	Good	95	Moderate	6	Good	150	Unhealthy
26	Brazil	Pinheiral	154	Unhealthy	5	Good	0	Good	13	Good	154	Unhealthy
27	India	Piravam	81	Moderate	1	Good	24	Good	1	Good	81	Moderate
28	Colombia	Plato	67	Moderate	1	Good	16	Good	2	Good	67	Moderate
29	Romania	Poiana Ma	62	Moderate	1	Good	37	Good	1	Good	62	Moderate

Figure 2.1

Dataset collection in Kaggle ensures access to a reliable dataset that aligns the objectives of the project. The dataset's nature, encompassing various aspects through analysis represented within the Global air pollution list.

By leveraging the dataset, the project aims to uncover insights into the factors effecting the pollution and provide information about the health conditions of the people.

2.2 Data Analysis Methodology Overview

Data analysis methodologies encompass a wide range of techniques and approaches used to analyze and interpret data to derive meaningful insights. Here's an overview of some common methodologies:

1. Descriptive Analysis: This involves summarizing and describing the main features of a dataset, such as its central tendency, variability, and distribution. Techniques include measures of central tendency (mean, median, mode), measures of dispersion (standard deviation, variance), and graphical representations (histograms, box plots).

2. Inferential Analysis: This involves making inferences or predictions about a population based on a sample of data. Techniques include hypothesis testing, confidence intervals, and regression analysis.

3. Exploratory Data Analysis (EDA): EDA is an approach to analyzing datasets to summarize their main characteristics, often with visual methods, to uncover patterns, anomalies, and relationships that may not have been initially apparent. Techniques include scatter plots, histograms, and correlation analysis.

4. Predictive Analytics: This involves using statistical or machine learning techniques to build models that predict future outcomes based on historical data. Techniques include regression analysis, decision trees, and neural networks.

5. Prescriptive Analytics: This goes beyond predicting outcomes by recommending actions to achieve desired outcomes. It often involves optimization techniques to identify the best course of action given certain constraints and objectives.

6. Time Series Analysis: This focuses on analyzing data collected over time to identify patterns, trends, and seasonality. Techniques include autoregressive integrated moving average (ARIMA) models, exponential smoothing, and Fourier analysis.

7. Cluster Analysis: This involves grouping similar data points together into clusters based on their characteristics or attributes. Techniques include k-means clustering, hierarchical clustering, and density-based clustering.

8. Dimensionality Reduction: This involves reducing the number of variables in a dataset while preserving its essential features. Techniques include principal component analysis (PCA), t-distributed stochastic neighbour embedding (t-SNE), and factor analysis.

9. Text Analysis: This involves analyzing unstructured text data to extract insights or patterns. Techniques include sentiment analysis, topic modelling, and natural language processing (NLP).

10. Network Analysis: This involves analyzing the relationships and interactions between entities in a network. Techniques include social network analysis (SNA), network centrality measures, and community detection algorithms.

CHAPTER 3

System Architecture and Mathematical Analysis

3.1 Google Colab

Google Colaboratory, commonly known as Google Colab, is a free online cloud-based Jupyter notebook environment tailored for training machine learn- ing and deep learning models. This article explores the functionalities, benefits, and features of Google Colab, elucidating its significance in the realm of data science and machine learning.

Figure 3.1: Google colab environment

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df=pd.read_csv('/content/global air pollution dataset (1).csv')
```

