



# **AIR QUALITY DATA IN INDIA (2015 - 2020)**

# Dataset



## Air Quality Data in India (2015 - 2020)

Air Quality Index (AQI) and hourly data across stations and cities in India

 [kaggle.com](https://www.kaggle.com)

# Dataset

city\_day

29,531 rows

city\_hour

707,875 rows

stations

station\_day

108,035 rows

station\_hour

2,589,083 rows

# Dataset

city\_day

29,531 rows

city\_hour

707,875 rows

stations

station\_day

108,035 rows

station\_hour

2,589,083 rows

## city\_hour

707,875 rows

City	Datetime	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
Ahmedabad	2015-01-01 01:00:00			1.0	40.01	36.37		1.0	122.07		0.0	0.0	0.0		
Ahmedabad	2015-01-01 02:00:00			0.02	27.75	19.73		0.02	85.9		0.0	0.0	0.0		
Ahmedabad	2015-01-01 03:00:00			0.08	19.32	11.08		0.08	52.83		0.0	0.0	0.0		
Ahmedabad	2015-01-01 04:00:00			0.3	16.45	9.2		0.3	39.53	153.58	0.0	0.0	0.0		
Ahmedabad	2015-01-01 05:00:00			0.12	14.9	7.85		0.12	32.63		0.0	0.0	0.0		
Ahmedabad	2015-01-01 06:00:00			0.33	15.95	10.82		0.33	29.87	64.25	0.0	0.0	0.0		
Delhi	2019-04-09 00:00:00	75.27	202.58	29.34	62.58	60.97	36.48	1.42	20.31	26.65	4.65	30.29	0.74	250.1	Poor
Delhi	2019-04-09 01:00:00	81.27	207.83	34.6	56.81	62.14	38.13	1.3	23.94	22.1	4.06	29.55	1.02	249.1	Poor
Delhi	2019-04-09 02:00:00	91.15	219.33	37.76	55.11	63.64	41.47	1.25	26.74	21.34	5.16	28.97	0.94	251.1	Poor
Delhi	2019-04-09 03:00:00	102.59	221.95	40.22	50.86	63.98	41.08	1.43	24.71	26.08	6.08	31.63	1.01	252.1	Poor
Delhi	2019-04-09 04:00:00	111.71	227.68	43.64	51.33	67.49	40.16	1.26	25.91	30.69	6.99	31.96	1.13	252.1	Poor
Delhi	2019-04-09 05:00:00	120.5	227.96	45.36	51.97	69.67	38.96	1.22	26.43	27.62	6.77	29.27	0.54	255.1	Poor
Delhi	2019-04-09 06:00:00	111.28	211.44	44.22	50.14	67.36	38.31	1.26	26.94	29.78	5.98	30.58	0.21	254.1	Poor

# 26 Cities

Ahmedabad	Guwahati
Aizawl	Hyderabad
Amaravati	Jaipur
Amritsar	Jorapokhar
Bengaluru	Kochi
Bhopal	Kolkata
Brajrajnagar	Lucknow
Chandigarh	Mumbai
Chennai	Patna
Coimbatore	Shillong
DelhiE rnakulam	Talcher
Gurugram	Thiruvananthapuram



เมืองอาห์มาดาบัด  
(Ahmedabad)

เมืองเดลี  
(Delhi)



# Dataset

City

Datetime

PM2.5

PM10

NO

NO2

NOx

NH3

CO

SO2

O3

Benzene

Toluene

Xylene

AQI

AQI\_Bucket

# 12 Attributes

**PM2.5**

Particulate matter with diameter of less than 2.5 micron  
ฝุ่นละอองขนาดจิ๋ว (เส้นผ่าศูนย์กลาง **ไม่เกิน 2.5 ไมครอน**)

**PM10**

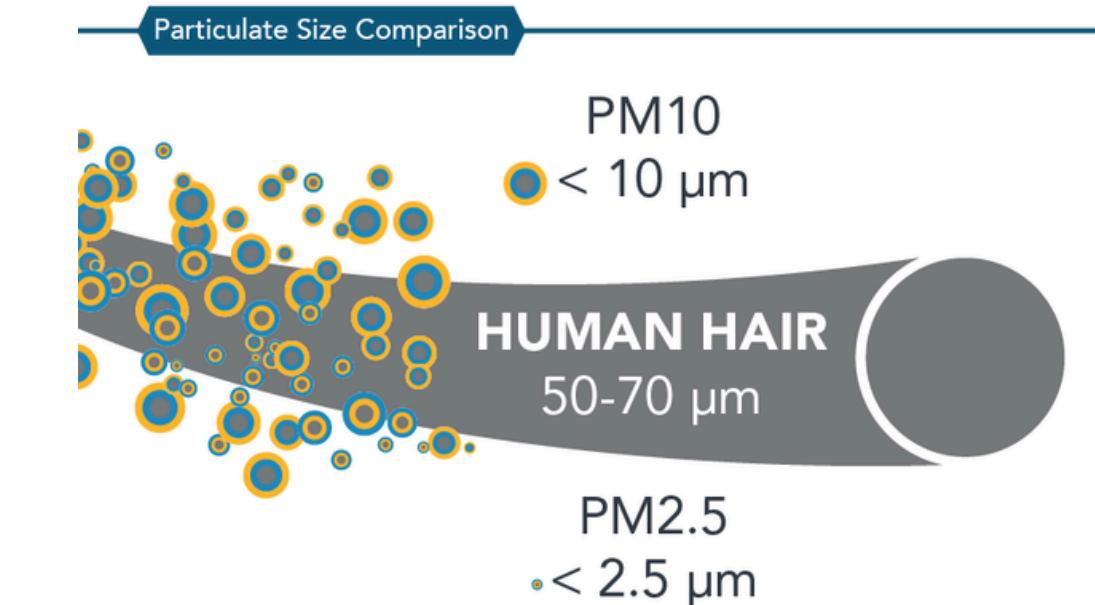
Particulate matter-10 micron  
ฝุ่นหยาบ (เส้นผ่านศูนย์กลาง **2.5 – 10 ไมครอน**)

**NO**

Nitric Oxide  
ก๊าซเฉื่อยมีคุณสมบัติเป็นยาสลบ เป็นก๊าซ ไม่มีสีและไม่มีกลิ่น  
ในธรรมชาติทั่วไปพบในปริมาณ**น้อยกว่า 0.5 ppm.**

**NO<sub>2</sub>**

Nitrogen Dioxide  
ก๊าซที่เกิดจากการรวมตัวของ NO กับ O<sub>2</sub> ในอากาศ ไม่มีสีและกลิ่น ละลายน้ำได้เล็กน้อย  
มีอยู่ทั่วไปในธรรมชาติ หรือเกิดจากการกระทำของมนุษย์ เช่น การเผาไหม้เชื้อเพลิง โดยเฉพาะเชื้อเพลิงฟอสซิล  
เป็นก๊าซที่อันตรายต่อมนุษย์ในระบบทางเดินหายใจ ระคายเคือง หายใจลำบาก กระตุ้นโรคหอบหืด เพิ่มความเสี่ยงของโรคปอด



# 12 Attributes

NOx

Nitrogen Oxide (NOx) ใช้เรียกกลุ่มของก๊าซที่ประกอบด้วย NO และ NO<sub>2</sub> ไม่มีสีและกลิ่น เมื่อรวมตัวกับอนุภาคอื่นในอากาศจะเห็นเป็นชั้นสีดำตากแดด เกิดขึ้นจากการเผาไหม้เชื้อเพลิงฟอสซิล เช่น ถ่านหิน ห้ามัน ก๊าซธรรมชาติ (มีเทน)

NH<sub>3</sub>

Ammonia  
เกิดขึ้นได้จากการเกษตรกรรม การหมัก หรือจากการเผาไหม้ในภาคอุตสาหกรรม แอมโมเนียมสามารถระคายเคืองต่องานการเดินหายใจ และทำให้เกิดฝุ่นละอองในอากาศ

CO

Carbon Monoxide  
เกิดจากการเผาไหม้ไม่สมบูรณ์ เช่น จากรถยนต์และการเผาไหม้ของเชื้อเพลิง เมื่อหายใจเข้าไป CO สามารถลดปริมาณออกซิเจนในเลือด ทำให้เกิดอาการมึนงงและอาจเป็นอันตรายถึงชีวิต

SO<sub>2</sub>

Sulfur Dioxide  
เกิดจากการเผาไหม้ของเชื้อเพลิงที่มีซัลเฟอร์ เช่น ถ่านหินและห้ามัน ทำให้เกิดการระคายเคืองในระบบทางเดินหายใจ และทำให้เกิดฝุ่นละอองในอากาศ



# 12 Attributes

O<sub>3</sub>

มีบทบาทในการกรองรังสีอัลตราไวโอเลตจากดวงอาทิตย์  
แต่ในระดับพื้นผิวโลก โอโซนสามารถเป็นมลพิษทางอากาศที่ทำให้เกิดปัญหาทางสุขภาพ  
เช่น การระคายเคืองในระบบทางเดินหายใจและการหายใจลำบาก

