

ANALISIS FEATURE IMPORTANCE DALAM STUDI KASUS TINGKAT PENCEMARAN UDARA DI KOTA KOLKATA DAN VISAKHAPATNAM, INDIA MENGGUNAKAN ALGORITMA XGBOOST DAN RANDOM FOREST

Final Project | Group Fourtunate



ABOUT US



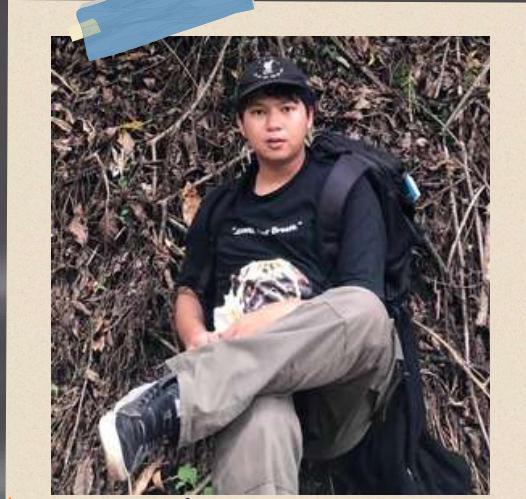
Project Leader

Adinda Nisrina P.H
Universitas Indonesia



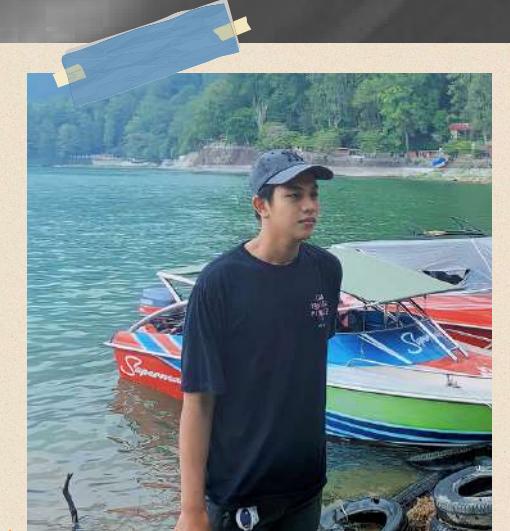
Visualization

Grace Susan Marien
Universitas Negeri Jakarta



Analysts

M. Dhimas Nugraha
Universitas Widyatama



Visualization

Gilang Enggar
Universitas Pembangunan
Nasional Veteran



Analysts

Azzahra Khairunisa
Universitas Negeri Jakarta

PROJECT

Overview



1. Business Understanding
2. Analysis Framework
3. Data Understanding
4. Data Preprocessing
5. Modelling & Evaluation
6. Deployment



BUSINESS UNDERSTANDING



ABOUT DATASET

Kota Kolkata dan Visahapatnam merupakan kota yang terdapat di India. Sebagai negara dengan jumlah penduduk mencapai 1,396 miliar per tahun 2020, tidak mengherankan bahwa permasalahan lingkungan seperti polusi udara yang timbul akibat aktivitas manusia juga turut semakin tinggi. Dimana, di tahun 2020 tingkat polusi udara di beberapa kota di India sudah mencapai tingkat yang sangat mengkhawatirkan. Selain itu, kondisi iklim khususnya cuaca juga memiliki pengaruh yang signifikan terhadap penyebaran polusi udara.

BBC NEWS INDONESIA

Banta Pemilu 2024 Indonesia Dunia Viral Liputan Mendalam Majalah

Polusi udara: Harapan hidup jutaan orang di India bisa berkurang sembilan tahun, warga Jakarta hampir lima tahun



ABOUT DATASET

Sebagai kota yang berada di India, Kolkata dan Visakhapatnam juga memiliki potensi mempunyai tingkat pencemaran udara yang tinggi akibat aktivitas manusia maupun iklim. Sehingga, dalam dataset ini terdapat data pengukuran beberapa parameter pencemar udara (PM 2.5, NO2, NH3, CO, SO2, VOC, dan O3) dan juga data terkait dengan musim dilakukannya pengukuran. Dimana, data parameter dan musim tersebut akan berperan sebagai variabel independen.

Sedangkan, untuk variabel dependen berupa AQI Level.

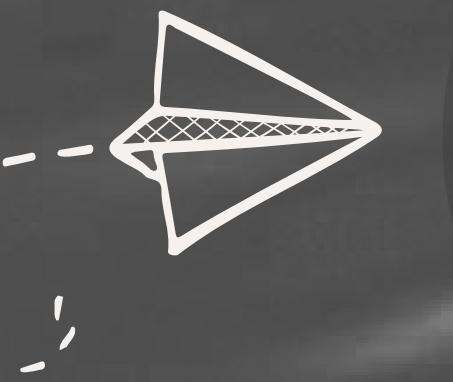
Note : Dataset menggunakan data yang tidak nyata / buatan author

WNI di India ungkap Delhi 'diselimuti kabut asap', aktivitas di luar rumah dikurangi, sekolah dan kampus ditutup



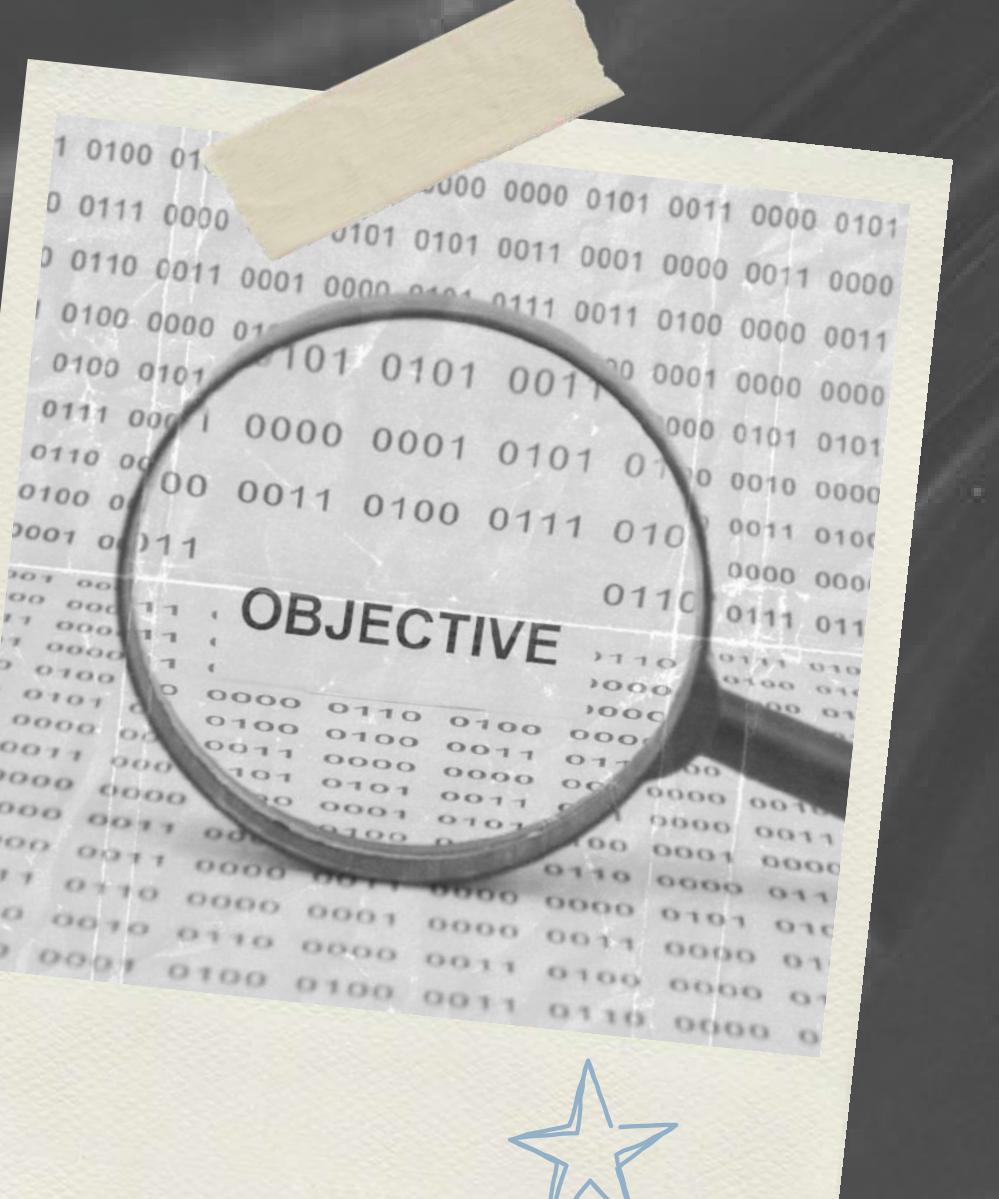
O

BJECTIVE



Menganalisis tingkat pencemaran udara di kota Kolkata dan Visakhapatnam, India, dengan mempertimbangkan beberapa parameter pencemaran udara dan musim dilakukanya pengukuran

Memprediksi feature importance dari parameter pencemar udara dan musim terhadap nilai AQI, yang nantinya hasil prediksi dapat dijadikan acuan dalam merumuskan rekomendasi penanganan yang lebih spesifik terhadap parameter tersebut.



BENEFIT

Mengurangi tingkat pencemaran udara di kedua kota tersebut dengan penanganan yang tepat dan spesifik terhadap sumber



Menjadi referensi bagi peneliti terkait masalah lingkungan polusi udara di India

RESULT

1

Hasil prediksi feature importance (parameter pencemar dan musim) yang berpengaruh terhadap AQI di masing-masing kota

2

Hasil analisis sumber penyebab dari feature importance

3

Dashboard interaktif yang berisi informasi berkaitan dengan data hasil pengukuran parameter pencemar udara di masing-masing kota

4

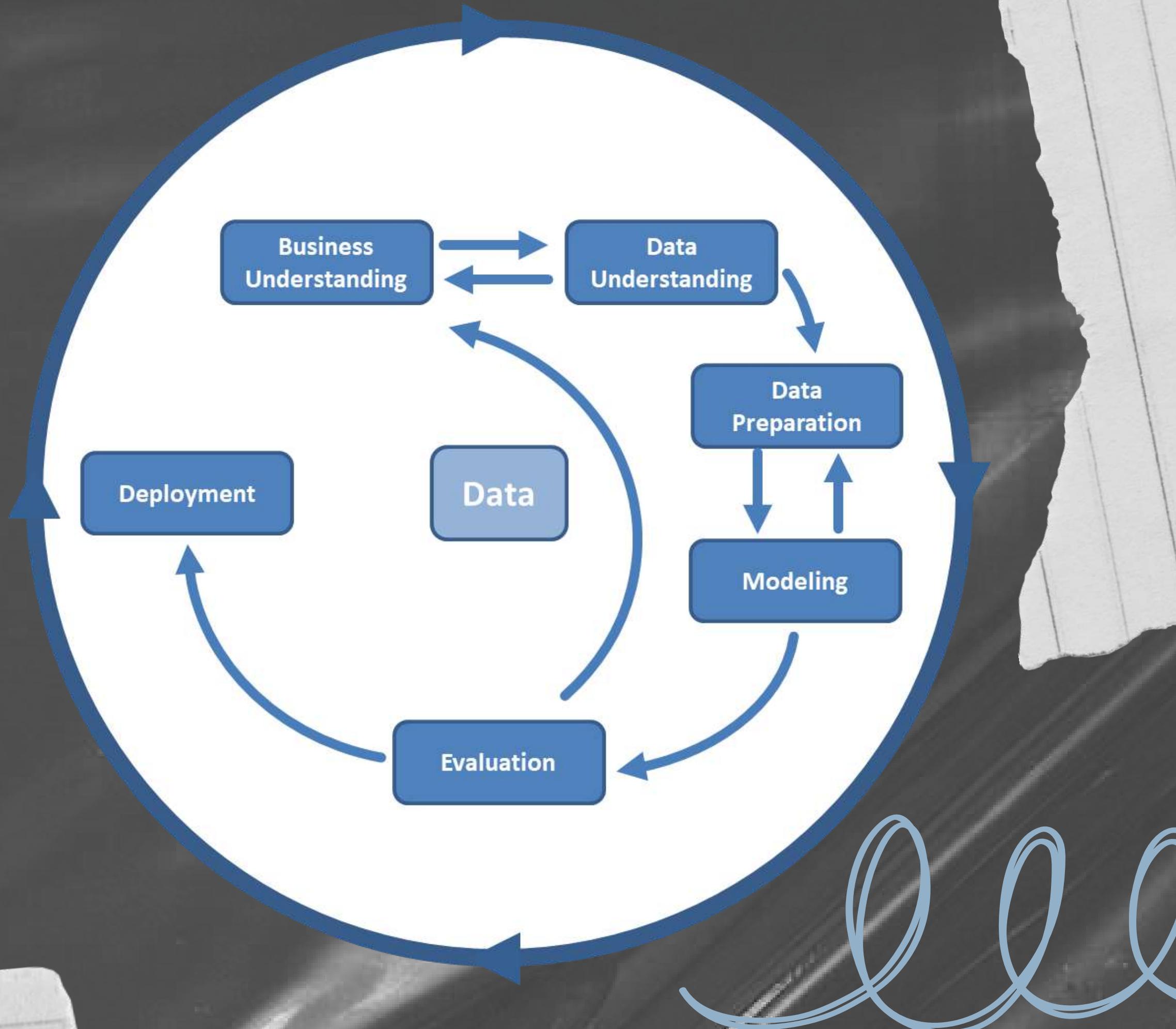
Merumuskan solusi / rekomendasi bagi pemerintah setempat dalam upaya mengurangi dampak pencemaran udara



A N A L Y T I C F R A M E W O

C R I S p - D m

PROCESS DIAGRAM



NO	CRISP - DM	TAHAPAN
1	Business Understanding	<ul style="list-style-type: none"> • Memilih dataset yang berkaitan dengan Air Quality • Menganalisis situasi yang tergambar secara umum dalam dataset • Merumuskan tujuan dan manfaat yang akan dicapai
2	Data Understanding	Menelaah isi data dalam dataset secara kritis dan pahami situasi umum yang digambarkan dalam kumpulan data
3	Data Preparation	<ul style="list-style-type: none"> • Menghilangkan data kota chennai • Menghilangkan data yang tidak diperlukan • Mengisi data null AQI dan AQI level • Encoding season dan AQI Level
4	Modelling & Evaluation	<ul style="list-style-type: none"> • Menerapkan algoritma seperti supervised learning • Mengevaluasi hasil prediksi dan mengidentifikasi faktor - faktor yang paling berpengaruh
5	Deployment	Membuat dashboard interaktif berisi informasi terkait hasil analisis

Signature

DEVELOPMENT TOOLS

kaggle

Tools untuk mencari dataset yang relevan



Tools untuk melakukan data cleansing, data transformation, analysis data, modelling dan evaluation



Tools untuk membuat visualisasi interaktif



DATA UNDERSTANDING

A hand-drawn style graphic where the word 'DATA' is in orange and 'UNDERSTANDING' is in white. A blue wavy line starts from the top of the 'U' and ends at the end of the 'D'. An orange arrow points from the 'D' towards the 'DATA' text.