```
df
```

	Country	City	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
0	Russian Federation	Praskoveya	51	Moderate	1	Good	36	Good	0	Good	51	Moderate
1	Brazil	Presidente Dutra	41	Good	1	Good	5	Good	1	Good	41	Good
2	Italy	Priolo Gargallo	66	Moderate	1	Good	39	Good	2	Good	66	Moderate
3	Poland	Przasnysz	34	Good	1	Good	34	Good	0	Good	20	Good
4	France	Punaauia	22	Good	0	Good	22	Good	0	Good	6	Good
...	...	...	...	...	...	...	...	...	...	...	...	...
23458	India	Gursahaiganj	184	Unhealthy	3	Good	154	Unhealthy	2	Good	184	Unhealthy
23459	France	Sceaux	50	Good	1	Good	20	Good	5	Good	50	Good

3.1.1 What is Google Colab?

Google Colab offers a cloud-based environment accessible via any web browser, eliminating the need for local software installation. Users can leverage its computing resources, including CPUs, GPUs, and TPUs, facilitating efficient model training and execution.

3.2 Benefits of Google Colab

Accessibility: Users can access Google Colab from any location with internet connectivity, streamlining collaboration and workflow.

Power: The platform provides access to potent computing resources like GPUs and TPUs, enabling swift and effective model training.

Collaboration: Google Colab simplifies collaborative efforts by allowing realtime editing and sharing of notebooks among team members.

Education: It serves as an invaluable educational tool for learning about machine learning and data science, offering a plethora of tutorials and resources.

3.2.1 Why Choose Google Colab?

Google Colab stands out as an ideal choice for students, data scientists, researchers, and enthusiasts due to it's:

Ease of Use: With no setup requirements, users can swiftly start coding after creating an account.

Affordability: The platform is largely free to use, with paid plans available for more demanding tasks.

Flexibility: Users can seamlessly train models, process data, create visualizations, and collaborate with others, making it a versatile tool for various applications.

3.2.2 Notebook in Google Colab

In Google Colab, a notebook serves as a web-based environment for code creation and execution. Notebooks offer several advantages, including real-time code execution and visualization, support for markdown for documentation,
and collaboration features, making them indispensable for data scientists and machine learning practitioners.

3.2.3 Google Colab Features

Google Colab boasts several features that enhance its usability and effectiveness:

Free Access to GPUs and TPUs: Users can leverage powerful computing resources without any additional cost.

Web-based Interface: The intuitive and user-friendly interface eliminates the need for local software installation.

Collaboration Tools: Multiple users can collaborate on the same notebook simultaneously, streamlining teamwork.

Markdown Support: Notebooks support markdown, enabling users to include formatted text, equations, and images alongside their code.

Pre-installed Libraries: Google Colab comes pre-installed with popular libraries and tools for machine learning and deep learning, such as TensorFlow and PyTorch, saving time on setup and configuration.

Google Colab emerges as a versatile and indispensable tool for machine learning and data science tasks, offering accessibility, power, and flexibility. Its user-friendly interface, collaborative features, and integration with powerful computing resources make it an invaluable asset for individuals and teams alike, driving innovation and progress in the field of machine learning and beyond.

3.3 Mathematical Analysis

A thorough analysis of the IMDb Top 250 Movies dataset was conducted, following a structured approach encompassing data loading, cleaning, exploration, visualization, and additional operations. We meticulously performed each step of the

data analysis process to ensure a comprehensive understanding of the dataset and to extract valuable insights.

This involved loading the dataset from its source, meticulously cleaning it to remove any inconsistencies or missing values, exploring the data to uncover patterns and trends, visualizing key aspects to facilitate interpretation, and conducting additional operations to enhance the depth of our analysis. Through this systematic approach, we were able to gain meaningful insights into the IMDb Top 250 Movies dataset, empowering us to draw informed conclusions and make data-driven decisions

Data Loading and Inspection:

We used Pandas'read_csv()' function to load the dataset and inspected the first and last 10 rows using `head()` and `tail()` functions. The `shape` attribute was used to determine the dataset's dimensions, and the `info()` function provided insights into data types and non-null values.

Data Cleaning:

We identified missing values with the `isnull()` function and ensured data integrity by dropping rows with missing data using `dropna()`. Duplicate rows were identified and removed with the `duplicated()` function to eliminate redundancy.

Data Exploration:

Descriptive statistics were computed using the `describe()` function, revealing key metrics such as mean, median, and quartiles. We analyzed the top 10 movies with the highest ratings and categorized movies into "Excellent," "Good," and "Average" based on custom-defined criteria. Genre-specific exploration was conducted by filtering the dataset based on genre and counting occurrences.

Data Visualization:

We created histograms to visualize distributions of movie ratings and release years. Bar charts displayed the top 10 movie genres based on movie counts, while scatter plots explored the relationship between ratings and box office earnings. Pie charts illustrated genre distribution, and line graphs depicted average movie ratings over time.