Benzene

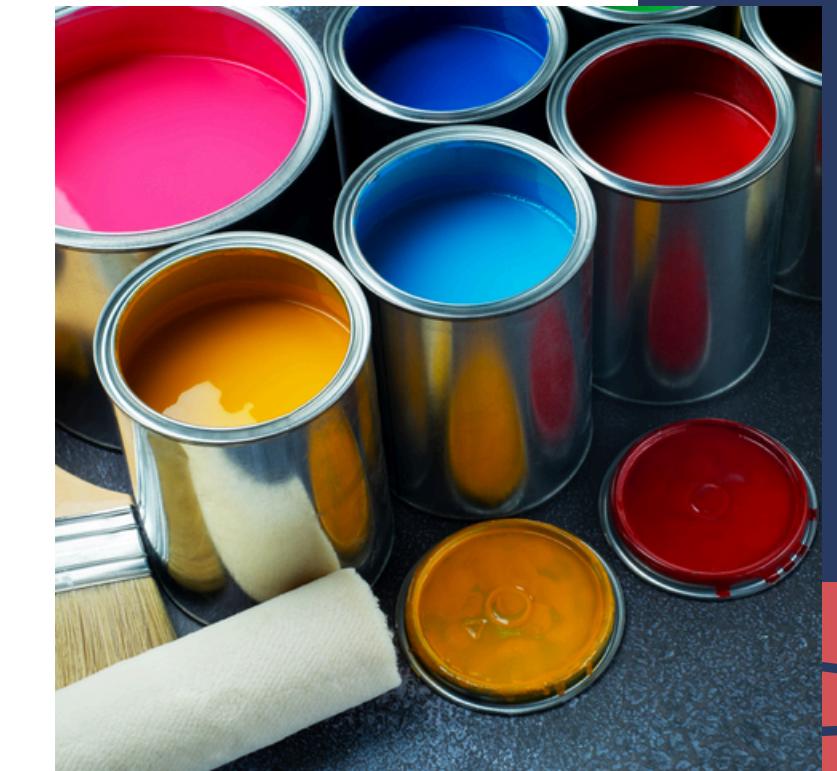
สารเคมีที่พบในมลพิษทางอากาศจากการเผาไหม้ที่ไม่สมบูรณ์หรือในผลิตภัณฑ์ที่เกี่ยวข้องกับปetroเคมี  
เป็นสารก่อมะเร็งที่มีผลกระทบต่อระบบเลือดและสามารถทำให้เกิดโรคมะเร็ง

Toluene

สารประกอบทางเคมีที่ใช้ในอุตสาหกรรมต่าง ๆ เช่น การทำสีหรือสารกำลังล้าย  
มีผลกระทบต่อระบบประสาทส่วนกลาง ทำให้เกิดอาการเวียนหัว มึนงง และหายใจลำบาก

Xylene

สารเคมีในกลุ่มอารีโมาไตร์ (aromatic hydrocarbons) มีหลายรูปแบบ เช่น ortho-xylene, meta-xylene  
มักใช้ในอุตสาหกรรมทำสี น้ำมัน สารเคมี และทำเป็นสารกำลังล้ายต่าง ๆ  
การสูดหายใจเข้าไปในระยะยาวสามารถส่งผลต่อระบบประสาทส่วนกลาง ทำให้รู้สึกมึนงง เวียนหัว หรือท้องเสีย



# 2 Targets

AQI

AQI\_Bucket

Prediction

Classification

# 2 Targets

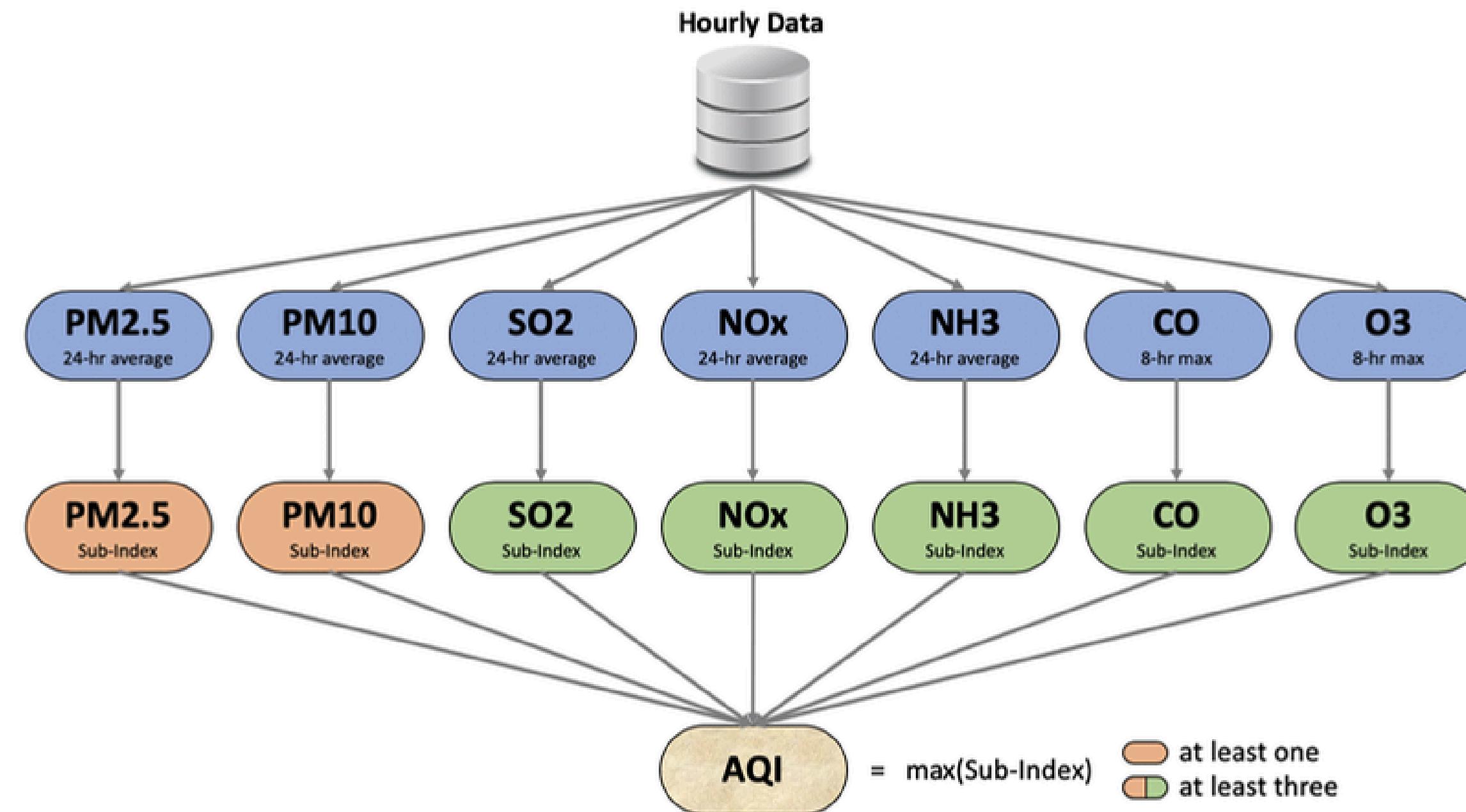
AQI

Prediction

AQI\_Bucket

Classification

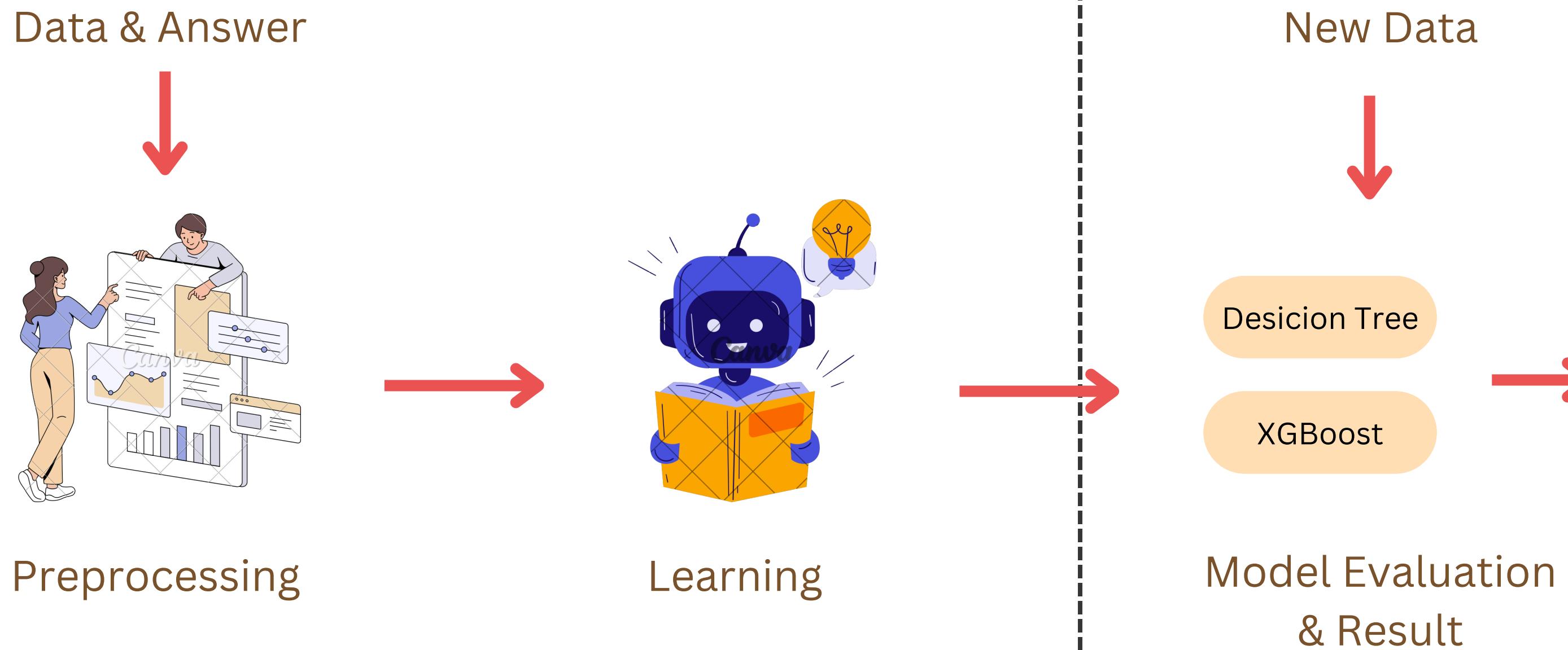
# AQI (Air Quality Index)



# AQI\_Bucket

<b>Good (0–50)</b>	Minimal Impact	<b>Poor (201–300)</b>	Breathing discomfort to people on prolonged exposure
<b>Satisfactory (51–100)</b>	Minor breathing discomfort to sensitive people	<b>Very Poor (301–400)</b>	Respiratory illness to the people on prolonged exposure
<b>Moderate (101–200)</b>	Breathing discomfort to the people with lung, heart disease, children and older adults	<b>Severe (&gt;401)</b>	Respiratory effects even on healthy people

