Artificial Intelligence and Machine Learning

Continuous Assessment

Problem Statement: Exploratory Data Analysis & Visualization of Air Quality Data

Understanding air quality trends and factors affecting it is crucial for public health and environmental policy. Visualizing air quality data can help identify patterns, compare different locations, and communicate findings effectively.

This project aims to analyze and visualize an air quality dataset to:

1. **Explore relationships:** Identify correlations between different air pollutants and the overall Air Quality Index (AQI).
2. **Compare locations:** Analyze and visualize air quality differences across various cities and countries.
3. **Identify trends:** Discover temporal patterns in air quality data.
4. **Communicate insights:** Use visualizations to effectively convey key findings regarding air quality.

Implementation:

The project utilizes Python libraries like Pandas for data manipulation, Matplotlib, and Seaborn for creating visualizations. The implementation involves:

1. **Data loading and cleaning:** Importing the dataset, handling missing values, and converting data types if necessary.
2. **Exploratory data analysis:** Calculating descriptive statistics, identifying outliers, and examining distributions of variables.
3. **Data visualization:** Generating various visualizations like scatter plots, bar charts, line graphs, pie charts, and box plots to represent the data and extract meaningful insights.

Dataset Info:

The air quality dataset contains information such as:

- City: Name of the city where the measurement was taken.
- Country: Name of the country where the measurement was taken.
- AQI Value: Overall Air Quality Index.

- Pollutant AQI Values: Individual AQI values for pollutants like PM2.5, NO2, etc.
- AQI Categories: Categorical representation of AQI values (e.g., Good, Moderate, Unhealthy).

country_name	city_name	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
Russian Federat	Praskoveya	51	Moderate	1	Good	36	Good	0	Good	51	Moderate
Brazil	Presidente I	41	Good	1	Good	5	Good	1	Good	41	Good
Italy	Priolo Garg	66	Moderate	1	Good	39	Good	2	Good	66	Moderate
Poland	Przasnysz	34	Good	1	Good	34	Good	0	Good	20	Good
France	Punaauia	22	Good	0	Good	22	Good	0	Good	6	Good
United States o	Punta Gord	54	Moderate	1	Good	14	Good	11	Good	54	Moderate
Germany	Puttlingen	62	Moderate	1	Good	35	Good	3	Good	62	Moderate
Belgium	Puurs	64	Moderate	1	Good	29	Good	7	Good	64	Moderate
Russian Federat	Pyatigorsk	54	Moderate	1	Good	41	Good	1	Good	54	Moderate
Egypt	Qalyub	142	Unhealthy for	3	Good	89	Moderate	9	Good	142	Unhealthy for Sensitive Groups
China	Qinzhou	68	Moderate	2	Good	68	Moderate	1	Good	58	Moderate
Netherlands	Raalte	41	Good	1	Good	24	Good	6	Good	41	Good
India	Radaur	158	Unhealthy	3	Good	139	Unhealthy for Sensiti	1	Good	158	Unhealthy
Pakistan	Radhan	158	Unhealthy	1	Good	50	Good	1	Good	158	Unhealthy
Republic of Nor	Radovis	83	Moderate	1	Good	46	Good	0	Good	83	Moderate
France	Raismes	59	Moderate	1	Good	30	Good	4	Good	59	Moderate
India	Rajgir	154	Unhealthy	3	Good	100	Unhealthy for Sensiti	2	Good	154	Unhealthy
Italy	Ramacca	55	Moderate	1	Good	47	Good	0	Good	55	Moderate

Code:

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

# Load the dataset
df = pd.read_csv("C:\\Users\\MADHURA GOLATKAR\\Downloads\\Air.csv")

# Create a scatter plot to visualize the relationship between aqi_value and
pm2.5_aqi_value
plt.figure(figsize=(10, 6))
sns.scatterplot(x='aqi_value', y='pm2.5_aqi_value', data=df)
plt.title('Relationship between AQI Value and PM2.5 AQI Value')
plt.xlabel('AQI Value')
plt.ylabel('PM2.5 AQI Value')
plt.show()

