Introduction to Data Science - 21KHDL1

FINAL PROJECT

Subject: Analyzing air pollution in Viet Nam

Instructors:

Nguyen Thi Thu Hang

Nguyen Bao Long

Le Ngoc Thanh

Nguyen Ngoc Thanh

Group ID: 2

Doan Anh Khoa - 21127076

Nguyen Phat Dat - 21127240

Nguyen Bao Tuan - 21127560

Nguyen Duc Tuan Dat - 21127590

Contents

Overview

Image source: VietnamPlus

Vietnam grapples with a pressing air pollution crisis, adversely affecting public health and the environment. The surge in industrialization, urban expansion, and increased vehicular traffic are primary factors behind the deteriorating air quality. Exploring diverse perspectives is crucial to comprehending Vietnam's current air quality challenges and devising effective solutions. This forms the focus of our group's final project.

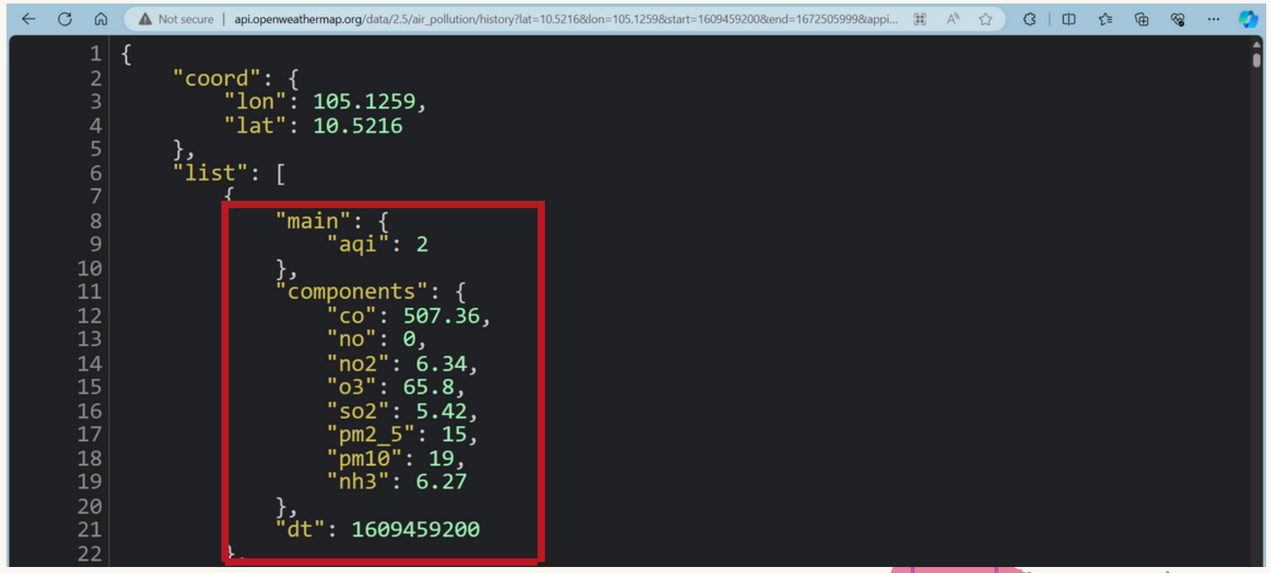
Data Collection

We aim to analyze the air population situation across different places in Viet Nam in 2021 - 2022.

First we will collect the air pollution data from the *OpenWeather API*:

- We will get coordinates of different places in Vietnam and store them to the following file: *VietNam_province_info.csv*.

- Then, we will send requests to get the data using the following url format: *https://api.openweathermap.org/data/2.5/air_pollution/history?lat={latitude}&lon={longitude}&start={starting time (unix)}&end={ending time (unix)}&appid={API_KEY}*



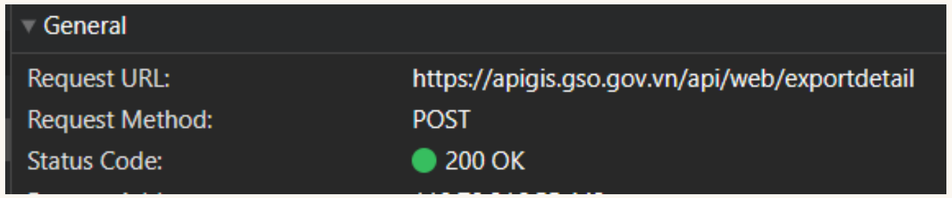
Data Collection

We aim to analyze the air population situation across different places in Viet Nam in 2021 - 2022.

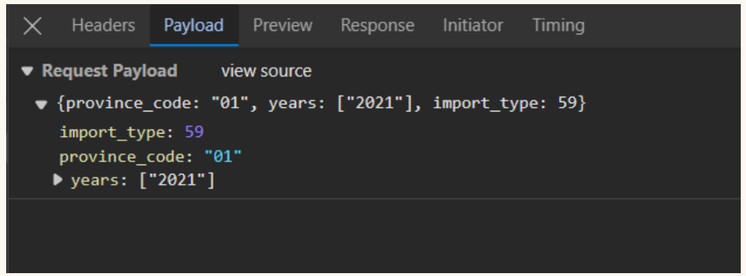
We will collect additional data from the *GIS on Population and Development* to get more information for our analysis, the new features include:

- *Total population in urban*
- *Total population in rural*
- *HDI*
- *Percentage of poor households*

1. We will post a request URL to crawl data



2. When posting, we will attach it a packet



3. Data will be responded

Data Preprocessing

The goal of this section is to format the dataset correctly for easy analysis of the problem.

