AIR QUALITY INDEX

GROUP 2

Samujitha Senaratne
Chamodi Siriwardhana
Senuri Perera

“Air pollution is not merely a nuisance and a threat to health. It is a reminder that our most celebrated technological achievements-the automobile, the jet plane, the power plant, industry in general, and indeed the modern city itself-are, in the environment, failures.”

- Barry Commoner

OVERVIEW


1
2
3
4
5
INTRODUCTION TO AQI
INTRODUCTION TO THE DATASET
DESCRIPTIVE ANALYSIS
FURTHER ANALYSIS
SUGGSETIONS FOR ADVANCED ANALYSIS

How does Air Quality Index Work ?

Air Quality Index

COLOMBO, SRI LANKA
9 DECEMBER, 2022

- An air quality index (AQI) is an indicator developed by government agencies to communicate to the public how polluted the air currently is or how polluted it is forecast to become.

- As air pollution levels rise, so does the AQI, along with the associated Public health risk.

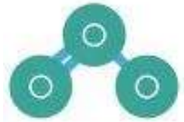
- The air in our atmosphere is mostly made up of two gases that are essential for life on Earth: nitrogen and oxygen. However, the air also contains smaller amounts of many other gases and particles.

- Air Quality can be affected by eight pollutants PM10, PM2.5, NO2, SO2, CO, O3, NH3, and Pb

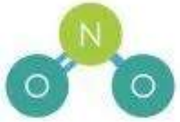
- The five major air pollutants:

Particulate Matter

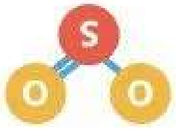
Carbon Monoxide



Ground Level Ozone



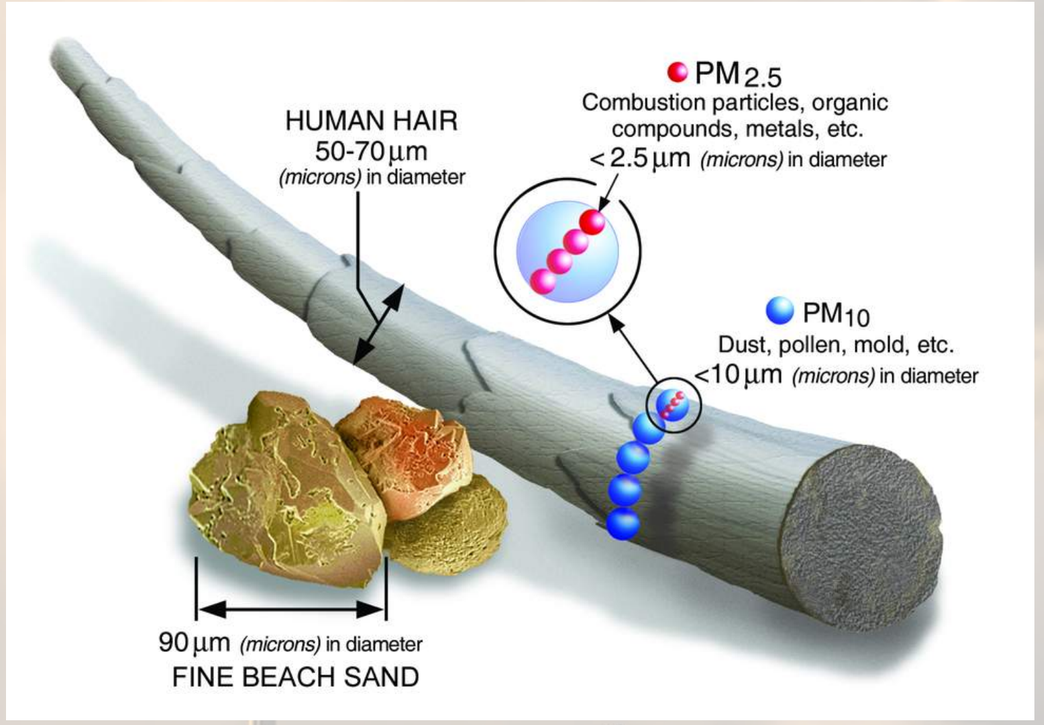
Nitrogen Dioxide



Sulphur Dioxide

MAJOR POLLUTANTS

PARTICULATE MATTER (PM 2.5)



- While all the forms of atmospheric pollution are a cause for concern, it's the smaller 2.5µm particles that get the most attention.
- For one, we can see visible evidence in the form of haze and smoke when PM2.5 levels increase.
- As well, these fine particles have a much easier time entering our bodies via breathing.

CARBON MONOXIDE

It is a colorless gas, released from automobile emissions, fires, industrial processes, gas stoves, kitchen chimneys, generators, wood-burning smoking, etc. into the atmosphere.

MAJOR POLLUTANTS

GROUND LEVEL OZONE

Ozone is composed of three oxygen atoms. It forms the protective layer which prevents entry of harmful ultraviolet radiation into the earth. The ground ozone is very harmful to human beings and the environment.

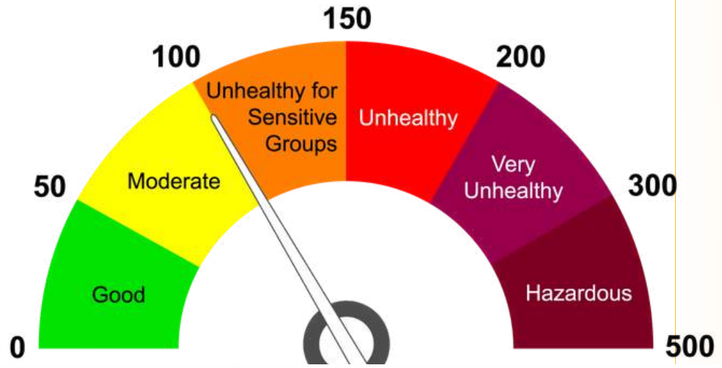
NITROGEN DIOXIDE

Nitrogen Dioxide is released into the environment from automobile emissions, generation of electricity, burning of fuel, combustion of fossil fuel, and different industrial processes.

SULFUR DIOXIDE

Sulfur dioxide is a colorless gas with a burnt odor and the chemical formula SO2. The gas is acidic & corrosive in nature and can react in the atmosphere with other compounds to form sulfuric acid and other oxides of sulfur.

Air Quality Index

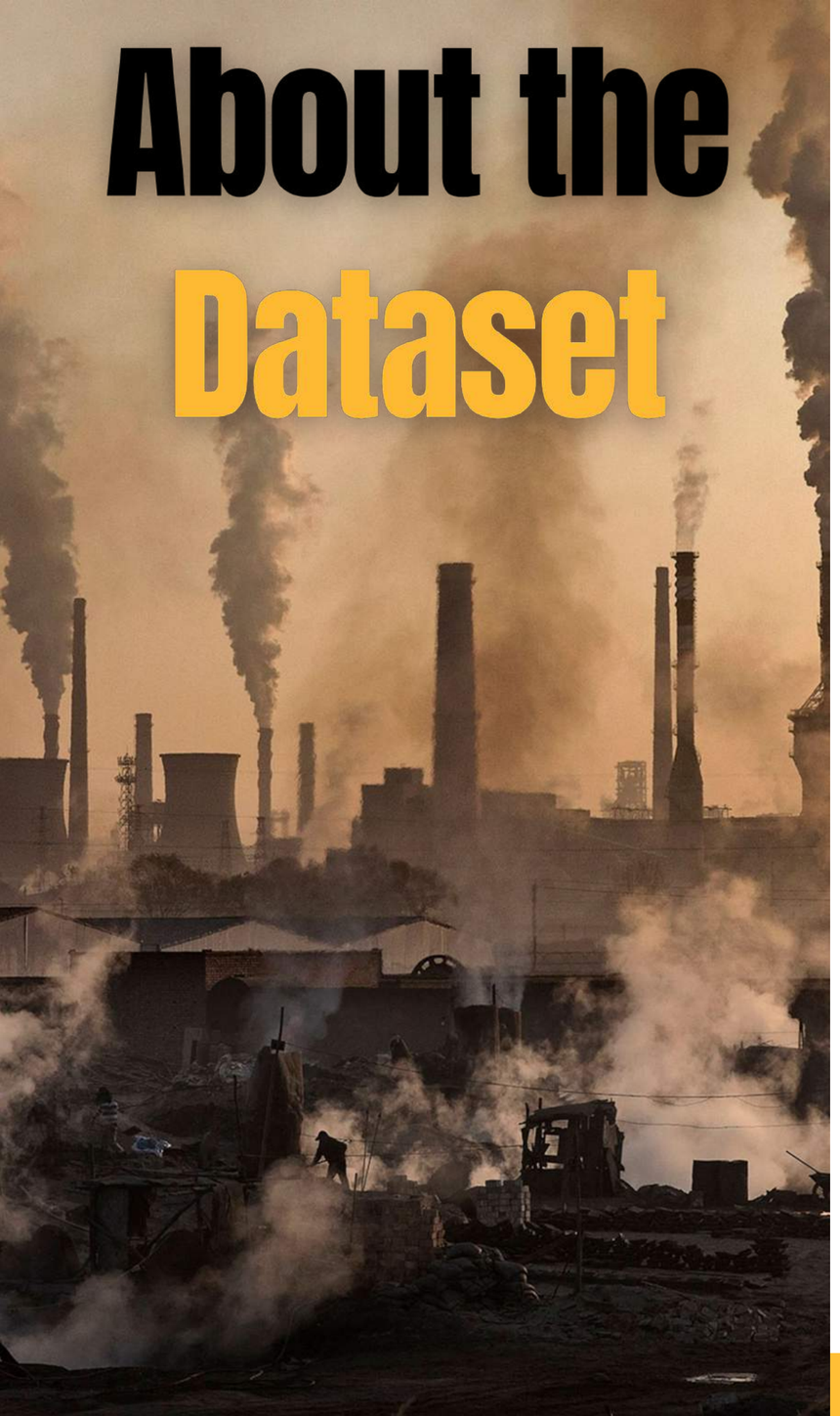


Air Quality Index	Actions to protect your health from air pollution
Good 0-50	None
Moderate 51-100	Usually sensitive people should consider reducing prolonged/heavy exertion
Unhealthy for sensitive groups 101-150	Following groups should reduce prolonged/heavy exertion: • People with heart or lung disease • Children and older adults
Unhealthy 151-200	Following groups should avoid prolonged/heavy exertion: • People with heart or lung disease • Children and older adults Everyone else should reduced prolonged/heavy exertion
Very Unhealthy 201-300	Following groups should avoid all physical activity outdoors: • People with heart or lung disease • Children or older adults Everyone else should avoid prolonged/heavy exertion
Hazardous >301	Avoid all physical activity outdoors Sensitive groups: remain indoors and keep activity levels low. Follow tips for keep particle levels low indoors.

- Different countries have their own air quality indices, corresponding to different national air quality standards. Some of these are Canada's Air Quality Health Index, Malaysia's Air Pollution Index, and Singapore's Pollutant Standards Index.

- The United States Environmental Protection Agency (EPA) has developed an Air Quality Index that is used to report air quality.

- This AQI is divided into six categories indicating increasing levels of health concern. An AQI value over 300 represents hazardous air quality and below 50 the air quality is good.



About the Dataset

- The ‘World Air Quality Index by City and Coordinates’ dataset was acquired from the Kaggle website.
- It contains 16695 records under 14 variables, where the response variable is ‘AQI Category’ (categorical).

Country	Character	Name of the Country
City	Character	Name of the City
CO AQI Value	Integer	The AQI value of Carbon Monoxide
CO AQI Category	Factor	The AQI category of Carbon Monoxide
Ozone AQI Value	Integer	The AQI value of Ozone
Ozone AQI Category	Factor	The AQI category of Ozone
NO2 AQI Value	Integer	The AQI value of Nitrogen Dioxide
NO2 AQI Category	Factor	The AQI category of Nitrogen Dioxide
PM2.5 AQI Value	Integer	Fine particulate matter less than 2.5 micrometers in diameter value
PM2.5 AQI Category	Factor	Fine particulate matter less than 2.5 micrometers in diameter category
lat	Float	Latitude value of the city
lng	Float	Longitude value of the city
AQI Value	Integer	Overall air quality index value
AQI Category	Factor	Overall air quality index category with respect to the AQI score range.