```
#Display first 15 rows of the dataset
df.head(15)
```

	Country	City	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
0	Russian Federation	Praskoveya	51	Moderate	1	Good	36	Good	0	Good	51	Moderate
1	Brazil	Presidente Dutra	41	Good	1	Good	5	Good	1	Good	41	Good
2	Italy	Priolo Gargallo	66	Moderate	1	Good	39	Good	2	Good	66	Moderate
3	Poland	Przasnysz	34	Good	1	Good	34	Good	0	Good	20	Good
4	France	Punaauia	22	Good	0	Good	22	Good	0	Good	6	Good
5	United States of America	Punta Gorda	54	Moderate	1	Good	14	Good	11	Good	54	Moderate
6	Germany	Puttlingen	62	Moderate	1	Good	35	Good	3	Good	62	Moderate
7	Belgium	Puurs	64	Moderate	1	Good	29	Good	7	Good	64	Moderate
8	Russian Federation	Pyatigorsk	54	Moderate	1	Good	41	Good	1	Good	54	Moderate
9	Egypt	Qalyub	142	Unhealthy for Sensitive Groups	3	Good	89	Moderate	9	Good	142	Unhealthy for Sensitive Groups
10	China	Qinzhou	68	Moderate	2	Good	68	Moderate	1	Good	58	Moderate
11	Netherlands	Raalte	41	Good	1	Good	24	Good	6	Good	41	Good

Figure 3.2

```
# Display last 15 rows of the dataset
df.tail(15)
```

	Country	City	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
23448	Colombia	Viterbo	55	Moderate	1	Good	24	Good	0	Good	55	Moderate
23449	China	Wangqing	101	Unhealthy for Sensitive Groups	3	Good	35	Good	2	Good	101	Unhealthy for Sensitive Groups
23450	United States of America	El Reno	39	Good	1	Good	39	Good	1	Good	35	Good
23451	Sri Lanka	Wattala	72	Moderate	1	Good	24	Good	1	Good	72	Moderate
23452	Pakistan	Havelian	124	Unhealthy for Sensitive Groups	1	Good	124	Unhealthy for Sensitive Groups	0	Good	85	Moderate
23453	United Kingdom of Great Britain and Northern I...	Urmston	33	Good	1	Good	30	Good	3	Good	33	Good
23454	India	Konnur	86	Moderate	0	Good	23	Good	0	Good	86	Moderate
23455	China	Shaoguan	160	Unhealthy	3	Good	160	Unhealthy	1	Good	79	Moderate
23456	United States of America	Highland Springs	54	Moderate	1	Good	34	Good	5	Good	54	Moderate
23457	Slovakia	Martin	71	Moderate	1	Good	39	Good	1	Good	71	Moderate

Figure 3.3

```
# find shape of our dataset(Number of rows and columns)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23463 entries, 0 to 23462
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Country            23036 non-null  object
 1   City               23462 non-null  object
 2   AQI Value          23463 non-null  int64
 3   AQI Category       23463 non-null  object
 4   CO AQI Value       23463 non-null  int64
 5   CO AQI Category    23463 non-null  object
```

Figure 3.4

```
# getting information about dataset like number of rows and columns,datatype of
# each column and memory requirement
df.info()
```

Figure 3.5

```
#checking missing values
print("Any value?",df.isnull().values.any)
df.isnull()
```

Figure 3.6

```
# drop all missing values
df.dropna(axis=0)
```

Figure 3.7

```
# check for duplicate data
dup_data=df.duplicated().any()
print("Are there any duplicate values:",dup_data)
```

```
Are there any duplicate values: False
```

Figure 3.8

```
# overall statistics about the dataframe
df.describe()
```

Figure 3.9

```
# highest average Value
df.columns
```

Figure 3.10

```
# number of country counts
df['Country'].value_counts()
```

Figure 3.11

```
# finding highest value
df['AQI Category'].max
df[df['AQI Category'].max()==df['AQI Category']]
```

Figure 3.12

```
# display top 15 hifhest AQI Values
top15_len = df.sort_values(by='AQI Value', ascending=False).head(15)[['City','AQI Value']].set_index('City')
print(top15_len)
```