<https://www.kaggle.com/code/rohanrao/calculating-aqi-air-quality-index-tutorial?scriptVersionId=41199538>



# Missing Value

	>Missing Values	> Total Null Values %
City	0	0.000000
Datetime	0	0.000000
PM2.5	145088	20.496274
PM10	296737	41.919407
NO	116632	16.476355
NO2	117122	16.545577
NOx	123224	17.407593
NH3	272542	38.501430
CO	86517	12.222073
SO2	130373	18.417517
O3	129208	18.252940
Benzene	163646	23.117923
Toluene	220607	31.164683
Xylene	455829	64.393996
AQI	129080	18.234858
AQI_Bucket	129080	18.234858

## Step 1

เนื่องจากมี row ที่ไม่มีข้อมูล AQI และ AQI\_Bucket ซึ่งเราจะใช้ในการดูว่าไม่เดาเราทำนายถูกหรือไม่

ดังนั้น เราจึง drop row ดังกล่าว ออกไปเหลือเพียงแค่ข้อมูลที่มี label

```
dataset = city_hour.dropna(subset=['AQI', 'AQI_Bucket'])

null_counts = dataset.isnull().sum()
null_counts
```

Ahmedabad	2017-02-23 23:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 00:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 01:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 02:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 03:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 04:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 05:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 06:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 07:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 08:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 09:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 10:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 11:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 12:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 13:00:00	0.0	0.0	0.0
Ahmedabad	2017-02-24 14:00:00	0.0	0.0	0.0

## Step 2

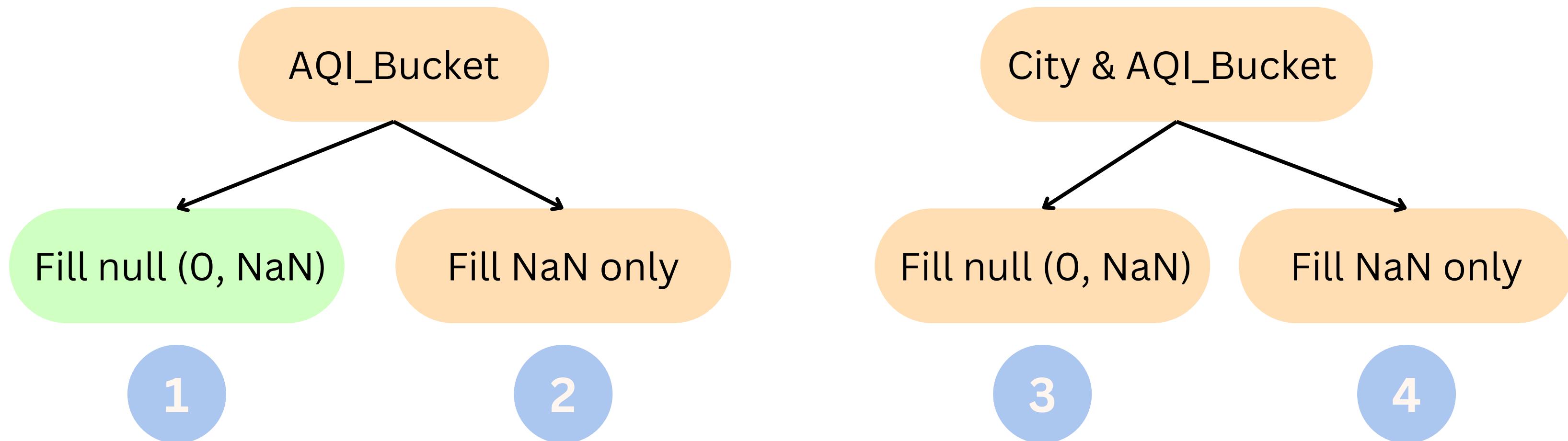
เราจะพบว่า **Xylene** มีข้อมูลเป็น null ถึง 371,700 ชั่งถือว่าเกิน 50% ของ dataset ของเราเลย ดังนั้นเราจะ drop feature นี้ออกไป

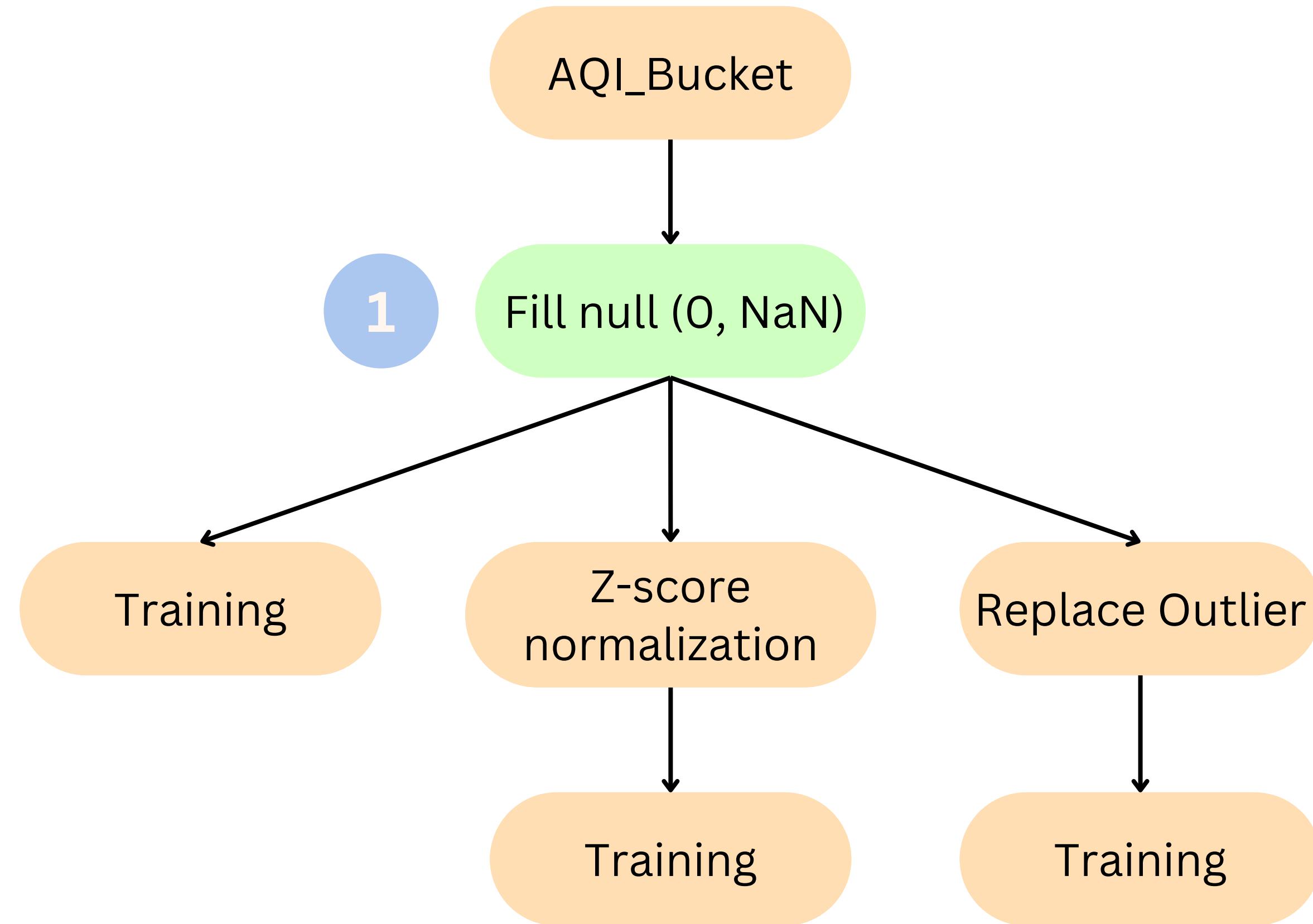
นอกจากนี้ เราจะขอ drop ในส่วนของ **Datetime** ออกไปด้วยเช่นกัน เพราะไม่ได้สำคัญกับการ train

0	City	578795	non-null	object	
1	PM2.5	547760	non-null	float64	
2	PM10	399932	non-null	float64	
3	N0	557314	non-null	float64	
4	N02	556528	non-null	float64	
5	N0x	526889	non-null	float64	
6	NH3	414801	non-null	float64	
7	C0	549649	non-null	float64	
8	S02	545737	non-null	float64	
9	O3	545423	non-null	float64	
10	Benzene	482430	non-null	float64	
11	Toluene	429794	non-null	float64	
12	AQI	578795	non-null	float64	
13	AQI_Bucket	578795	non-null	object	