ABOUT DATASET

Dalam dataset ini terdapat 29 kolom yang memiliki label dan 363 baris data

Air Quality Prediction Cleaned.csv (58.8 kB)

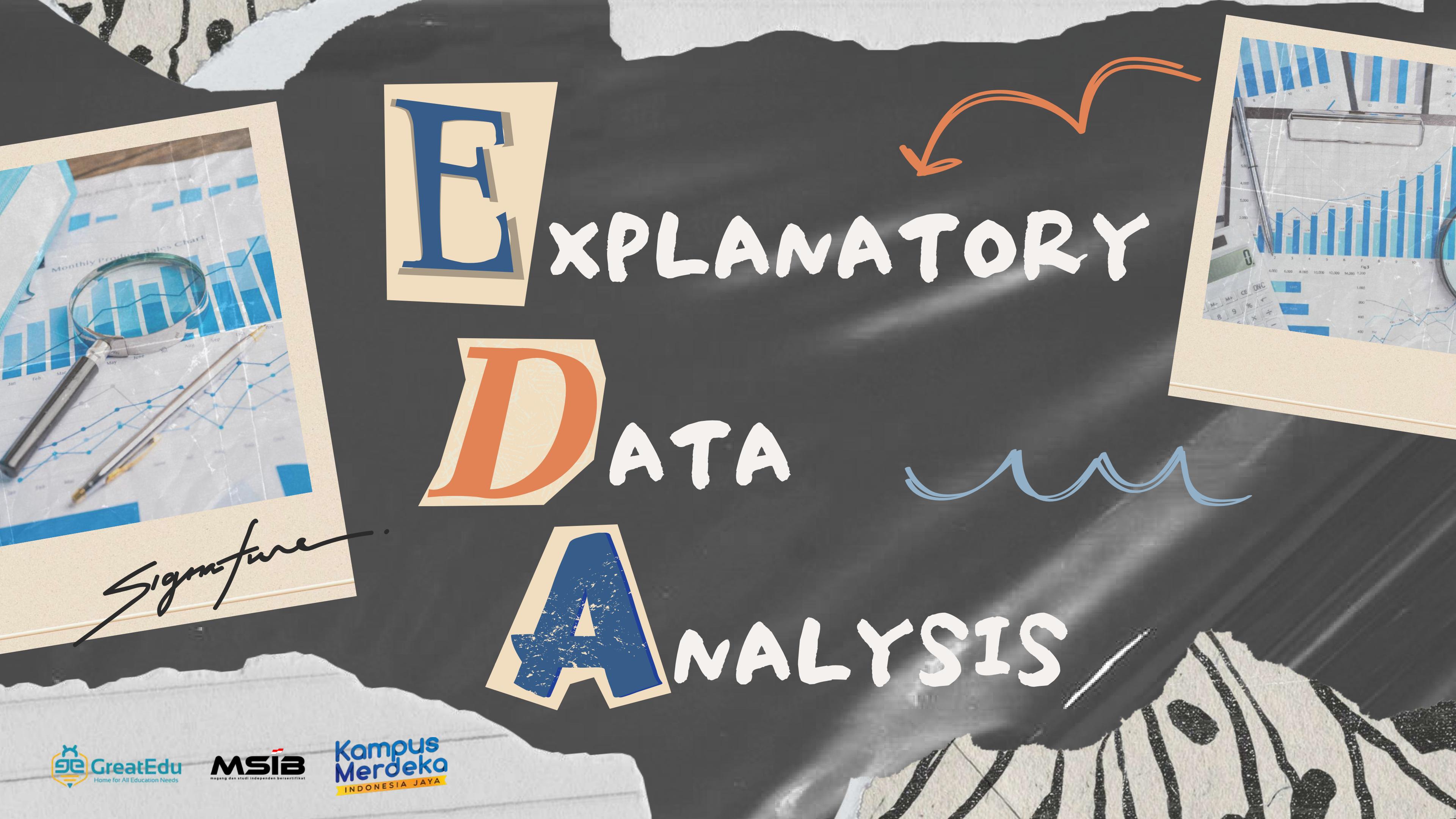
Detail Compact Column 7 of 29 columns

City	Season	Date	# PM2.5	# PM2.5 AQI	PM2.5 AQI
Chennai	spring	33%	242 unique values	6 325	POOR
	winter	17%			SATISFACTOR
	Other (181)	50%			Other (144)
Chennai	Spring	5/1/2022	6.55	27	GOOD
Chennai	Spring	5/2/2022	12.58	52	SATISFACTOR
Chennai	Spring	5/3/2022	25.98	80	SATISFACTOR
Chennai	Spring	5/4/2022	29.88	88	SATISFACTOR
Chennai	Spring	5/5/2022	35.93	102	MODERATE
Chennai	Spring	5/6/2022	37.59	106	MODERATE
Chennai	Spring	5/7/2022	43.69	121	MODERATE
Chennai	Spring	5/8/2022	52.36	145	MODERATE
Chennai	Spring	5/9/2022	58.98	153	POOR
Chennai	Spring	5/10/2022	62.58	155	POOR
Chennai	Spring	5/11/2022	69.12	158	POOR

No	Feature	Penjelasan	Tipe Data
1	S.no	Nomor data	integer
2	City	Kota tempat pengukuran kualitas udara	object
3	Season	Musim yang sedang terjadi pada waktu pengukuran	object
4	Date	Tanggal dilakukannya pengukuran	object
5	Ship Entry / Left	Jenis kedatangan kapal ke pelabuhan	object
6	Type of ship present	Kategori kapal yang datang	object
7	PM 2.5 (Particulate Matter 2.5)	Hasil pengukuran PM 2.5 (Particulate Matter 2.5) di udara	float
8	PM2.5 AQI	Hasil perhitungan nilai AQI untuk parameter PM 2.5 (Particulate Matter 2.5)	integer
9	PM2.5 AQI CAT	Kategorisasi nilai AQI parameter PM 2.5 (Particulate Matter 2.5)	object
10	NO2 (nitrogen dioksida)	Hasil pengukuran NO2(nitrogen dioksida) di udara	float
11	NO2 AQI	Hasil perhitungan nilai AQI untuk parameter NO2(nitrogen dioksida)	integer
12	NO2 AQI CAT	Kategorisasi nilai AQI parameter NO2 (nitrogen dioksida)	object
13	NH3 (Amonia)	Hasil pengukuran NH3 (Amonia) di udara	float
14	NH3 AQI	Hasil perhitungan nilai AQI untuk parameter NH3 (amonia)	integer
15	NH3 AQI CAT	Kategorisasi nilai AQI parameter NH3	object

No	Feature	Penjelasan	Tipe Data
16	CO	Hasil pengukuran konsentrasi CO (karbon monoksida) di udara	float
17	CO AQI	Hasil perhitungan nilai AQI untuk parameter CO (karbon monoksida)	integer
18	CO AQI CAT	Kategorisasi nilai AQI parameter CO (karbon monoksida)	object
19	SO2	Hasil pengukuran konsentrasi SO2 (sulfur dioksida) di udara	float
20	SO2 AQI	Hasil perhitungan nilai AQI untuk parameter SO2 (sulfur dioksida)	integer
21	SO2 AQI CAT	Kategorisasi nilai AQI parameter SO2 (sulfur dioksida)	object
22	O3	Hasil pengukuran konsentrasi O3 (ozon) di udara	float
23	O3 AQI	Hasil perhitungan nilai AQI untuk parameter O3 (ozon)	integer
24	O3 AQI CAT	Kategorisasi nilai AQI parameter O3 (ozon)	object
25	VOC	Hasil pengukuran konsentrasi VOC (volatile organic compound) di udara	float
26	VOC AQI	Hasil perhitungan nilai AQI untuk parameter VOC (volatile organic compound)	integer
27	VOC AQI CAT	Kategorisasi nilai AQI parameter VOC (volatile organic compound)	object
28	AQI	Hasil perhitungan nilai indeks kualitas udara berdasarkan 7 parameter	float
29	AQI LEVEL	Kategorisasi nilai indeks kualitas udara	object

Explanatory DATA ANALYSIS



No	EDA	REASON	JENIS VISUALISASI
1	Analisis deskriptif	Mengetahui analisis deskriptif (mean, median, min-max, standar deviasi) setiap feature	Tabel
2	Detection	Mengetahui apakah ada duplikat pada rows maupun kolom	Tabel
3	Rata-rata nilai AQI / Minggu / Kota	Melihat trend nilai AQI per minggu per kota	Line chart
4	Best AQI City	Mengetahui proporsi AQI Level per kota	Bar Chart
5	Heatmap correlation / feature	Mengetahui korelasi antar feature	Heatmap
	AQI Level per Musim	Mengetahui proporsi AQI level per musim	Bar Chart
7	Boxplot per parameter dan nilai AQI	Melihat nilai outlier	Boxplot

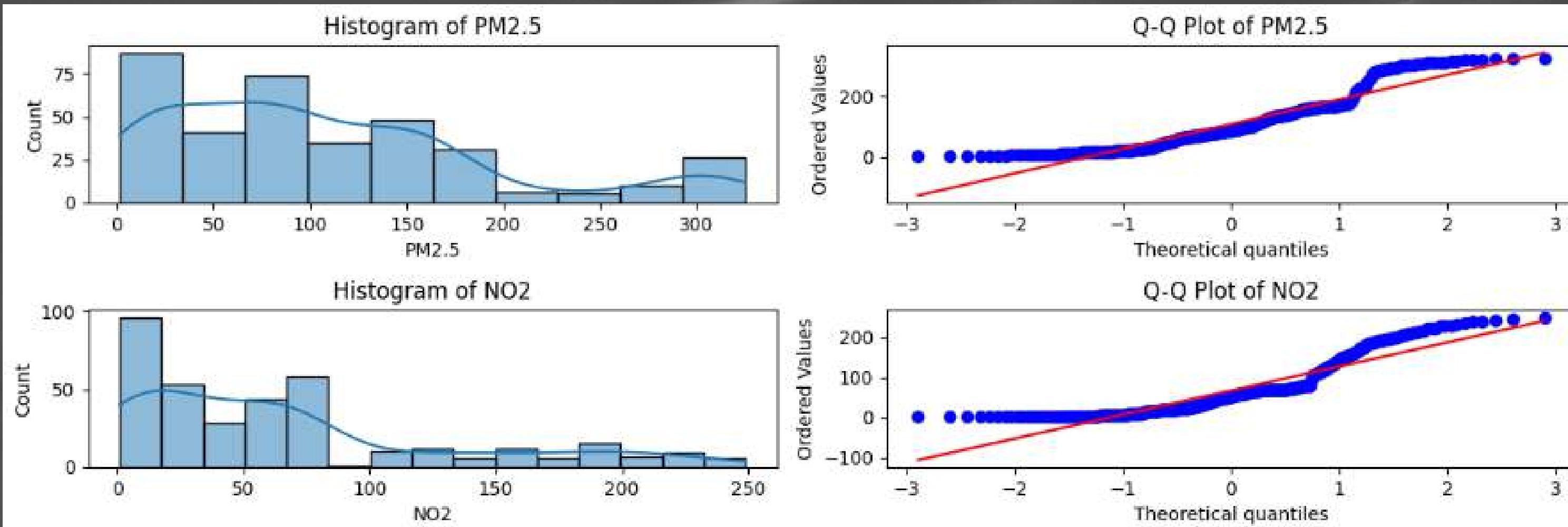
Analisis Deskriptif

	s.no	PM2.5	PM2.5 AQI	NO2	NO2 AQI	NH3	NH3 AQI	CO	CO AQI	S02	S02 AQI	O3	O3 AQI	VOC	VOC AQI	AQI
count	363.000000	363.000000	363.000000	363.000000	363.000000	363.000000	363.000000	363.000000	363.000000	363.000000	363.000000	363.000000	363.000000	363.000000	121.000000	
mean	182.000000	108.387201	168.101928	66.668485	52.374656	28.554559	77.242424	6.381006	65.831956	28.995612	39.787879	34.662391	36.625344	22.954656	204.247934	78.057851
std	104.933312	85.604263	85.577034	64.675558	41.116834	51.746795	79.502574	11.213948	75.064259	22.110893	30.726861	20.076841	30.384509	33.256577	151.497112	27.785032
min	1.000000	1.500000	6.000000	0.900000	1.000000	0.020000	1.000000	0.110000	1.000000	0.210000	0.000000	1.060000	1.000000	0.050000	1.000000	4.714286
25%	91.500000	38.500000	108.500000	16.695000	15.000000	0.570000	6.000000	1.800000	20.000000	12.365000	17.000000	19.010000	18.000000	4.340000	78.000000	59.571429
50%	182.000000	88.590000	168.000000	52.870000	49.000000	4.210000	47.000000	2.800000	33.000000	19.000000	27.000000	32.900000	30.000000	11.360000	166.000000	85.571429
75%	272.500000	156.995000	207.000000	77.945000	76.500000	21.170000	132.000000	7.700000	83.000000	40.215000	56.000000	47.190000	44.000000	33.330000	281.000000	102.714286
max	363.000000	325.400000	424.000000	249.000000	137.000000	236.960000	261.000000	175.000000	293.000000	69.990000	212.000000	83.400000	143.000000	485.000000	515.000000	107.857143

Analisis deskriptif yang disajikan memuat terkait dengan pengukuran karakteristik dataset, yang terdiri atas nilai count, mean, median, standar deviasi, min-max, dan nilai Q1-Q2-Q3.

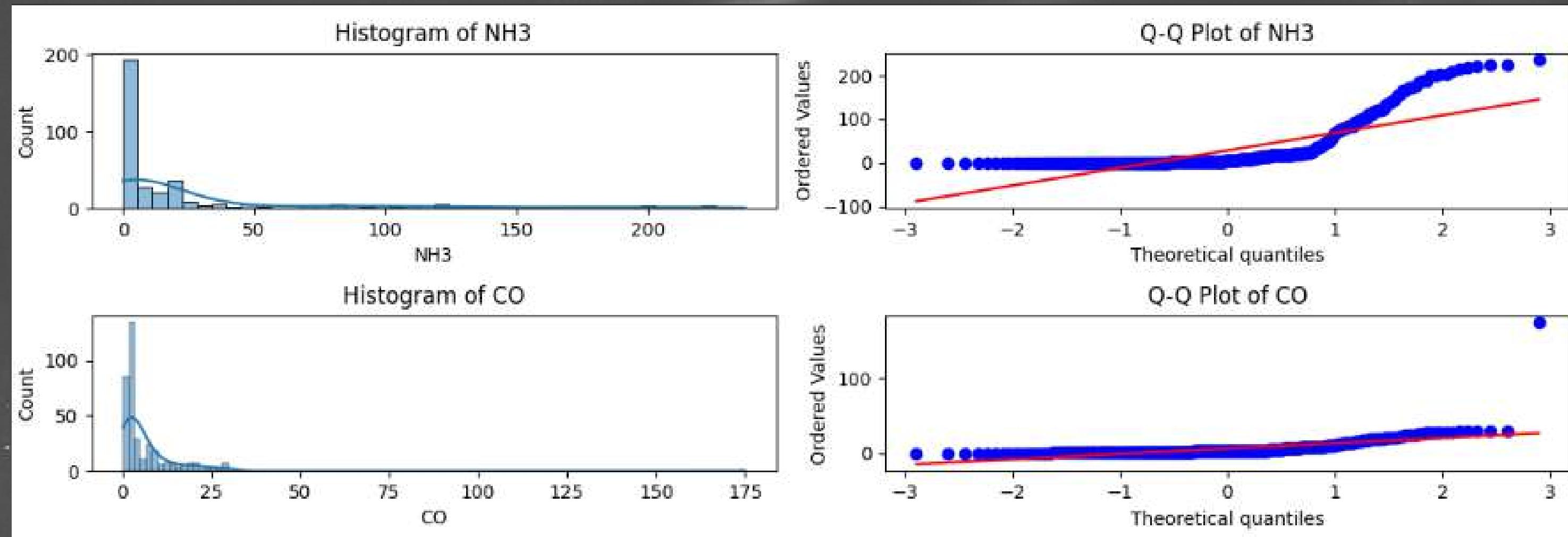
Pada bagian count, semua feature selain AQI memiliki nilai 363 yang mengartikan bahwa jumlah baris sebanyak 363. Sedangkan, pada AQI jumlah baris hanya 121 sehingga terdapat data null.

Analisis Normalitas features



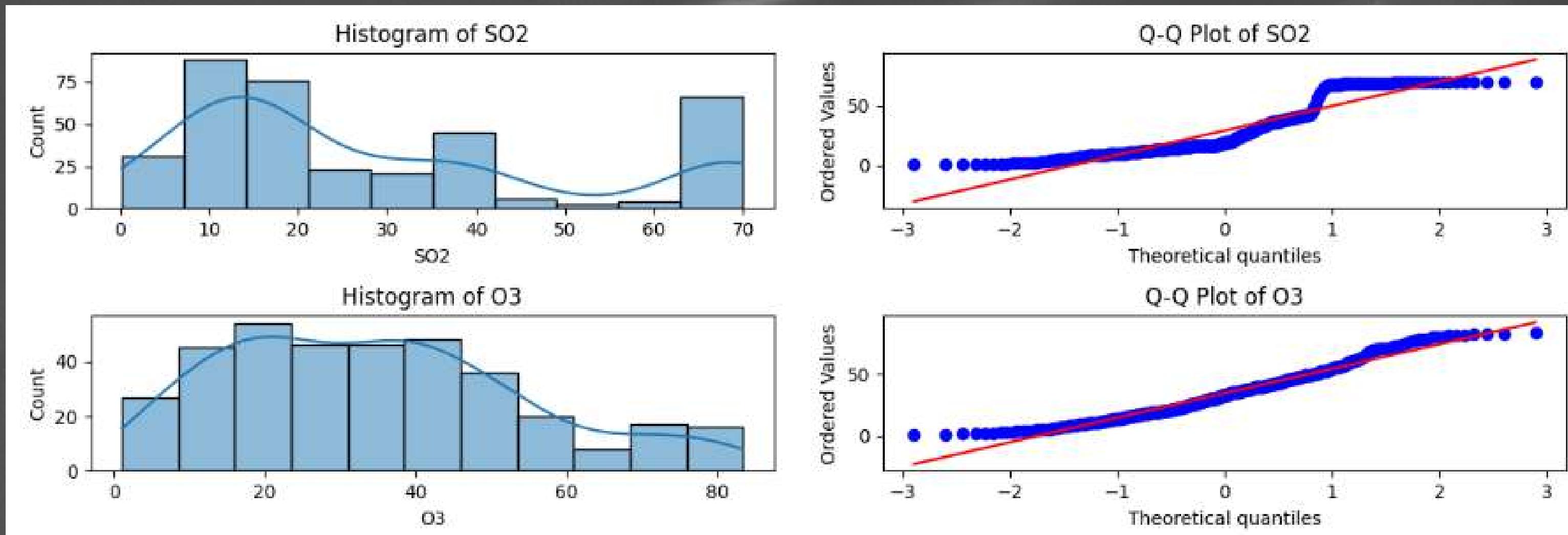
Distribusi data PM2.5 (Particulate Matter 2.5) dan NO2 (Nitrogen Dioksida) **tidak normal**, terlihat dari grafik histogram yang **condong ke kanan (Right-Skewed)** dan banyak titik pada Q-Q plot yang menjauhi garis lurus.

