MIT | Academy of Engineering
(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

Analysis On Air Quality Index

Guided By,
Mrs. Shubhangi Kale(SPK), MITAOE

Presented By :

- ***Sanket Kolhe(233)***
- ***Gopal Mankar(242)***
- ***Sairaj Gujar(223)***

Introduction

- The analysis of air quality index (AQI) data set aims to utilize data science techniques to gain insights into the quality of air in a specific area or region.
- By examining historical air quality data, applying data manipulation and visualization methods, and employing predictive techniques, this analysis provides a comprehensive understanding of factors influencing air quality.
- The analysis serves the purpose of assessing environmental impact, identifying health risks, predicting future air quality conditions, optimizing resource allocation for pollution control, and supporting evidence-based policy decisions.
- Through data-driven approaches, this analysis contributes to the mitigation of air pollution and the protection of public health.

Motivation

- Analyzing air quality index data is motivated by several factors. Firstly, it helps us understand the environmental impact of various sources, aiding in the development of effective pollution control strategies.
- Secondly, poor air quality directly affects public health, necessitating the identification of high pollution areas and associated health risks.
- Thirdly, analyzing air quality data informs policy development and implementation, enabling targeted interventions and emission standards.
- Additionally, it supports urban planning and infrastructure decisions for healthier and sustainable cities. Lastly, sharing air quality analysis results raises community awareness and fosters engagement in environmental initiatives.

Details of Dataset

Data Set Name: AQI and Lat Long of Countries

No. of Features:- 14 (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, lat, lng)

No. of Records: 4007

Country	City	AQI Value	AQI Categ	CO AQI V	CO AQI C	Ozone AQ	Ozone AQ	NO2 AQI V	NO2 AQI Cat	PM2.5 AQI \	PM2.5 AC	lat	lng
Russian Fe	Praskovey	51	Moderate	1	Good	36	Good	0	Good	51	Moderate	44.7444	44.2031
Brazil	Presidente	41	Good	1	Good	5	Good	1	Good	41	Good	-5.29	-44.49
Brazil	Presidente	41	Good	1	Good	5	Good	1	Good	41	Good	-11.2958	-41.9869
Italy	Priolo Gar	66	Moderate	1	Good	39	Good	2	Good	66	Moderate	37.1667	15.1833
Poland	Przasnysz	34	Good	1	Good	34	Good	0	Good	20	Good	53.0167	20.8833
United Sta	Punta Gor	54	Moderate	1	Good	14	Good	11	Good	54	Moderate	16.1005	-88.8074
United Sta	Punta Gor	54	Moderate	1	Good	14	Good	11	Good	54	Moderate	26.8941	-82.0513
Belgium	Puurs	64	Moderate	1	Good	29	Good	7	Good	64	Moderate	51.0761	4.2803
Russian Fe	Pyatigorsk	54	Moderate	1	Good	41	Good	1	Good	54	Moderate	44.05	43.0667
China	Qinzhou	68	Moderate	2	Good	68	Moderate	1	Good	58	Moderate	21.95	108.6167
Netherlan	Raalte	41	Good	1	Good	24	Good	6	Good	41	Good	52.3833	6.2667
France	Raismes	59	Moderate	1	Good	30	Good	4	Good	59	Moderate	50.3892	3.4858
Italy	Ramacca	55	Moderate	1	Good	47	Good	0	Good	55	Moderate	37.3833	14.7
United Sta	Phoenix	72	Moderate	1	Good	4	Good	23	Good	72	Moderate	33.5722	-112.089
Poland	Piaseczno	28	Good	1	Good	28	Good	2	Good	28	Good	52.0667	21.0167
Brazil	Pinheiral	154	Unhealthy	5	Good	0	Good	13	Good	154	Unhealthy	-22.5128	-44.0008
Colombia	Plato	67	Moderate	1	Good	16	Good	2	Good	67	Moderate	9.7919	-74.7872
Romania	Poiana Ma	62	Moderate	1	Good	37	Good	1	Good	62	Moderate	43.9333	23.0833
Russian Fe	Polevskoy	31	Good	1	Good	31	Good	0	Good	17	Good	56.45	60.1833