# Create a bar chart to compare the mean aqi_value for different countries
country_means = df.groupby('country_name')['aqi_value'].mean().reset_index()
plt.figure(figsize=(10, 6))
sns.barplot(x='country_name', y='aqi_value', data=country_means)
plt.title('Mean AQI Value by Country')
plt.xlabel('Country')
plt.ylabel('Mean AQI Value')
plt.xticks(rotation=90)
plt.show()
```

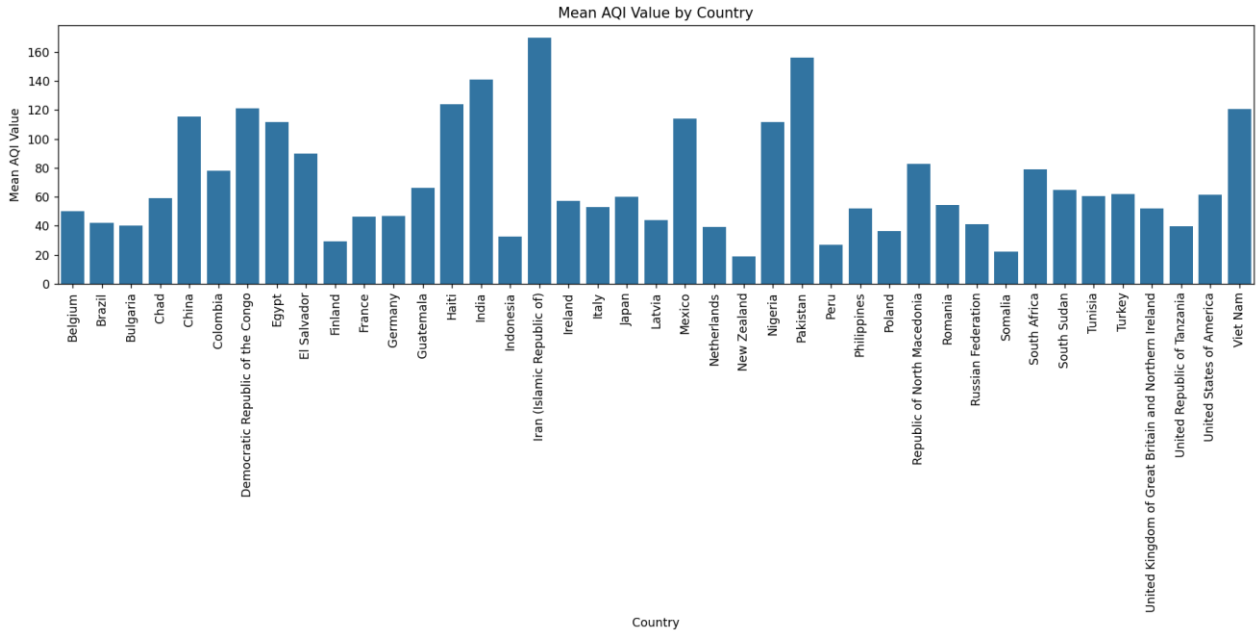
```
# Create a pie chart to show the distribution of aqi_category
aqi_categories = df['aqi_category'].value_counts().reset_index()
aqi_categories.columns = ['aqi_category', 'count']
plt.figure(figsize=(10, 6))
plt.pie(aqi_categories['count'], labels=aqi_categories['aqi_category'],
autopct='%1.1f%%')
plt.title('Distribution of AQI Category')
plt.show()