- **Step 1:** Reading the raw data from ‘air_raw.csv’.

	location	dt	co	no	no2	o3	so2	pm2_5	pm10	nh3	aqi
0	An Giang	1609459200	507.36	0.00	6.34	65.80	5.42	15.00	19.00	6.27	2
1	An Giang	1609545600	400.54	0.02	5.66	60.08	4.83	14.48	18.26	6.08	2
2	An Giang	1609632000	500.68	0.01	7.97	42.20	3.82	19.29	23.05	8.04	2
3	An Giang	1609718400	654.22	0.06	12.51	24.68	4.35	23.56	29.31	9.63	3
4	An Giang	1609804800	714.30	0.04	14.22	18.95	4.23	22.98	25.93	4.88	3

- **Step 2:** Convert the 'dt' attribute from Unix time format to datetime and rename it to 'datetime'.

dt
1609459200
1609545600
1609632000
1609718400
1609804800
...
1672099200
1672185600
1672272000
1672358400
1672444800

date
2021-01-01
2021-01-02
2021-01-03
2021-01-04
2021-01-05
...
2022-12-27
2022-12-28
2022-12-29
2022-12-30
2022-12-31

Data Preprocessing

The goal of this section is to format the dataset correctly for easy analysis of the problem.

- **Step 3:** Once we have obtained a good dataset from the first source, we will collect data from GIS on Population and Development spanning two years.

First year:

	Total population in urban	Total population in rural	HDI	Percentage of poor households
Ha Noi	4095366.0	4235468.0	0.81	0.4
Ha Giang	140327.0	746759.0	0.59	25.0
Cao Bang	138178.0	404039.0	0.65	24.5
Bac Kan	73114.0	250598.0	0.68	20.6
Tuyen Quang	NaN	NaN	NaN	NaN
...	...	...	...	...

Second year:

	Total population in urban	Total population in rural	HDI	Percentage of poor households
Ha Noi	4138505.0	4297147.0	0.82	0.1
Ha Giang	142345.0	750378.0	0.60	31.6
Cao Bang	138465.0	404587.0	0.66	23.6
Bac Kan	73565.0	250788.0	0.69	20.1
Tuyen Quang	NaN	NaN	NaN	NaN
...	...	...	...	...

Data Preprocessing

The goal of this section is to format the dataset correctly for easy analysis of the problem.

- **Step 4:** After having processed data from the two sources, it’s time to merge the data from the OpenWeather API with the data that we've just scraped from the GIS website.

	location	date	co	no	no2	o3	so2	pm2_5	pm10	nh3	aqi	Total population in urban	Total population in rural	HDI	Percentage of poor households
0	An Giang	2021-01-01	507.36	0.00	6.34	65.80	5.42	15.00	19.00	6.27	2	646021.0	1263485.0	0.66	3.5
1	An Giang	2021-01-02	400.54	0.02	5.66	60.08	4.83	14.48	18.26	6.08	2	646021.0	1263485.0	0.66	3.5
2	An Giang	2021-01-03	500.68	0.01	7.97	42.20	3.82	19.29	23.05	8.04	2	646021.0	1263485.0	0.66	3.5
3	An Giang	2021-01-04	654.22	0.06	12.51	24.68	4.35	23.56	29.31	9.63	3	646021.0	1263485.0	0.66	3.5
4	An Giang	2021-01-05	714.30	0.04	14.22	18.95	4.23	22.98	25.93	4.88	3	646021.0	1263485.0	0.66	3.5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Data Preprocessing

The goal of this section is to format the dataset correctly for easy analysis of the problem.

- **Step 5:** We will verify the presence of missing values in the dataset and determine which locations have missing data and how many are missing.

```
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   location                      45864 non-null  object
 1   date                          45864 non-null  datetime64[ns]
 2   co                            45864 non-null  float64
 3   no                            45864 non-null  float64
 4   no2                           45864 non-null  float64
 5   o3                            45864 non-null  float64
 6   so2                           45864 non-null  float64
 7   pm2_5                         45864 non-null  float64
 8   pm10                          45864 non-null  float64
 9   nh3                           45864 non-null  float64
 10  aqi                           45864 non-null  int64
 11  Total population in urban     45136 non-null  float64
 12  Total population in rural     45136 non-null  float64
 13  HDI                           45136 non-null  float64
 14  Percentage of poor households 45136 non-null  float64
dtypes: datetime64[ns](1), float64(12), int64(1), object(1)
memory usage: 5.2+ MB
```

After checking which locations have missing values

```
array(['Tuyen Quang'], dtype=object)
```

Data Preprocessing

The goal of this section is to format the dataset correctly for easy analysis of the problem.

	location	Total population in urban	Total population in rural	HDI	Percentage of poor households
42952	Tuyen Quang	NaN	NaN	NaN	NaN
42953	Tuyen Quang	NaN	NaN	NaN	NaN
42954	Tuyen Quang	NaN	NaN	NaN	NaN
42955	Tuyen Quang	NaN	NaN	NaN	NaN
42956	Tuyen Quang	NaN	NaN	NaN	NaN
...	...	...	...	...	...
43675	Tuyen Quang	NaN	NaN	NaN	NaN
43676	Tuyen Quang	NaN	NaN	NaN	NaN
43677	Tuyen Quang	NaN	NaN	NaN	NaN
43678	Tuyen Quang	NaN	NaN	NaN	NaN
43679	Tuyen Quang	NaN	NaN	NaN	NaN