Obtaining the SO_2 data through Web-Scraping

- Given the significance of sulfur dioxide (SO2) as a principal dictator of air quality, the inclusion of SO2 data for each observation becomes very important.

- A web-scraping technique was employed as an effective means of data extraction.

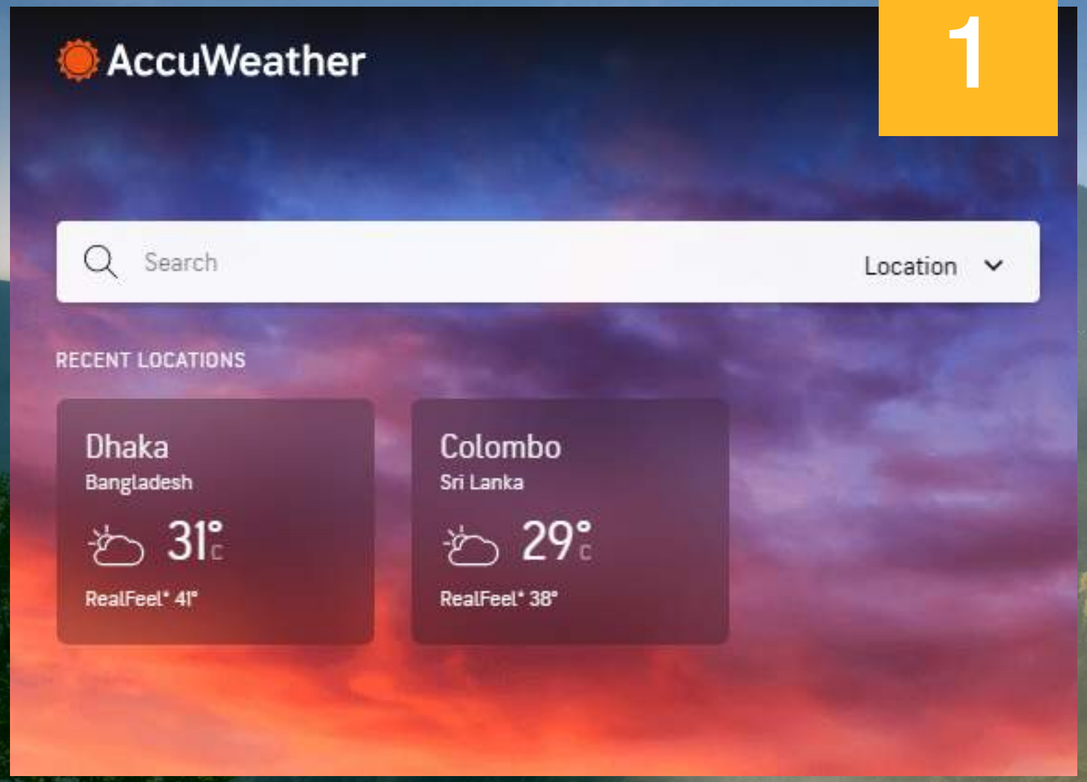
- AccuWeather, which is a reputable source of weather and air quality information, was utilized for the web-scraping task.

- Selenium, a popular tool in web automation, was utilized to interact with the web page dynamically.

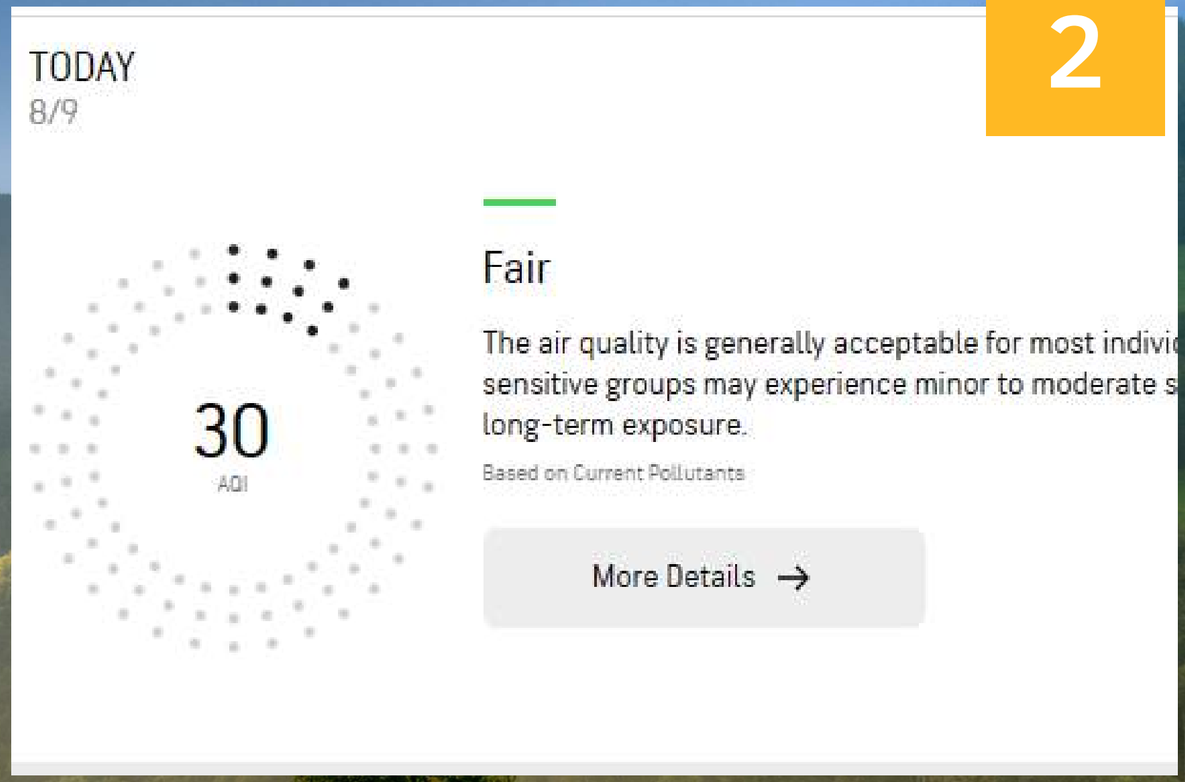
- XPath, on the other hand, is a query language used to navigate through the structure of an XML or HTML document.

WEB-SCRAPING

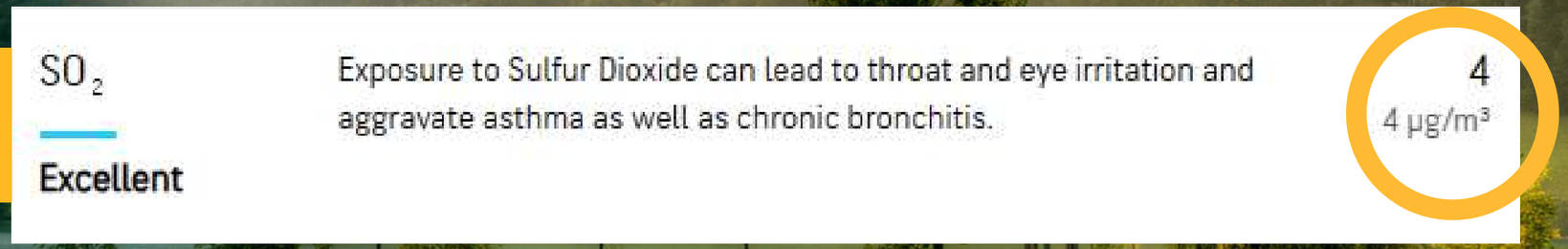
1



2



3



- **Note that the dataset that we originally obtained, has not mentioned any date of data collection.**
- **Thus, by the process of including the SO2 data, we generalize the dataset to the present.**

OBJECTIVES

Gain a comprehensive understanding of the dataset's characteristics and underlying patterns. By thoroughly exploring the dataset, uncover valuable insights that will guide subsequent analysis and modeling.

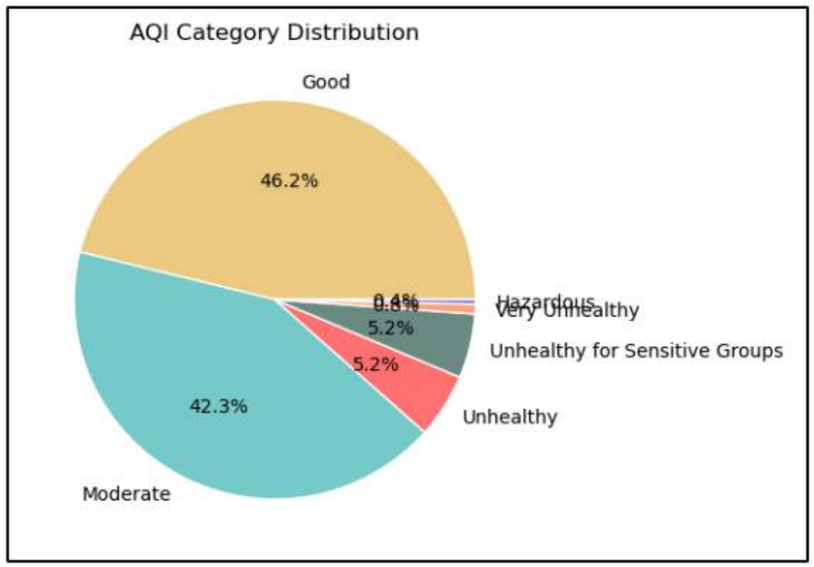
By developing an accurate AQI prediction model,

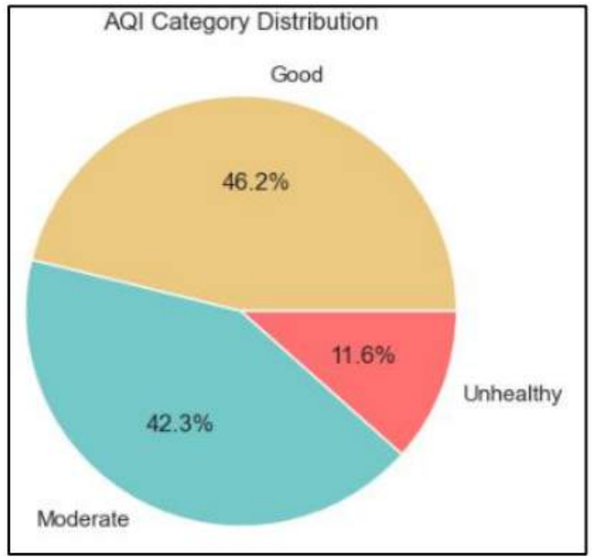
contribute to the detection and assessment of air quality hazards.

This predictive capability empowers individuals to make informed decisions about outdoor activities, adopt protective measures, and adjust daily routines to mitigate potential health risks posed by varying air quality levels.

Data Preprocessing

- Since there was a limited number of observations within the categories "Unhealthy for Sensitive Groups," "Very Unhealthy," "Unhealthy," and "Hazardous", these categories were lumped together and were named as “Unhealthy”.





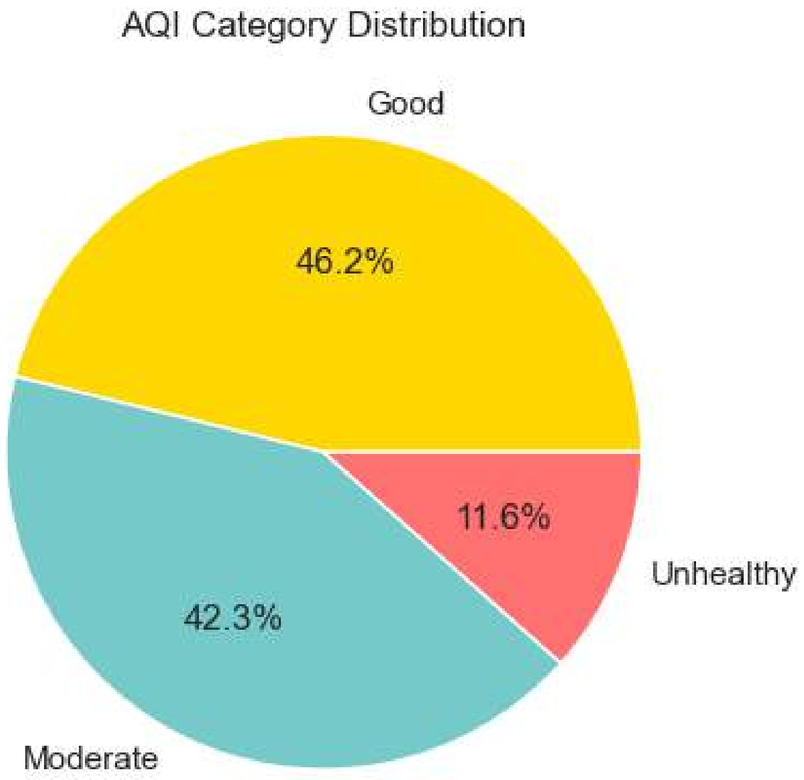
- There were 302 missing values in the ‘Country’ variable. It was imputed using the city name from ‘City’ variable and the Geopy library in Python with the help of GeoNames web database.