```
             AQI Value
City
Haldaur            500
Mahendragarh       500
Barkhera           500
Khetri             500
Jahangirpur        500
Phalauda           500
Patiala            500
Kakrala            500
Kandhla            500
Hasanpur           500
Dhuri              500
Lachhmangarh       500
Malaut             500
Padampur           500
Jhunjhunun         500
```

Figure 3.13

```
# classifying condition based on city[moderate,good,unhealthy]
def City(City):
    if City >= 65:
        return 'Moderate'
    elif City >= 40:
        return 'Good'
    else:
        return 'Unhealthy'
```

Figure 3.14

MATPLOTLIB

```
# histogram
df['Country'] = df['Country'].astype('category')
df['Country'] = df['Country'].cat.codes

plt.figure(figsize=(10,6))
plt.hist(df['Country'],bins=20, edgecolor='black')
plt.title('Names of cities')
plt.xlabel=('Country')
plt.ylabel('frequency')
plt.show()
```

Figure 3.15

```
# bar
plt.figure(figsize=(10,6))
Country_counts=df['Country'].value_counts().head(15)
Country_counts.plot(kind='bar',color='skyblue')
plt.title('Top 15 countries')
plt.xlabel('Country')
plt.ylabel('Number of City')
plt.xticks(rotation=45)  # Fix the typo here
plt.show()
```

Figure 3.16

```
#scatter
plt.figure(figsize=(8, 6))
plt.scatter(df['AQI Value'], df['Ozone AQI Value'], color='yellow', alpha=0.5)
plt.title('AQI Value vs Ozone AQI Value')
plt.xlabel('AQI Value')
plt.ylabel('Ozone AQI Value')
plt.show()
```

Figure 3.17

```python
#pie chart
import matplotlib.pyplot as plt

Country = ['Brazil', 'Italy', 'India', 'France', 'United States of America', 'Egypt', 'Pakistan', 'Belgium']
Country_counts = [41, 66, 22, 54, 142, 158, 64]

# Add a missing value to Country_counts
Country_counts.append(0)

plt.figure(figsize=(8, 8))
plt.pie(Country_counts, labels=Country, autopct='%1.1f%%', startangle=140,
        colors=['lightcoral', 'green', 'pink', 'yellow', 'red', 'blue', 'purple', 'orange'])
plt.title('Countries')
plt.show()
```

Figure 3.18

```python
#line
plt.figure(figsize=(10,6))
Country_AQI_Value=df.groupby('Country')['AQI Value'].mean()
Country_AQI_Value.plot(color='green')
plt.title('Average Country AQI Value')
plt.xlabel=('Country')
plt.ylabel=('Average AQI Value')
plt.show()
```

Figure 3.19

CHAPTER 4

Result Analysis

```
# histogram
df['Country'] = df['Country'].astype('category')
df['Country'] = df['Country'].cat.codes