Analisis Normalitas features



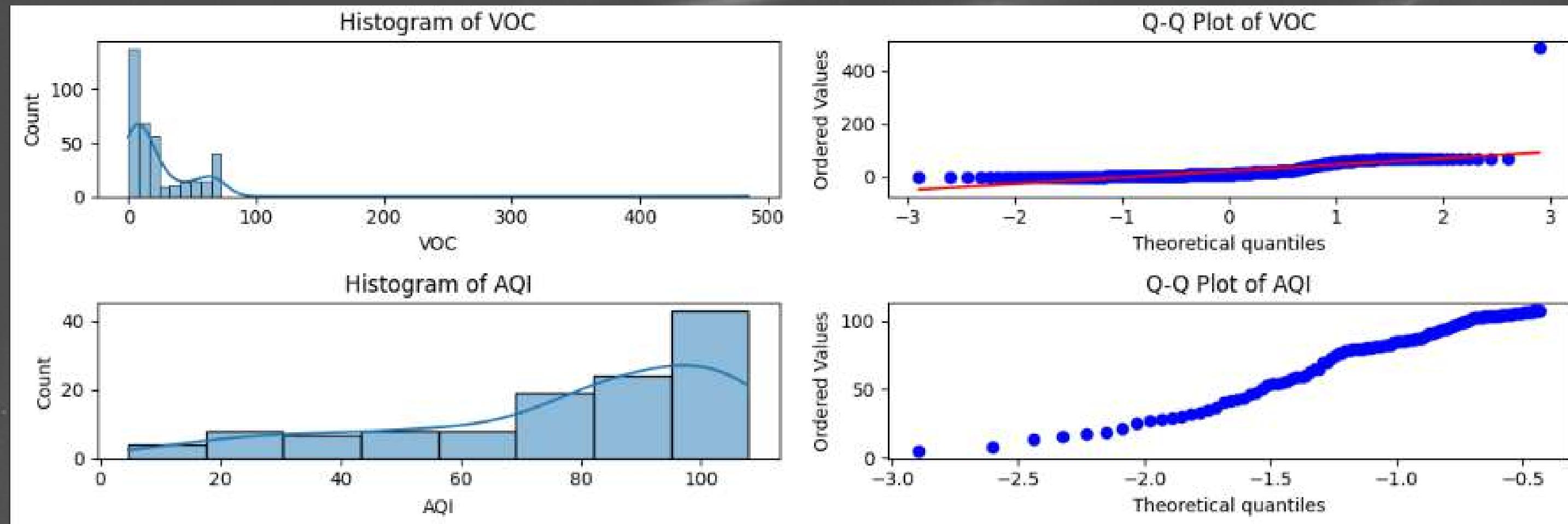
Distribusi data NH3 (Amonia) dan CO (Karbon Monoksida) **tidak normal**, terlihat dari grafik histogram yang **condong ke kanan (Right-Skewed)** dan banyak titik pada Q-Q plot yang menjauhi garis lurus.

Analisis Normalitas features



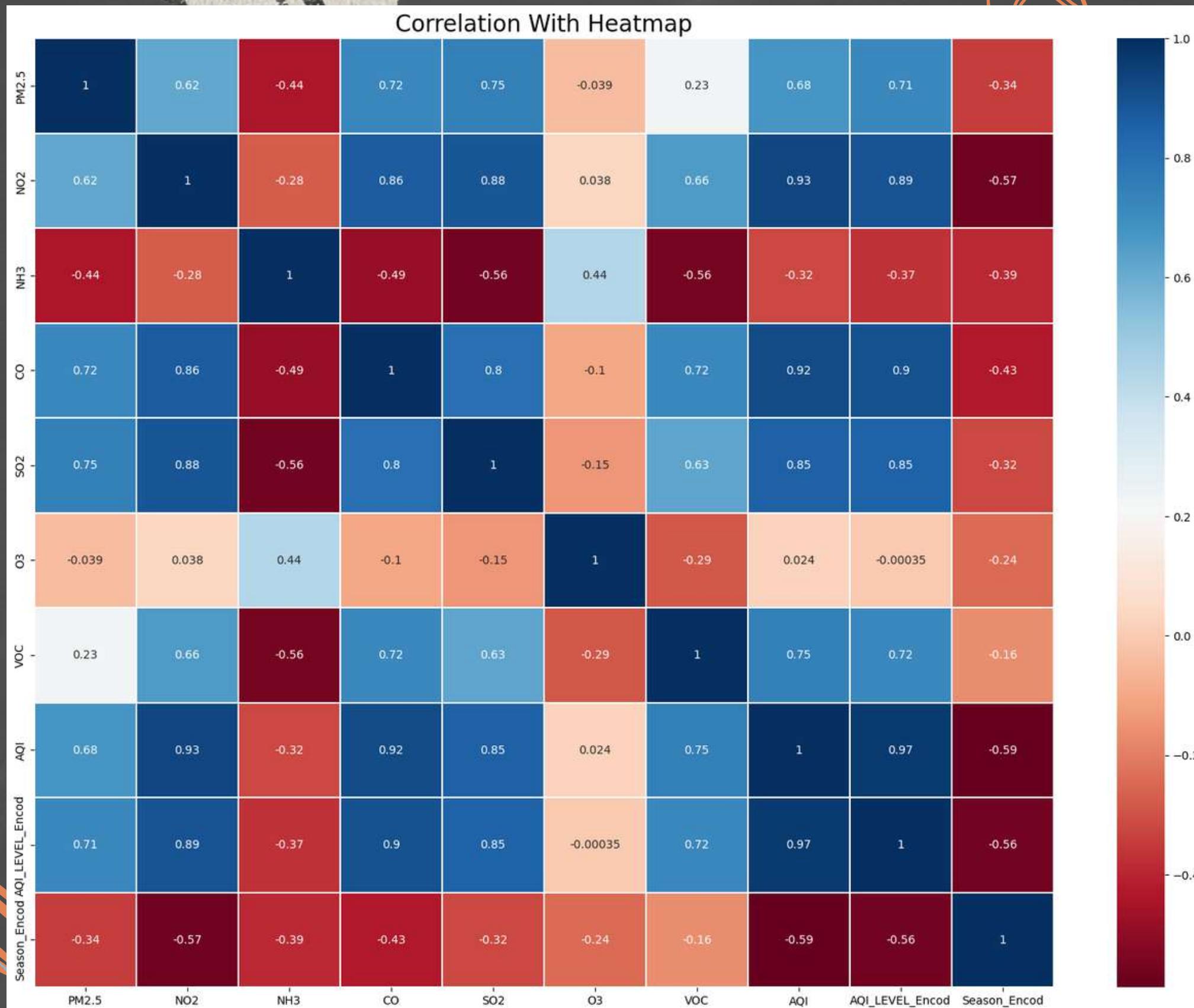
Distribusi data SO2(Sulfur Dioksida) dan O3(Ozon) **tidak normal** karena dilihat dari grafik histogram banyak data yang condong ke kanan (**Right-Skewed**), selain itu juga banyak titik pada Q-Q plot yang menjauhi garis lurus.

Analisis Normalitas features



Distribusi data VOC (Volatile Organic Compounds) **tidak normal**, karena pada grafik histogram banyak data yang **condong ke kanan (Right-Skewed)**, selain itu juga banyak titik pada Q-Q plot yang menjauhi garis lurus. Sedangkan pada AQI (Air Quality Index), distribusi data juga **tidak normal** karena pada grafik histogram banyak data yang **condong ke kiri (Left-Skewed)**, selain itu juga banyak titik pada Q-Q plot yang menjauhi garis lurus.

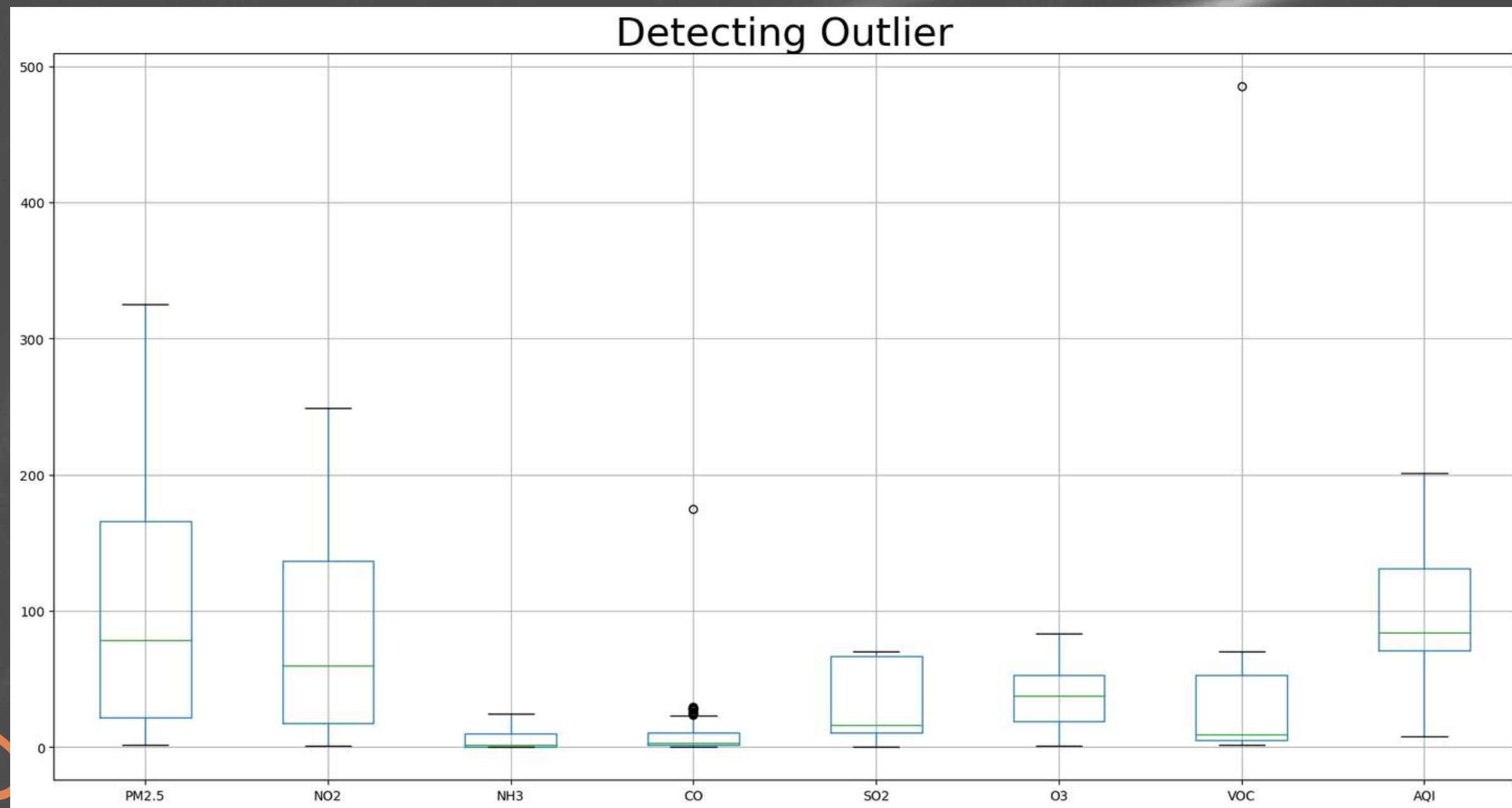
Heatmap Korelasi Antar Features



Feature yang memiliki korelasi positif sangat kuat terhadap variabel target (AQI Level) adalah AQI (0,97) dan CO (0,90)

Feature yang memiliki korelasi negatif terhadap variabel target (AQI Level) adalah Season (-0,56) dan NH3 (-0,32)

Boxplot Deteksi Outlier



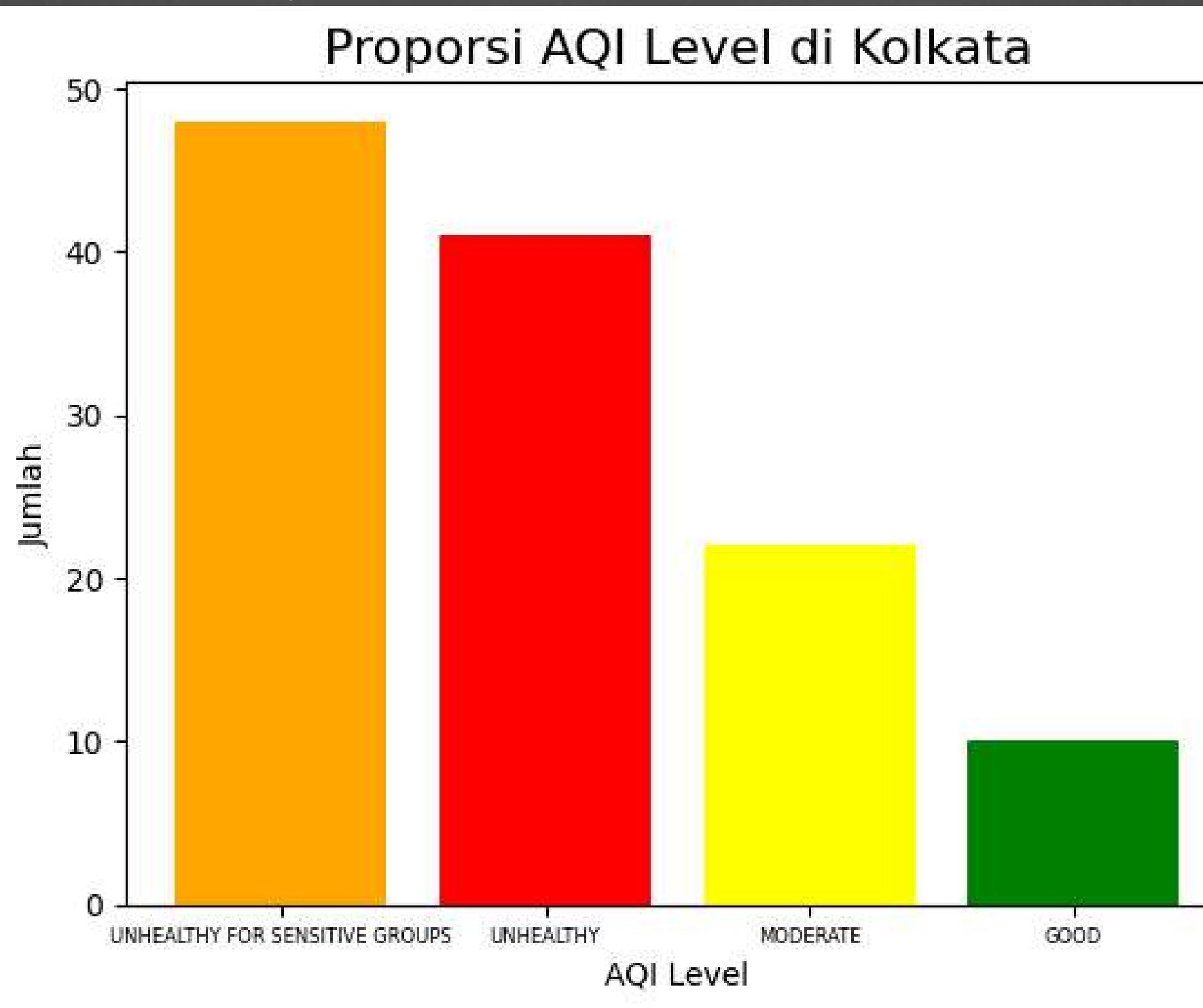
Feature yang memiliki nilai outlier adalah CO dan VOC, namun setelah dianalisis adanya nilai outlier tersebut akibat kesalahan penginputan data

AQI Category

Air Quality Index (AQI) Values	Levels of Health Concern	Colors
<i>When the AQI is in this range:</i>	<i>...air quality conditions are:</i>	<i>...as symbolized by this color:</i>
0-50	Good	Green
51-100	Moderate	Yellow
101-150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

U.S. Environmental Protection Agency (no date) Air Data Basic Information | US EPA . Available at: <https://www.epa.gov/outdoor-air-quality-data/air-data-basic-information> (Accessed: 09 November 2023).