- A new variable named ‘Continent’ was created by categorizing each country to its respective Continent.

Descriptive Analysis

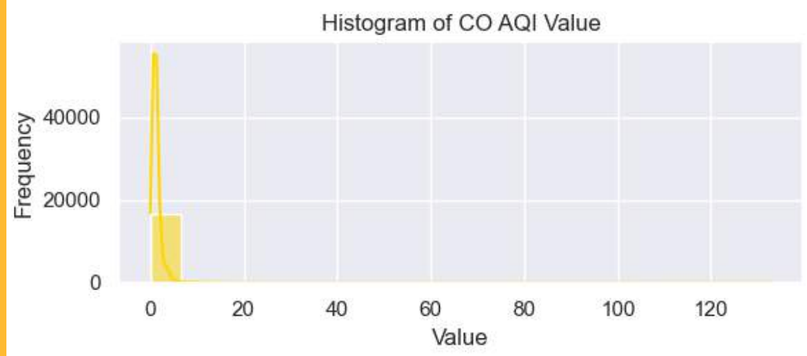
Univariate Analysis

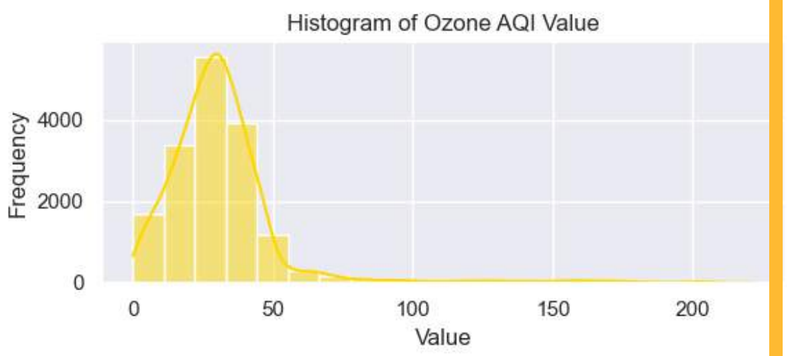
Response Variable :
" AQI Category "

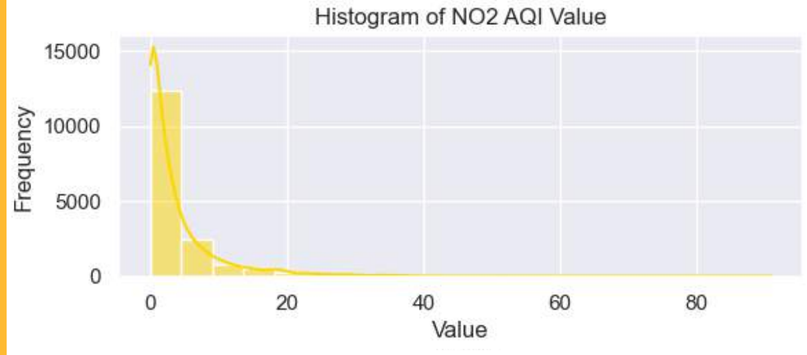


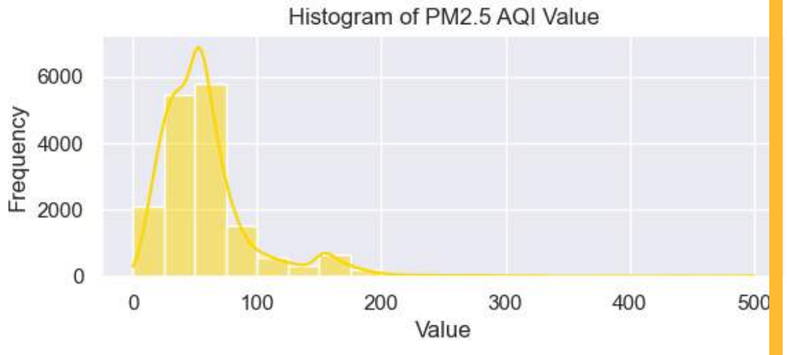
- **According to the World Air Quality Report by IQAir, only 0.18% of the world's land area has good air quality**
- **Our data set does not represent the real world senario.**
- **This may lead to biased results in predictions.**

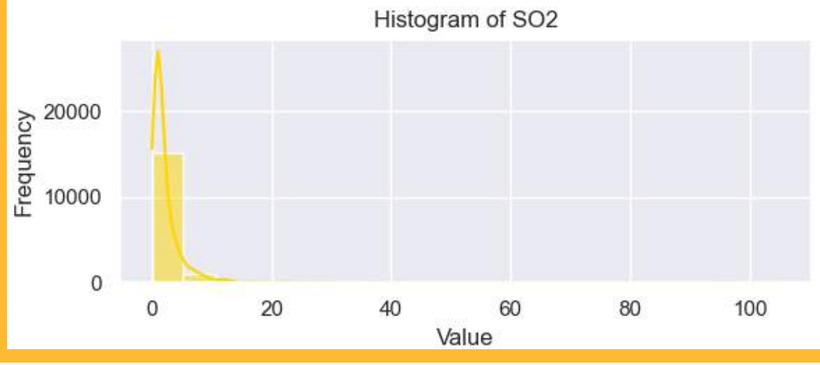
Univariate Analysis









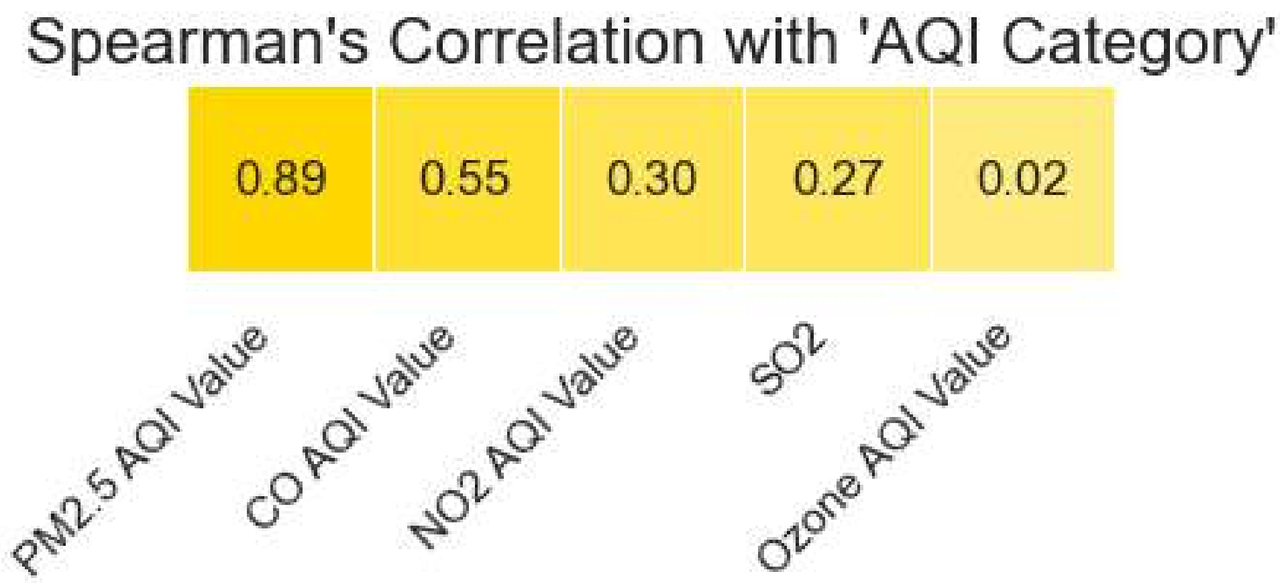


- **CO AQI Value, NO2 AQI Value, and SO2 AQI Value are highly skewed to the right**
  -Tally with global air distribution parameters given by World Health Organization -
- **To ensure accurate insights and interpretations,**
  Can Apply logarithmic transformation to reduce the impact of extreme values

Bivariate Analysis

Correlation Between Categorical Response Variable and Continuous Variables

Spearman's Correlation

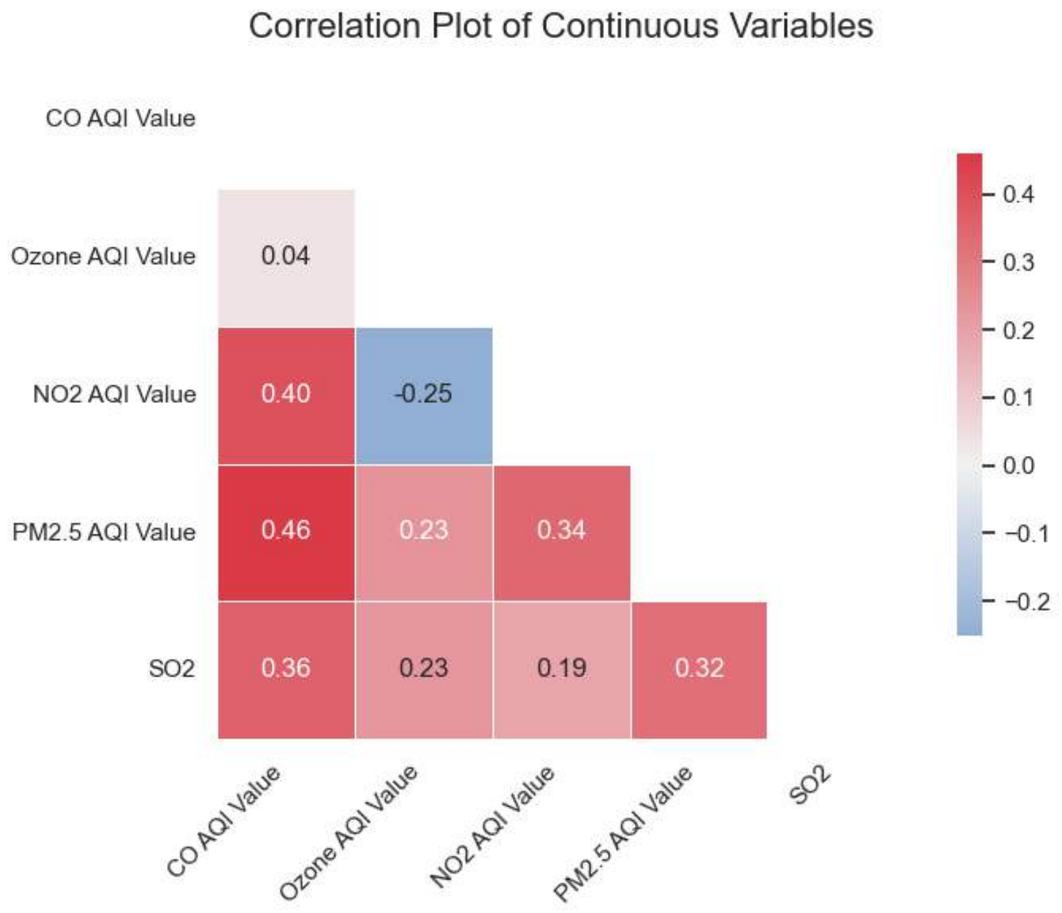


The notably strong positive correlation between 'PM2.5 AQI Value' and AQI Category underscores its substantial influence on air quality categorization.