plt.figure(figsize=(10,6))
plt.hist(df['Country'],bins=20, edgecolor='black')
plt.title('Names of cities')
plt.xlabel=('Country')
plt.ylabel('frequency')
plt.show()
```

Figure 4.1

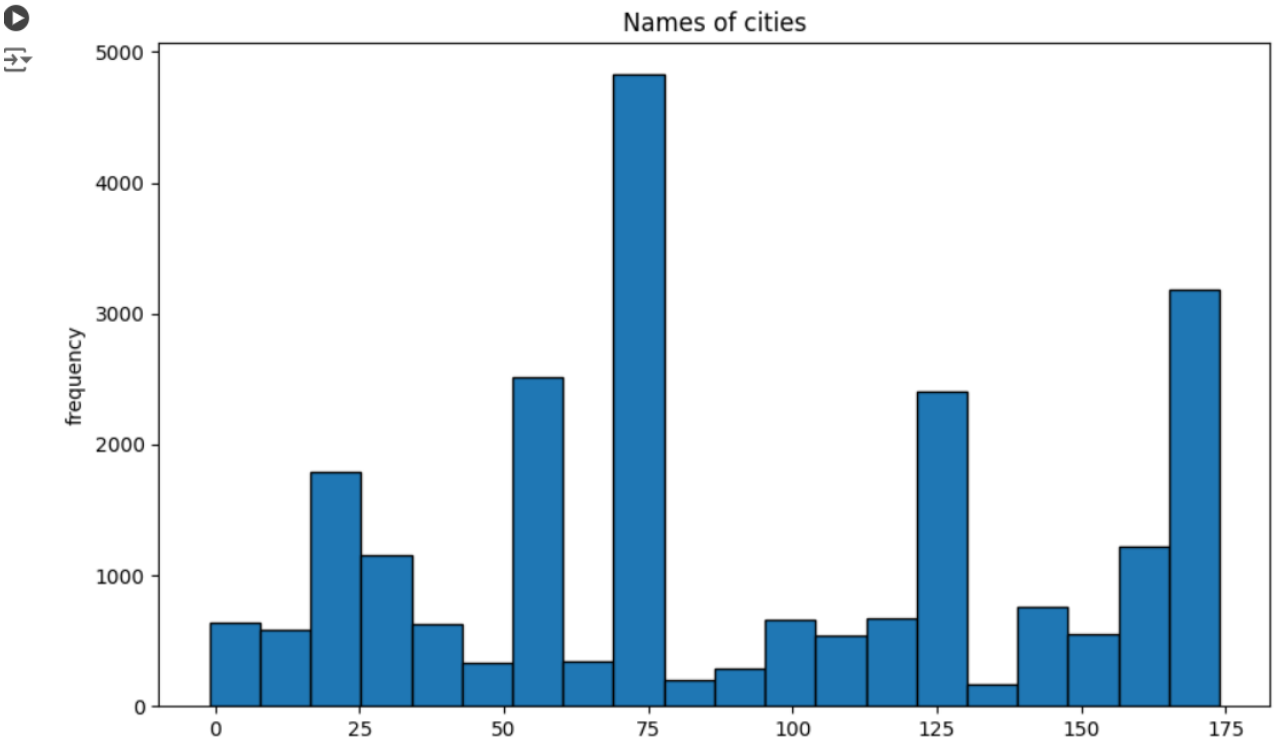

Figure 4.2

```
# bar
plt.figure(figsize=(10,6))
Country_counts=df['Country'].value_counts().head(15)
Country_counts.plot(kind='bar',color='skyblue')
plt.title('Top 15 countries')
plt.xlabel('Country')
plt.ylabel('Number of City')
plt.xticks(rotation=45)
plt.show()
```

Figure 4.3

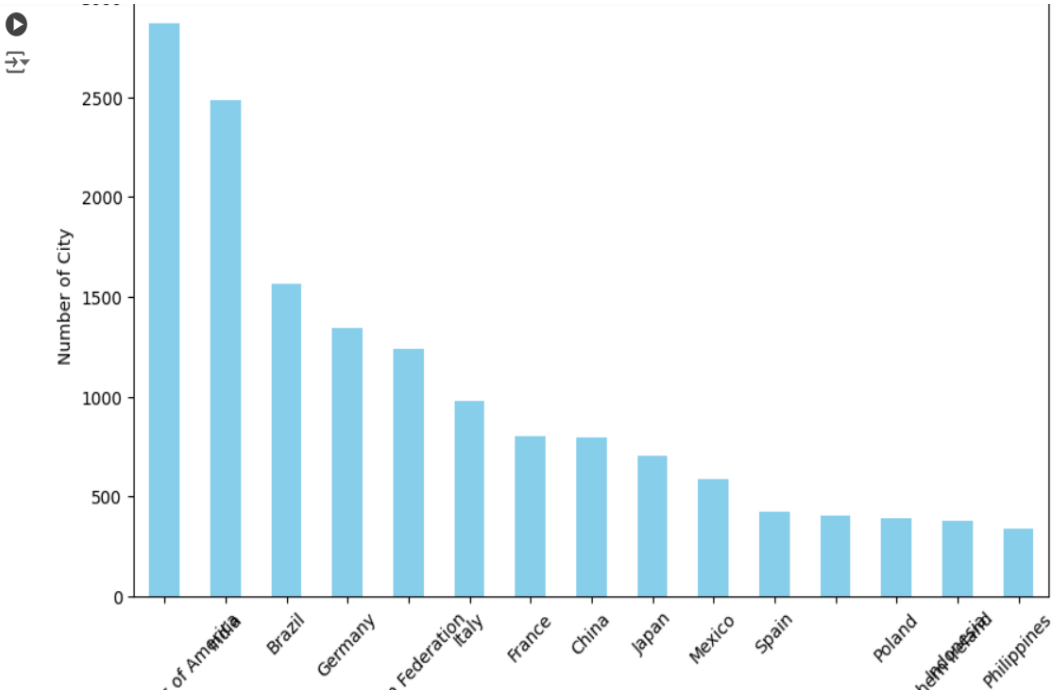

Figure 4.4