Proporsi AQI Level di Kota Kolkata



Tingkat AQI level di kota Kolkata memiliki kualitas "Unhealthy for sensitive groups" dan "unhealthy" yang sangat tinggi, dimana jika pada 2 kategori itu memiliki nilai tinggi bisa dipastikan kualitas udara pada kota Kolkata buruk



Kolkata



India

Proporsi AQI Level di Kota Visakhapatnam

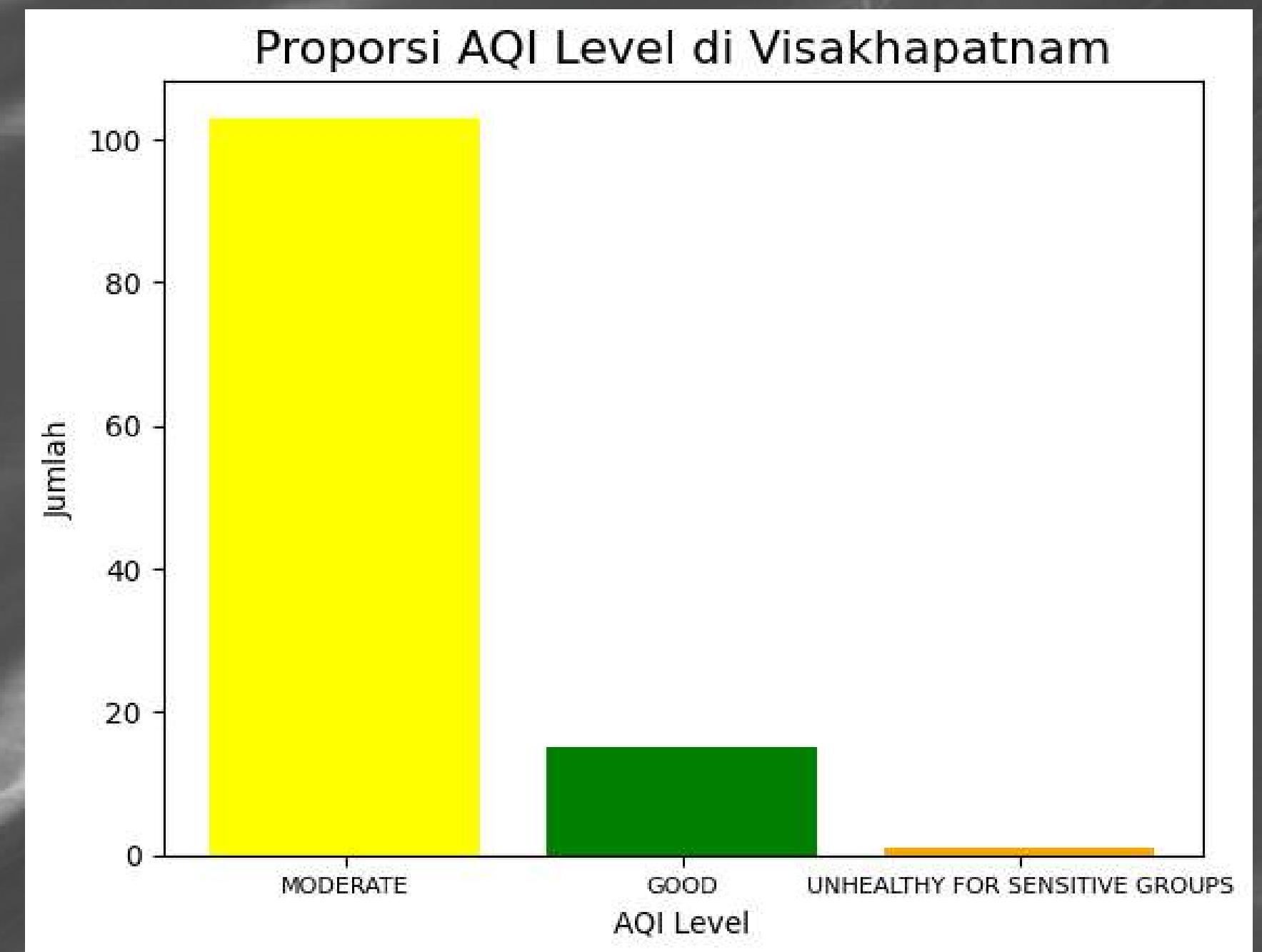
Pada bar chart AQI Level di samping menampilkan bar 'Moderate' sangat tinggi, tingkat ini masih dapat diterima namun sedikit memberikan risiko, disarankan untuk menggunakan masker bila beraktivitas di luar ruangan



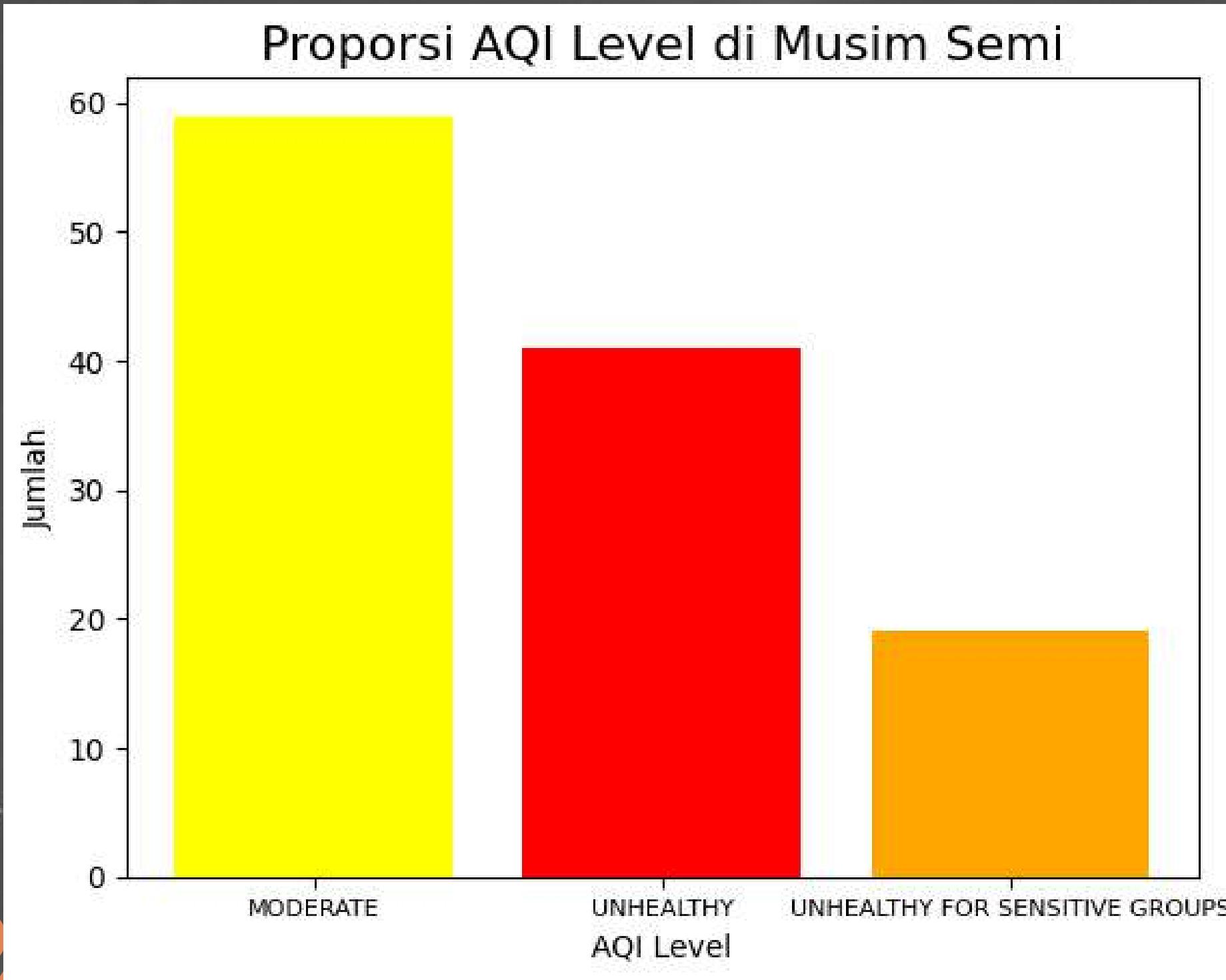
Visakhapatnam



India



Proporsi AQI Level Pada Musim Semi

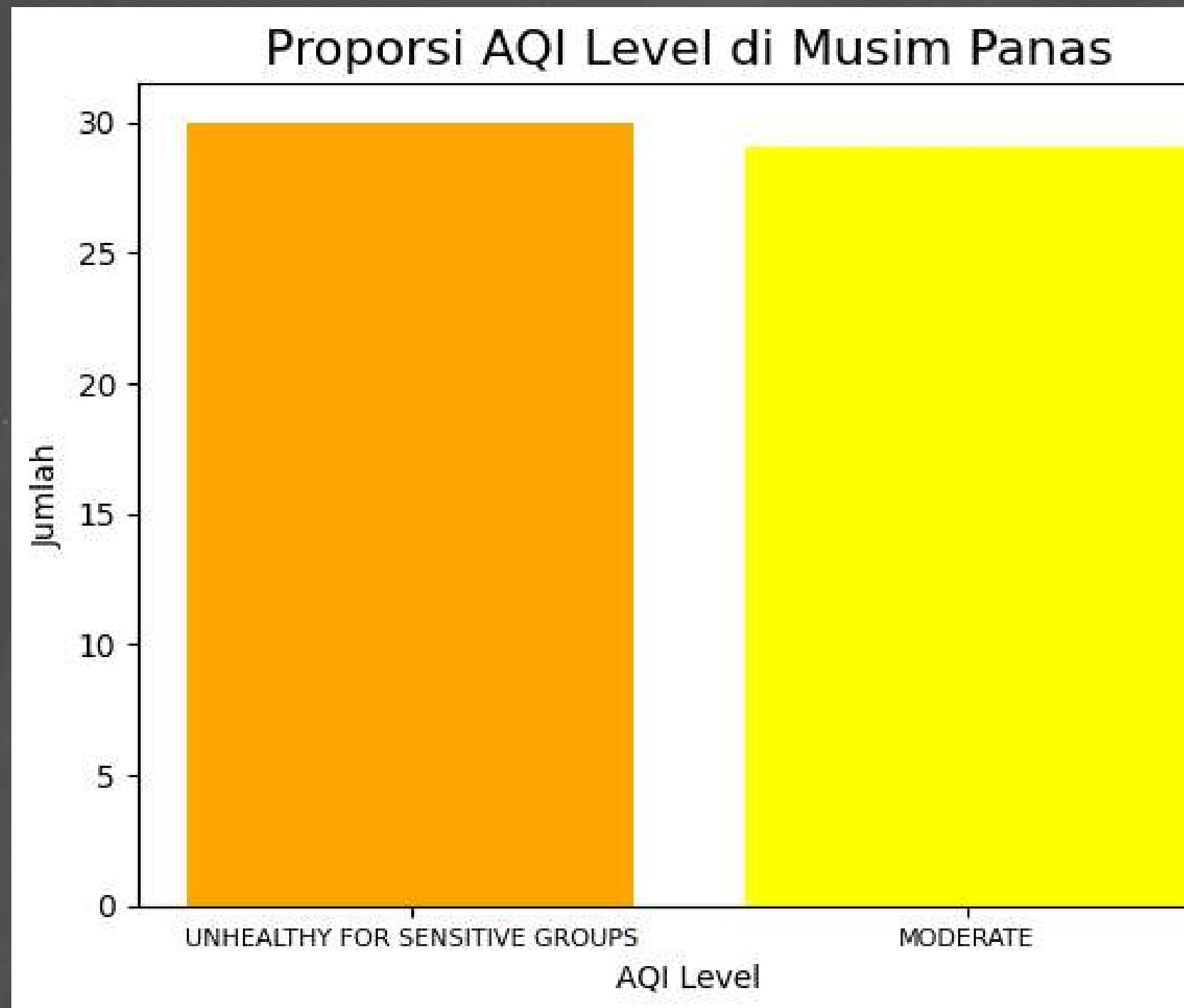


Cuaca kering dan hangat dapat meningkatkan polusi udara. Partikel kecil lebih mudah terangkat, dan praktik pertanian, pembakaran sampah, serta pola angin dapat memperburuk kualitas udara



Musim semi

Proporsi AQI Level Pada Musim Panas

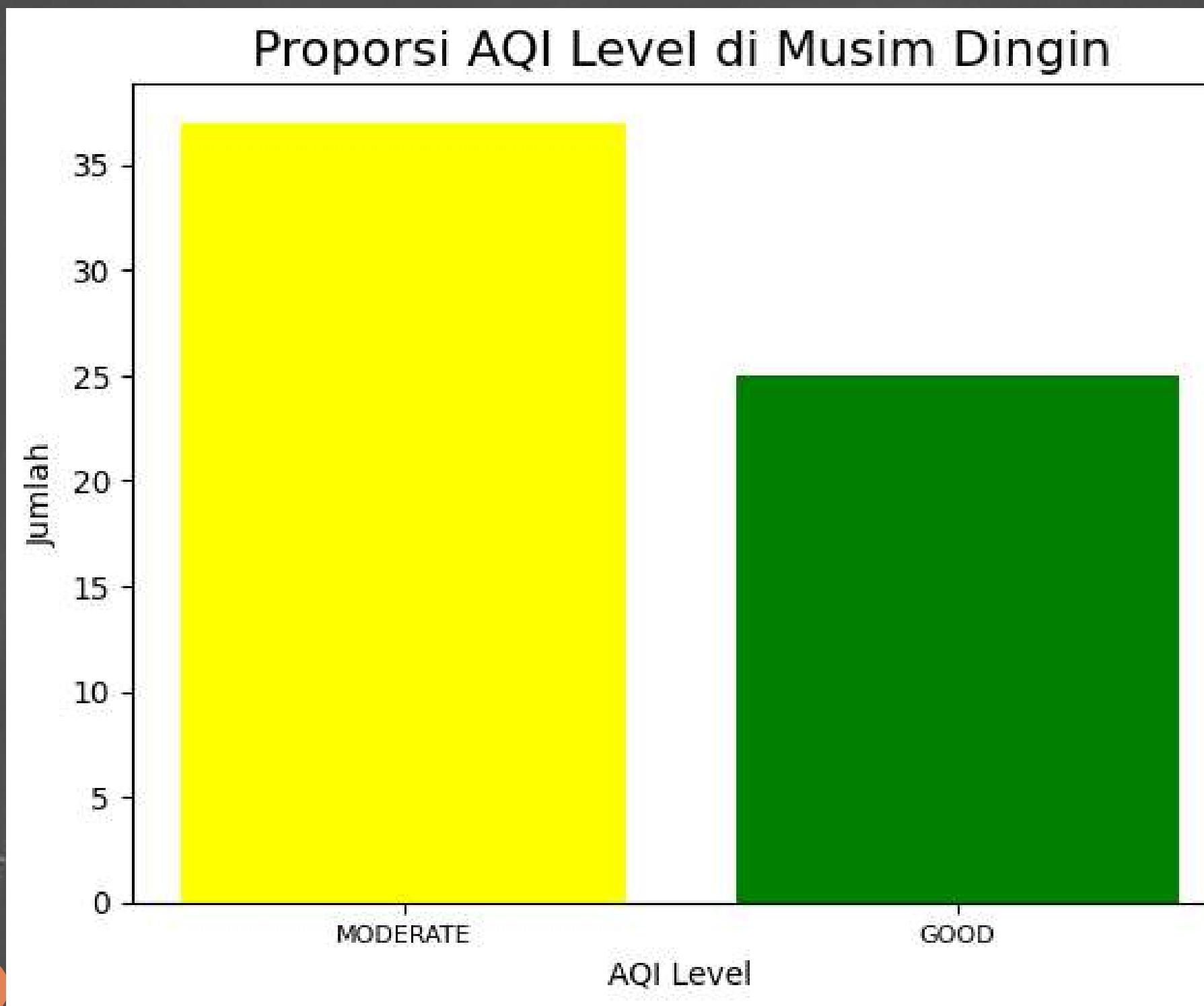


Musim panas memiliki keadaan hawa yang sangat panas dan memiliki sedikit angin dan curah hujan, suhu ini dapat meningkatkan dispersi polusi udara ke atmosfer, mengurangi tingkat pencemaran udara dibanding musim semi



Musim panas

Proporsi AQI Level Pada Musim Dingin

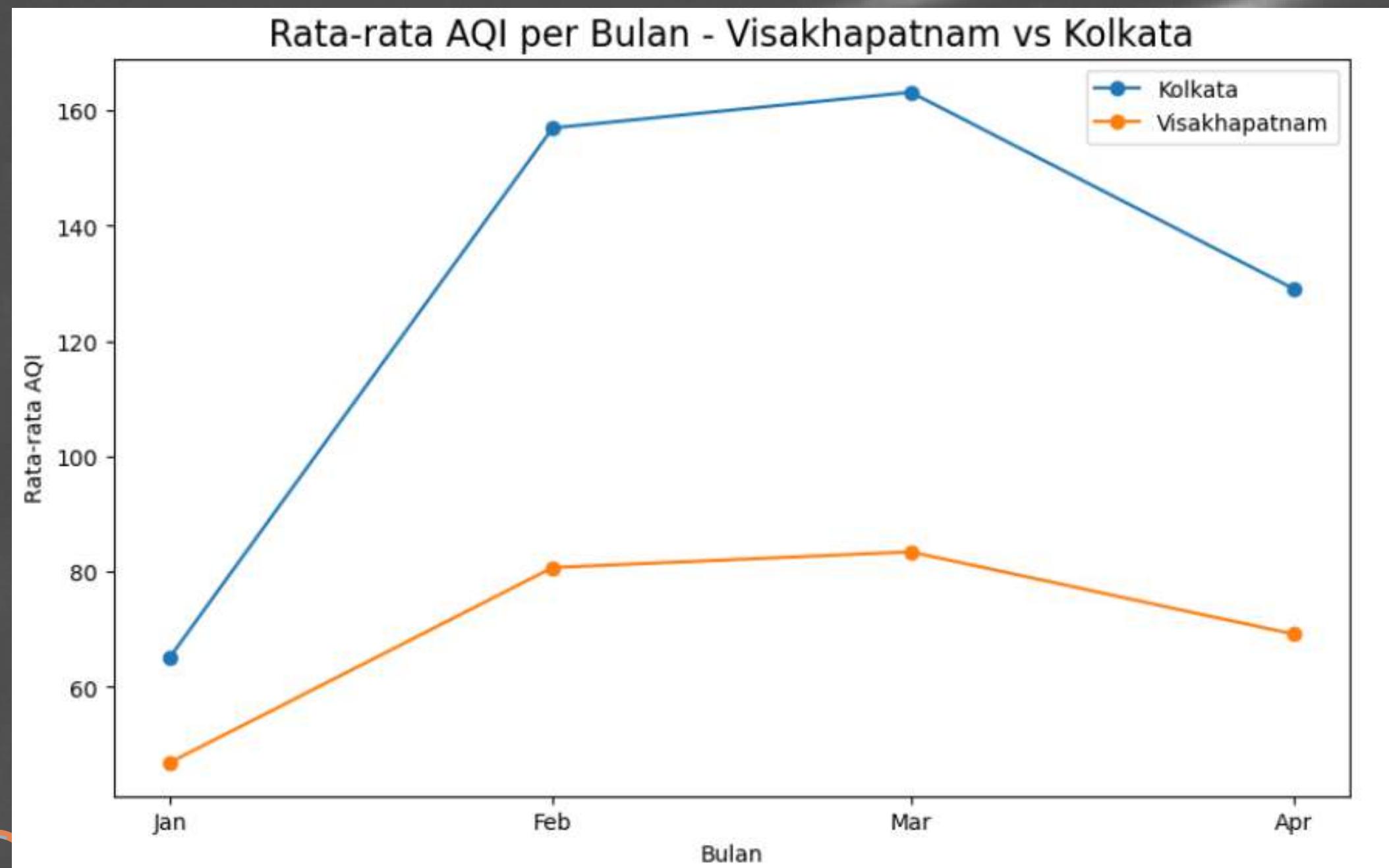


Musim dingin memiliki kondisi cuaca yang stabil, yang dapat menyebabkan akumulasi polutan. Hawa sejuk dan adanya sedikit hujan dapat membersihkan atmosfer dari polutan.