'CO AQI Value',' NO2 AQI Value', and 'SO2', exhibit moderate positive correlations with AQI Categories, while 'Ozone AQI Value' shows weaker associations

Correlation Between Continuous Variables

Pearson Correlation



Bivariate Analysis

Positive correlation between several variables :

indicating the presence of multicollinearity.

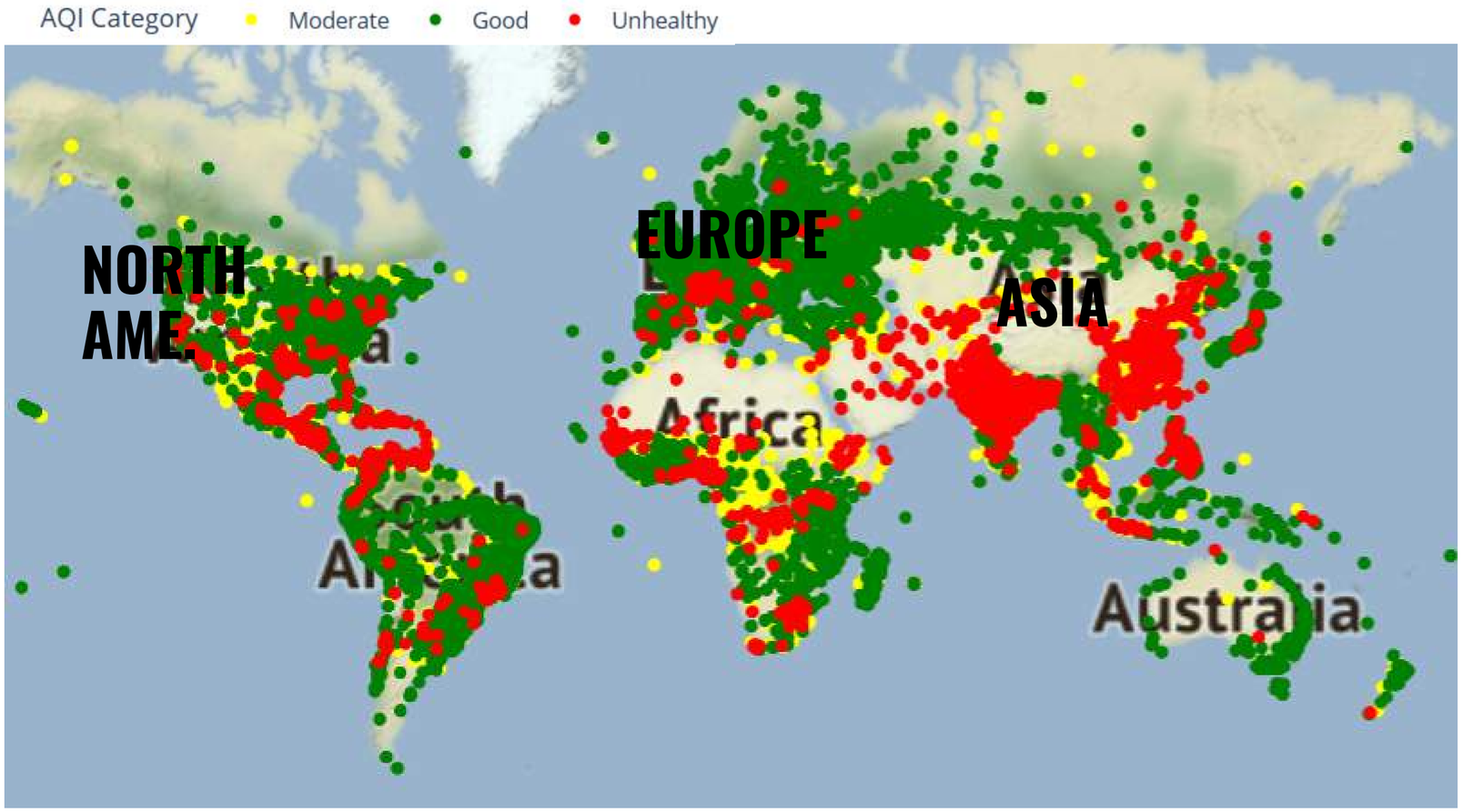
The reason for that is pollutants such as CO, NO2, PM2.5, and SO2 are all emitted from similar sources, such as car engines, power plants, and industrial facilities.

This means that they are often found in the same places, and they can be correlated with each other

Bivariate Analysis

Correlation Between Categorical Response Variable and Categorical variable Continent

Chi-square Test



p-value 9.2644e-27 p-value < 0.05 indicate an association between the AQI Category and Continent.

Eye inspection

ASIA > AFRICA > NOR. AMERI. > EUROPE > SOUTH AMERI.

Confirming our test results

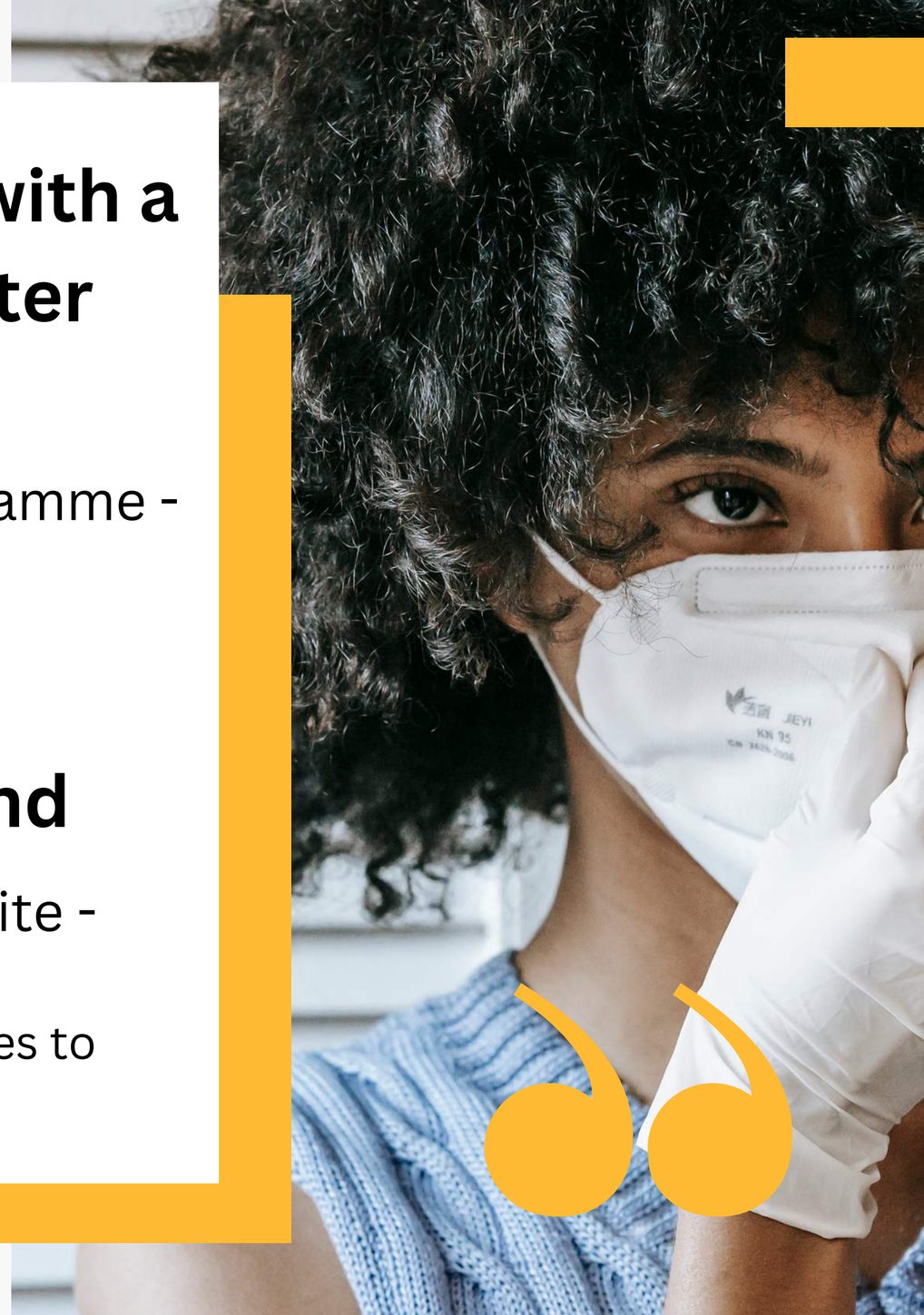
"Asia is the most polluted continent, with a high concentration of particulate matter (PM2.5) and ground-level ozone "

-United Nations Environment Programme -

" Africa is the second most polluted continent, with high levels of PM2.5 and Ozone "

-Health policy watch web site -

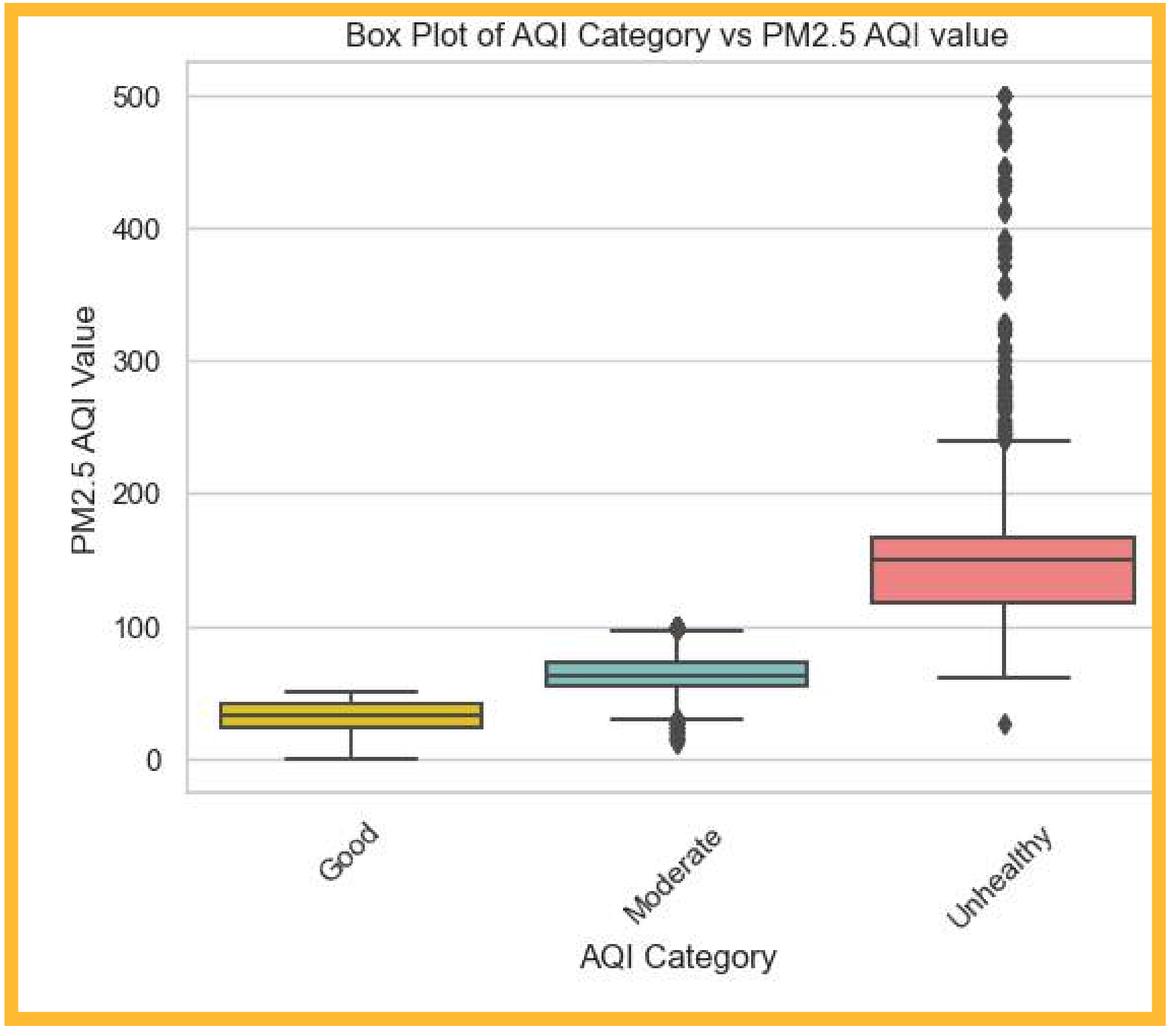
- underscore the urgency of implementing effective policies to improve air quality in these continen

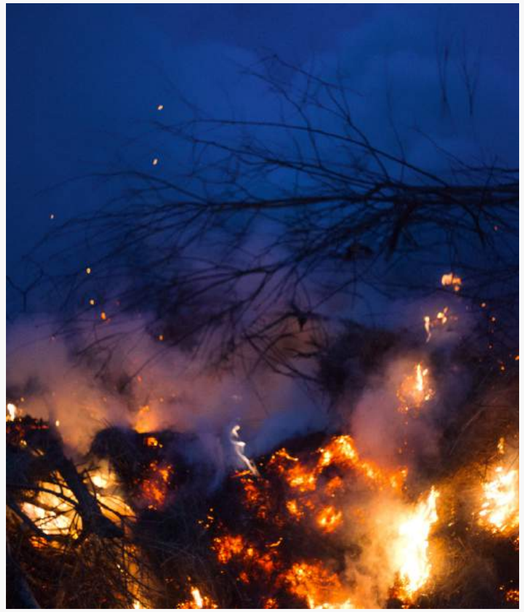


PM 2.5 AQI VALUE VS AQI CATEGORY

When the AQI category goes from Good to Unhealthy;

The median level of PM 2.5 AQI Value increases as the Spearman correlation value indicates the significant relationship.