Data Manipulation

```
print(df1.describe())
```

```
         AQI Value
count  4007.000000
mean     63.183928
std      46.127164
min      10.000000
25%      38.000000
50%      52.000000
75%      69.000000
max     500.000000
```

```
print(df1.max())
```

```
City                                        Zwettl
AQI Value                                      500
AQI Category                        Very Unhealthy
CO AQI Value                                   133
CO AQI Category     Unhealthy for Sensitive Groups
Ozone AQI Value                                206
Ozone AQI Category                  Very Unhealthy
NO2 AQI Value                                   91
NO2 AQI Category                          Moderate
PM2.5 AQI Value                                500
PM2.5 AQI Category                  Very Unhealthy
lat                                        69.4167
lng                                       176.1667
dtype: object
```

```
print(df1['AQI Value'].max())
```

```
500
```

```
print(df1['AQI Value'].describe())
```

```
count    4007.000000
mean       63.183928
std        46.127164
min        10.000000
25%        38.000000
50%        52.000000
75%        69.000000
max       500.000000
Name: AQI Value, dtype: float64
```

```
print(df1['AQI Value'].min())
```

```
10
```

```
print(df1.count())
```

```
Country      3929
City         4007
AQI Value    4007
dtype: int64
```

```
print(df.shape)
```

```
(4007, 14)
```

```
print(df1['AQI Value'].mean())
```

```
63.183928125779886
```

```
print(df1.mean)
```

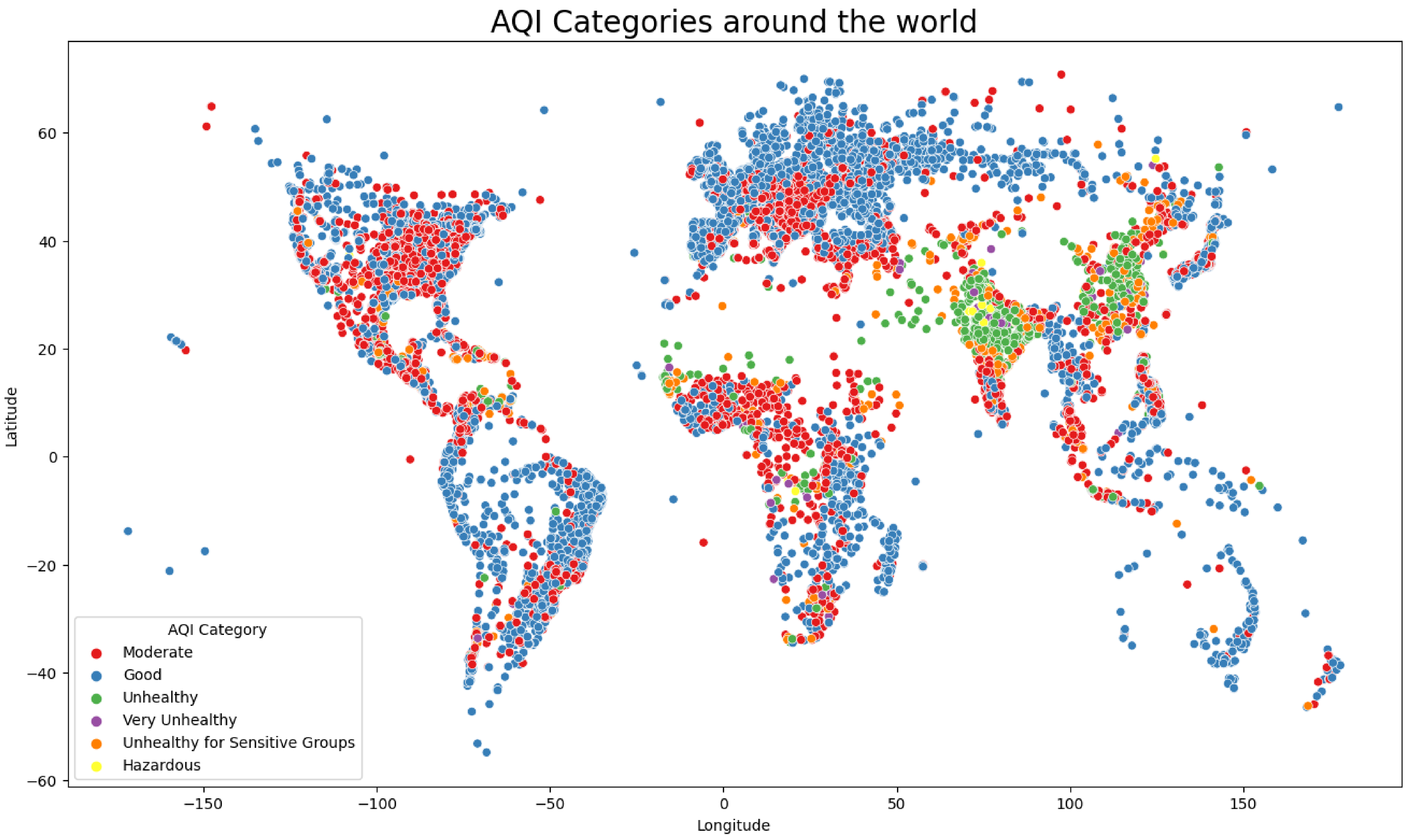
```
<bound method NDFrame._add_numeric_operations.<locals>.mean of
0     Russian Federation        Praskoveya        51
1                 Brazil  Presidente Dutra        41
2                 Brazil  Presidente Dutra        41
3                  Italy   Priolo Gargallo        66
4                 Poland         Przasnysz        34
...                  ...               ...       ...
4002        South Africa            Benoni       123
4003         Netherlands            Bergen        39
4004         Netherlands            Bergen        39
4005         Netherlands            Bergen        39
4006         Netherlands            Bergen        39

                        AQI Category  CO AQI Value CO AQI Category  \
0                           Moderate             1            Good
1                               Good             1            Good
2                               Good             1            Good
3                           Moderate             1            Good
4                               Good             1            Good
...                              ...           ...             ...
4002  Unhealthy for Sensitive Groups            15            Good
4003                            Good             1            Good
4004                            Good             1            Good
4005                            Good             1            Good
4006                            Good             1            Good
```