- **Step 6:** We have identified 'Tuyen Quang' as a location with missing values. Our solution is to fill the missing value attributes in 'Tuyen Quang' with their respective mean values.

	location	Total population in urban	Total population in rural	HDI	Percentage of poor households
42952	Tuyen Quang	594203.343141	989429.788617	0.708209	6.87578
42953	Tuyen Quang	594203.343141	989429.788617	0.708209	6.87578
42954	Tuyen Quang	594203.343141	989429.788617	0.708209	6.87578
42955	Tuyen Quang	594203.343141	989429.788617	0.708209	6.87578
42956	Tuyen Quang	594203.343141	989429.788617	0.708209	6.87578
...	...	...	...	...	...
43675	Tuyen Quang	594203.343141	989429.788617	0.708209	6.87578
43676	Tuyen Quang	594203.343141	989429.788617	0.708209	6.87578
43677	Tuyen Quang	594203.343141	989429.788617	0.708209	6.87578
43678	Tuyen Quang	594203.343141	989429.788617	0.708209	6.87578
43679	Tuyen Quang	594203.343141	989429.788617	0.708209	6.87578

Data Preprocessing

The goal of this section is to format the dataset correctly for easy analysis of the problem.

- **Step 7:** Finally, we will fill in the missing value for each day by using accumulated value: [**(next year - previous year) / (number of days in that year)**] and save to ‘air_pollution.csv’.

	location	date	co	no	no2	o3	so2	pm2_5	pm10	nh3	aqi	Total population in urban	Total population in rural	HDI	Percentage of poor households
0	An Giang	2021-01-01	507.36	0.00	6.34	65.80	5.42	15.00	19.00	6.27	2	646021.000000	1.263485e+06	0.660000	3.5
1	An Giang	2021-01-02	400.54	0.02	5.66	60.08	4.83	14.48	18.26	6.08	2	646019.283356	1.263481e+06	0.660000	3.5
2	An Giang	2021-01-03	500.68	0.01	7.97	42.20	3.82	19.29	23.05	8.04	2	646017.566713	1.263477e+06	0.660000	3.5
3	An Giang	2021-01-04	654.22	0.06	12.51	24.68	4.35	23.56	29.31	9.63	3	646015.850069	1.263474e+06	0.660000	3.5
4	An Giang	2021-01-05	714.30	0.04	14.22	18.95	4.23	22.98	25.93	4.88	3	646014.133425	1.263470e+06	0.660000	3.5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45859	Yen Bai	2022-12-27	714.30	0.00	13.19	55.07	12.52	138.08	156.96	0.94	5	176638.436039	6.705814e+05	0.659945	16.6
45860	Yen Bai	2022-12-28	607.49	0.00	3.60	72.24	11.80	124.45	126.85	0.66	5	176641.077029	6.705850e+05	0.659959	16.6
45861	Yen Bai	2022-12-29	487.33	0.00	3.60	41.13	2.44	26.11	27.11	1.44	3	176643.718019	6.705887e+05	0.659972	16.6
45862	Yen Bai	2022-12-30	407.22	0.00	1.97	48.64	1.64	18.41	19.14	1.41	2	176646.359010	6.705923e+05	0.659986	16.6
45863	Yen Bai	2022-12-31	567.44	1.06	5.18	1.65	1.12	36.95	43.53	8.11	3	176649.000000	6.705960e+05	0.660000	16.6

Data Exploration

- The number of rows and columns in our data are respectively (45864, 15)
- The meaning of each row is that it indicates Air pollution level information and other pieces of information about economy, population and development in a certain province/city on a certain date.

location	date	co	no	no2	o3	so2	pm2_5	pm10	nh3	aqi	Total population in urban	Total population in rural	HDI	Percentage of poor households
Tra Vinh	2022-07-09	620.84	0.27	9.42	13.77	2.56	22.56	26.30	5.64	3	183313.184319	8.357963e+05	0.687620	7.0

=> Every row indicates information about a distinct area on a specific date. Therefore, our data do not have duplicate rows as expected.

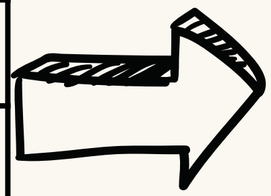
- The meaning of each column is that it indicates:

Column	Description
`location`	Name of province/city.
`date`	Specific time in the format **year-month-day**, providing information for each province/city.
`co`, `no`, `no2`, `o3`, `so2`, `pm2_5`, `pm10`, `nh3`	Concentrations of polluting gases: Carbon monoxide (CO), Nitrogen monoxide (NO), Nitrogen dioxide (NO_2), Ozone (O_3), Sulphur dioxide (SO_2), Ammonia (NH_3), and particulates ($PM_{2.5}$ and PM_{10}). Units: $\mu g/m^3$.
`aqi`	Air Quality Index, representing air pollution severity for the general public, ranging from 1 to 5 (Good, Fair, Moderate, Poor, Very Poor).
`Total population in urban`	Total urban population in each province/city.
`Total population in rural`	Total rural population in each province/city.
`HDI`	Human Development Index, a comparative, quantitative index of Income, Education, and Life Expectancy in provinces/cities in Vietnam.
`Percentage of poor households`	Proportion or percentage of poor households within a given population in a specific area.