PM 2.5 AQI VALUE VS AQI CATEGORY

The report also states that **"there is a clear relationship between PM2.5 levels and air quality, with higher PM2.5 levels associated with worse air quality."**

- World Health Organization (WHO) -

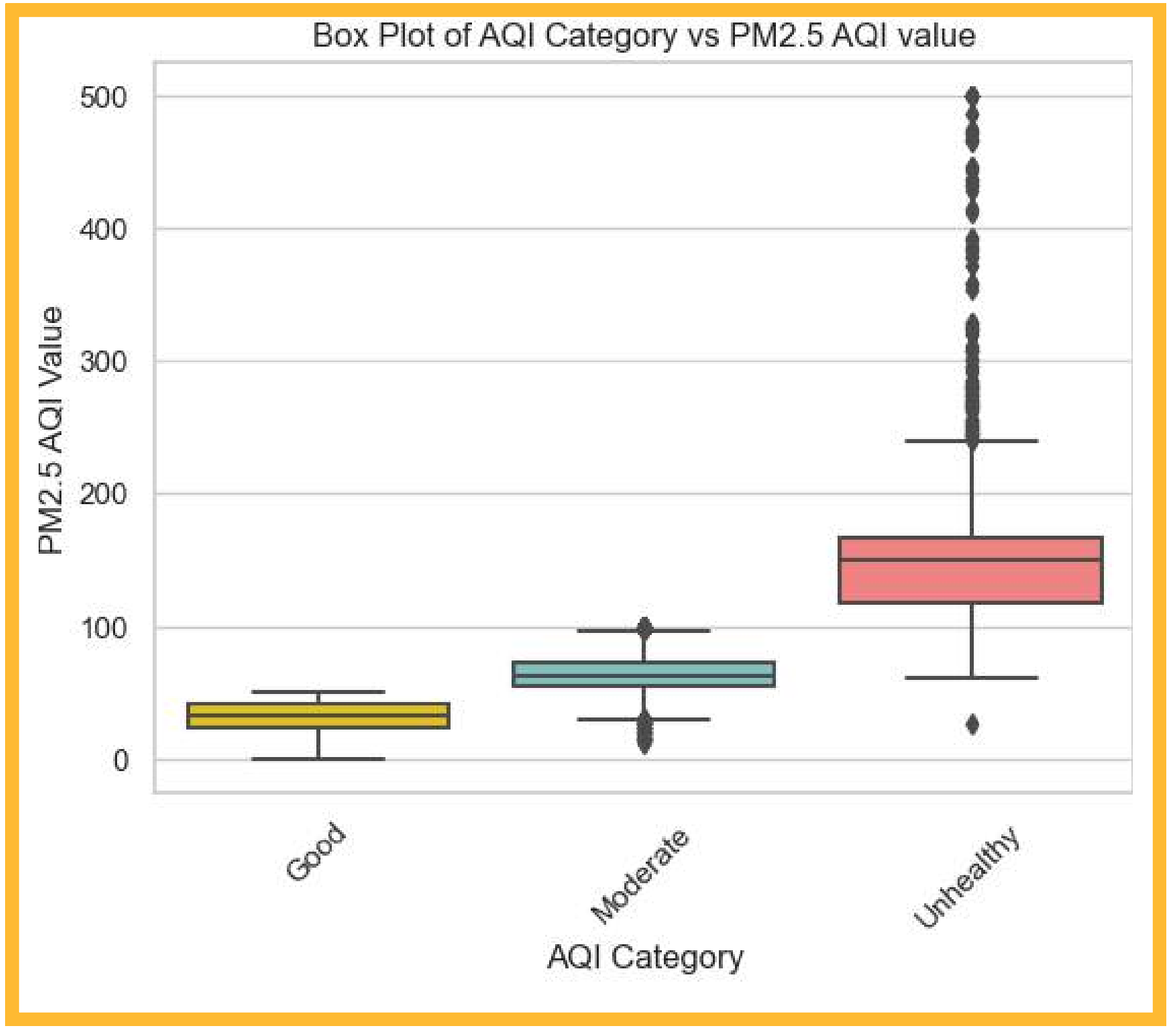
Sensitive groups, such as children and the elderly, should take steps to reduce their exposure, like staying indoors on days when there is Hazy or smoggy air because it may be due to high PM 2.5 levels

PM 2.5 AQI VALUE VS AQI CATEGORY

- **Lot of outliers in Unhealthy Category**
- **Investigate & Identified**
  Around 60% of those outliers are from India.

Air pollution is a major problem in India, and it is caused by a variety of factors, including industrial pollution, traffic pollution, and agricultural pollution.

- Wikipedia -



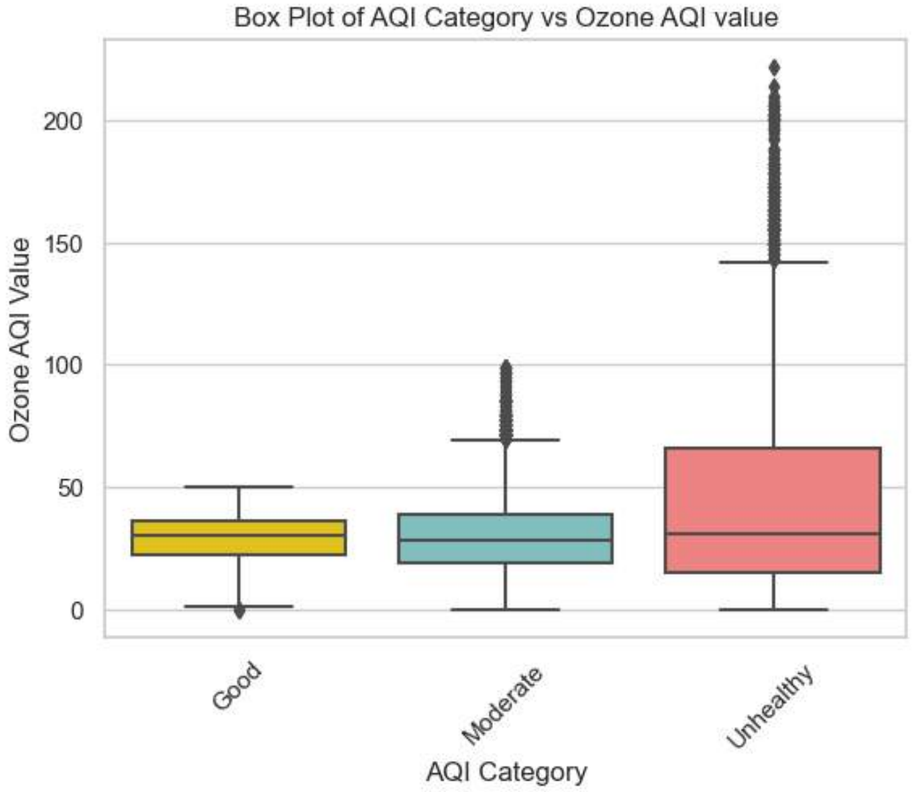
Ozone AQI VALUE VS AQI CATEGORY

Cannot see significant differences in the median level of the Ozone AQI Value for 3 AQI categories like in previous.

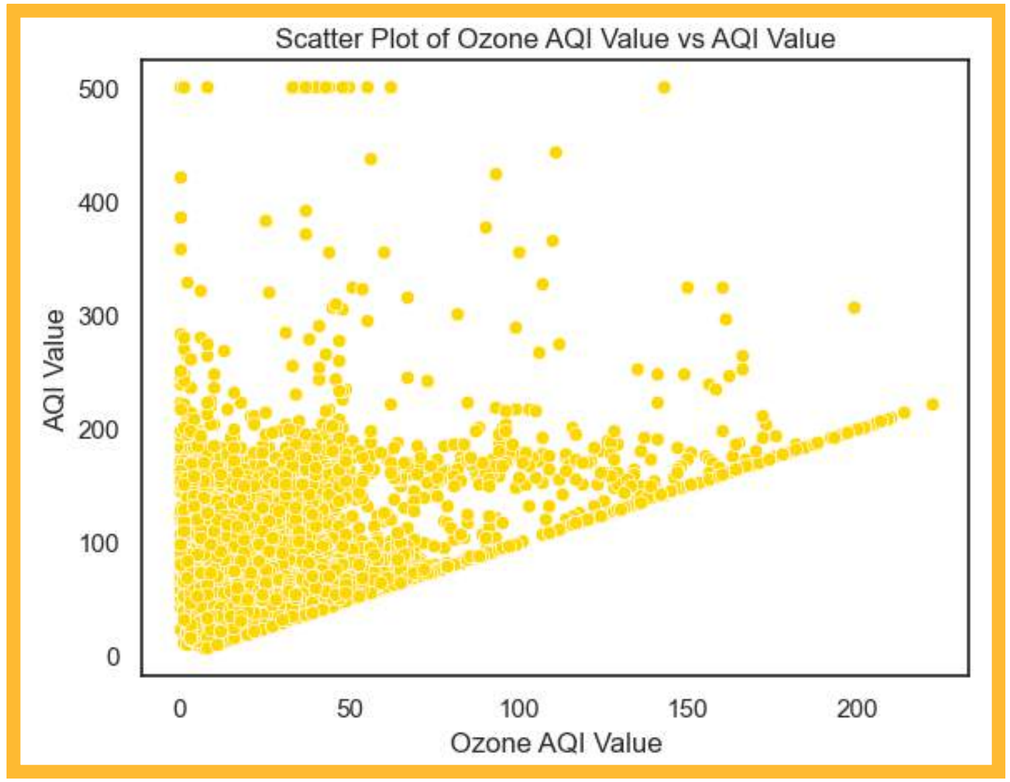
Ground-level ozone (ozone) is the main ingredient in smog. Breathing in unhealthy levels of ozone can increase the risk of health problems.