# Handle Missing Value by filling Mean

19

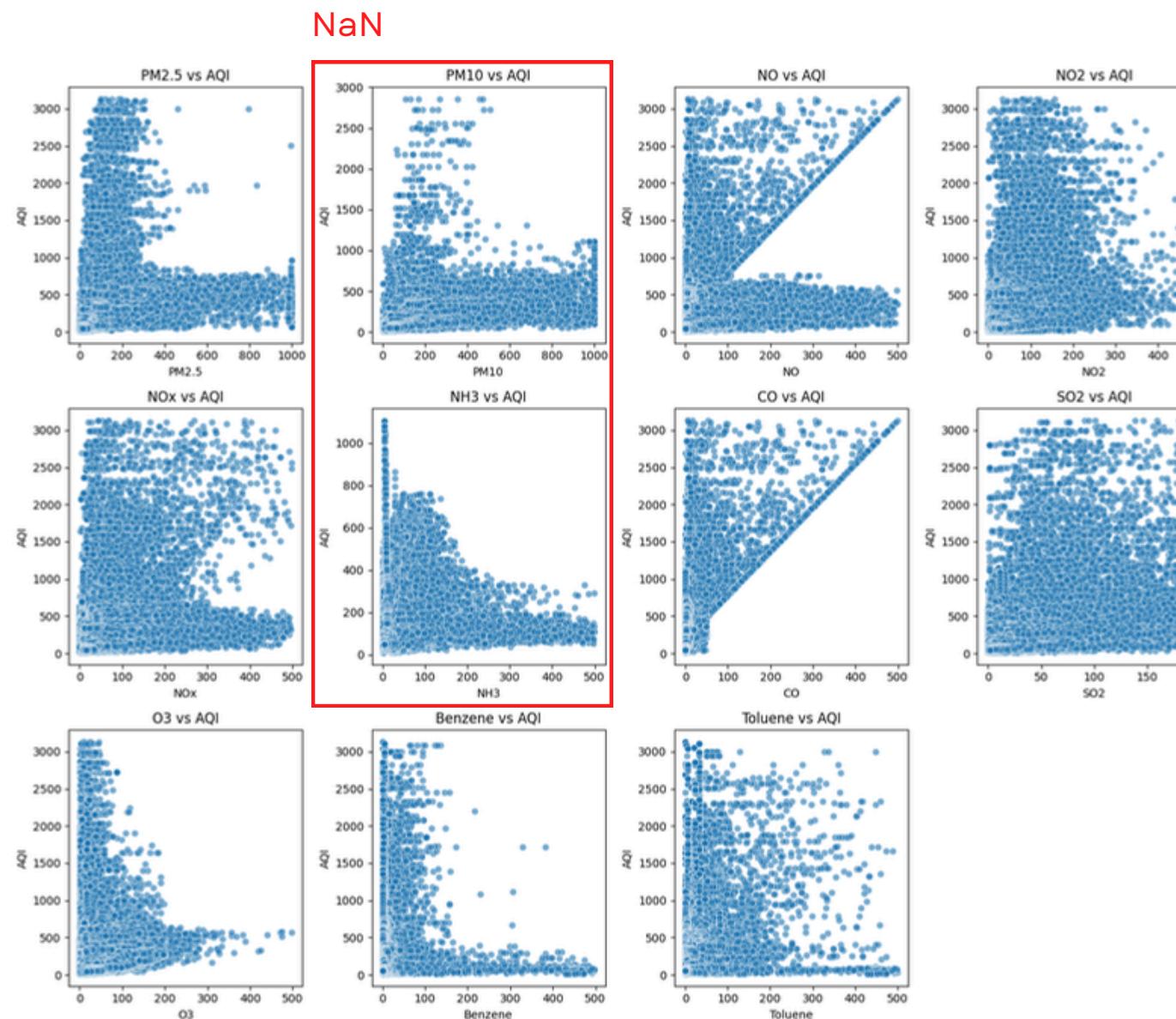




1

# Comparison

SCATTER PLOT OF EACH ATTRIBUTE - AQI



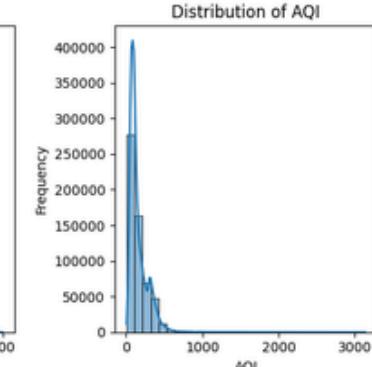
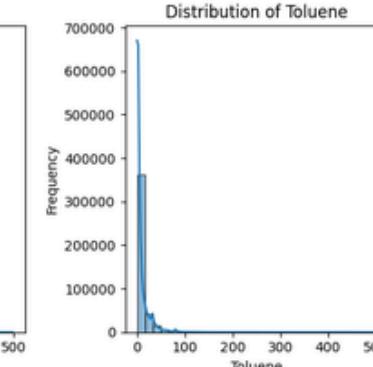
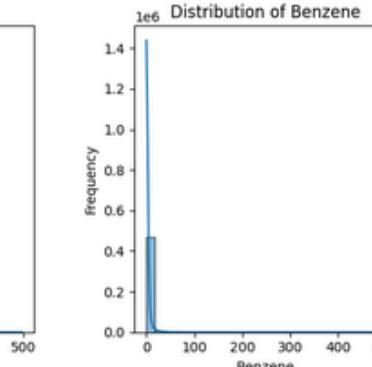
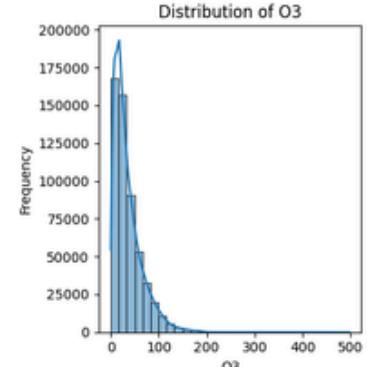
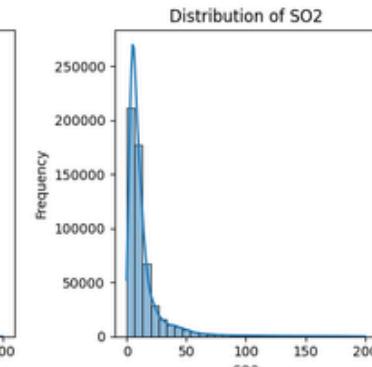
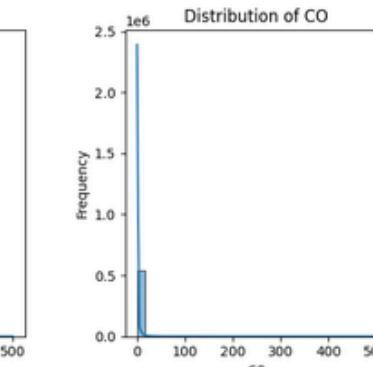
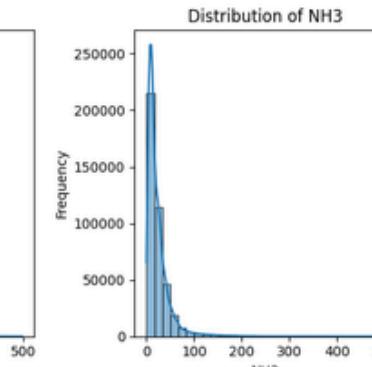
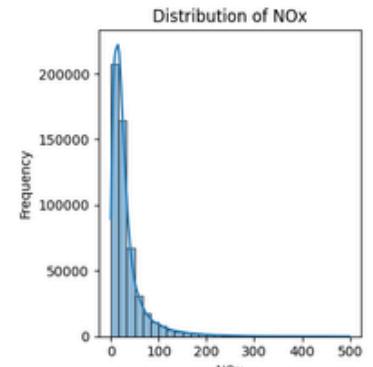
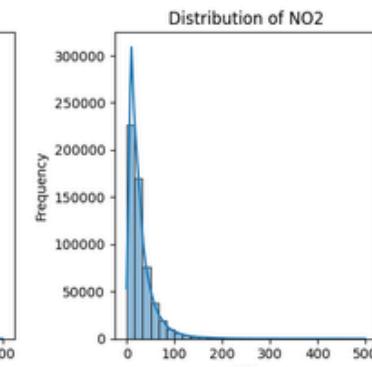
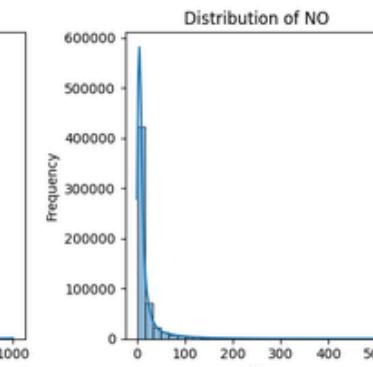
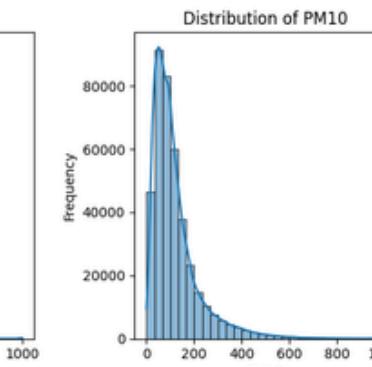
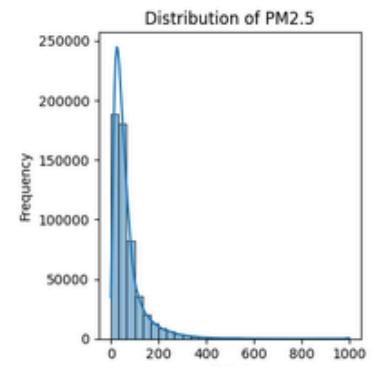
BEFORE CLEANING