Data Exploration

- The data type of each column:

Column	Null Count	Datatype
location	0	object
date	0	datetime
co, no, no2, o3, so2, pm2_5, pm10, nh3	0	float
aqi	0	int
Total population in urban	0	float
Total population in rural	0	float
HDI	0	float
Percentage of poor households	0	float



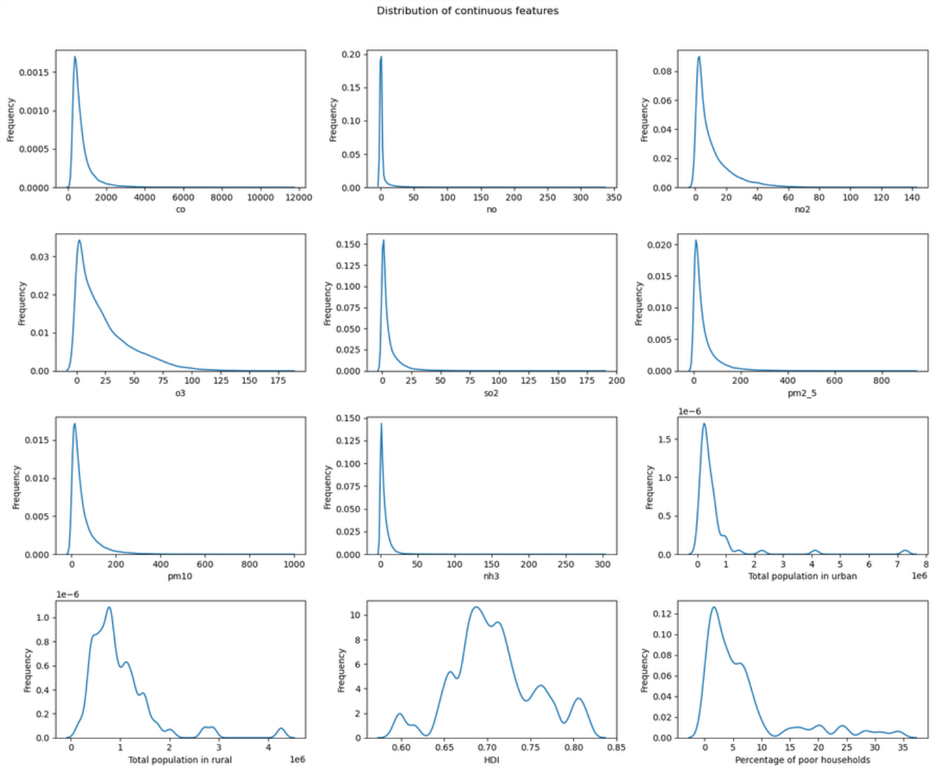
Total population in urban and **Total population in rural** seem to have an inappropriate data type
=> We convert them to integer data type

Data Exploration

- The distribution of the data in each column:

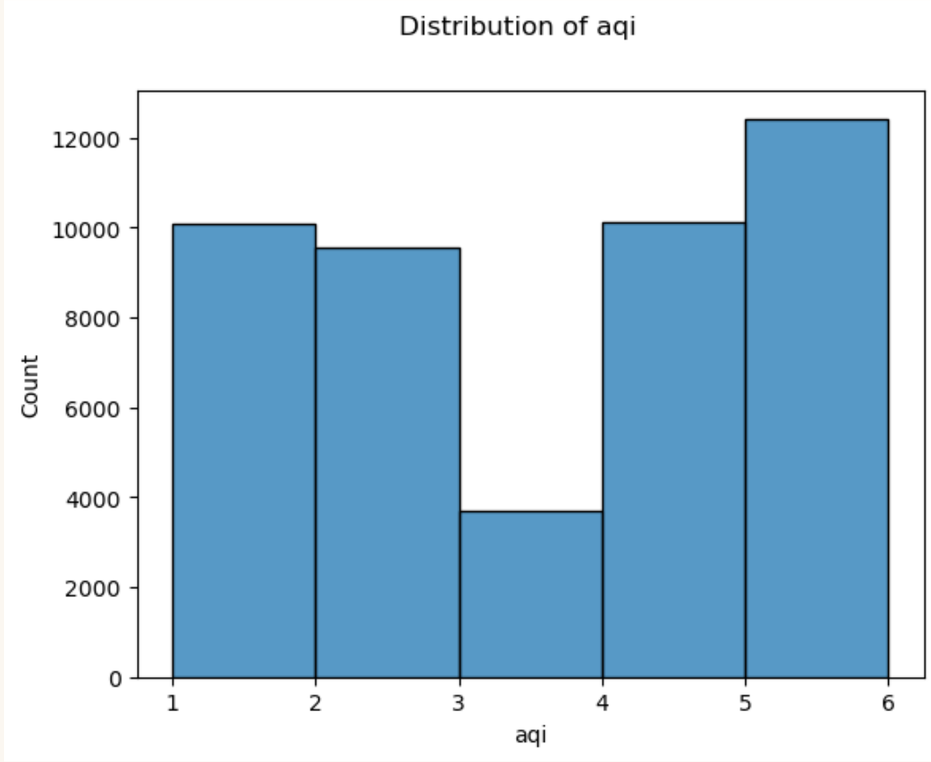
Continuous numerical column(s)

Using Kernel Distribution Estimation Plot (Kdeplot) to depict the probability density function of the continuous or non-parametric data variables