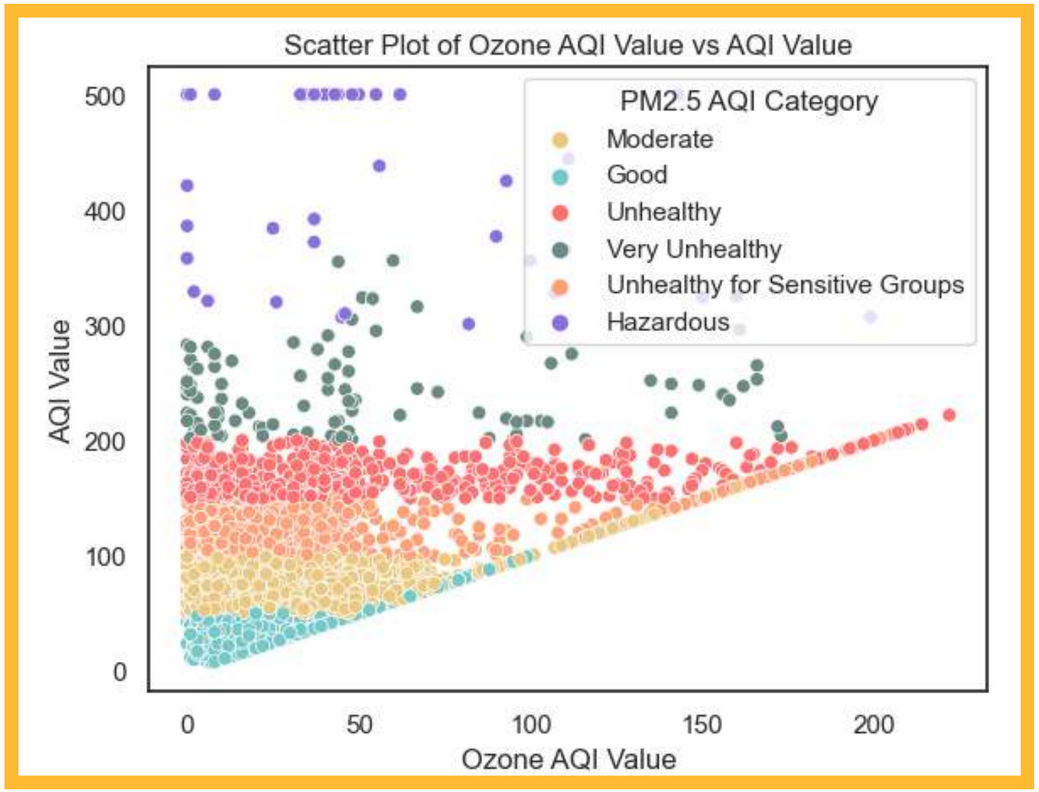
- New York State Department of Health -

Our findings do not match with the research papers!!



Ozone AQI VALUE VS AQI CATEGORY : Deeper Inside

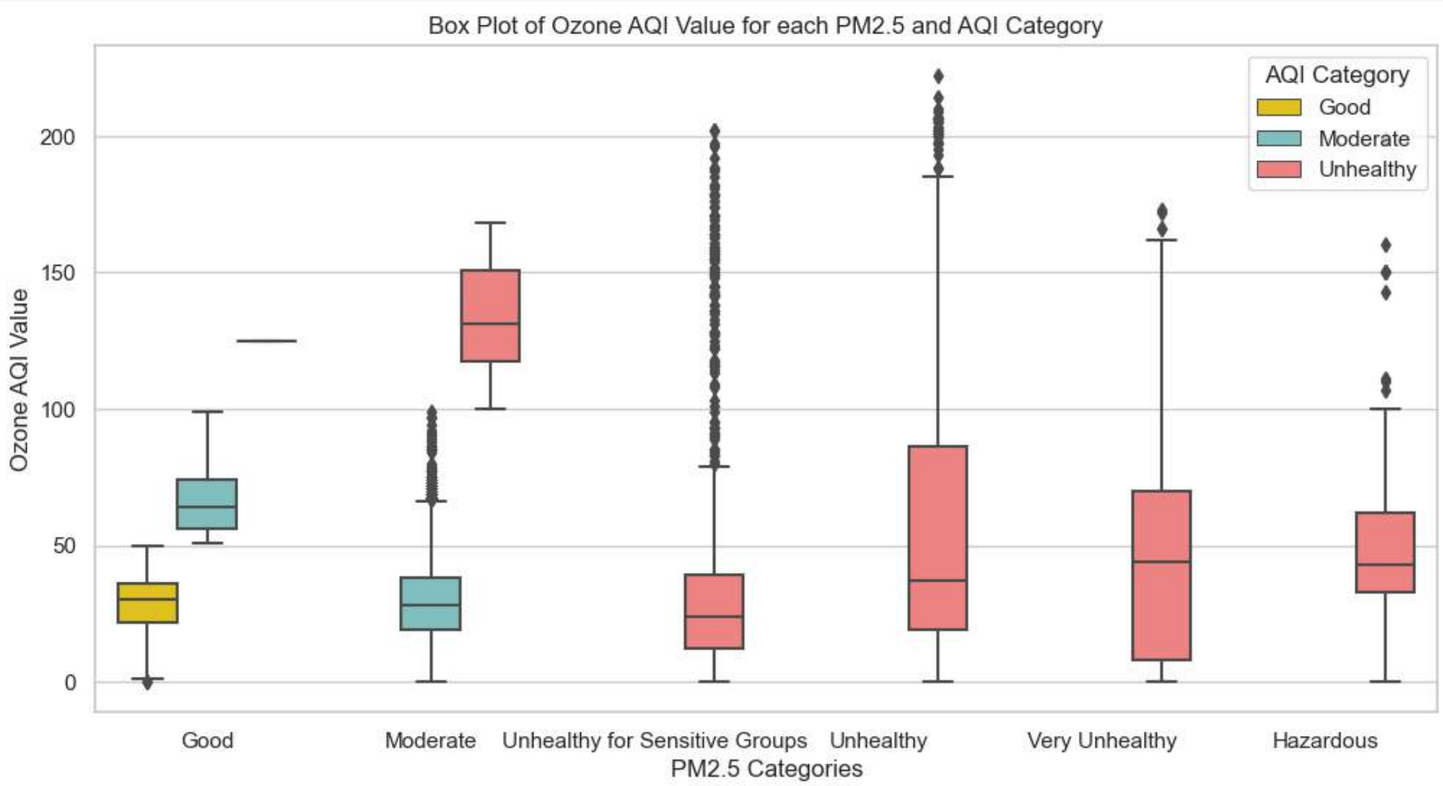




- Although some observations have very low Ozone AQI values they have very high AQI Values which reflects the Unhealthy Air category

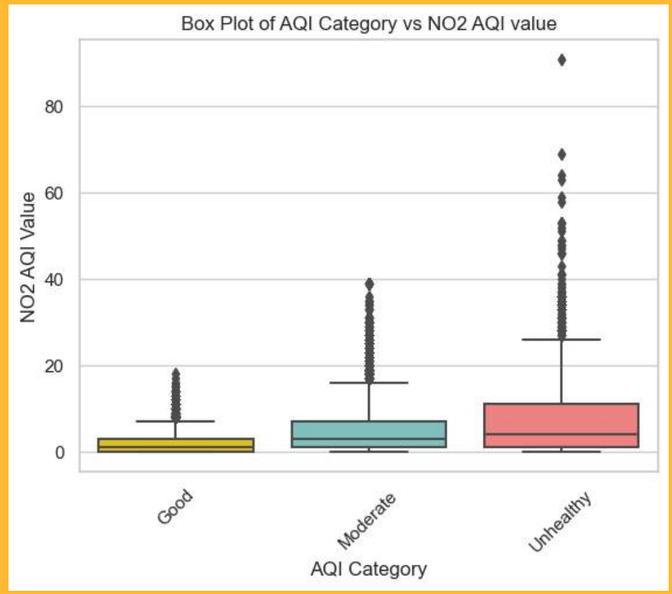
- Although the observations belong to very low Ozone AQI values they are having unhealthy PM2.5 levels

Ozone AQI VALUE VS AQI CATEGORY : Deeper Inside

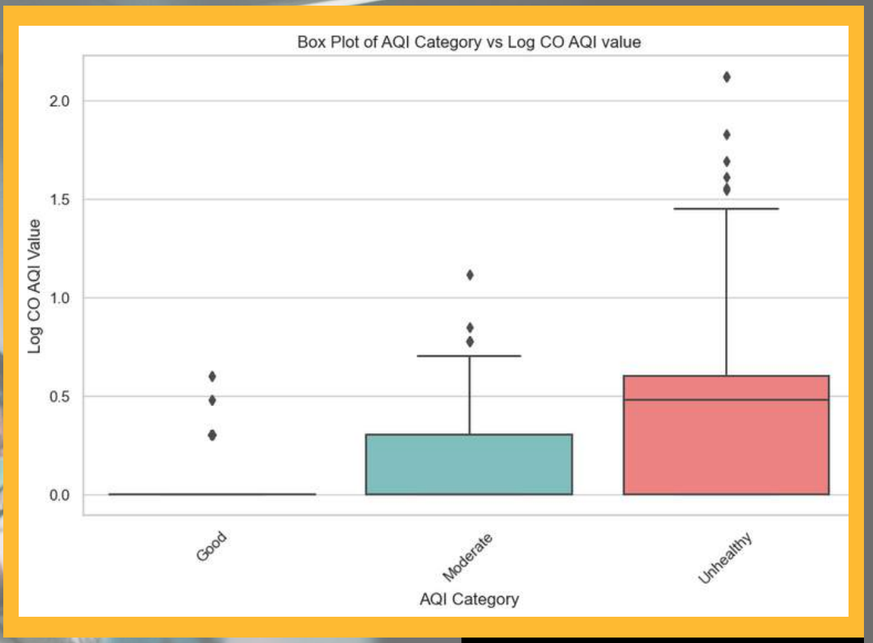


- For Good and Moderate PM 2.5 categories when the AQI Category goes from Good to Bad the median value of Ozone is also Increases.
- However, for higher levels of PM2.5, all observations belong to the Unhealthy AQI category due to the stronger relationship between PM2.5 and AQI category.

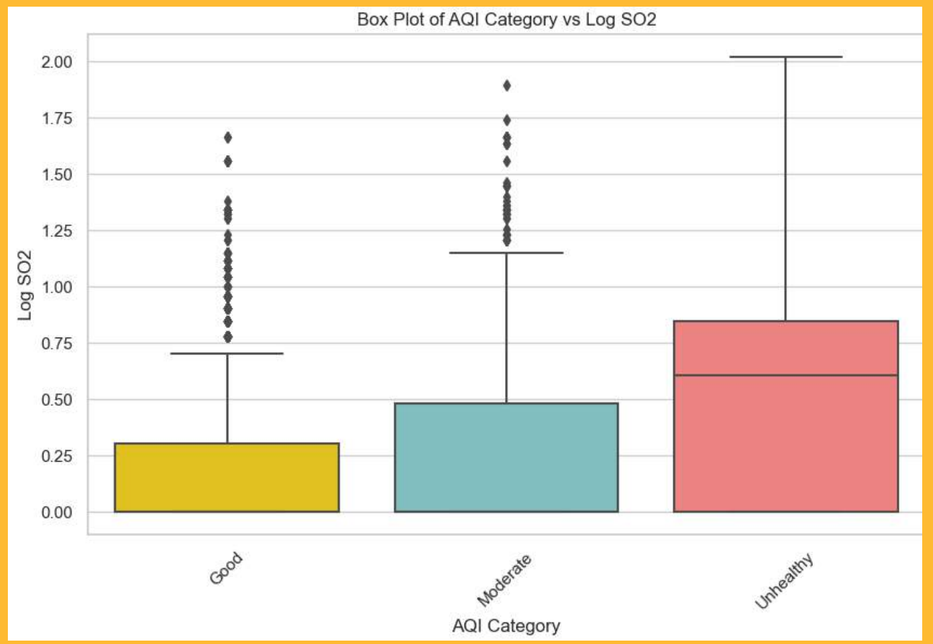
SIMILAR KIND OF RESULTS FOR OTHER VARIABLES...