Musim dingin

Rata-Rata nilai AQI/Bulan/Kota



	City	Month	AQI
0	Kolkata	April	129.109524
1	Kolkata	February	156.970443
2	Kolkata	January	65.184332
3	Kolkata	March	163.170507
4	Visakhapatnam	April	69.180952
5	Visakhapatnam	February	80.724138
6	Visakhapatnam	January	46.912442
7	Visakhapatnam	March	83.447005

Grafik tersebut menyertakan pada bulan januari memiliki rata-rata aqi rendah 65,184 (Kolkata), 46,912 (Visakhapatnam), dan paling tinggi berada di 163,17 untuk kolkata, 83,447 untuk visakhapatnam masing-masing pada bulan Maret

DATA REPROCESSING

Signature

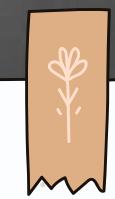
Drop Rows

Before:

Out[3]:

s.no	City	Season	Date	Ship Entry/Left	Type of ship present	PM2.5	PM2.5 AQI	PM2.5 AQI CAT	NO2	NO2 AQI	NO2 AQI CAT	NH3	NH3 AQI	NH3 AQI CAT	
0	1	Chennai	Spring	5/1/2022	Yes/Entry	Coal	6.550	27	GOOD	0.90	1	GOOD	0.020	1	GOOD
1	2	Chennai	Spring	5/2/2022	No	Coal	12.580	52	SATISFACTORY	1.25	1	GOOD	0.090	1	GOOD
2	3	Chennai	Spring	5/3/2022	No	Coal	25.980	80	SATISFACTORY	1.60	1	GOOD	0.210	3	GOOD
3	4	Chennai	Spring	5/4/2022	No	Coal	29.880	88	SATISFACTORY	1.56	1	GOOD	0.260	3	GOOD
4	5	Chennai	Spring	5/5/2022	No	Coal	35.930	102	MODERATE	2.65	2	GOOD	0.340	4	GOOD
5	6	Chennai	Spring	5/6/2022	No	Coal	37.590	106	MODERATE	2.30	2	GOOD	0.380	4	GOOD
6	7	Chennai	Spring	5/7/2022	No	Coal	43.690	121	MODERATE	1.90	1	GOOD	0.390	5	GOOD
7	8	Chennai	Spring	5/8/2022	No	Coal	52.360	145	MODERATE	2.10	2	GOOD	0.410	5	GOOD
8	9	Chennai	Spring	5/9/2022	No	Coal	58.980	153	POOR	1.89	1	GOOD	0.420	5	GOOD
9	10	Chennai	Spring	5/10/2022	No	Coal	62.580	155	POOR	2.65	2	GOOD	0.430	5	GOOD

In [4]: df.shape
 Out[4]: (363, 29)



Pada bagian ini dilakukan drop baris pada kota Chennai, karena kota Chennai memiliki data 'Date' pada tahun 2022 yang tidak selaras dengan dua kota lainnya yang memiliki data 'Date' tahun 2020. Sehingga semula dataset memiliki 363 baris menjadi 242 baris.

After:

Out[5]:

s.no	City	Season	Date	Ship Entry/Left	Type of ship present	PM2.5	PM2.5 AQI	PM2.5 AQI CAT	NO2	NO2 AQI	NO2 AQI CAT	NH3	NH3 AQI	NH3 AQI CAT	CO	
121	122	Visakapatnam	winter	01-01-2020	Yes/Entry	Coal	1.500	6	GOOD	1.00	1	GOOD	0.50	6	GOOD	0.110
122	123	Visakapatnam	winter	02-01-2020	No	Coal	2.600	11	GOOD	2.20	2	GOOD	0.66	7	GOOD	0.220
123	124	Visakapatnam	winter	03-01-2020	No	Coal	5.800	24	GOOD	2.30	2	GOOD	1.20	14	GOOD	0.340
124	125	Visakapatnam	winter	04-01-2020	No	Coal	6.900	29	GOOD	2.70	2	GOOD	2.30	26	GOOD	0.450
125	126	Visakapatnam	winter	05-01-2020	No	Coal	8.200	34	GOOD	3.40	3	GOOD	2.59	28	GOOD	0.670

In [12]: df1.shape
 Out[12]: (242, 29)

DROP COLUMNS

Before:

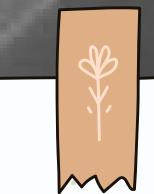
s.no	City	Season	Date	Ship Entry/Left	Type of ship present	PM2.5	PM2.5 AQI	PM2.5 AQI CAT	NO2	NO2 AQI	NO2 AQI CAT	NH3	NH3 AQI	NH3 AQI CAT	CO	
121	122	Visakapatnam	winter	01-01-2020	Yes/Entry	Coal	1.500	6	GOOD	1.00	1	GOOD	0.50	6	GOOD	0.110
122	123	Visakapatnam	winter	02-01-2020	No	Coal	2.600	11	GOOD	2.20	2	GOOD	0.66	7	GOOD	0.220
123	124	Visakapatnam	winter	03-01-2020	No	Coal	5.800	24	GOOD	2.30	2	GOOD	1.20	14	GOOD	0.340
124	125	Visakapatnam	winter	04-01-2020	No	Coal	6.900	29	GOOD	2.70	2	GOOD	2.30	26	GOOD	0.450
125	126	Visakapatnam	winter	05-01-2020	No	Coal	8.200	34	GOOD	3.40	3	GOOD	2.59	28	GOOD	0.670

After:

City	Season	Date	PM2.5	PM2.5 AQI	PM2.5 AQI CAT	NO2	NO2 AQI	NO2 AQI CAT	NH3	NH3 AQI	NH3 AQI CAT	CO	CO AQI	CO AQI CAT	SO	
121	Visakapatnam	winter	01-01-2020	1.500	6	GOOD	1.00	1	GOOD	0.50	6	GOOD	0.110	1	GOOD	0.21
122	Visakapatnam	winter	02-01-2020	2.600	11	GOOD	2.20	2	GOOD	0.66	7	GOOD	0.220	2	GOOD	0.22
123	Visakapatnam	winter	03-01-2020	5.800	24	GOOD	2.30	2	GOOD	1.20	14	GOOD	0.340	3	GOOD	0.34
124	Visakapatnam	winter	04-01-2020	6.900	29	GOOD	2.70	2	GOOD	2.30	26	GOOD	0.450	5	GOOD	0.45
125	Visakapatnam	winter	05-01-2020	8.200	34	GOOD	3.40	3	GOOD	2.59	28	GOOD	0.670	7	GOOD	0.67

In [12]: df1.shape

Out[12]: (242, 29)



Pada bagian ini dilakukan drop kolom pada s.no, Ship Entry / Left, dan Type of Ship Present karena kolom tersebut tidak digunakan dan tidak berkaitan dalam analisis ini.

Sehingga semula dataset memiliki 29 kolom menjadi 26 kolom.

In [93]: df2.shape

Out[93]: (242, 26)

R ESET INDEX

Before:

	City	Season	Date	PM2.5	PM2.5 AQI	PM2.5 AQI CAT	NO2	NO2 AQI	NO2 AQI CAT	NH3	NH3 AQI	NH3 AQI CAT	CO	CO AQI	CO AQI CAT	SO
121	Visakapatnam	winter	01-01-2020	1.500	6	GOOD	1.00	1	GOOD	0.50	6	GOOD	0.110	1	GOOD	0.21
122	Visakapatnam	winter	02-01-2020	2.600	11	GOOD	2.20	2	GOOD	0.66	7	GOOD	0.220	2	GOOD	0.22
123	Visakapatnam	winter	03-01-2020	5.800	24	GOOD	2.30	2	GOOD	1.20	14	GOOD	0.340	3	GOOD	0.34
124	Visakapatnam	winter	04-01-2020	6.900	29	GOOD	2.70	2	GOOD	2.30	26	GOOD	0.450	5	GOOD	0.45
125	Visakapatnam	winter	05-01-2020	8.200	34	GOOD	3.40	3	GOOD	2.59	28	GOOD	0.670	7	GOOD	0.67

After:

	City	Season	Date	PM2.5	PM2.5 AQI	PM2.5 AQI CAT	NO2	NO2 AQI	NO2 AQI CAT	NH3	NH3 AQI	NH3 AQI CAT	CO	CO AQI	CO AQI CAT
0	Visakapatnam	winter	01-01-2020	1.500	6	GOOD	1.00	1	GOOD	0.50	6	GOOD	0.110	1	GOOD
1	Visakapatnam	winter	02-01-2020	2.600	11	GOOD	2.20	2	GOOD	0.66	7	GOOD	0.220	2	GOOD
2	Visakapatnam	winter	03-01-2020	5.800	24	GOOD	2.30	2	GOOD	1.20	14	GOOD	0.340	3	GOOD
3	Visakapatnam	winter	04-01-2020	6.900	29	GOOD	2.70	2	GOOD	2.30	26	GOOD	0.450	5	GOOD
4	Visakapatnam	winter	05-01-2020	8.200	34	GOOD	3.40	3	GOOD	2.59	28	GOOD	0.670	7	GOOD

Pada bagian ini dilakukan
reset index dimulai kembali
dari 0 sehingga dataset
kembali konsisten.



HANDLING INCONSISTENCY

Before:

	City	Season	Date	PM2.5	PM2.5 AQI	PM2.5 AQI CAT	NO2	NO2 AQI	NO2 AQI CAT	NH3	NH3 AQI	NH3 AQI CAT	CO	CO AQI	CO AQI CAT
0	Visakapatnam	winter	01-01-2020	1.500	6	GOOD	1.00	1	GOOD	0.50	6	GOOD	0.110	1	GOOD
1	Visakapatnam	winter	02-01-2020	2.600	11	GOOD	2.20	2	GOOD	0.66	7	GOOD	0.220	2	GOOD
2	Visakapatnam	winter	03-01-2020	5.800	24	GOOD	2.30	2	GOOD	1.20	14	GOOD	0.340	3	GOOD
3	Visakapatnam	winter	04-01-2020	6.900	29	GOOD	2.70	2	GOOD	2.30	26	GOOD	0.450	5	GOOD
4	Visakapatnam	winter	05-01-2020	8.200	34	GOOD	3.40	3	GOOD	2.59	28	GOOD	0.670	7	GOOD

After:

	City	Season	Date	PM2.5	PM2.5 AQI	PM2.5 AQI CAT	NO2	NO2 AQI	NO2 AQI CAT	NH3	NH3 AQI	NH3 AQI CAT	CO	CO AQI	CO AQI CAT
0	Visakhapatnam	winter	01-01-2020	1.500	6	GOOD	1.00	1	GOOD	0.50	6	GOOD	0.110	1	GOOD
1	Visakhapatnam	winter	02-01-2020	2.600	11	GOOD	2.20	2	GOOD	0.66	7	GOOD	0.220	2	GOOD
2	Visakhapatnam	winter	03-01-2020	5.800	24	GOOD	2.30	2	GOOD	1.20	14	GOOD	0.340	3	GOOD
3	Visakhapatnam	winter	04-01-2020	6.900	29	GOOD	2.70	2	GOOD	2.30	26	GOOD	0.450	5	GOOD
4	Visakhapatnam	winter	05-01-2020	8.200	34	GOOD	3.40	3	GOOD	2.59	28	GOOD	0.670	7	GOOD

Pada bagian ini dilakukan penanganan kesalahan penggunaan kata pada kota Visakhapatnam sesuai dengan literatur. Hal tersebut bertujuan untuk menghindari terjadinya kesalahpahaman pembaca.

H

ANDLING MISSING VALUE

Before:

CO2 AQI CAT	NH3	NH3 AQI	NH3 AQI CAT	CO	CO AQI	CO AQI CAT	SO2	SO2 AQI	SO2 AQI CAT	O3	O3 AQI	O3 AQI CAT	VOC	VOC AQI	VOC AQI CAT	AQI	AQI_LVL
GOOD	0.50	6	GOOD	0.110	1	GOOD	0.210	0	GOOD	6.430	6	GOOD	1.54	33	GOOD	NaN	NaN
GOOD	0.66	7	GOOD	0.220	2	GOOD	0.220	0	GOOD	7.880	6	GOOD	1.56	36	GOOD	NaN	NaN
GOOD	1.20	14	GOOD	0.340	3	GOOD	0.340	0	GOOD	8.210	7	GOOD	1.60	39	GOOD	NaN	NaN
GOOD	2.30	26	GOOD	0.450	5	GOOD	0.450	0	GOOD	9.870	9	GOOD	1.74	42	GOOD	NaN	NaN
GOOD	2.59	28	GOOD	0.670	7	GOOD	0.670	0	GOOD	15.000	14	GOOD	1.86	47	GOOD	NaN	NaN

After:

ICAT	NH3	NH3 AQI	NH3 AQI CAT	CO	CO AQI	CO AQI CAT	SO2	SO2 AQI	SO2 AQI CAT	O3	O3 AQI	O3 AQI CAT	VOC	VOC AQI	VOC AQI CAT	AQI	AQI_LVL
GOOD	0.50	6	GOOD	0.110	1	GOOD	0.210	0	GOOD	6.430	6	GOOD	1.54	33	GOOD	7.571429	NaN
GOOD	0.66	7	GOOD	0.220	2	GOOD	0.220	0	GOOD	7.880	6	GOOD	1.56	36	GOOD	9.142857	NaN
GOOD	1.20	14	GOOD	0.340	3	GOOD	0.340	0	GOOD	8.210	7	GOOD	1.60	39	GOOD	12.714286	NaN
GOOD	2.30	26	GOOD	0.450	5	GOOD	0.450	0	GOOD	9.870	9	GOOD	1.74	42	GOOD	16.142857	NaN
GOOD	2.59	28	GOOD	0.670	7	GOOD	0.670	0	GOOD	15.000	14	GOOD	1.86	47	GOOD	19.000000	NaN



Pada bagian ini dilakukan penanganan nilai null dengan pengisian nilai null AQI menggunakan hasil perhitungan AQI dengan rumus (total nilai AQI per parameter / total parameter).

HANDLING MISSING VALUE

Before:

ICAT	NH3	NH3 AQI	NH3 AQI CAT	CO	CO AQI	CO AQI CAT	SO2	SO2 AQI	SO2 AQI CAT	O3	O3 AQI	O3 AQI CAT	VOC	VOC AQI	VOC AQI CAT	AQI	AQI_LVL
GOOD	0.50	6	GOOD	0.110	1	GOOD	0.210	0	GOOD	6.430	6	GOOD	1.54	33	GOOD	7.571429	NaN
GOOD	0.66	7	GOOD	0.220	2	GOOD	0.220	0	GOOD	7.880	6	GOOD	1.56	36	GOOD	9.142857	NaN
GOOD	1.20	14	GOOD	0.340	3	GOOD	0.340	0	GOOD	8.210	7	GOOD	1.60	39	GOOD	12.714286	NaN
GOOD	2.30	26	GOOD	0.450	5	GOOD	0.450	0	GOOD	9.870	9	GOOD	1.74	42	GOOD	16.142857	NaN
GOOD	2.59	28	GOOD	0.670	7	GOOD	0.670	0	GOOD	15.000	14	GOOD	1.86	47	GOOD	19.000000	NaN

Levels of Concern	Values of Index
Good	0 to 50
Moderate	51 to 100
Unhealthy for Sensitive Groups	101 to 150
Unhealthy	151 to 200
Very Unhealthy	201 to 300
Hazardous	301 and higher

After:

NH3	NH3 AQI	NH3 AQI CAT	CO	CO AQI	CO AQI CAT	SO2	SO2 AQI	SO2 AQI CAT	O3	O3 AQI	O3 AQI CAT	VOC	VOC AQI	VOC AQI CAT	AQI	AQI_LEVEL
0.50	6	GOOD	0.110	1	GOOD	0.210	0	GOOD	6.430	6	GOOD	1.54	33	GOOD	7.571429	GOOD
0.66	7	GOOD	0.220	2	GOOD	0.220	0	GOOD	7.880	6	GOOD	1.56	36	GOOD	9.142857	GOOD
1.20	14	GOOD	0.340	3	GOOD	0.340	0	GOOD	8.210	7	GOOD	1.60	39	GOOD	12.714286	GOOD
2.30	26	GOOD	0.450	5	GOOD	0.450	0	GOOD	9.870	9	GOOD	1.74	42	GOOD	16.142857	GOOD
2.59	28	GOOD	0.670	7	GOOD	0.670	0	GOOD	15.000	14	GOOD	1.86	47	GOOD	19.000000	GOOD

Pada bagian ini dilakukan penanganan nilai null dengan pengisian nilai null AQI_LVL menggunakan kategorisasi AQI yang bersumber dari US EPA.
[\(https://www.airnow.gov/aqi/aqi-basics/\)](https://www.airnow.gov/aqi/aqi-basics/)

H ANDLING OUTLIER

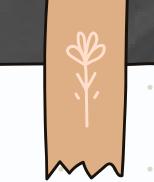
Before:

	df5[df5['CO'] > MAX_IQR_CO]
77	Visakhapatnam spring 18-03-2020 19.67 67 SATISFACTORY 60.12 57 MODERATE 17.62 215 VERY POOR 175.00 19

	df5[df5['VOC'] > MAX_IQR_VOC]
166.0	218 VERY POOR 16.0 15 GOOD 5.55 61 SATISFACTORY 0.43 5 GOOD 16.68 23 GOOD 19.76 18 GOOD 485.0 79

After:

```
df6 = df5[(df5['CO'] != 175.00) & (df5['VOC'] != 485.00)]
df6
```



Pada bagian ini dilakukan penanganan nilai outlier pada kolom "CO" dan "VOC" dengan cara dihapus karena adanya kesalahan input data.