# Create a box plot to compare the distribution of aqi_value for different countries
plt.figure(figsize=(10, 6))
sns.boxplot(x='country_name', y='aqi_value', data=df)
plt.title('Distribution of AQI Value by Country')
plt.xlabel('Country')
plt.ylabel('AQI Value')
plt.xticks(rotation=90)
plt.show()
```

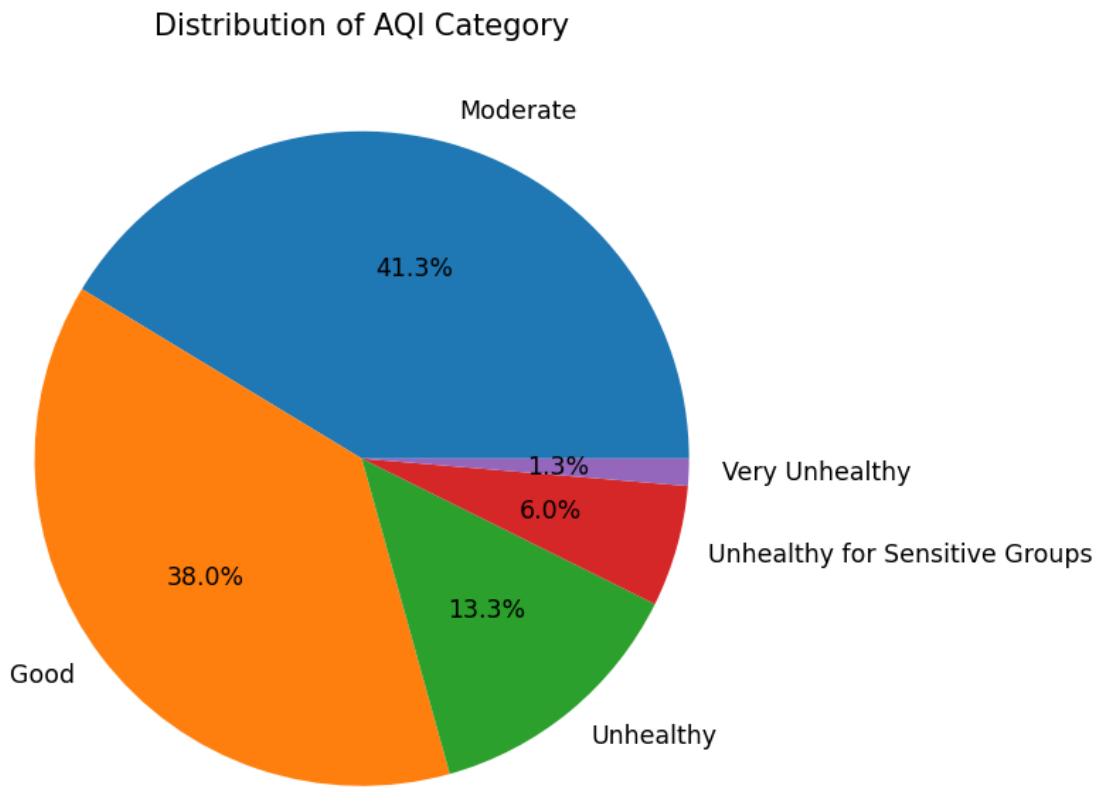
Results:

1. Scatter plot to visualize the relationship between aqi_value and pm2.5_aqi_value

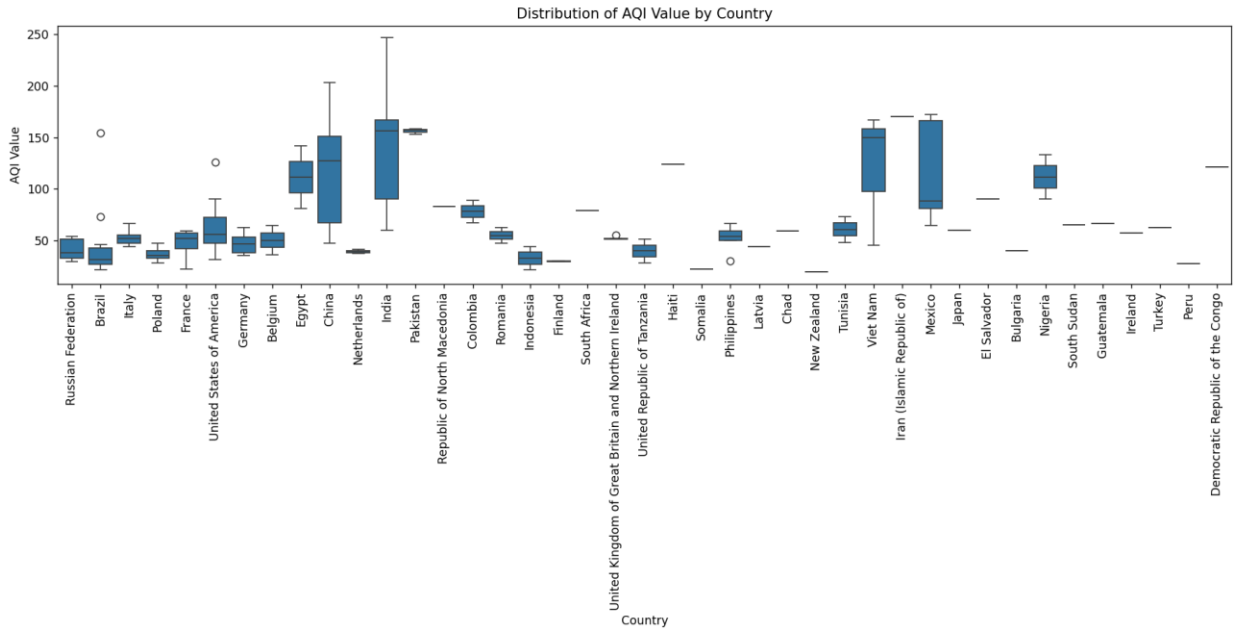
2. Bar chart to compare the mean aqi_value for different countries



3. Pie chart to show the distribution of aqi_category



4. Box plot to compare the distribution of aqi_value for different countries



Analysis:

1. **Distribution of AQI Category (Pie Chart)**

- Moderate air quality is the most prevalent, making up 41.3% of the data.
- Good air quality is close behind, comprising 38% of the data.
- Unhealthy air quality is observed at 13.3%.
- Unhealthy for Sensitive Groups accounts for 6%.
- The least represented categories are Very Unhealthy (1.3%).

A large portion of the air quality falls within the moderate and good categories, but there's still a significant percentage of unhealthy air quality, affecting public health.

2. **Distribution of AQI Value by Country (Box Plot)**

- Countries like India, China, and Pakistan have high AQI values, indicating poor air quality. The median AQI in these countries hovers between 150-200, with some extreme outliers (above 200).
- United States, Europe (Italy, Poland, France, Germany, etc.), and Brazil exhibit lower AQI values, with median values between 50-100.
- Netherlands shows an especially low and consistent AQI, with minimal variation, indicating stable and good air quality.
- Some countries have minimal variations, suggesting more consistent air quality, while others have a wider range, indicating fluctuating air quality.

Asian countries like India and China suffer from significantly worse air quality than European and American countries, which generally maintain better air quality.

3. **Relationship between AQI Value and PM2.5 AQI Value (Scatter Plot)**

- A clear positive correlation between AQI and PM2.5 AQI values can be observed. As the AQI value increases, PM2.5 concentrations also rise, showing a direct relationship between general air quality and PM2.5 particulate pollution.