Data Visualization

(Scatter Plot)

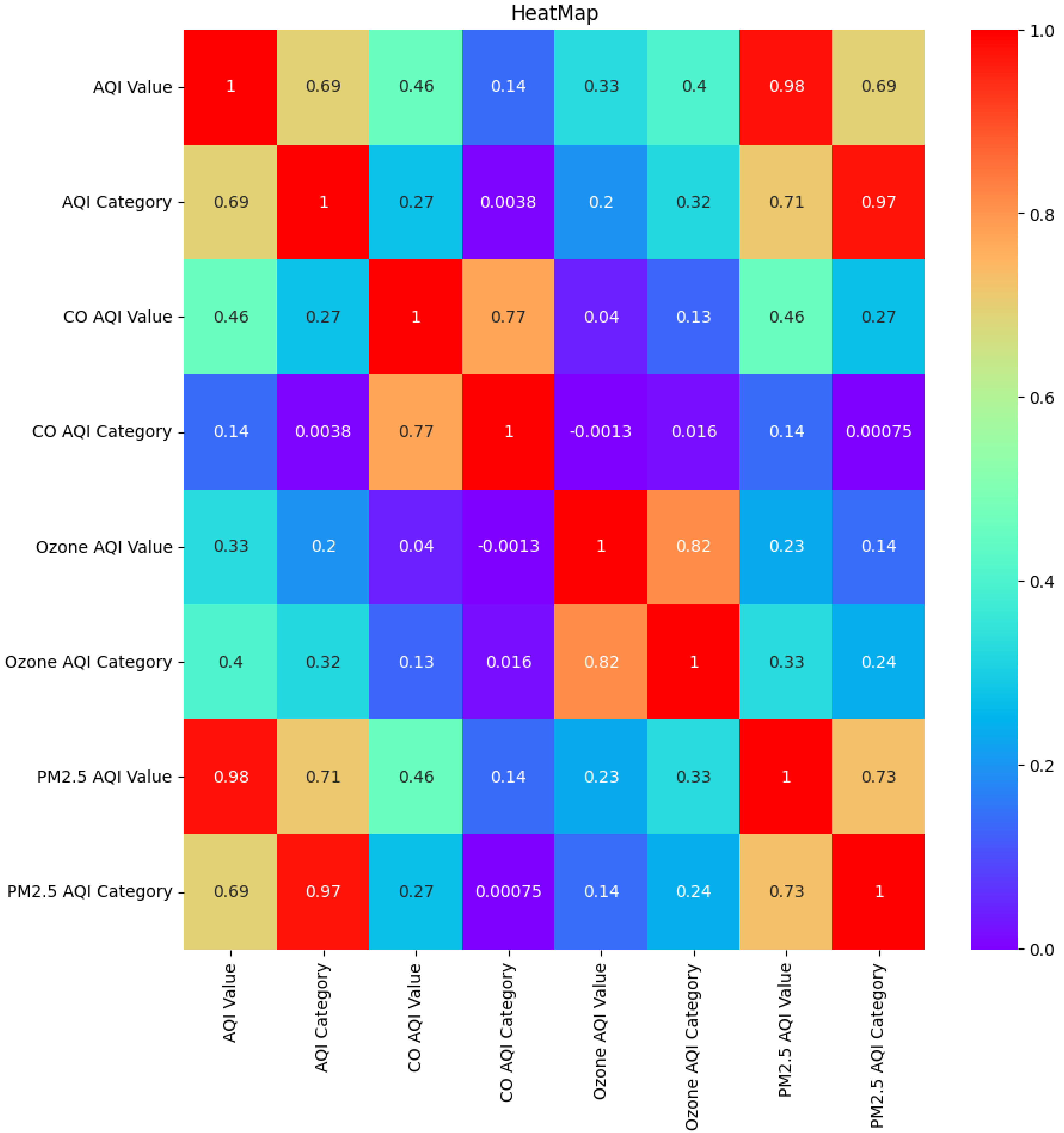
```
import matplotlib.pyplot as plt
```

```
plt.figure(figsize=(16,9))
sns.scatterplot(data=df,x='lng',y='lat',hue='AQI Category',
palette='Set1')
plt.title('AQI Categories around the world',fontsize=20)
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.show()
```