F E A T U R E E N G I N E E R I N G

```
encoding_dict_aqilevel = {
    'GOOD': 0,
    'MODERATE': 1,
    'UNHEALTHY FOR SENSITIVE GROUPS': 2,
    'UNHEALTHY': 3,
    'VERY UNHEALTHY': 4,
    'HAZARDOUS': 5
}
encoding_dict_season = {
    'spring': 0,
    'summer': 1,
    'winter': 2,
}
```

CO	CO AQI	CO AQI CAT	SO2	SO2 AQI	SO2 AQI CAT	O3	O3 AQI	O3 AQI CAT	VOC	VOC AQI	VOC AQI CAT	AQI	AQI_LEVEL	AQI_LEVEL_Encod	Season_Encod
0.110	1	GOOD	0.210	0	GOOD	6.430	6	GOOD	1.54	33	GOOD	7.571429	GOOD	0	2
0.220	2	GOOD	0.220	0	GOOD	7.880	6	GOOD	1.56	36	GOOD	9.142857	GOOD	0	2
0.340	3	GOOD	0.340	0	GOOD	8.210	7	GOOD	1.60	39	GOOD	12.714286	GOOD	0	2
0.450	5	GOOD	0.450	0	GOOD	9.870	9	GOOD	1.74	42	GOOD	16.142857	GOOD	0	2
0.670	7	GOOD	0.670	0	GOOD	15.000	14	GOOD	1.86	47	GOOD	19.000000	GOOD	0	2

Pada bagian dilakukan tahap feature engineering berupa label encoding pada kolom Season dan AQI_LEVEL. Hal tersebut bertujuan agar algoritma dapat memproses dan menganalisis data dengan lebih efektif dikarenakan model yang digunakan merupakan klasifikasi yang mengharuskan tipe data output berupa nominal

label encoding adalah teknik untuk mengubah variabel kategori ke dalam format numerik, sehingga memudahkan algoritma pembelajaran mesin untuk memproses data dan membuat prediksi yang akurat

Signature

MODELLING & EVALUATION

MODELLING

Modelling Klasifikasi adalah proses membangun model matematis atau statistik yang dapat mengelompokkan atau mengidentifikasi data ke dalam kategori atau kelas berdasarkan ciri – ciri tertentu

A LGORITMA

Random Forest

Random Forest adalah algoritma ensemble yang terdiri dari sejumlah besar pohon keputusan yang bekerja bersama untuk meningkatkan kinerja dan ketahanan terhadap overfitting.

XGBoost

XGBoost (Extreme Gradient Boosting) adalah algoritma ensemble yang memperluas konsep decision tree menjadi model yang lebih kompleks dan kuat.

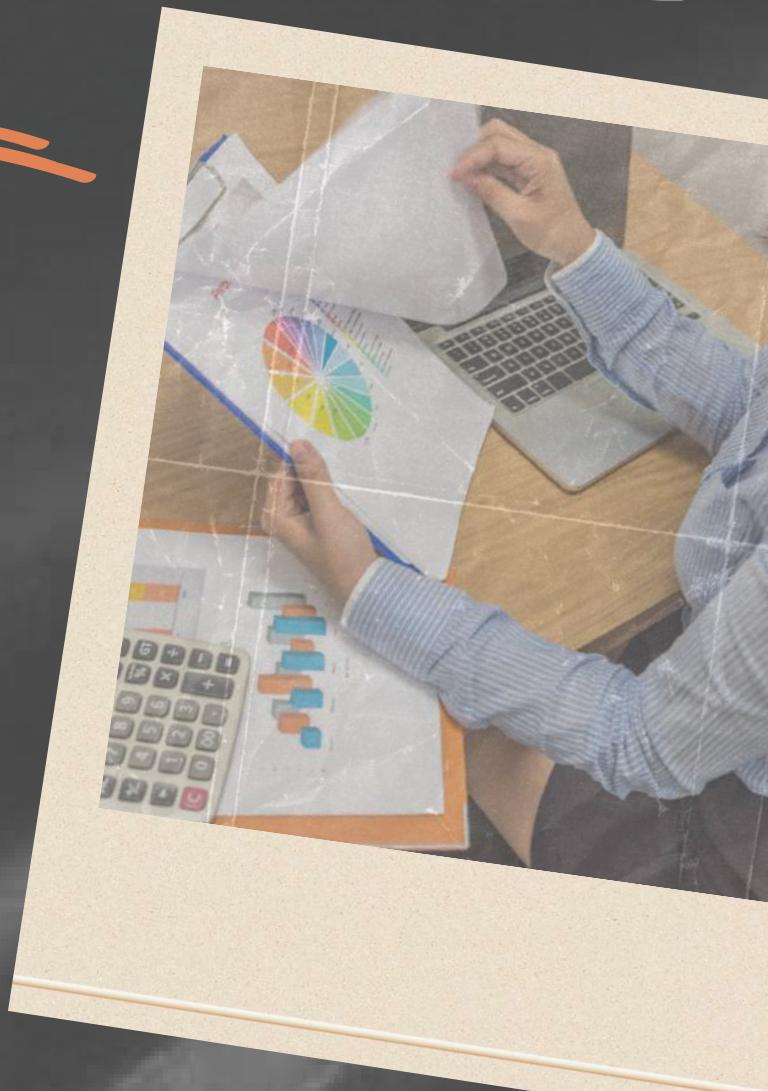


Signature



MODELLING

Include
AQI





AHAPAN

Include AQI

Mendefinisikan X dan Y

```
x = df_a[['PM2.5', 'NO2', 'NH3', 'CO', 'SO2', 'O3', 'VOC', 'AQI', 'Season_Encoder']]  
y = df_a['AQI_LEVEL_Encoder']
```



Rescaler menggunakan Min-Max Scaler

```
from sklearn.preprocessing import MinMaxScaler  
  
scaler = MinMaxScaler()  
x_standar = scaler.fit_transform(x)  
x_standar
```

Train-Test Data Splitting

```
#Teknik pemilihan data random sampling  
random_seed = 42  
x_train, x_test, y_train, y_test = train_test_split(x_standar, y, test_size=0.2, random_state=random_seed)
```



Melatih dan Memprediksi Dengan Model

```
#Random Forest  
rf = RandomForestClassifier()  
rf.fit(x_train, y_train)  
rf_pred = rf.predict(x_test)
```

```
#XGBoost  
xgb_mod = xgb.XGBClassifier()  
xgb_mod.fit(x_train, y_train)  
xgb_pred = xgb_mod.predict(x_test)
```

Memeriksa indikasi overfitting

RF on the training dataset : 100.00%

RF on the test dataset : 95.83%

XGB on the training dataset : 100.00%

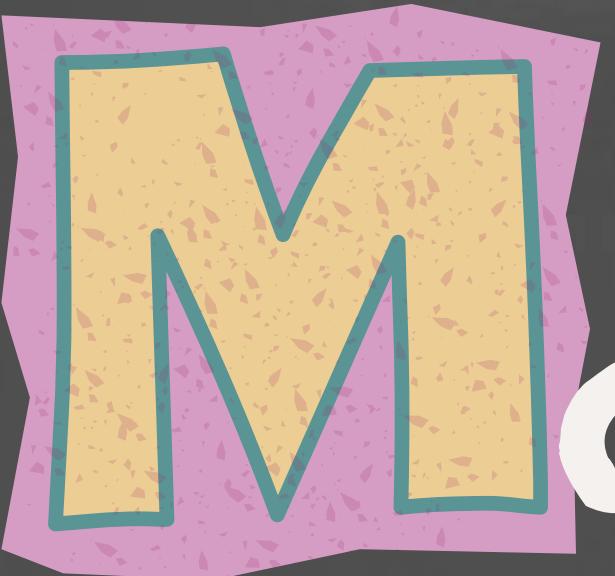
XGB on the test dataset : 97.92%

- Overfitting adalah kondisi dimana nilai akurasi pada data training tinggi dengan nilai akurasi data testing rendah.
- Best fitting merupakan kondisi yang ideal antara nilai akurasi pada data testing dan data training sama-sama tinggi.





Signature



MODELLING

Exclude AQI





AHAPAN

Exclude AQI

Mendefinisikan X dan Y

```
x2 = df_a[['PM2.5', 'NO2', 'NH3', 'CO', 'SO2', 'O3', 'VOC', 'Season_Encod']]  
y2 = df_a['AQI_LEVEL_Encod']
```



Rescaler menggunakan Min-Max Scaler

```
from sklearn.preprocessing import MinMaxScaler  
  
scaler = MinMaxScaler()  
x2_standar = scaler.fit_transform(x2)  
x2_standar
```

Train-Test Data Splitting

```
#Teknik pemilihan data random sampling  
random_seed = 42  
x2_train, x2_test, y2_train, y2_test = train_test_split(x2_standar, y2, test_size=0.2, random_state=random_seed)
```



Melatih dan Memprediksi Dengan Model

```
#Random Forest  
rf2 = RandomForestClassifier()  
rf2.fit(x2_train, y2_train)  
rf2_pred = rf2.predict(x2_test)
```

```
#XGBoost  
xgb2_mod = xgb.XGBClassifier()  
xgb2_mod.fit(x2_train, y2_train)  
xgb2_pred = xgb2_mod.predict(x2_test)
```

Memeriksa indikasi overfitting

RF on the training dataset : 100.00%

RF on the test dataset : 97.92%

XGB on the training dataset : 100.00%

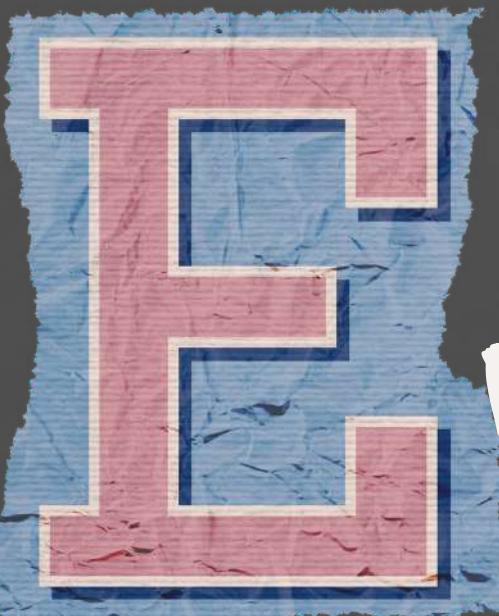
XGB on the test dataset : 97.92%

- Overfitting adalah kondisi dimana nilai akurasi pada data training tinggi dengan nilai akurasi data testing rendah.
- Best fitting merupakan kondisi yang ideal antara nilai akurasi pada data testing dan data training sama-sama tinggi.





EVALUATION



EVALUATION



Include AQI

ALGORITMA	HASIL EVALUASI AKURASI	HASIL EVALUASI RECALL	HASIL EVALUASI PRECISION
Random Forest	95.8 %	95.8 %	96.4 %
XGBoost	97.9 %	97.9 %	98.2 %

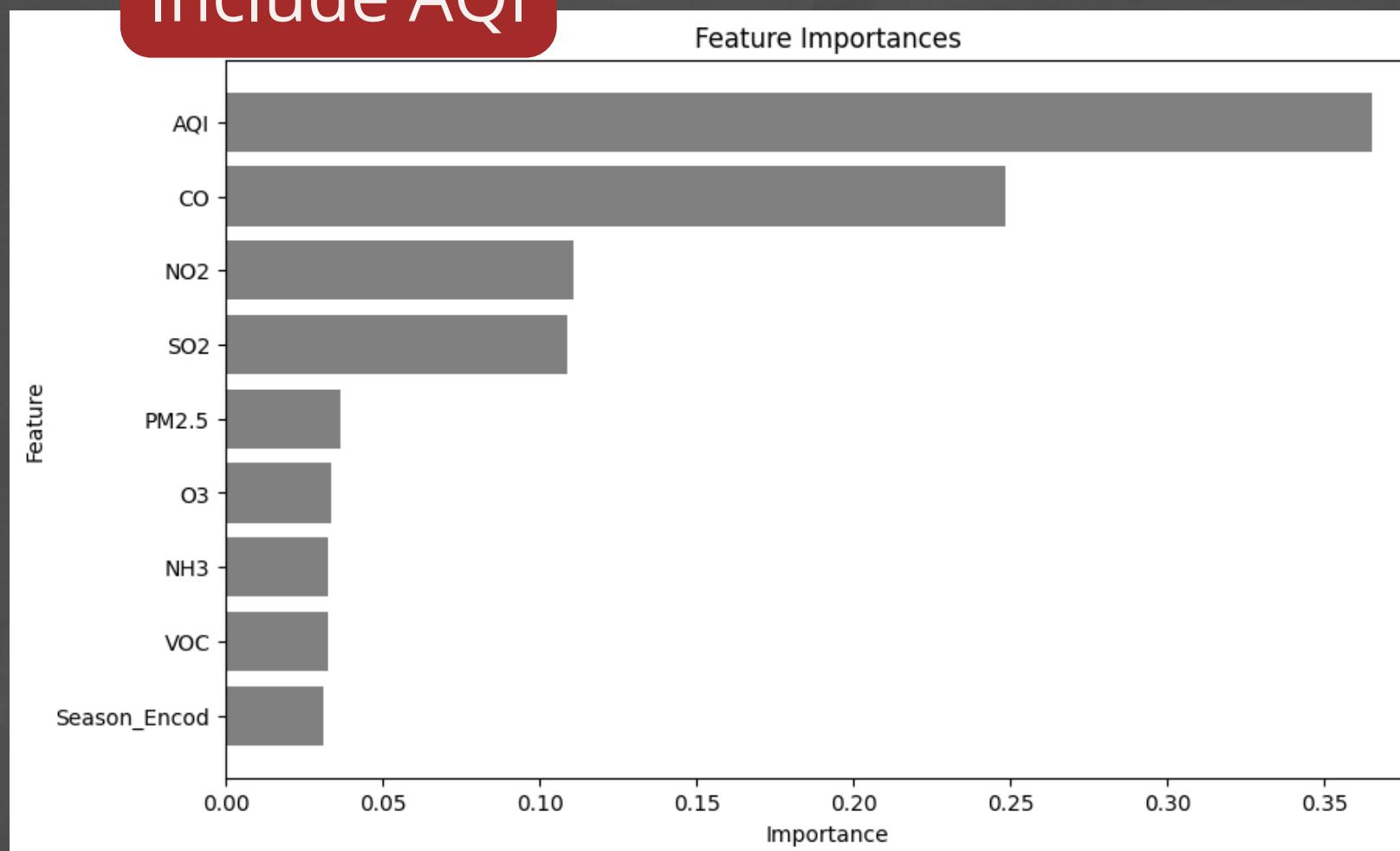
H HASIL

Exclude AQI

ALGORITMA	HASIL EVALUASI AKURASI	HASIL EVALUASI RECALL	HASIL EVALUASI PRECISION
Random Forest	97.9 %	97.9 %	98 %
XGBoost	97.9 %	97.9 %	98 %