```python
#pie chart
import matplotlib.pyplot as plt

Country = ['Brazil', 'Italy', 'India', 'France', 'United States of America', 'Egypt', 'Pakistan', 'Belgium']
Country_counts = [41, 66, 22, 54, 142, 158, 64]

# Add a missing value to Country_counts
Country_counts.append(0)

plt.figure(figsize=(8, 8))
plt.pie(Country_counts, labels=Country, autopct='%1.1f%%', startangle=140,
        colors=['lightcoral', 'green', 'pink', 'yellow', 'red', 'blue', 'purple', 'orange'])
plt.title('Countries')
plt.show()
```

Figure 4.5

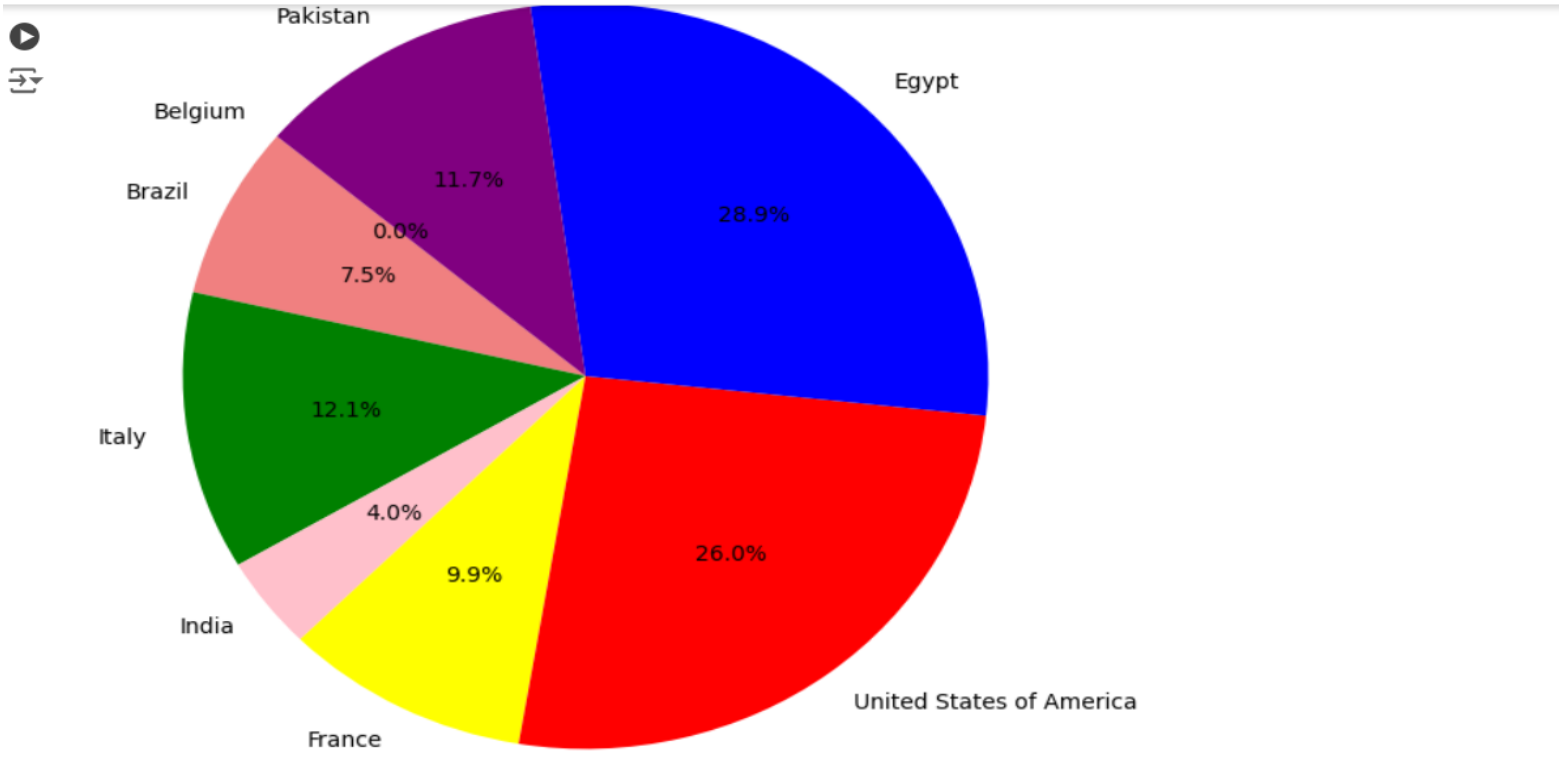

Figure 4.6