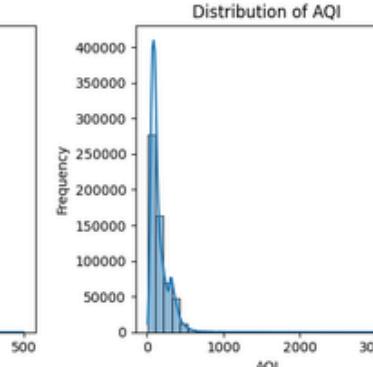
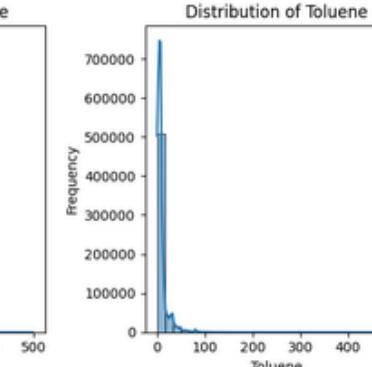
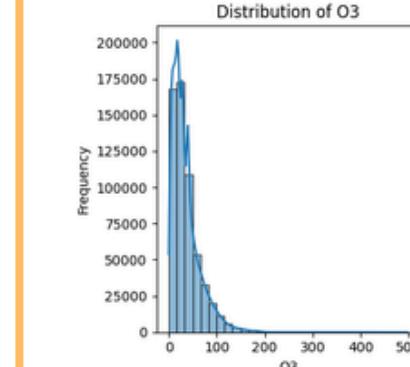
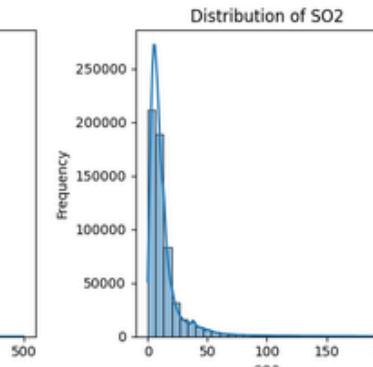
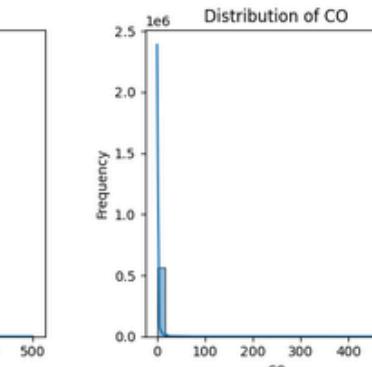
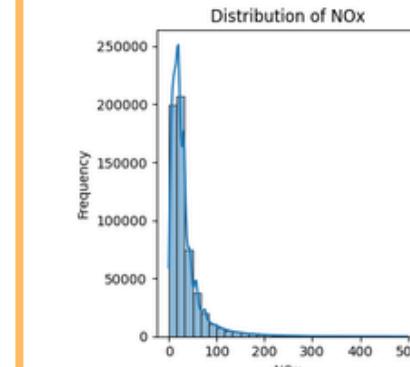
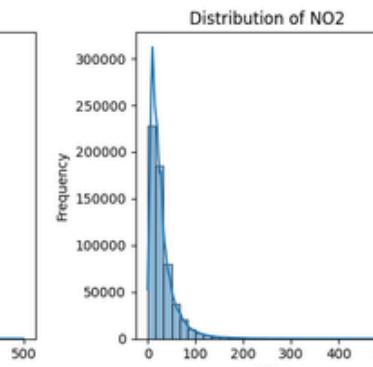
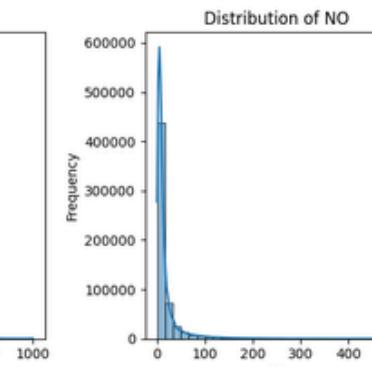
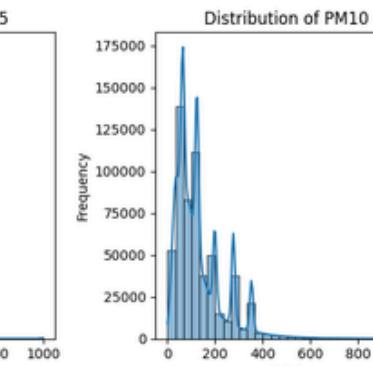
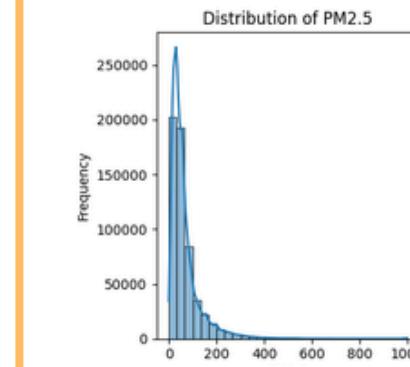
AFTER CLEANING

# Comparison

## DISTRIBUTION OF ATTRIBUTE



BEFORE CLEANING



AFTER CLEANING

# Z-score normalization

23

เนื่องจากมีความแตกต่างของ range ของข้อมูลสูง  
เช่น

NO2 มีค่าตั้งแต่ 0.01 - 499.51

PM2.5 มีค่าตั้งแต่ 0.01 - 999.99

```
# เตรียมข้อมูล (แยก features และ target)
X = filled_dataset.drop(columns=['AQI_Bucket', 'AQI', 'City']) # features (ลบ 'City' ออก)
y = filled_dataset['AQI_Bucket'] # target

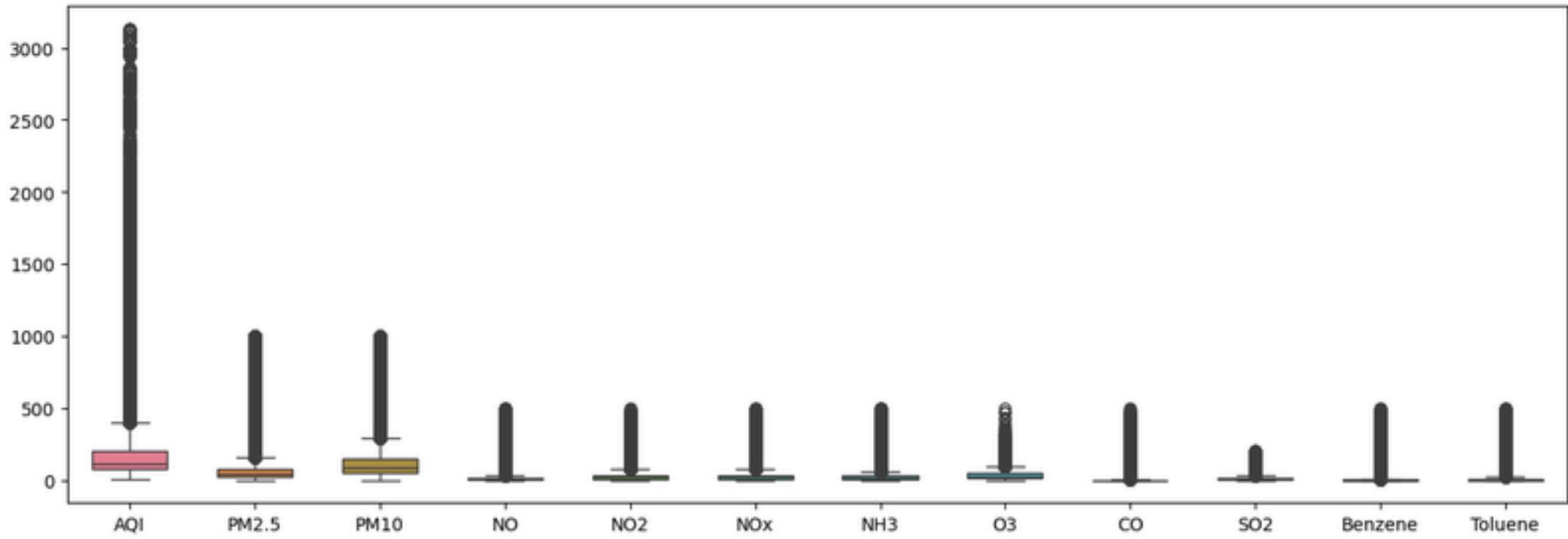
# Z-score normalization (ทำ normalization ก่อนแบ่งข้อมูล)
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

# Comparison

IQR OF ATTRIBUTE

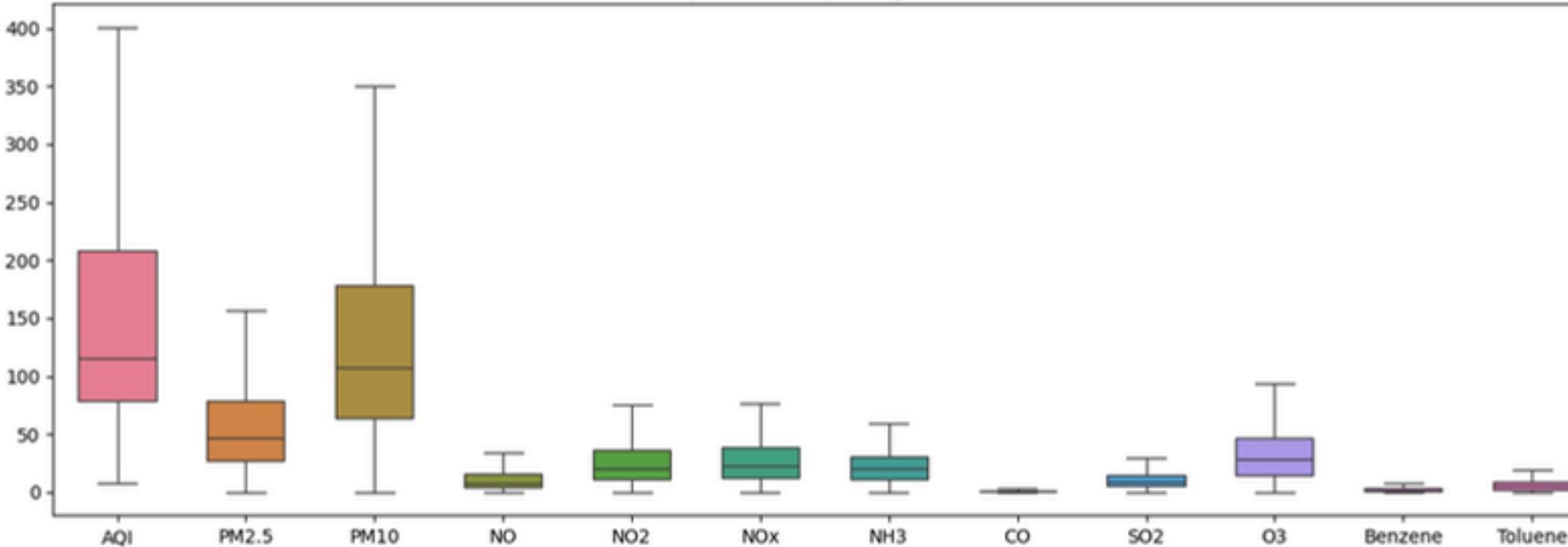
24

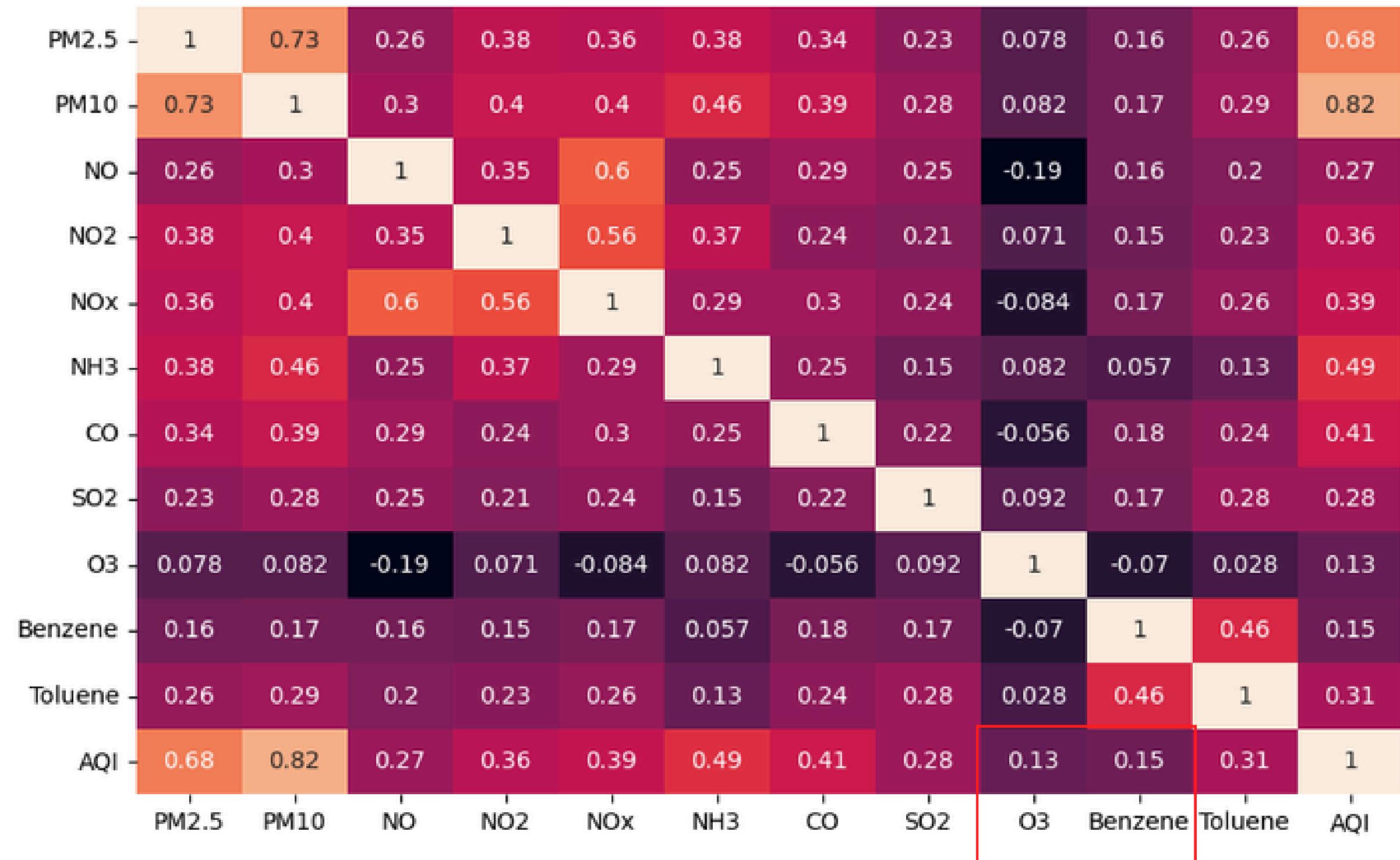
BEFORE CLEANING



AFTER CLEANING

Boxplot After Replacing Outliers





สำหรับวิธี Outlier เราจะทำการตัด O3 และ Benzene ออกไป  
เนื่องจากมีความสัมพันธ์ (correlation) กับค่า AQI ที่ต่ำ

# Replace Outlier

26

ประโยชน์ของวิธีนี้  
ลดผลกระทบจากค่าผิดปกติ โดยไม่ต้องลบข้อมูล ใช้ได้กับชุดข้อมูลที่ใช้ใน  
Regression และ Classification

คำนวณค่า Q1 และ Q3

- Q1 (25th percentile): ค่าแบ่งข้อมูล 25% ล่าง
- Q3 (75th percentile): ค่าแบ่งข้อมูล 75% ล่าง

หาช่วง IQR (Interquartile Range)

- $IQR = Q3 - Q1$

# Replace Outlier

27

คำนวณขอบเขตการตรวจจับ Outlier

- ขอบเขตล่าง:  $Q1 - 1.5 \times IQR$
- ขอบเขตบน:  $Q3 + 1.5 \times IQR$

แทนค่าที่ผิดปกติ / outlier

- ค่าที่น้อยกว่าขอบเขตล่าง  $\rightarrow$  แทนด้วย  $Q1$
- ค่าที่มากกว่าขอบเขตบน  $\rightarrow$  แทนด้วย  $Q3$

# Cross-validation

28

เราใช้การ cross-validation แบบ 5-fold เพื่อตรวจสอบว่า  
โมเดลทำงานได้ดีในชุดข้อมูลที่แตกต่างกันอย่างไร