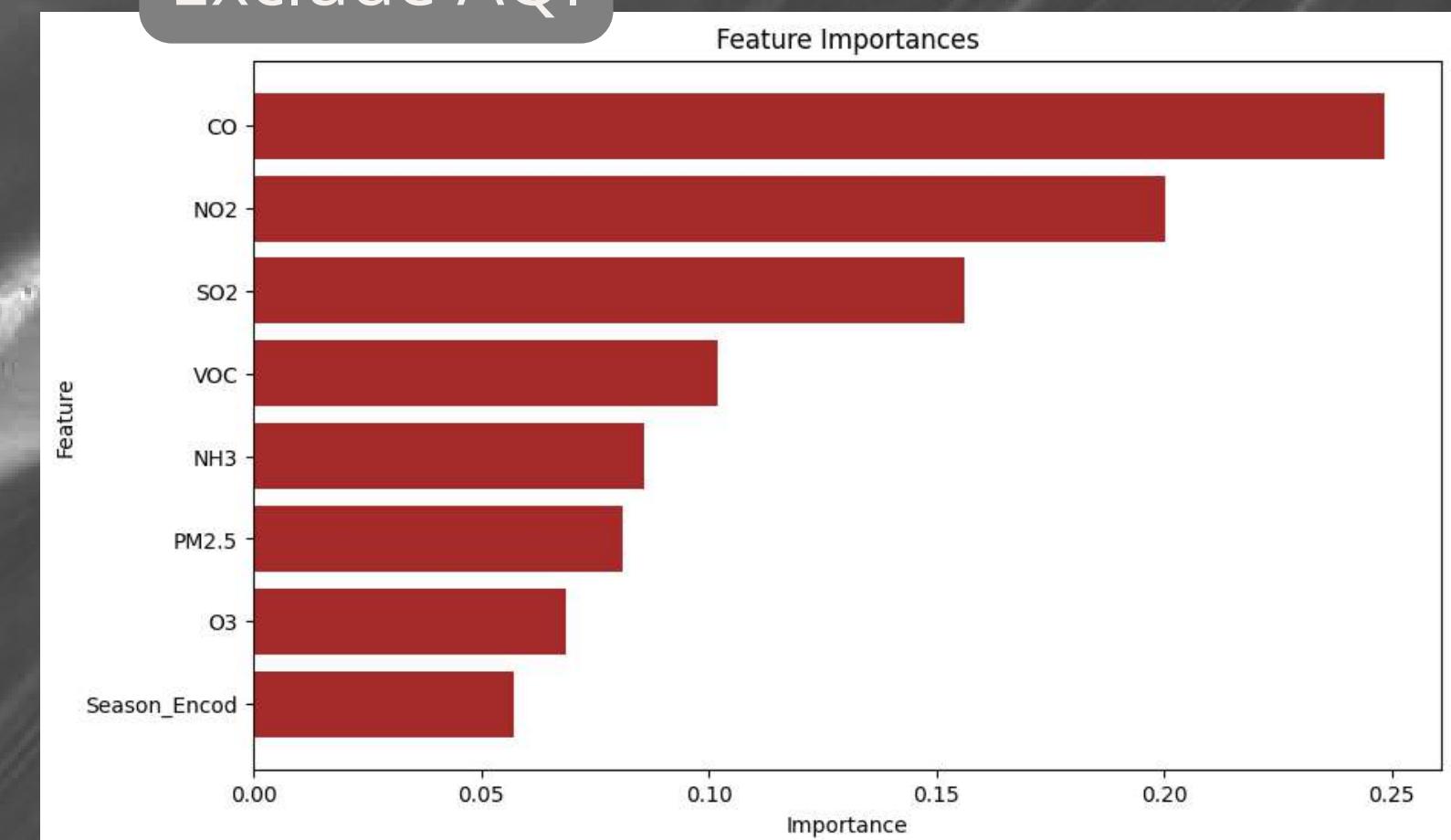
EVALUASI

Include AQI



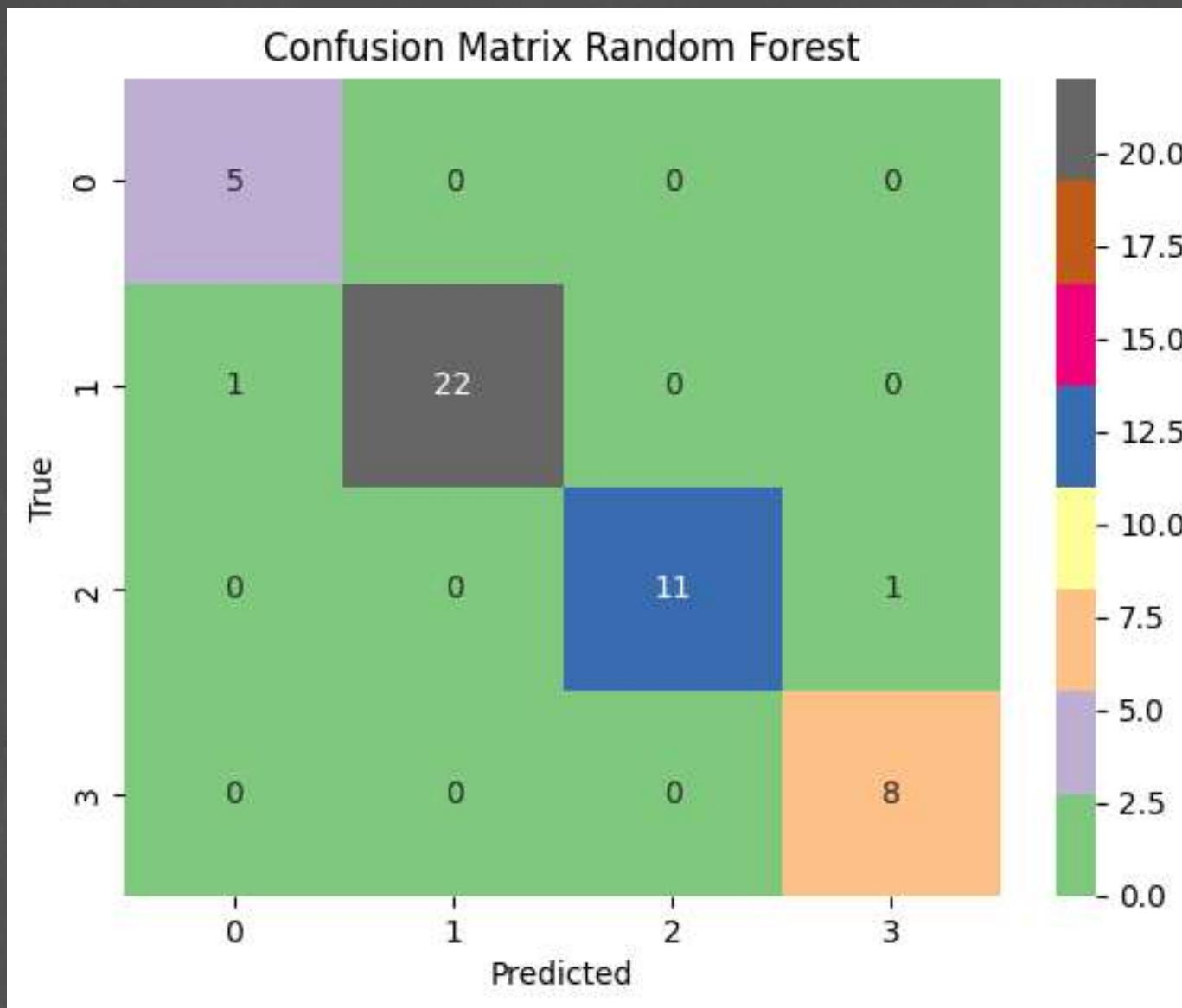
F E A T U R E

Exclude AQI



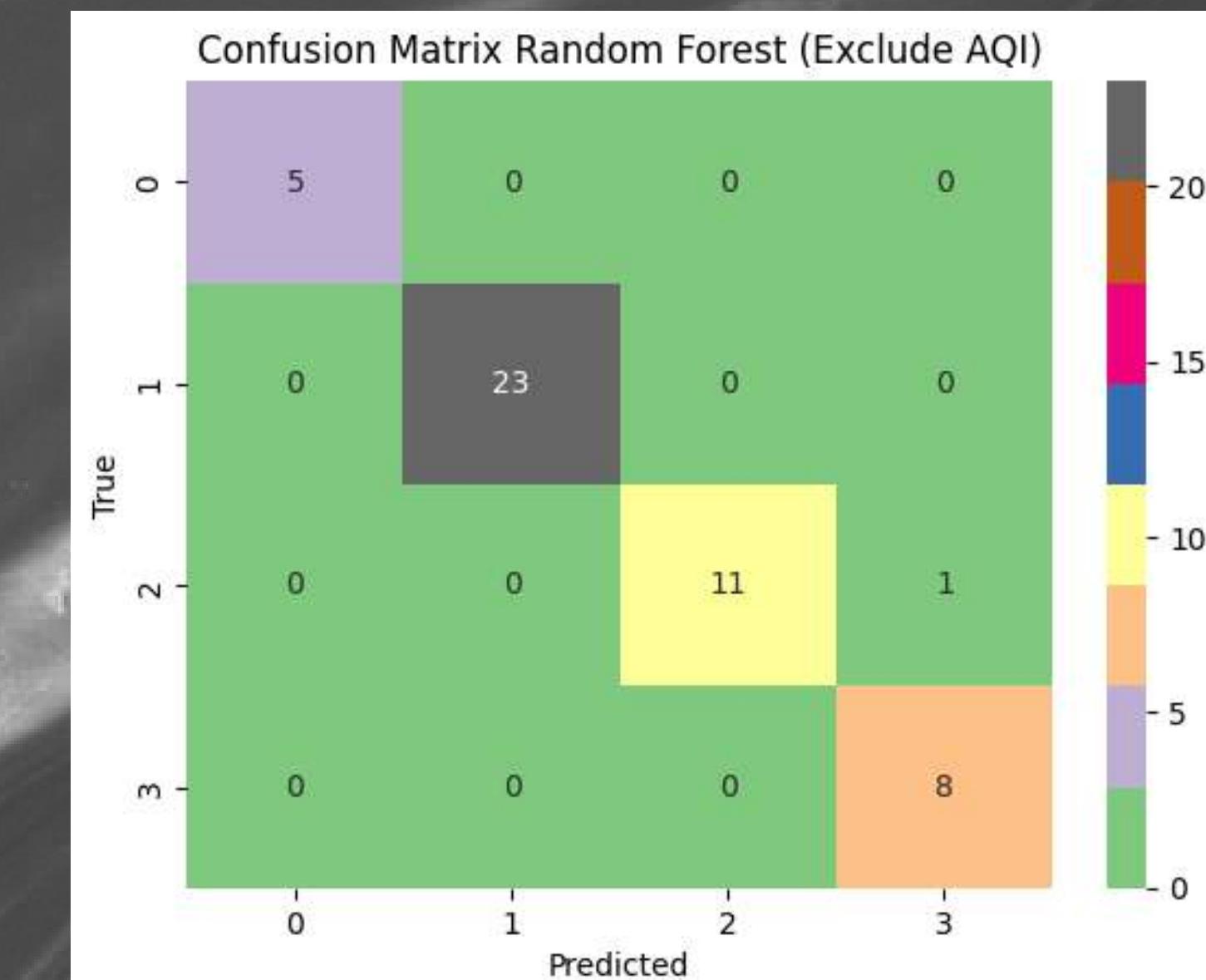
i IMPORTANCE

Confusion Matrix Random Forest



MATRIX

Confusion Matrix Random Forest (Exclude AQI)

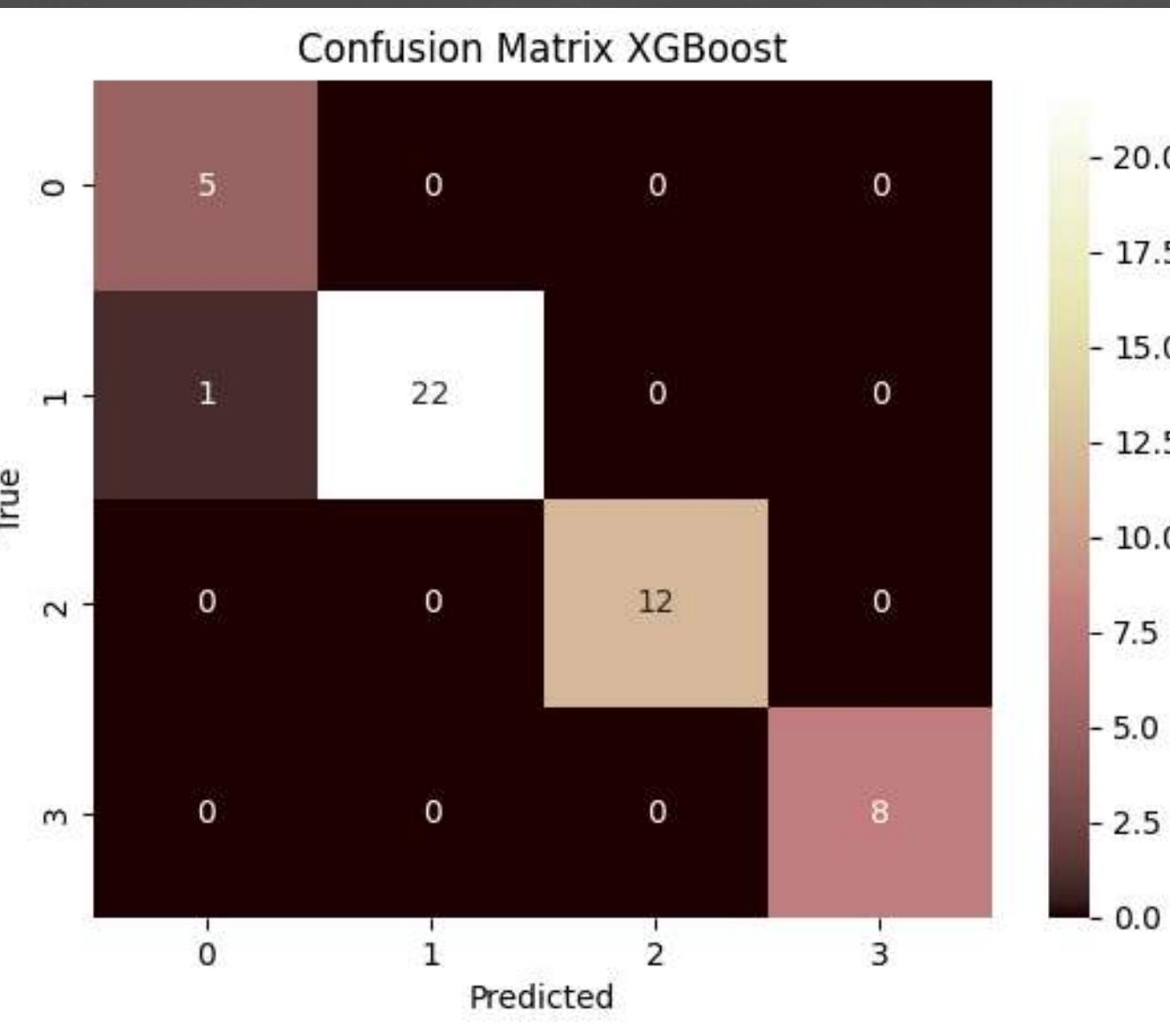


CONFUSION

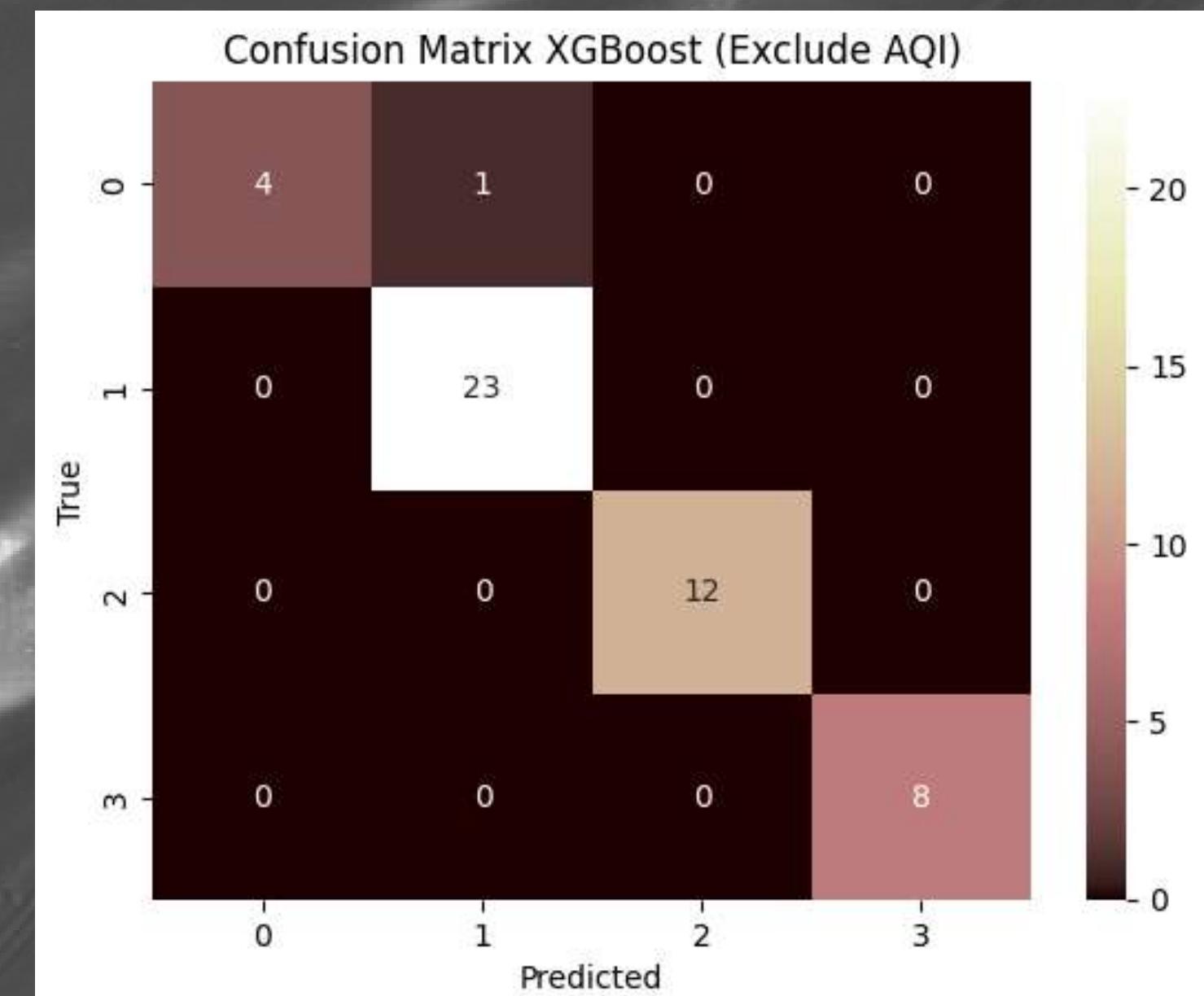


C

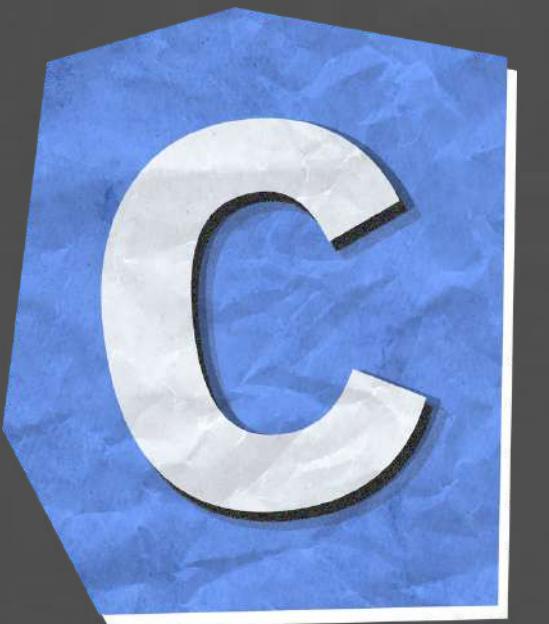
CONFUSION



MATRIX



CONCLUSION



Dari kedua algoritma yang digunakan, yaitu Random Forest dan XGBoost, keduanya memiliki performa yang baik dan sangat cocok untuk digunakan dalam memprediksi feature importance yang berkaitan dengan Air Quality Index Level. Sehingga, kelebihannya ketika ingin melakukan hal yang sama maka dapat memilih diantara kedua algoritma tersebut untuk dapat digunakan.