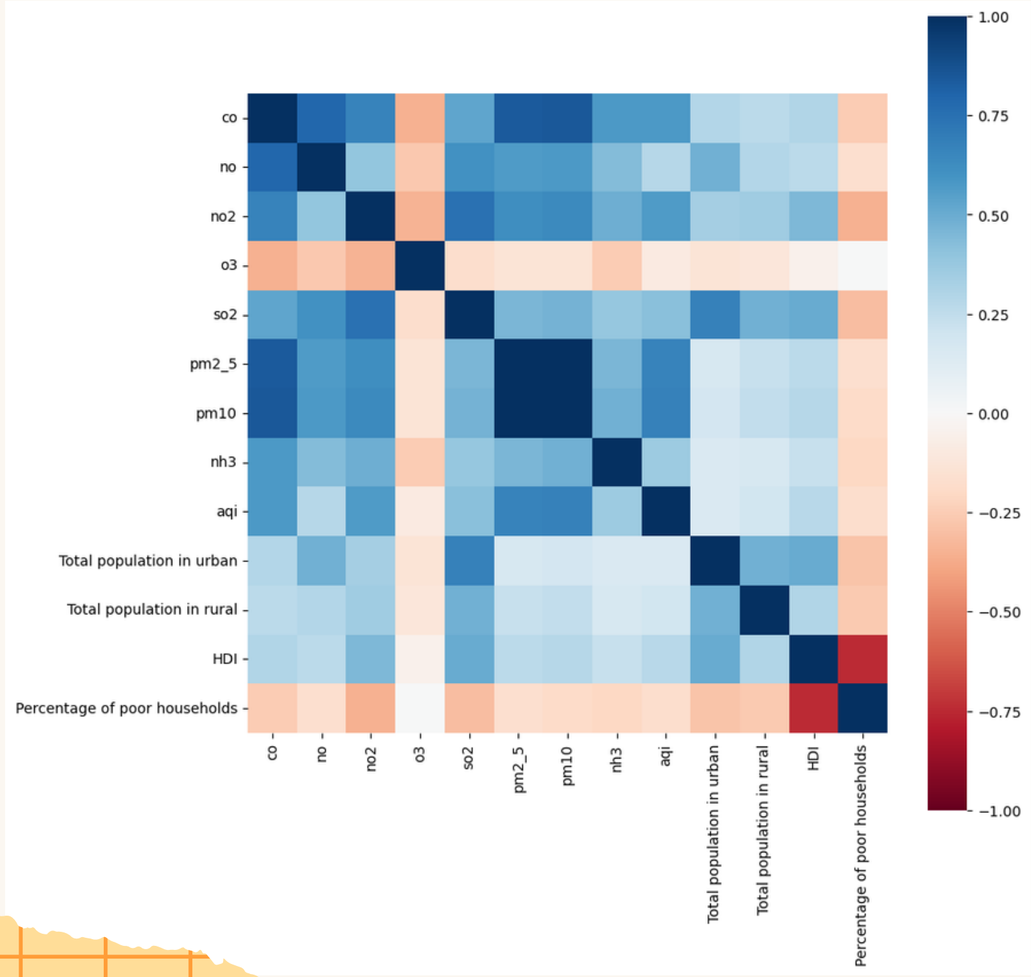
Categorical column(s)

Using Histogram plot to show frequency distributions.



Data Exploration

- The correlation among features:
  - Calculating the correlation coefficient among features.
  - Using Heat map to show relationships of features.



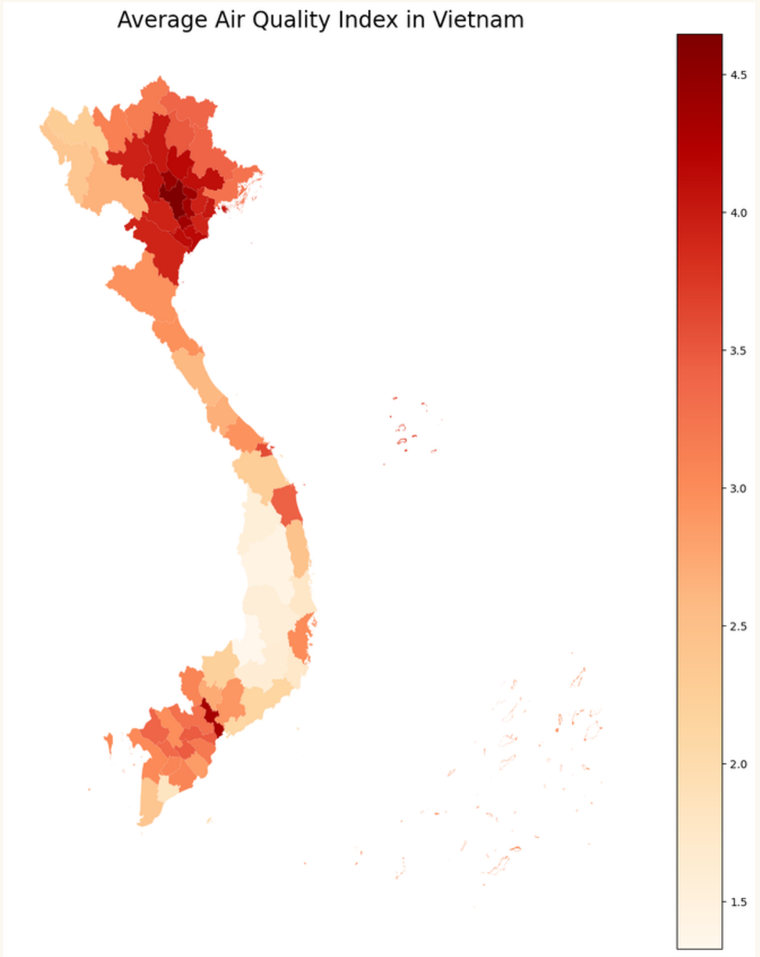
Data Exploration

Question 1: What is the situation about the air quality from all the provinces and cities in Viet Nam?

- How we answer this question is to visualize the average air quality of different locations in Vietnam using a choropleth map.