```
cv_scores = cross_val_score(dt_model, X_train, y_train, cv=5)
print(f"Cross-validation scores: {cv_scores}")
print(f"Mean cross-validation score: {cv_scores.mean()}")
```

→ Cross-validation scores: [0.85504492 0.8560368 0.85769974 0.85748378 0.85568046]  
Mean cross-validation score: 0.8563891389529505  
Test accuracy: 0.8607883620280065

# Training

From fillna mean by AQI\_Bucket

# Decision Tree

30

- เตรียมข้อมูล



แยก target กับ features

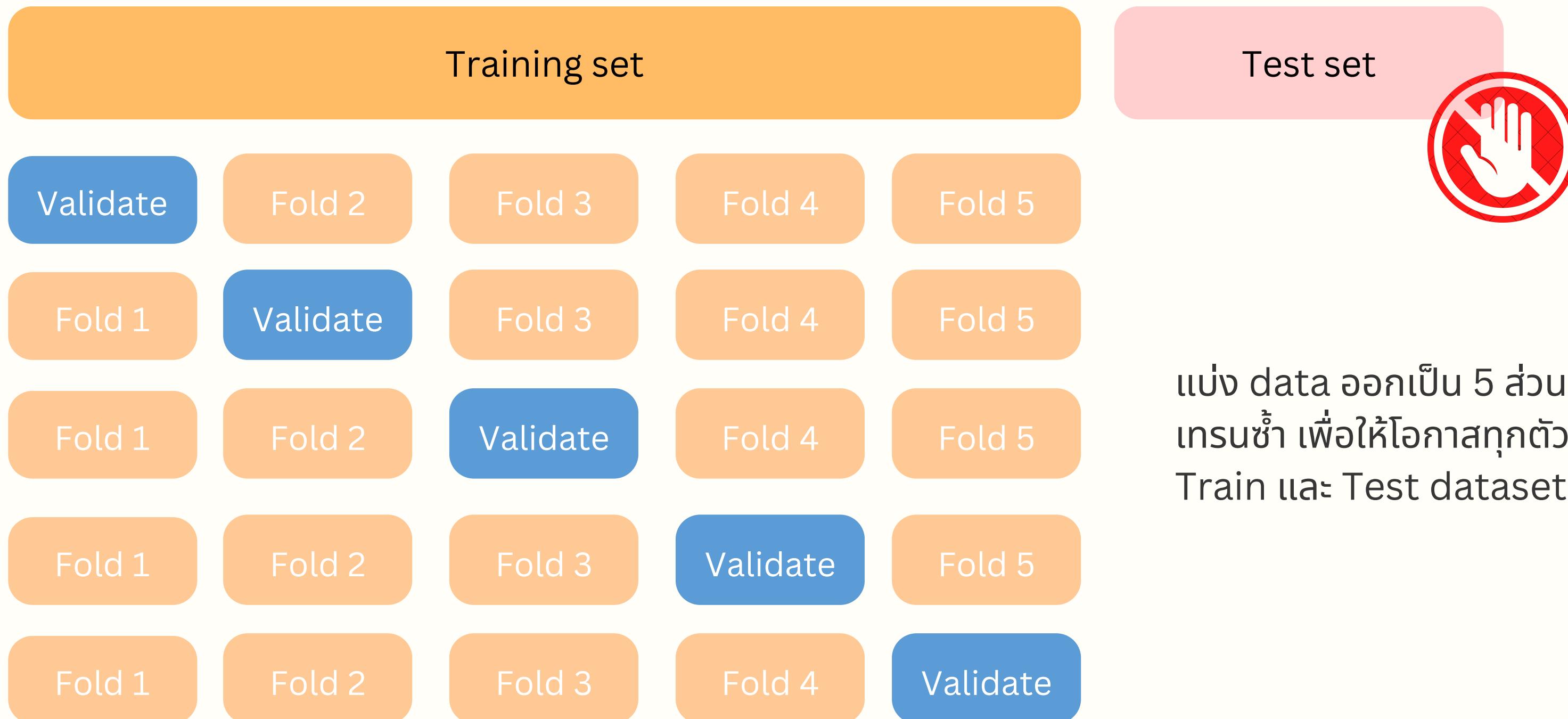
- features (X) คือปัจจัยที่มีผลต่อคุณภาพอากาศ เราจะlab column 'AQI\_Bucket', 'AQI', 'City' ทิ้งไป
- Target (y) ก็คือ column 'AQI\_Bucket'

# Decision Tree

- แบ่ง Dataset