Selain itu, dapat disimpulkan berdasarkan hasil prediksi yang ada, 3 parameter yang memiliki peran penting dalam menentukan Level AQI adalah CO (Karbon Monoksida), NO₂ (Nitrogen Dioksida), dan SO₂ (Sulfur Dioksida)



Kolkata dan Visakhapatnam merupakan kota yang dikenal sebagai metropolitan city di India. Dimana, hal tersebut mengindikasikan kedua kota tersebut merupakan kota dengan aktivitas manusia yang tergolong tinggi.

Berdasarkan beberapa literatur, sumber utama dari parameter CO, SO₂, dan NO₂ di kedua kota tersebut berasal dari 2 sektor utama, yaitu transportasi dan pengelolaan sampah yang belum memadai.

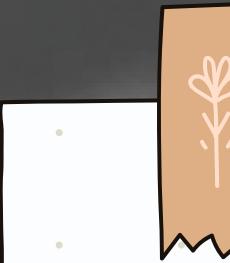


Kesibukan yang tinggi dan padat dalam aktivitas transportasi menjadi ciri khas metropolitan city, dengan lalu lintas jalan raya yang ramai, menciptakan suatu dinamika kota yang tak pernah berhenti. Hasil dari aktivitas transportasi ini berupa emisi kendaraan yang tinggi dan mencemari udara di kedua kota tersebut. Dimana, parameter CO, SO₂, dan NO₂ yang terlepas di udara memang merupakan emisi hasil pembakaran bahan bakar kendaraan.



Permasalahan yang terjadi dari sektor pengelolaan sampah yang belum memadai diakibatkan oleh sistem / alur pengelolaan yang tidak terintegrasi dengan baik. Hal tersebut mengakibatkan praktik membakar sampah masih tergolong tinggi di kalangan masyarakat di kedua kota. Dimana, praktik tersebut menyebabkan timbulnya emisi hasil pembakaran yang juga mencakup parameter CO, SO₂, dan NO₂. Emisi SO₂ biasanya dihasilkan oleh pembakaran sampah anorganik seperti plastik dan kertas, sedangkan untuk emisi NO₂ biasanya dihasilkan oleh pembakaran sampah organik

PARAMETER PENCEMAR	STANDART	KONSENTRASI TERTINGGI	RATA - RATA
CO, mg / m ³	6	29,4	7
SO ₂ , µg / m ³	80	69	30
NO ₂ , µg / m ³	80	249	81



Jika dibandingkan dengan India Air Quality Standart, konsentrasi dari parameter pencemar memiliki batas nilai standartnya tersendiri. Dari dataset kami ditemukan bahwa konsentrasi SO₂ di udara masih tergolong memenuhi standar dibandingkan dengan konsentrasi CO dan juga NO₂, dimana nilai maks dan rata - rata kedua parameter tersebut cenderung melebihi batas konsentrasi standar.



R ECOMENDATION

The word "RECOMENDATION" is written in white, bold, sans-serif capital letters. To its left, a large yellow letter "R" is positioned on a purple rectangular background. A blue wavy line is drawn below the main text. An orange arrow points from the top right towards the "E" in "RECOMENDATION".

1. Praktik uji emisi kendaraan di India masih terdapat anomali yang serius, akibat adanya ketidakpatuhan terhadap kode praktik, tidak tersedianya sertifikat alat uji, dan kondisi laboratorium yang tidak memadai. Selain itu, jenis gas yang diuji pada emisi kendaraan hanya mencakup CO₂, HC, O₂, dan CO. Sedangkan gas seperti SO₂ dan NO₂ yang juga merupakan emisi hasil pembakaran bahan bakar tidak dilakukan uji emisi. Sehingga rekomendasi yang dapat kami berikan adalah meninjau ulang praktik uji emisi di kedua kota dan melakukan peningkatan kualitas alat uji emisi dan standar uji emisi dengan menjadikan negara lain sebagai acuan.
2. Mengalihkan penggunaan bahan bakar kendaraan yang menghasilkan emisi tinggi menjadi bahan bakar dengan emisi yang lebih kecil dan ramah lingkungan khususnya pada kendaraan umum
3. Membuat regulasi terkait dengan upaya pengurangan jumlah kendaraan tua yang tidak memenuhi syarat uji emisi.
4. Memperbanyak ruang terbuka hijau di area jalan yang memiliki kepadatan lalu lintas yang tinggi seperti dengan membuat jalur hijau tepi atau tengah jalan dan taman pulau jalan, seperti yang sudah diterapkan di beberapa kota besar di Indonesia. Dimana, jalur hijau tersebut nantinya dapat ditanami beberapa tanaman pengikat polutan udara.
5. Melakukan peninjauan ulang terkait dengan alur pengelolaan sampah rumah tangga di kedua kota, dan menggunakan teknologi plasma dengan prinsip waste to energy sebagai salah satu teknologi pengelolaan sampah yang ramah lingkungan.



DAFTAR

Pustaka



AQI Category

<https://www.epa.gov/outdoor-air-quality-data/air-data-basic-information>

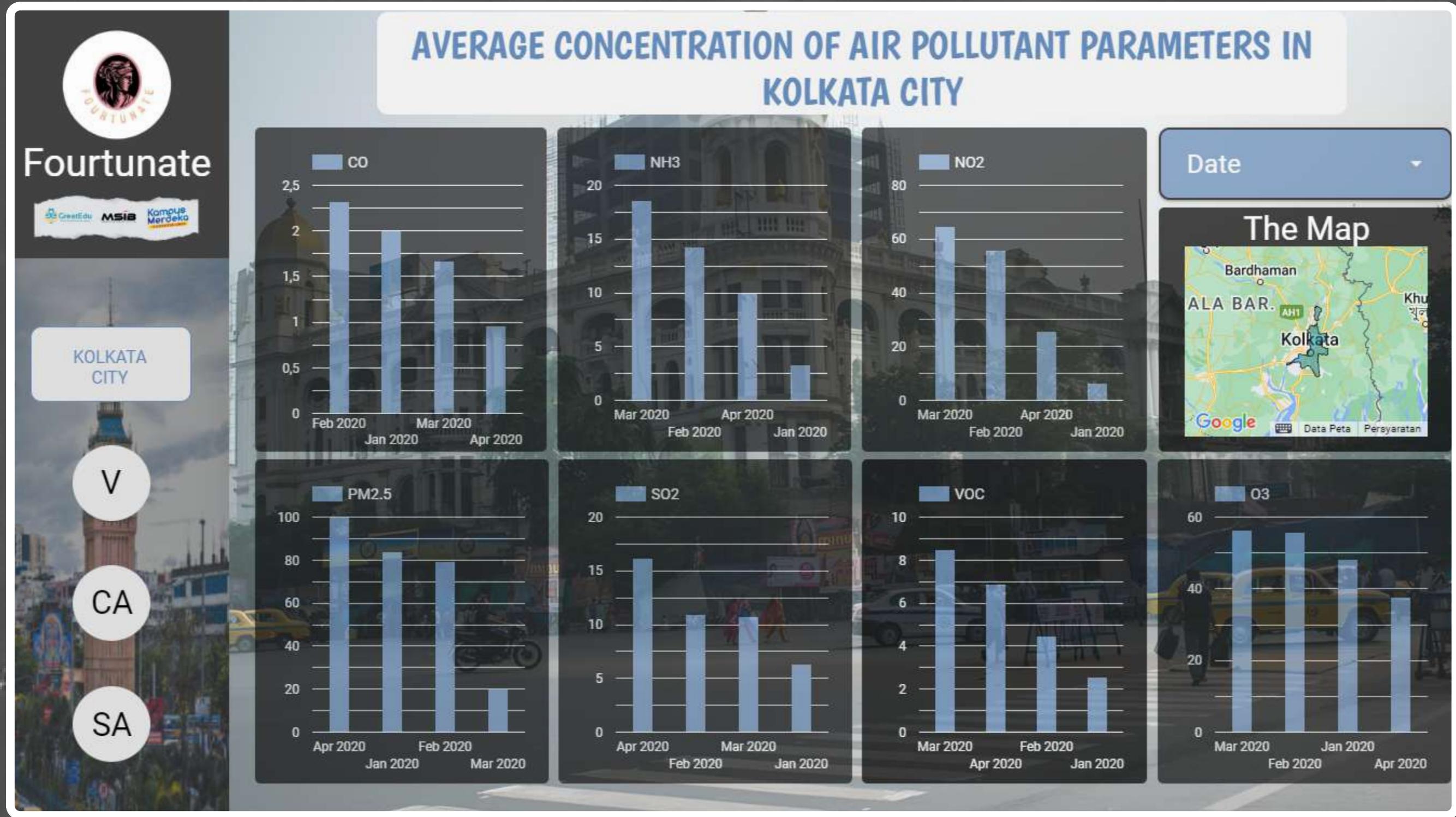
India: Air Quality Standard

<https://www.transportpolicy.net/standard/india-air-quality-standards/>

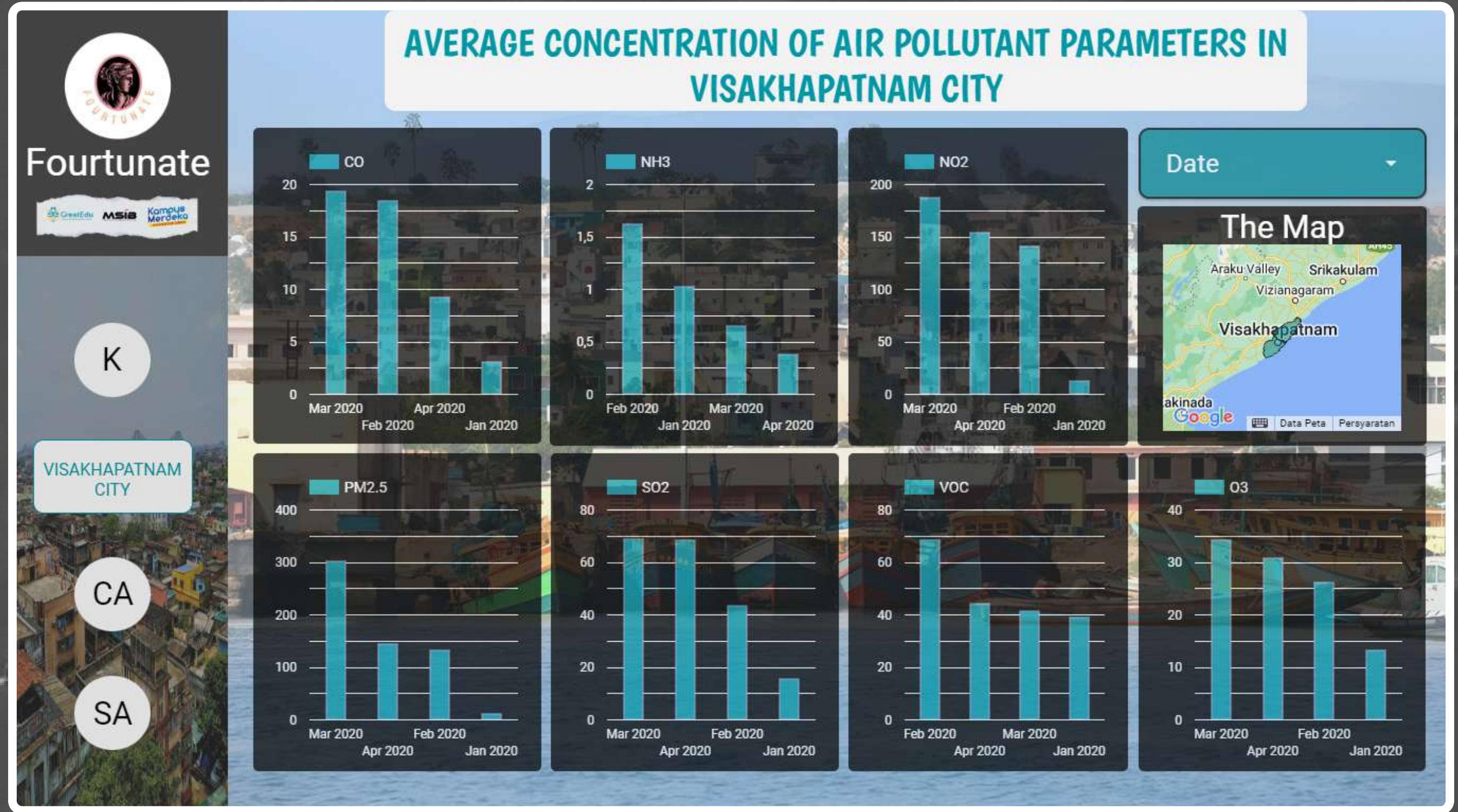
VISUALIZATION

With Looker Studio

Dashboard 1 Average Air Pollutant in Kolkata City



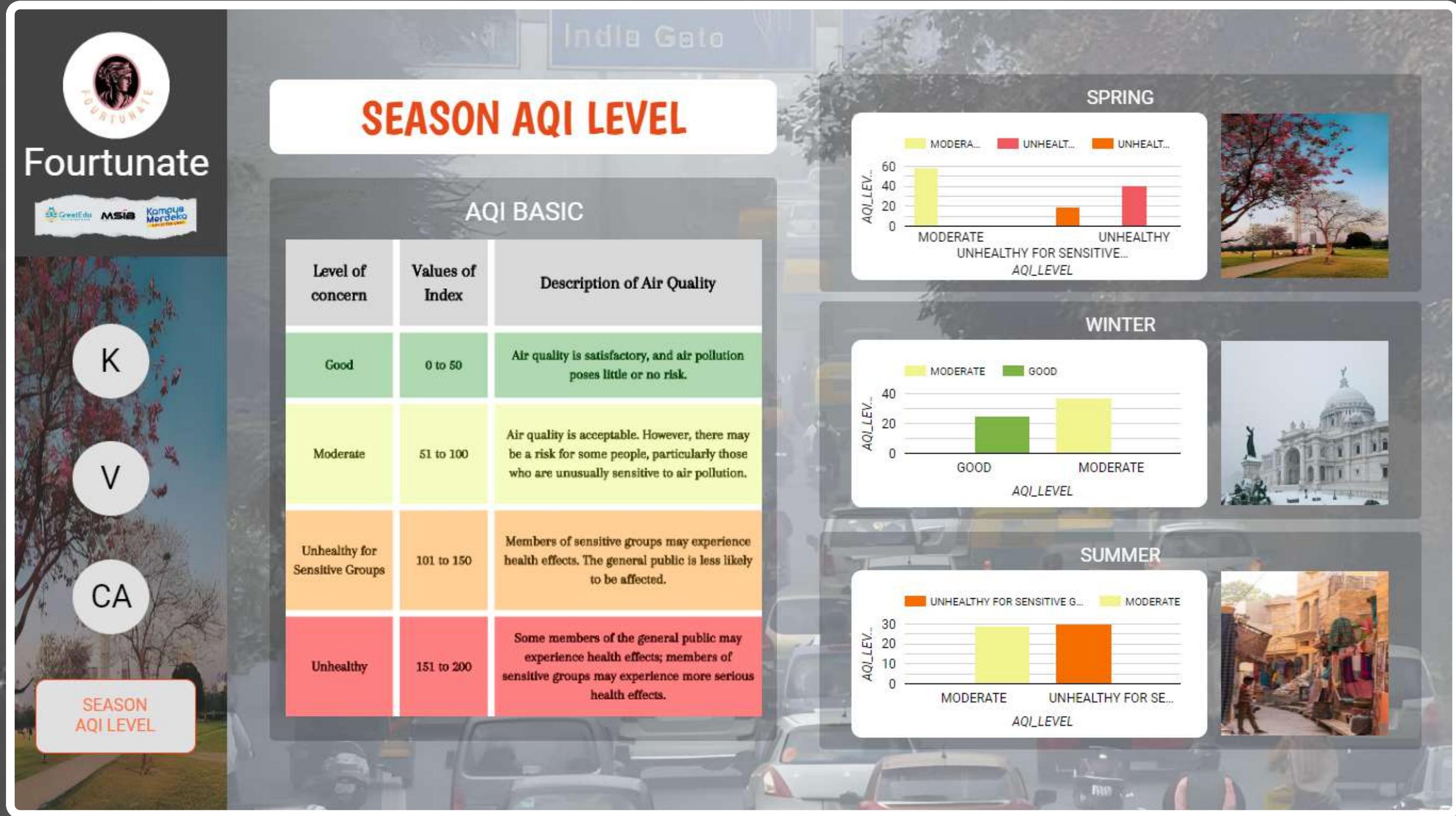
Dashboard 2 Average Air Pollutant in Visakhapatnam City

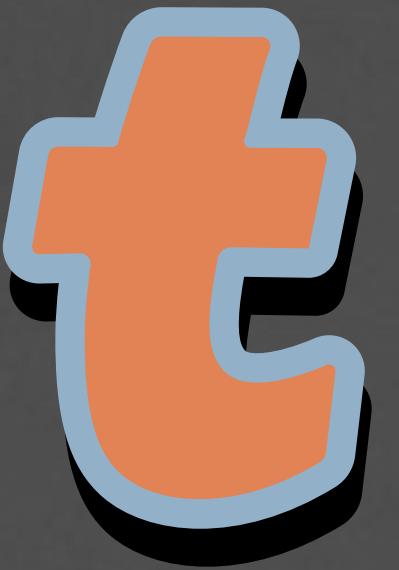


Dashboard 3 City AQI Level



Dashboard 4 Season AQI Level





THANK YOU

R