- This question will give us some insights into the areas having bad or good quality.

=> It helps us concentrate on some specific areas to resolve air pollution.

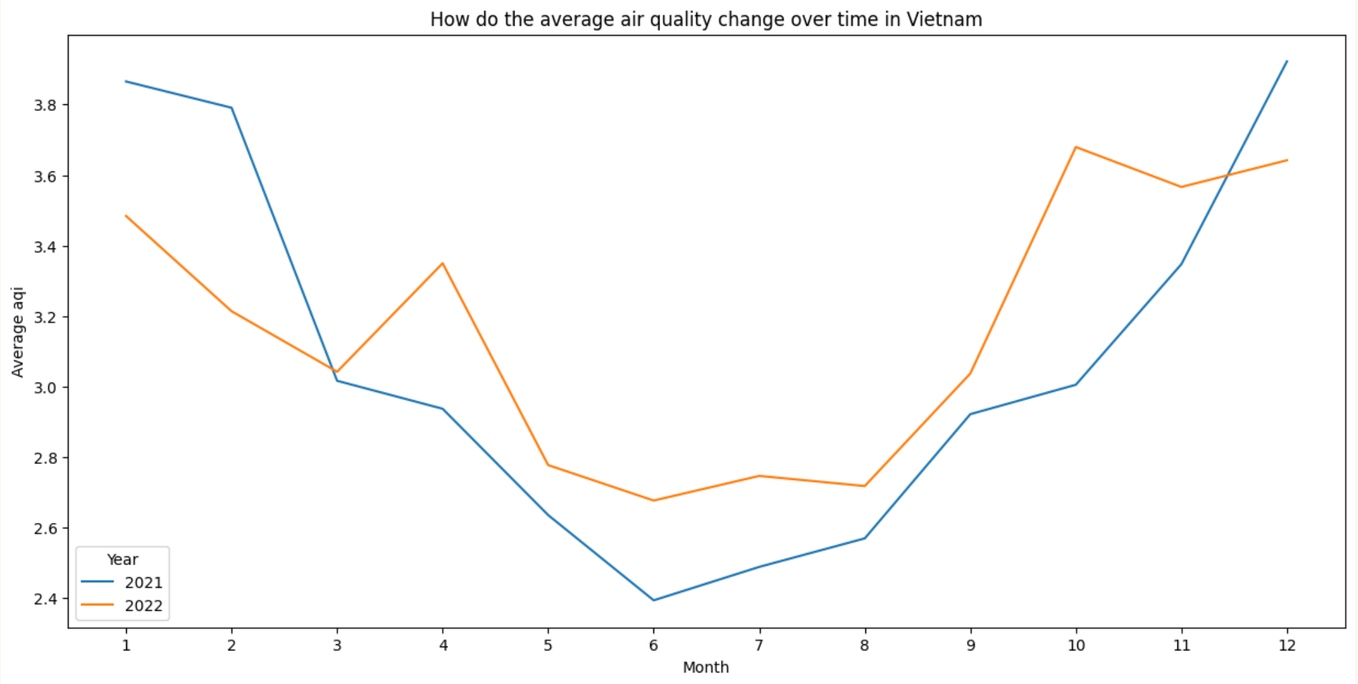


Data Exploration

Question 2: At what time of the year does air quality become poor or good in Vietnam?

- How we answer this question is to visualize the average air quality by month using a line graph showing the changes over the period.

- Knowing when the air gets better or worse in Vietnam helps us plan things better. For example:
  - Applying timely environmental policies.
  - Promoting public awareness of environmental protection.
  - Adopting multiple measures for environmental preservation are crucial steps.

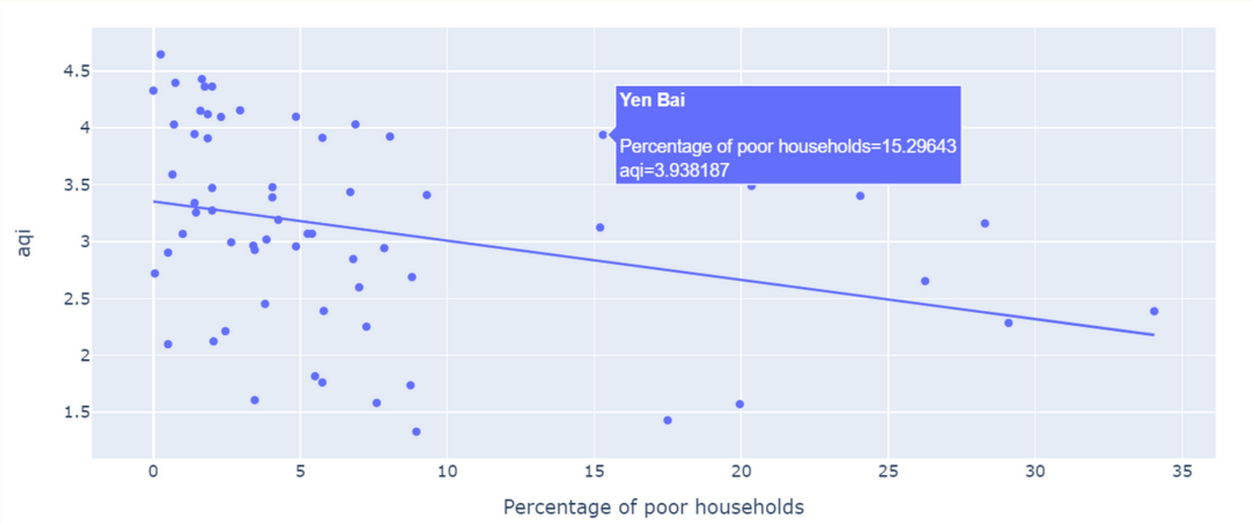


Data Exploration

Question 3: Is there a connection between the percentage of poor households and pollution levels?