```
9 # แบ่งข้อมูลเป็น train และ test (80:20)
10 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
11
```



แบ่ง data ออกเป็น 5 ส่วน ทำการ  
เทรนช้า เพื่อให้โอกาสทุกตัวเป็น<sup>↑</sup>  
Train และ Test dataset

5-fold cross validation

# Decision Tree

33

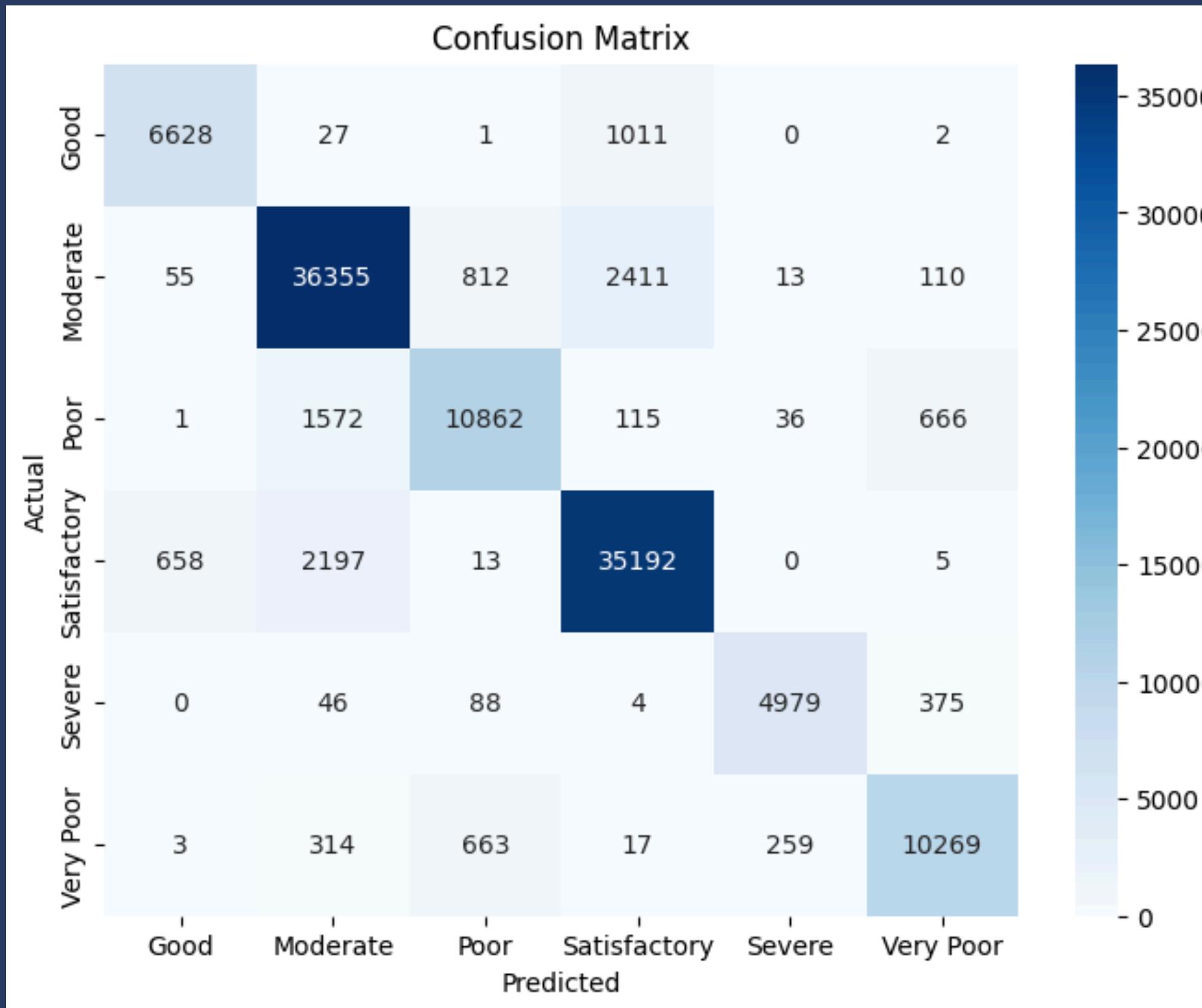
- Train model

เราจะฝึกโมเดลด้วยข้อมูลฝึก โดยใช้ `dt_model.fit()` และใช้ข้อมูล `test` ในการคำนวณผล และใช้ `accuracy_score` เพื่อประเมินว่าโมเดลสามารถคำนวณได้แม่นยำแค่ไหน