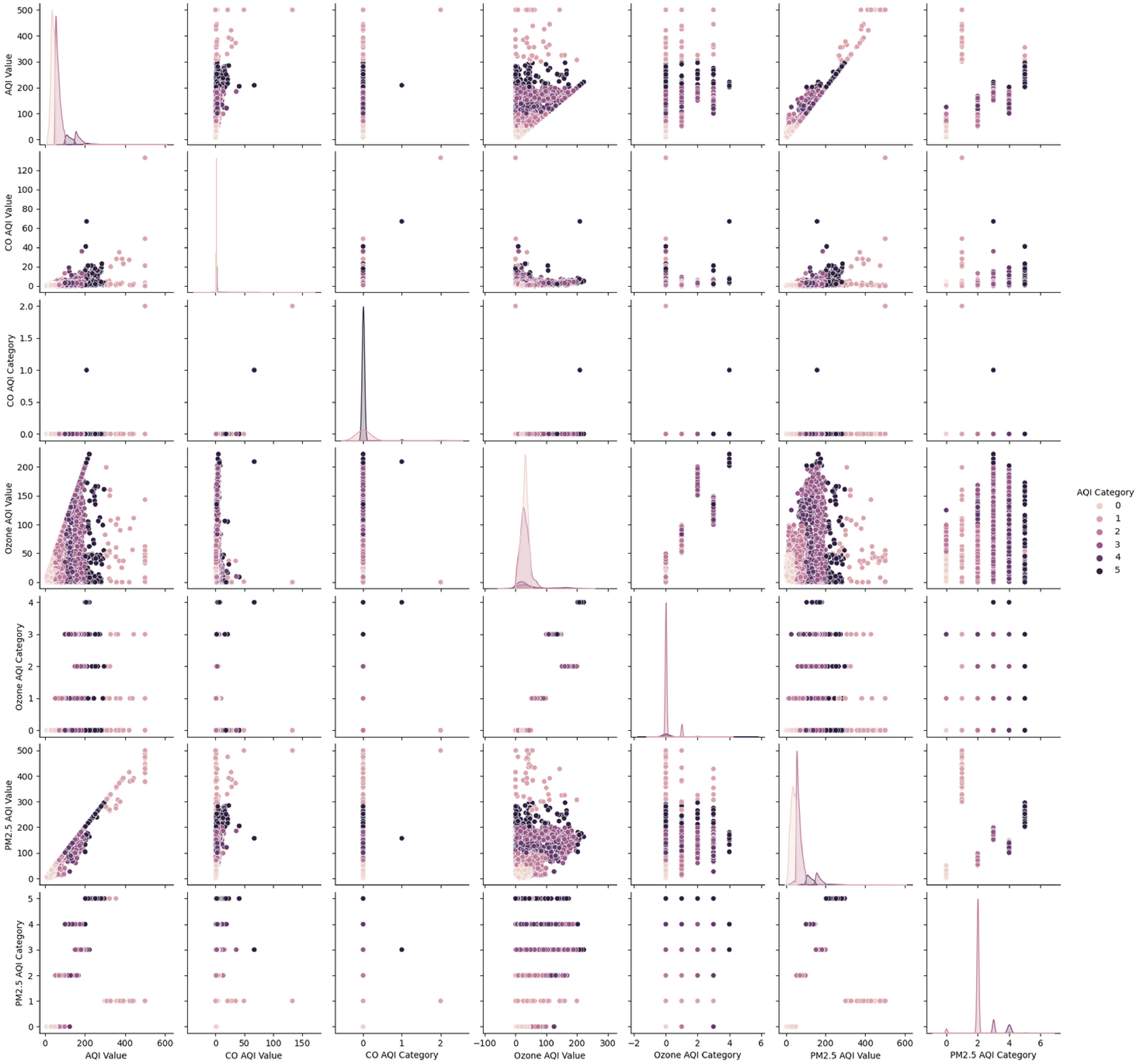
(Heat Map)

```
import seaborn as sns

Corr = df.corr()

plt.figure(figsize=(10,10))
sns.heatmap(Corr,annot=True,cmap='rainbow')
plt.title('HeatMap')
plt.show()
```