```
#line
plt.figure(figsize=(10,6))
Country_AQI_Value=df.groupby('Country')['AQI Value'].mean()
Country_AQI_Value.plot(color='green')
plt.title('Average Country AQI Value')
plt.xlabel=('Country')
plt.ylabel=('Average AQI Value')
plt.show()
```

Figure 4.7

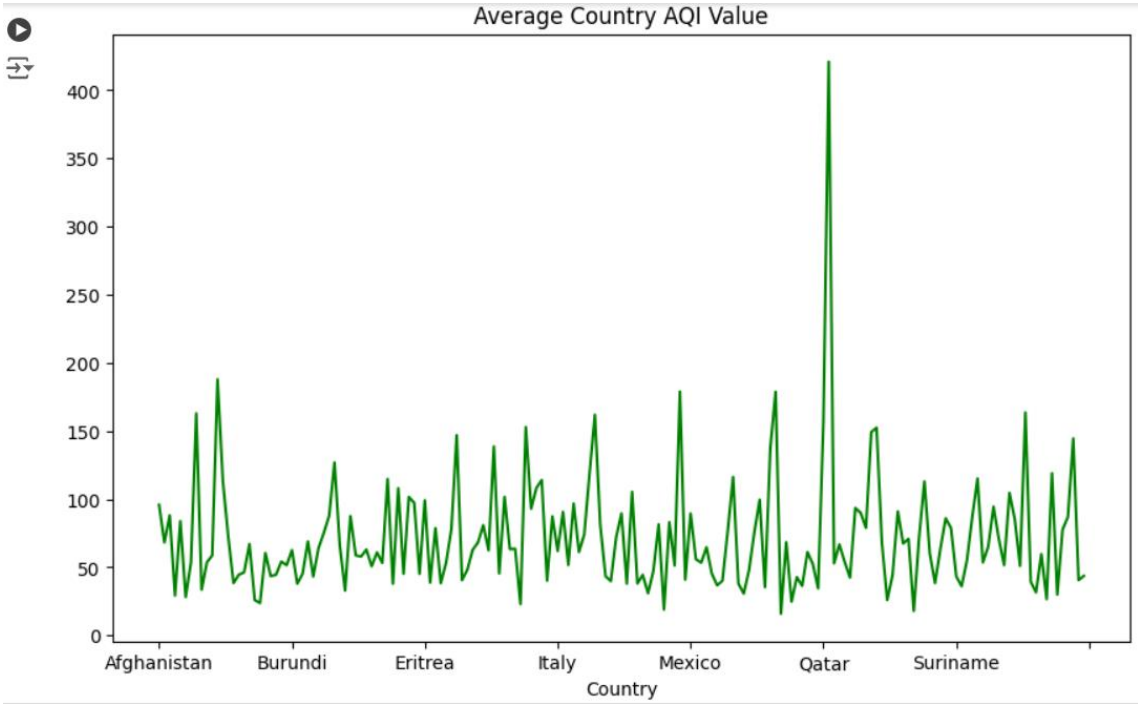

Figure 4.8