- The linear trend indicates that PM2.5 is a major contributor to overall AQI, especially at higher values (150 and above).

The presence of fine particulate matter (PM2.5) is a strong determinant of poor air quality. Areas with high AQI values are also likely to experience higher levels of PM2.5 pollution, which is a critical health concern.

4. **Mean AQI Values Across Countries (Bar Chart Analysis)**

- The chart displays the mean AQI (Air Quality Index) values for different countries, giving a comparative view of average air pollution levels.
- India, China, and Iran have notably higher mean AQI values, indicating poorer air quality. These countries are among those with the highest levels of pollution, likely driven by industrial emissions, traffic, and urbanization.
- Countries like Iraq, Nepal, and Pakistan also show elevated AQI values, suggesting significant air quality issues.
- On the opposite end, countries such as Finland, Norway, Australia, and New Zealand have relatively low mean AQI values, indicating much better air quality. These countries likely benefit from stricter environmental regulations, less industrial activity, and more natural green spaces.
- United States and some European countries like Germany and France sit in the middle range, with moderate AQI values, showing neither severe nor perfect air quality.

Countries in Asia and the Middle East exhibit higher mean AQI values, reflecting poorer air quality and greater pollution. Meanwhile, countries in Europe, Oceania, and North America tend to have lower AQI values, suggesting better management of air quality and cleaner environmental conditions.

Conclusion:

We implemented various graphs to visualize data patterns using Python libraries such as Pandas, Matplotlib, and Seaborn. A **scatter plot** was used to explore the relationship between overall AQI and PM2.5 values, showing a clear positive correlation, while a **bar chart** compared mean AQI values across countries, highlighting differences in air quality. A **pie chart** illustrated the distribution of AQI categories, revealing that most regions fall under "Moderate" or "Good" air quality, but some face unhealthy conditions. Finally, a **box plot** was used to examine AQI value distributions across countries, identifying both stable and fluctuating air quality patterns. These visualizations helped us uncover significant insights into air quality trends, pollutants, and geographical disparities.