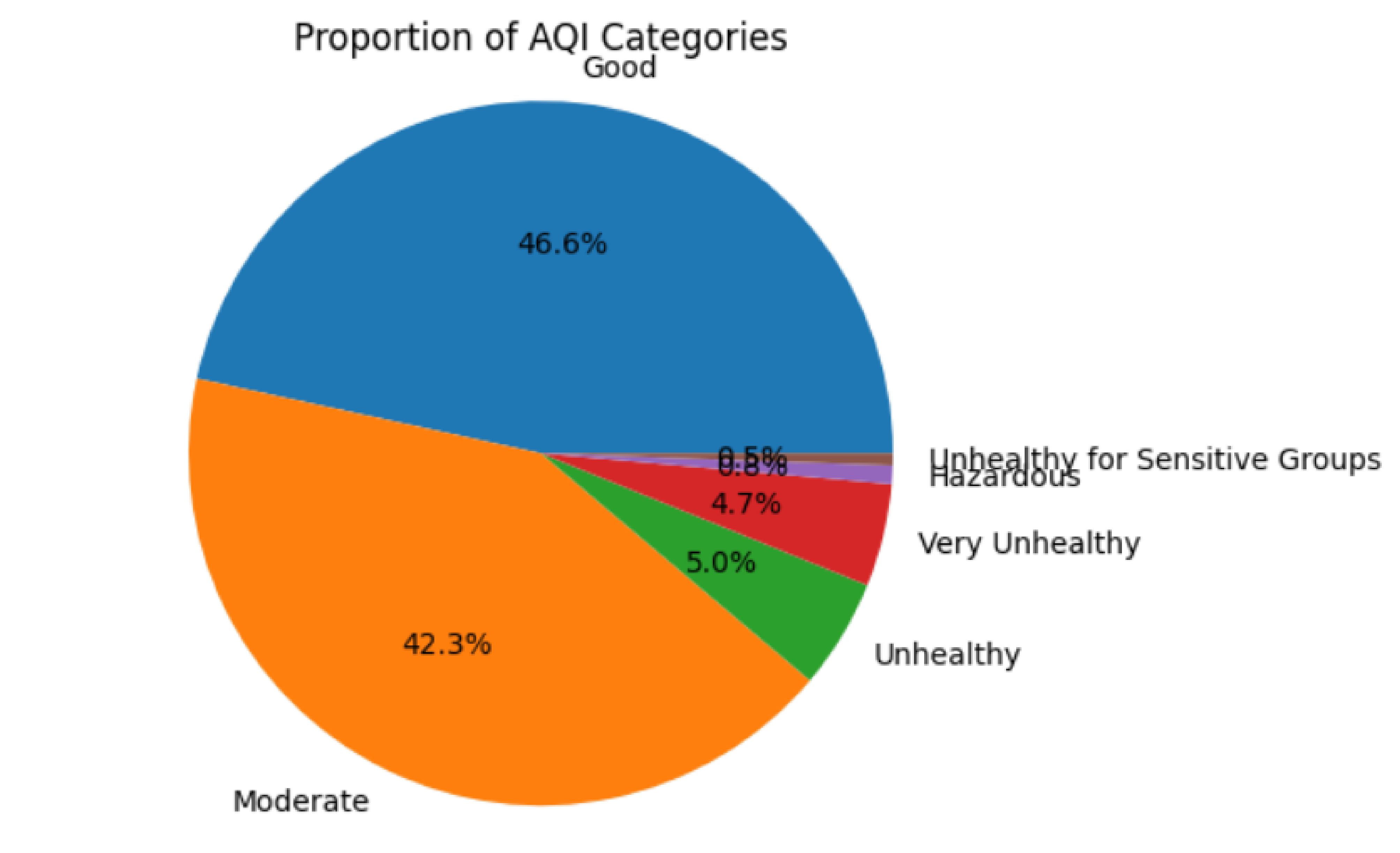
(Pair Plot)

```
sns.pairplot(df,hue='AQI Category')
plt.show()
```