- How we answer this question: We can analyze the correlation between the percentage of poor households and air pollution indicators **aqi** using a scatter plot.

- This question gives us:
  - Understanding this connection is crucial for identifying environmental justice issues
  - Formulating targeted policies to address the impact of pollution on vulnerable populations.

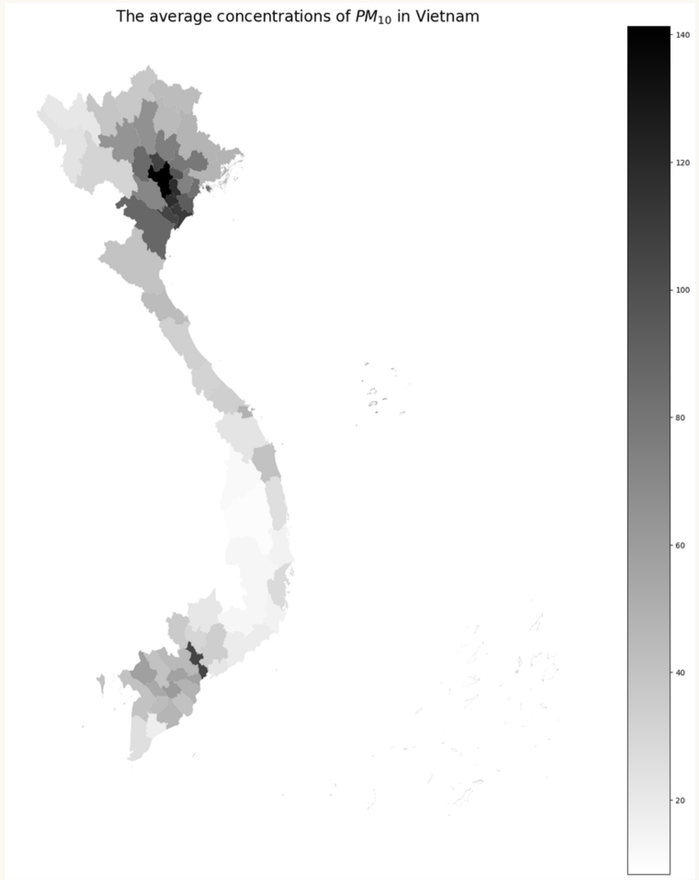


Data Exploration

Question 4: How is the average concentration of PM10 distributed in provinces/cities in Vietnam?

- How we answer this question is to visualize the average concentration in different provinces/cities in Vietnam using a choropleth map.

- This question helps us:
  - Knowing the difference in the average concentration of pm in different areas, thereby tells us what the cause is for high or low concentrations, or specifically geographical differences.
  - From there, we will promote and propagate to residents to pay attention to protecting their health and protecting the environment.

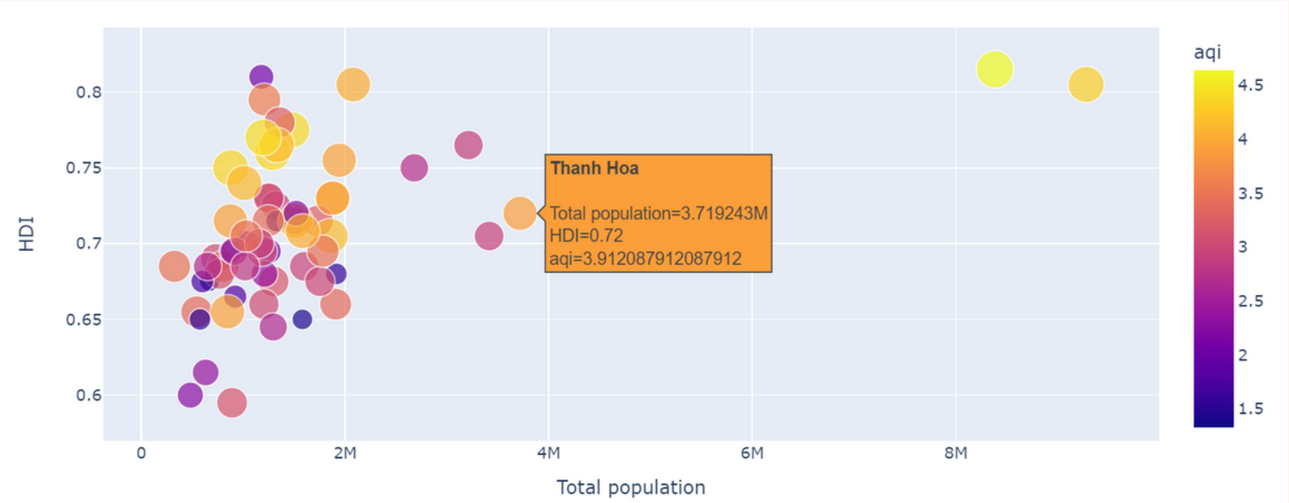


Data Exploration

Question 5: How does the air quality vary across different human development features like HDI and total population?

