AQI VS DEATH RATE PREDICTION

NATTAWAREE PIYARAT

BRAINSTATION – DATA SCIENCE BOOTCAMP

PROBLEM STATEMENT

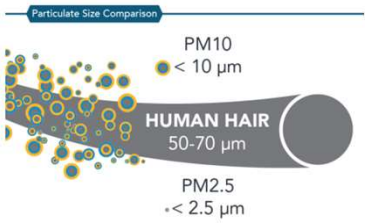
Can we use machine learning to predict how will the air quality index be in the next 5 years?

Is there any relationship between air quality and health (death rate) and what are the factors that impact air quality? Nowadays, we know what will impact air pollution (wildfire, dust, car, manufacturing)

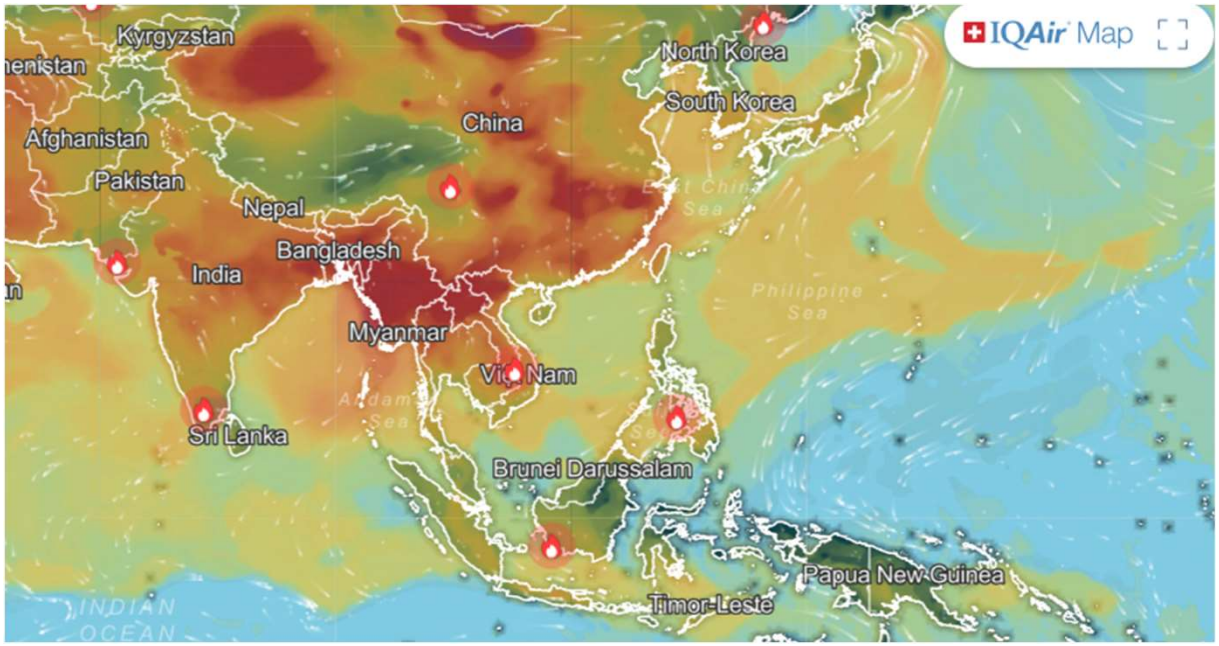
Focus group AQI and number of deaths in **selected Asian countries** and compare between low AQI and high AQI areas.

- Bangladesh
- Bhutan
- India
- Indonesia
- Sri Lanka
- Thailand

AQI (AIR QUALITY INDEX)



US AQI Level	PM2.5 (µg/m³)	Health Recommendation (for 24 hour exposure)
WHO PM2.5 (µg/m³) Recommended Guidelines as of September 22, 2021: **0-5.0**		
Good 0-50	0-12.0	Air quality is satisfactory and poses little or no risk.
Moderate 51-100	12.1-35.4	Sensitive individuals should avoid outdoor activity as they may experience respiratory symptoms.
Unhealthy for Sensitive Groups 101-150	35.5-55.4	General public and sensitive individuals in particular are at risk to experience irritation and respiratory problems.
Unhealthy 151-200	55.5-150.4	Increased likelihood of adverse effects and aggravation to the heart and lungs among general public.
Very Unhealthy 201-300	150.5-250.4	General public will be noticeably affected. Sensitive groups should restrict outdoor activities.
Hazardous 301+	250.5+	General public at high risk of experiencing strong irritations and adverse health effects. Should avoid outdoor activities.