NO2



CO



SO2

MAJOR FINDINGS OF DESCRIPTIVE ANALYSIS

ASIA AND AFRICA CONTINENTS HAVE WORST AIR QUALITY

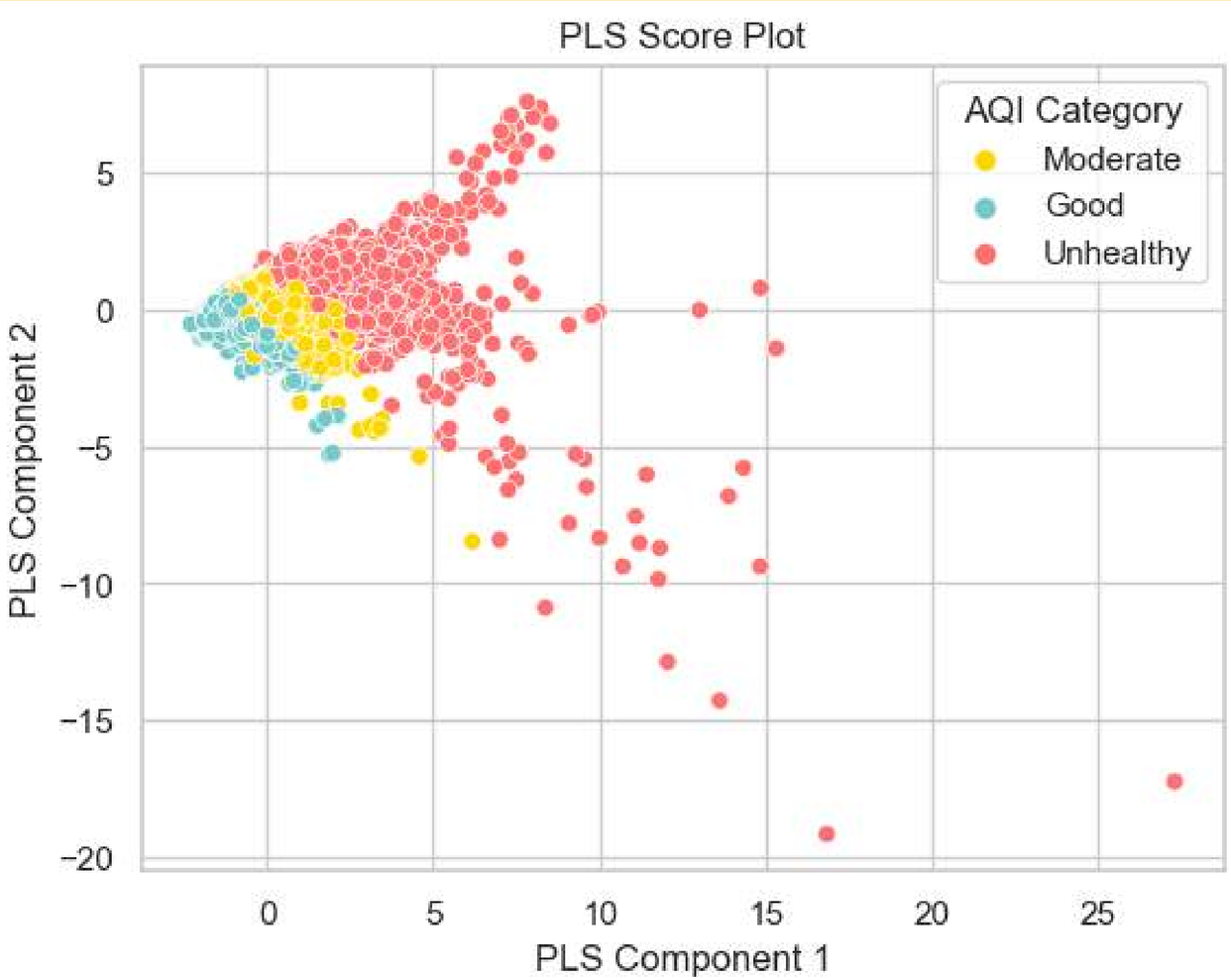
PM2.5 AQI VALUE HIGHLY CORRELATED

WITH AQI CATEGORY

PRESENCE OF MULTICOLINEARITY

FURTHER ANALYSIS

Partial Least Squares Analysis



- **49.3% of the variance is explained by the first two PLS components.**
- **the observations in the score plot were colored with respect to AQI category.**
- **moderately seperable clusters were identified.**
- **Can be considered some linear classification algorithms.**

Partial Least Squares Analysis

Predictors which show a significant association with AQI category

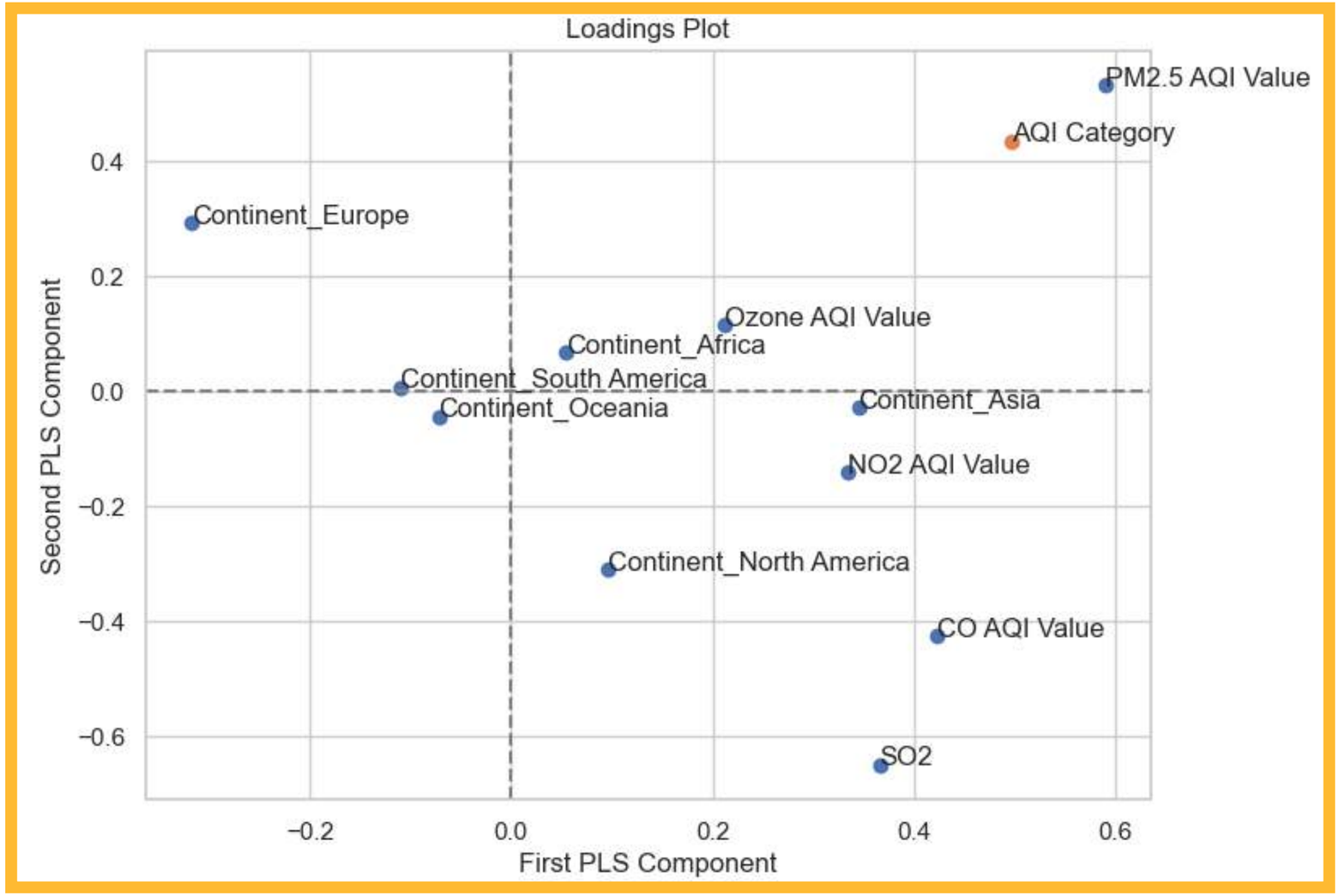
- **PM2.5 AQI value**

This has proven by the descriptive analysis.

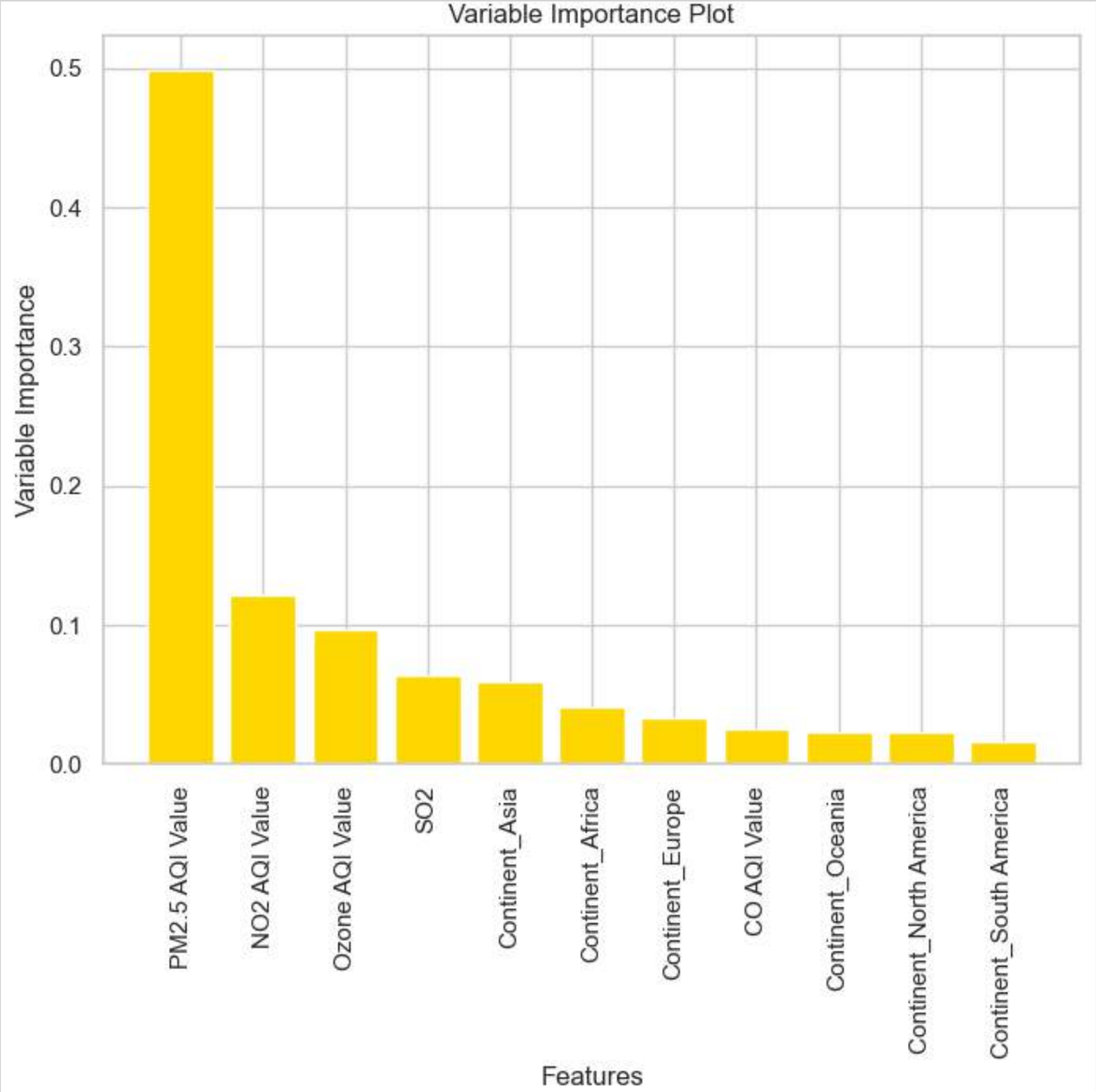
- **Ozone AQI value**

However, Spearman's correlation showed a very low value (0.02).

- **Continent Africa**
- **Continent Asia**



Variable Importance Plot



PM2.5 SHOWS A MASSIVE IMPORTANCE TO THE AQI CATEGORY.

This was proven by the descriptive analysis for multiple times.