(Pie Chart)

```
# Pie chart: Proportion of AQI categories
aqi_categories = ['Good', 'Moderate', 'Unhealthy', 'Very Unhealthy', 'Hazardous','Unhealthy for Sensitive Groups']
aqi_counts = df['AQI Category'].value_counts()
plt.pie(aqi_counts, labels=aqi_categories, autopct='%1.1f%%')
plt.title('Proportion of AQI Categories')
plt.axis('equal')
plt.show()
```



Predictive Technique (LR/KNN/KMeans)

```
#KNN
import pandas as pd
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score

data = pd.read_csv('/content/AQI and Lat Long of Countries2.csv')
X = data[['AQI Value', 'CO AQI Value', 'Ozone AQI Value', 'NO2 AQI Value', 'PM2.5 AQI Value']]
y = data['AQI Category']
le = LabelEncoder()
y = le.fit_transform(y)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

k = 3
knn = KNeighborsClassifier(n_neighbors=k)

knn.fit(X_train, y_train)
y_pred = knn.predict(X_test)

y_pred = le.inverse_transform(y_pred)

accuracy = accuracy_score(le.inverse_transform(y_test), y_pred)
print(f"Accuracy: {accuracy}")
example_data = [[60, 2, 35, 1, 40]]
predicted_category = le.inverse_transform(knn.predict(example_data))
print(f"Predicted AQI Category: {predicted_category}")
```

Output

```
Accuracy: 0.9850374064837906
Predicted AQI Category: ['Good']
```

Predictive Technique (KMeans)

```
#Linear Regression
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
data = pd.read_csv('/content/AQI and Lat Long of Countries2.csv')
X = data[['lat', 'lng']]
y = data['AQI Value']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
rmse = mean_squared_error(y_test, y_pred, squared=False)
print(f"RMSE: {rmse}")
r2 = r2_score(y_test, y_pred)
print(f"R-squared: {r2}")
example_data = [[44.7444, 44.2031]]
predicted_aqi = model.predict(example_data)
print(f"Predicted AQI Value: {predicted_aqi}")
```

Output

```
RMSE: 45.3923032194421
R-squared: 0.02458718467384724
Predicted AQI Value: [65.34166434]
```

Application

- The main use of the Air Quality Index is to provide warnings to citizens, keeping them informed of the current conditions.
- This is particularly important for sensitive groups such as the elderly, children, or people with respiratory or heart problems. Such people should stay indoors when air quality is low.

Conclusion

- The Air Quality Index (AQI) data set provides valuable information about the air quality levels in different locations.
- It allows researchers, policymakers, and the general public to assess and monitor air pollution levels.
- The data set includes measurements of various pollutants, such as particulate matter (PM2.5 and PM10), ozone (O3), nitrogen dioxide (NO2), and sulfur dioxide (SO2).
- By analyzing the AQI data, trends and patterns can be identified, helping to understand the factors influencing air pollution and its impacts on human health and the environment.
- The data set enables comparisons between different regions and time periods, facilitating the identification of areas with poor air quality and the evaluation of air pollution control measures.
- It serves as a valuable resource for developing strategies to mitigate air pollution and improve air quality standards.