Picture from IQAir website

PM 2.5 meaning: Fine particulate matter is defined as particles that are 2.5 microns or less in diameter

DATA SCIENCE

- The machine learning approach will be using historical AQI data, and historical death causes data. And predict the relationship between air quality and the number of deaths which is assumed that caused by high AQI.
- Approach
  - Logistic Regression 
  - Time series

The results of this project should raise awareness of air pollution and reduce the number of deaths of people by at least 1%, especially in the concerned areas.

DATA SET

Dataset from WHO database (Air quality index, number of deaths by cause, pollution, dust, wildfire)

Currently, I reference to datasets below.

- Air quality index year 2000 – 2019 from WHO
  - Country, City
  - Year
  - Number of PM10, PM2.5, NO2
- Number of deaths by cause year 2016 – 2019 from WHO
  - Country
  - Year
  - Age group
  - Cause of death
  - Number of deaths, death rate
- Additional dataset

Study fact sheets and find some datasets about the causes of pollution and how they create bad air quality.

DATA DICTIONARY

who_ambient_air_quality

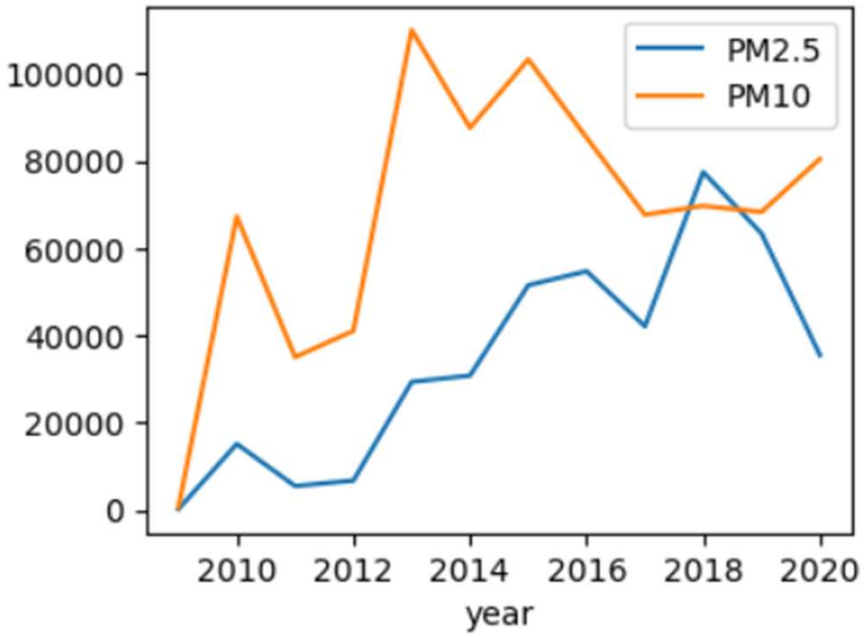
Column name	*Description*	*Datatype*
who_region	WHO region (1_Afr = African region, 2_Amr = Region of the Americas, 3_Sear = South-East Asian region, 4_Eur = European region, 5_Emr = Eastern Mediterranean region, 6_Wpr = Western Pacific region, 7_NonMS = non-member state) For more information: https://www.who.int/about/who-we-are/regional-offices	String
iso3	ISO country code. See https://www.iso.org/iso-3166-country-codes.html for details.	String
country_name	WHO Country Name	String
city	Name of the city	String
year	Year of the annual mean concentration	String
version	Version of the database when the data was collected and published for the first time (Version 2016 V3; version 2018 V4; version 2022 V5; version 2023 V6)	String
pm10_concentration	Annual mean concentration of particulate matter with diameter of 10 µm or less.	Number
pm25_concentration	Annual mean concentration of particulate matter with diameter of 2.5µm or less.	Number
no2_concentration	Annual mean concentration of nitrogne dioxide	Number
pm10_tempcov	Annual temporal coverage for PM10, on a 100 base (full year= 100; 1=1% of the year)	Number
pm25_tempcov	Annual temporal coverage for PM2.5, on a 100 base (full year= 100; 1=1% of the year)	Number
no2_tempcov	Annual temporal coverage for NO2, on a 100 base (full year= 100; 1=1% of the year)	Number
type_of_stations	Type as station as provided by the Member states, such as residential, background, etc.	String
reference	Originator of the data	String
who_ms	WHO member states (0 = no, 1 = yes)	Number