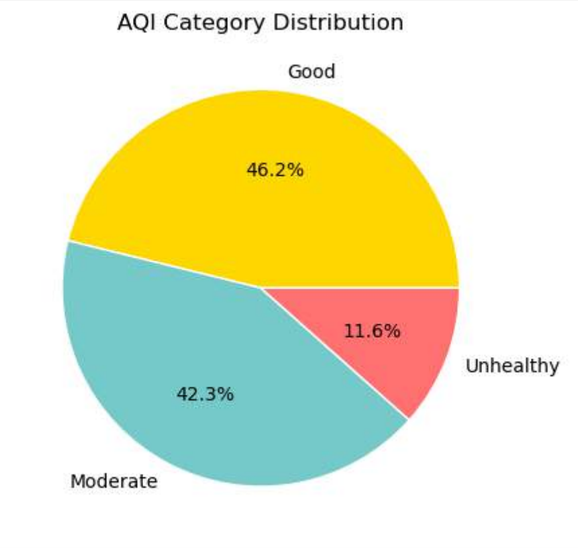
OZONE SHOWS A MODERATE IMPORTANCE TO THE CLASSIFICATION.

This was not identified from the Spearman's Correlation(0.02) but identified from the three variate boxplot graph.

CONTINENTS ASIA AND AFRICA SHOWS HIGHER IMPORTANCE THAN THE OTHER CONTINENTS.

This have been proved by the descriptive analysis also.

SUGGESTIONS FOR THE ADVANCED ANALYSIS



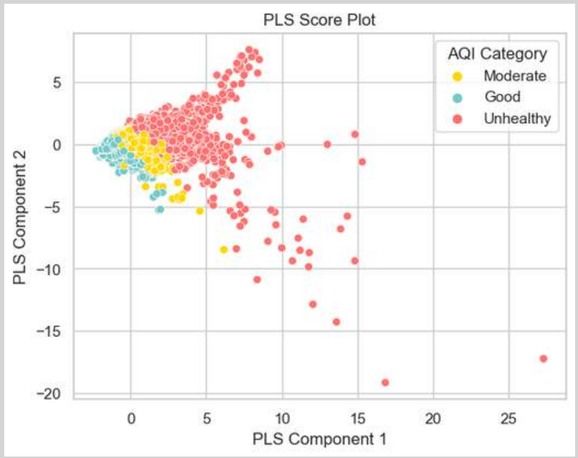
- Class **'Unhealthy'** constitutes a relatively lower percentage, accounting for only 11.6% of the total observations.
- To ensure a more balanced representation of classes;

  SMOTE,

  Upsampling

  Class Weighting

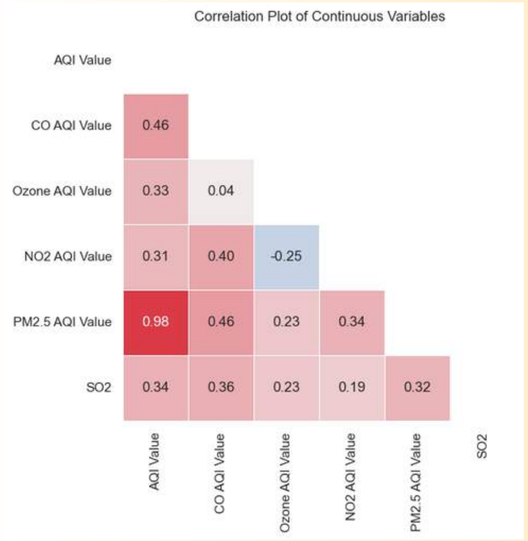
  can be employed.



- Separation between categories in the data is not very distinct.

- But still can consider some Linear classification algorithms, such as **ordinal logistic regression**, **ordinal support vector machines (SVM)**, and **linear discriminant analysis (LDA).**

SUGGESTIONS FOR THE ADVANCED ANALYSIS



- There exists moderate **multicollinearity** between several explanatory variables.
- Applying regularization techniques like **Lasso or Ridge to the ordinal logistic regression models** along with high-performing machine learning algorithms, such as **Random Forest** and the Boosting algorithm **XGBoost** can help mitigate multicollinearity issues.

- Evaluate non-linear algorithms like decision trees or neural networks if linear models are not sufficient.

THANK YOU!

Contribution

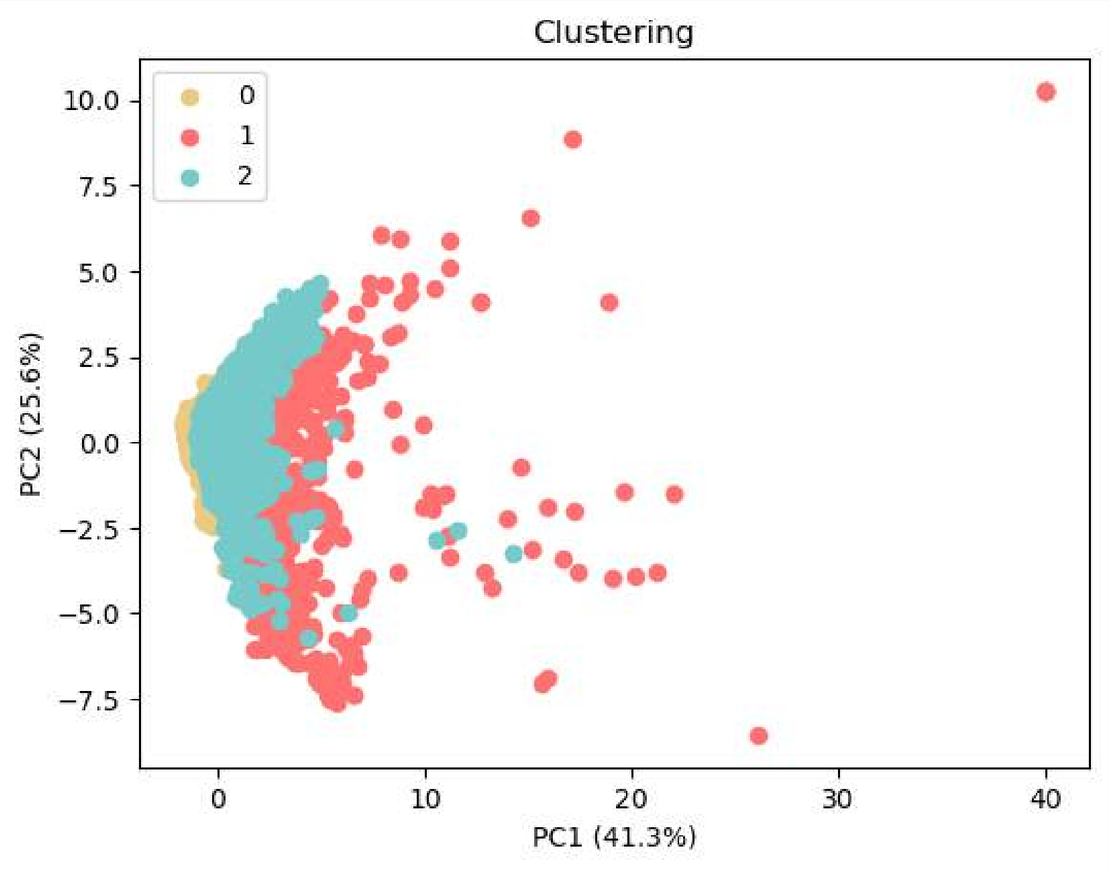
Samujitha	• Part of Descriptive Analysis • Web-Scrapping • Background Research	• **https://www.visualcapitalist.com/how-air-quality-index-works/** • **https://en.wikipedia.org/wiki/Air_quality_index** • **https://scrapfly.io/blog/web-scraping-with-selenium-and-python/** • **https://www.scrapingbee.com/blog/selenium-python/**
Chamodi	• Descriptive Analysis • Part of Further Analysis • Background Research	• **World Air Quality Report by IQAir** • **World Health Organization's official website** • **New York State Department of Health's official website** • **Wikipedia**
Senuri	• Part of Further Analysis • Preprocessing • Background Research	• **GeoNames WebServices Overview.** • **Available at: https://www.geonames.org/export/ws-overview.html (Accessed: 21 July 2023).** • **Zack Aboulazm (2023) Understanding how the Air Quality Index Works, Visual Capitalist** • **Available at: https://www.visualcapitalist.com/how-air-quality-index-works/ (Accessed: 21 July 2023).**

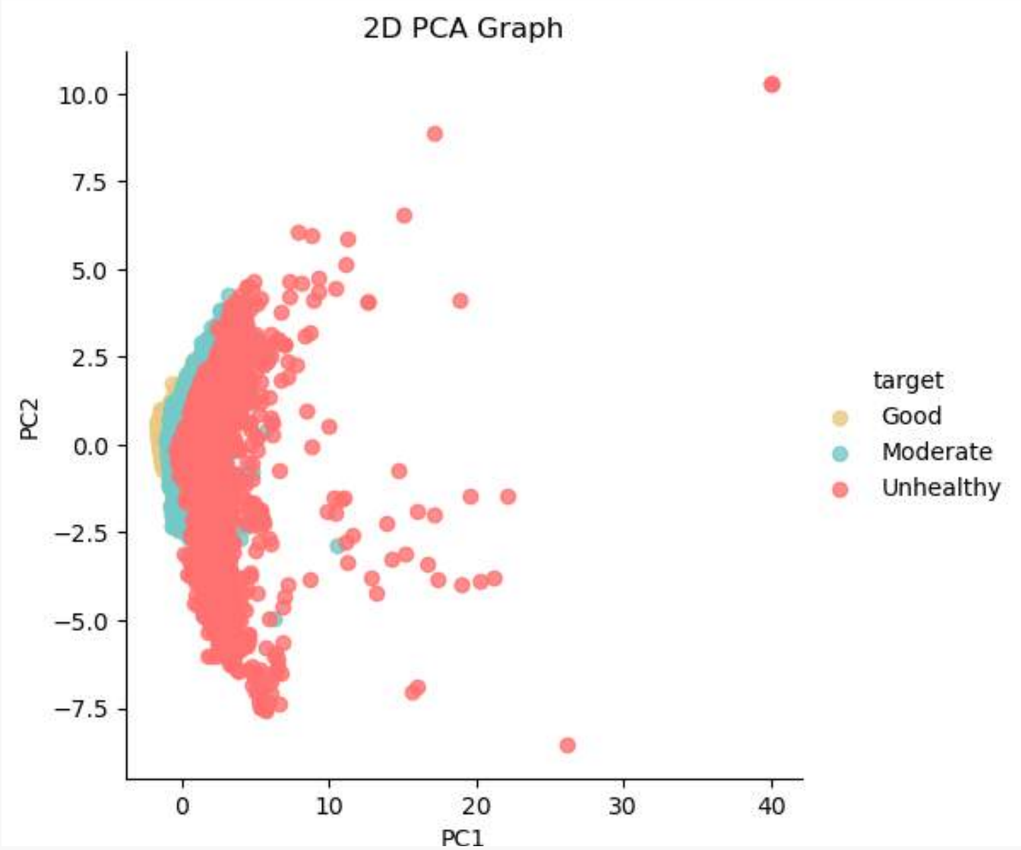
THE PURPOSE OF PREDICTING AQI ?

PORTABLE AIR QUALITY MONITOR

Low-cost & portable air quality devices Anyone can place these aesthetic monitors anywhere they desire or carry them for instant readings. Air Quality Index (AQI) is usually calculated through complex algorithms at an air quality monitoring station. However, the high-end air quality measuring devices condense an entire station into a portable device without punching a hole inside your pockets.

CLUSTERING & PCA





Cluster	0	1	2
AQI Category			
Good	7659	0	49
Moderate	464	0	6590
Unhealthy	1	1356	576

Obtaining the SO_2 data through Web-Scraping

- Given the significance of sulfur dioxide (SO2) as a principal dictator of air quality, the inclusion of SO2 data for each observation becomes very important.
- A web-scraping technique was employed as an effective means of data extraction.
- AccuWeather, which is a reputable source of weather and air quality information, was utilized for the web-scraping task.
- Selenium, a popular tool in web automation, was utilized to interact with the web page dynamically. It allows the program to navigate through the website, simulate user actions such as clicking buttons and filling out forms, and retrieve data from dynamically loaded or JavaScript-rendered elements.
- XPath, on the other hand, is a query language used to navigate through the structure of an XML or HTML document. It provides a way to specify the precise location of elements or data within the web page's HTML structure.
- By using XPath expressions, the web-scraping script can identify the exact location of the required SO2 AQI data on the webpage and extract it accordingly.

AccuWeather

Selenium