```python
#scatter
plt.figure(figsize=(8, 6))
plt.scatter(df['AQI Value'], df['Ozone AQI Value'], color='yellow', alpha=0.5)
plt.title('AQI Value vs Ozone AQI Value')
plt.xlabel('AQI Value')
plt.ylabel('Ozone AQI Value')
plt.show()
```

Figure 4.9

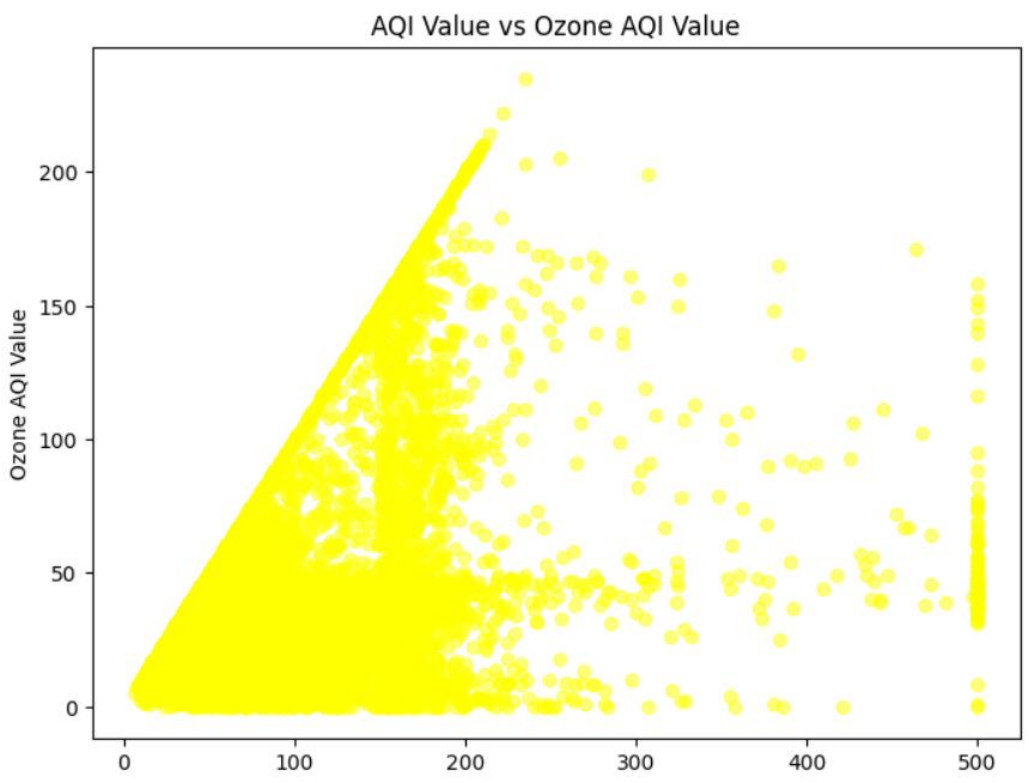

Figure 4.10

.

CHAPTER 5
Conclusion

In conclusion, the analysis of global air pollution reveals a complex and pressing environmental challenge that demands immediate and sustained attention. The multifaceted nature of air pollution, driven by industrial emissions, vehicular exhaust, deforestation, and agricultural practices, highlights the need for a coordinated international response. Mitigating the adverse health effects and environmental damage requires the implementation of stringent regulatory frameworks, the adoption of cleaner technologies, and the promotion of sustainable practices. Collaborative efforts among governments, businesses, and civil society are essential to address the disparities in pollution levels between developed and developing regions. As the world moves towards a more interconnected future, investing in clean energy, enhancing public awareness, and fostering innovation will be pivotal in reducing air pollution and ensuring a healthier planet for future generations.

Furthermore, the Global air pollution analysis is crucial for understanding the extent and impact of air pollution, guiding policies and actions to improve air quality and protect public health and the environment. With advancements in technology and a collaborative approach, significant progress can be made in reducing air pollution and its adverse effects worldwide of systematic approaches in extracting actionable insights from large datasets.

CHAPTER 6
References

[1] Matplotlib: A 2D Graphics Environment ,John D. Hunter

Computing in science & engineering (Print) 2007. 17993 Citations, 1 References.

Python Data Analytics: With Pandas, NumPy, and Matplotlib Fabio
Nelli 2023

[2] Research on Big Data Analysis Data Acquisition and Data Analysis
Hong Li
2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)