Death Rate

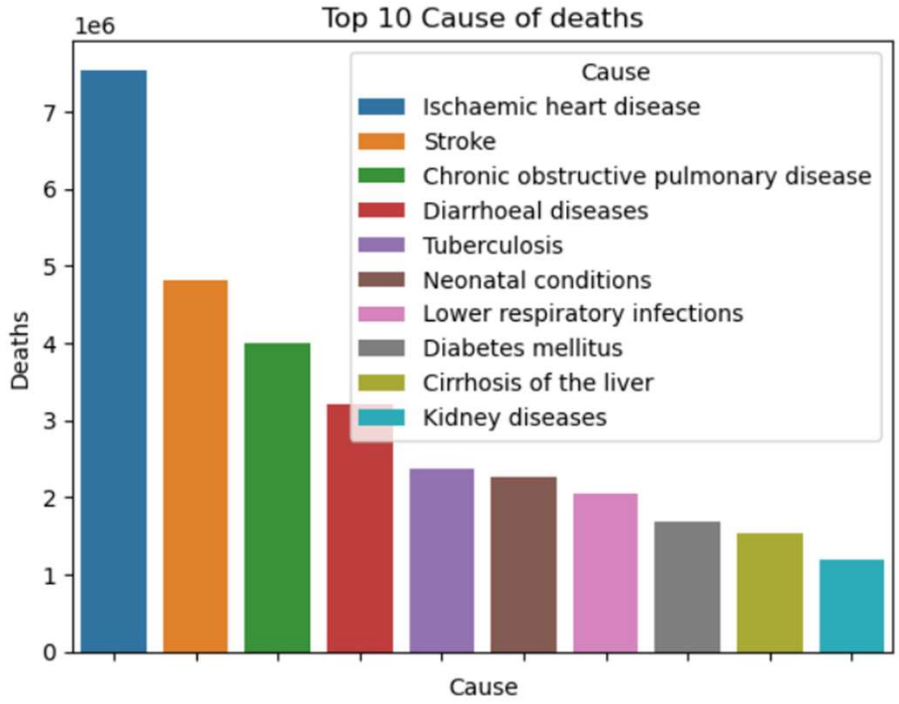
Column name	*Description*	*Datatype*
Code	Cause code	String
Cause	Decease	String
Country	Country	String
ISO3	Country code	String
Year	Year	String
Sex	Sex	String
Age Group	Age Group	String
Population	Population	Number
Deaths	Deaths	Number
Death rate per 100 000 population	Death rate per 100 000 population	Percentage
DALY	DALY	Number
DALY rate per 100 000 population	DALY rate per 100 000 population	Percentage

FINDING

PM 10 peaked during 2013 – 2015 and PM2.5 increased and peaked in 2018.



Chronic obstructive pulmonary might come from air quality also Ischaemic heart disease.



NEXT STEP

- Add more data points in the Deathrate dataset (year, countries)
- Add more datasets if any e.g. causes of pollution
- Using machine learning to predict AQI and Death rate
  - Logistic regression
  - Time series
  - Etc. (in case found proper other models)

REFERENCE

WHO. WHO Ambient Air Quality Database (update 2023). Version 6.0. Geneva, World Health Organization, 2023.