```
Cross-validation scores: [0.85503412 0.85592882 0.85771054 0.85744058 0.85570205]
Mean cross-validation score: 0.8563632230087956
Test accuracy: 0.8608747483996925
```

# Decision Tree

34

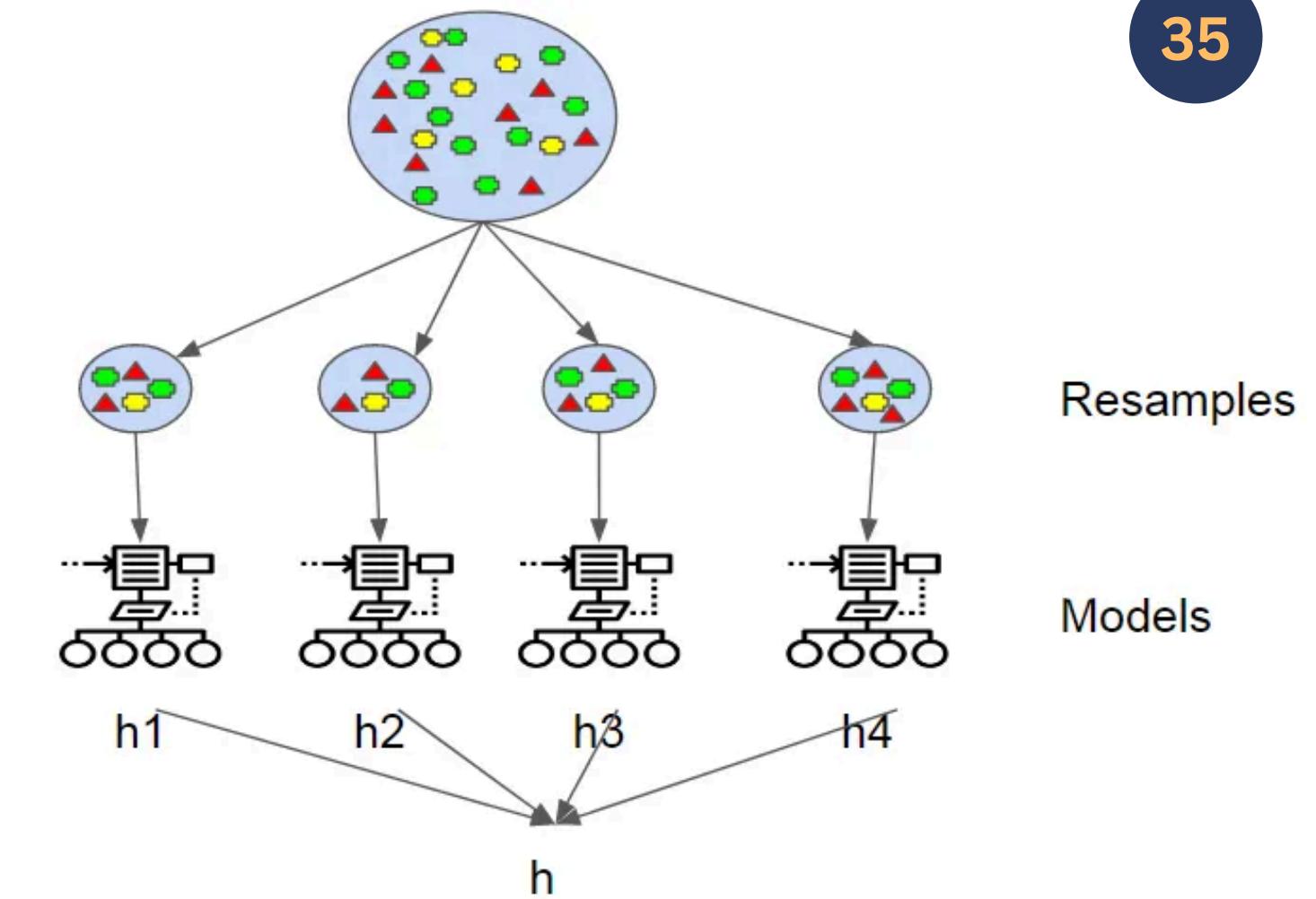


Classification Report:					
	precision	recall	f1-score	support	
Good	0.82	0.83	0.82	7669	
Moderate	0.87	0.87	0.87	39756	
Poor	0.80	0.81	0.80	13252	
Satisfactory	0.88	0.88	0.88	38065	
Severe	0.90	0.89	0.89	5492	
Very Poor	0.86	0.86	0.86	11525	
accuracy				0.86	115759
macro avg	0.85	0.85	0.85	115759	
weighted avg	0.86	0.86	0.86	115759	

# XGBoost

XGBoost เป็น ensemble learning method ซึ่งเป็นการเรียนรู้แบบ multiple-learners

- Bagging
- Stacking
- Boosting



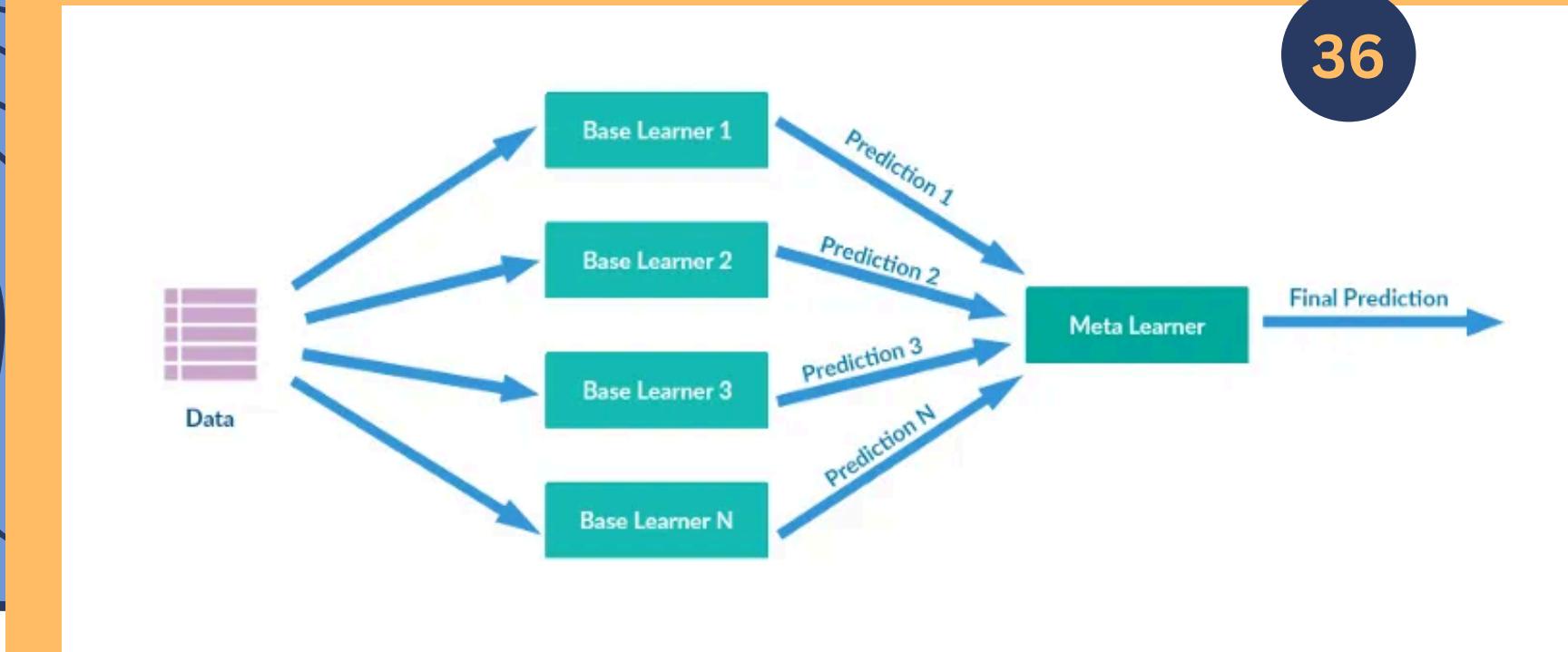
## Bagging

การสร้าง learner หลายตัว แล้วให้แต่ละตัวเรียน subset ของข้อมูลทั้งหมด หลังจากนั้นมาทำการ vote ว่าถ้ามีคำตามมา ดูซึ่งส่วนมากตอบอะไร หรือเฉลี่ยแล้วควรตอบอะไร และด้วยเทคนิคนี้เองทำให้เราสามารถลด variance ช่วยลดการเกิด overfit

# XGBoost

XGBoost เป็น ensemble learning method ซึ่งเป็นการเรียนรู้แบบ multiple-learners

- Bagging
- Stacking
- Boosting



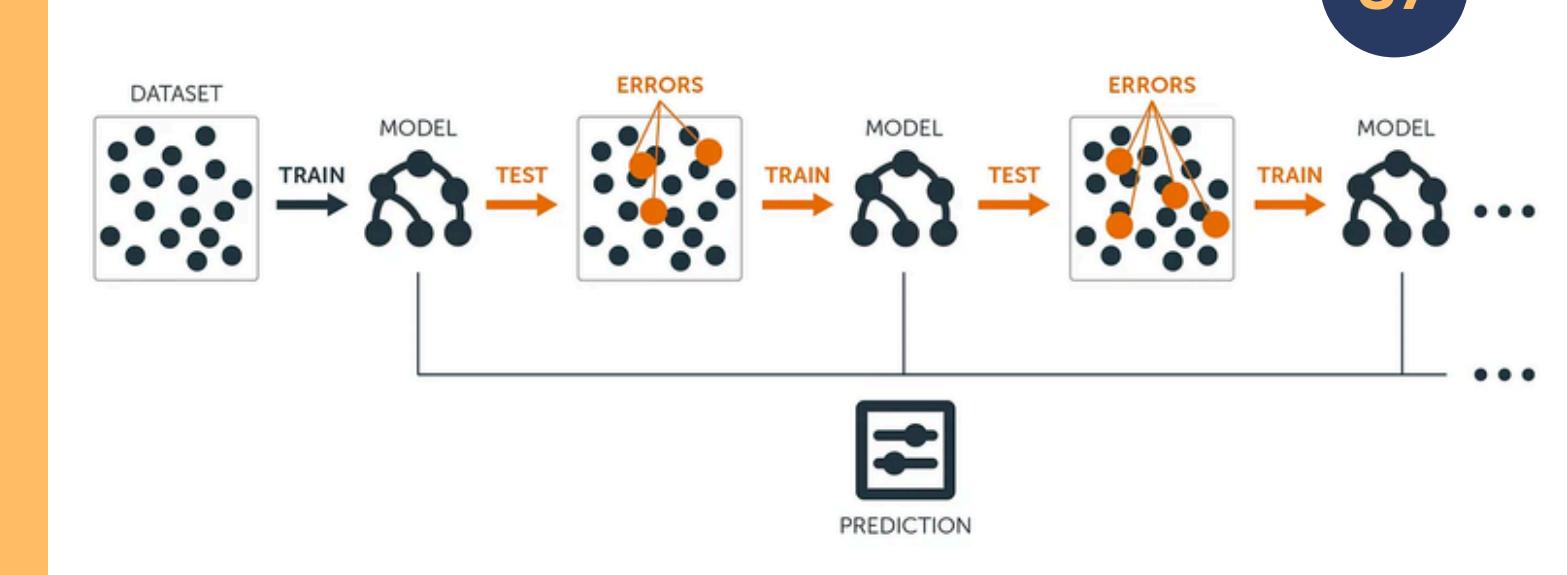
## Stacking

มีการแบ่ง learner เป็นหลายกลุ่ม และเอาข้อมูลทั้งหมดให้กลุ่มแรกเรียน และเอา "คำตอบ" ของกลุ่มแรกมารวมกันแล้ว ส่งต่อให้กลุ่มต่อมาเรียนต่อๆ กันไป

# XGBoost

XGBoost เป็น ensemble learning method ซึ่งเป็นการเรียนรู้แบบ multiple-learners

- Bagging
- Stacking
- Boosting



## Boosting

การเรียนแบบเป็นลำดับ โดย learner ก่อนหน้าเรียนแล้วนำเอา “ข้อผิดพลาด” ของตัวเอง มาปรับปรุง learner ต่อ ๆ ไป เพื่อลด error จาก learner ก่อนหน้า ส่งผลให้มี accuracy ที่ดีกว่า bagging (ลด bias) แต่ทำให้เกิด overfit ง่าย

# Result: Accuracy

38

	Used Data	Decision Tree	XGBoost
AQI_Bucket	Filled 0 & NaN	0.8609	0.9009
	Filled NaN only	0.8609	0.9009
	Filled 0 & NaN แล้วปรับโดย Z-score	0.8608	0.9009
AQI_Bucket & City	Filled 0 & NaN แล้ว replace Outlier	0.8525	0.8966
	Filled 0 & NaN	0.8401	0.8875
	Filled NaN only	0.8563	0.8885

# គ្របាថ្មីម្នូលែង

6434450523

នាយកវន្ធិត

ធម្មព័ន្ធ

6434457023

ន.ស. ជ័ទម្យ

សនលិននរុ

6434458623

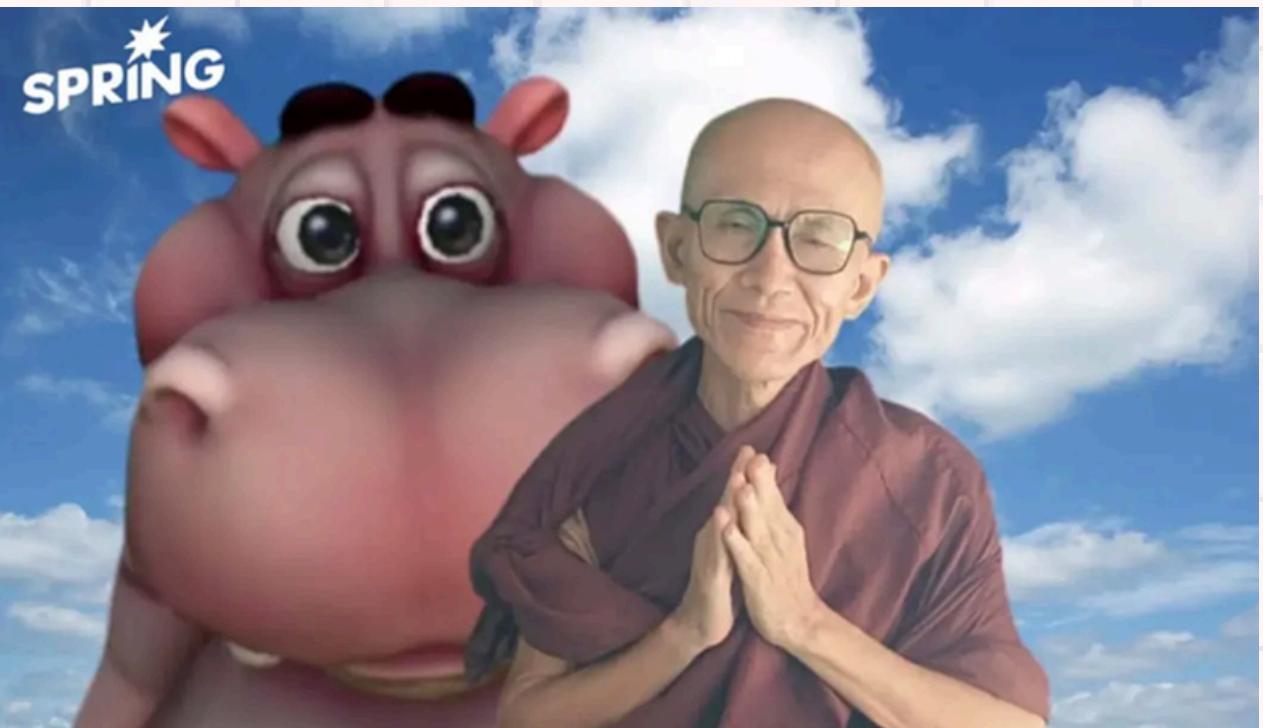
ន.ស. ផែននិត្តា

កំកង

6434466623

ន.ស. ពេរកង

ឡាកំ



Thank  
you!



ຂອບຂອບດຸຈະ

ກົກົມົງ