- How we answer this question: is to visualize a bubble graph about the average total population and GRDP in 2 axes and the bubble indicates the average air quality of 64 provinces

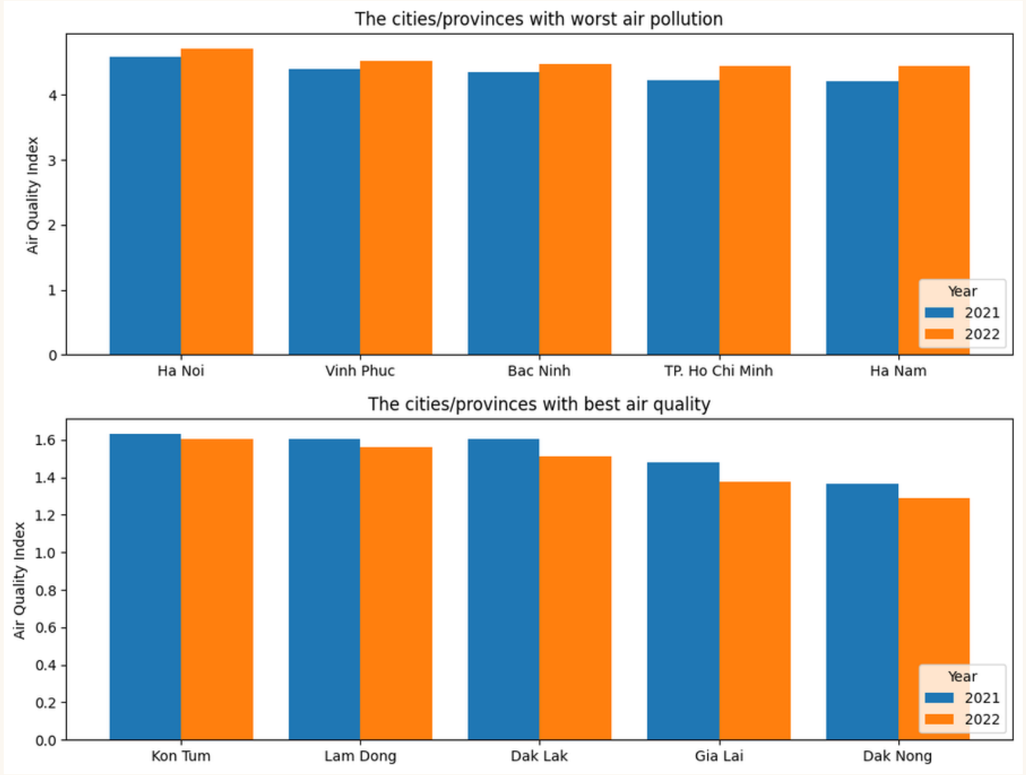
- This question gives us some information about the relationship between some population and development features across the provinces/cities in Vietnam and how those affect the air quality.



Data Exploration

Question 6: What are the changes in the air quality for some provinces or cities having good/bad air quality?

- How we answer this question: is to use the bar graphs to visualize how the average air quality of some provinces/cities having good/bad quality change for 2 consecutive years.
- This question gives us some insights into the efficiency of government policies in terms of protecting the environment.

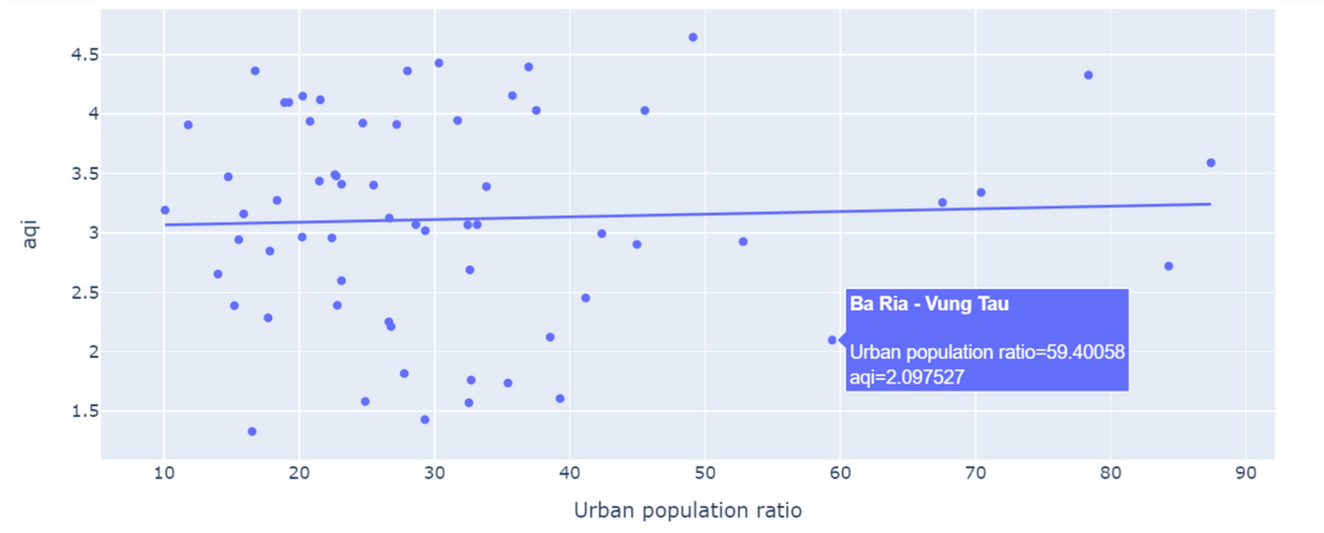


Data Exploration

Question 7: Do rural areas in Vietnam have better air quality?

- How we answer this question: is to explore whether areas with a higher ratio of urban population experience better or worse air quality.

- This question will
  - Benefits public health
  - Inform government policies and urban planning
  - Contributes to environmental conservation,
  - Promotes rural tourism, etc.



Data Modeling

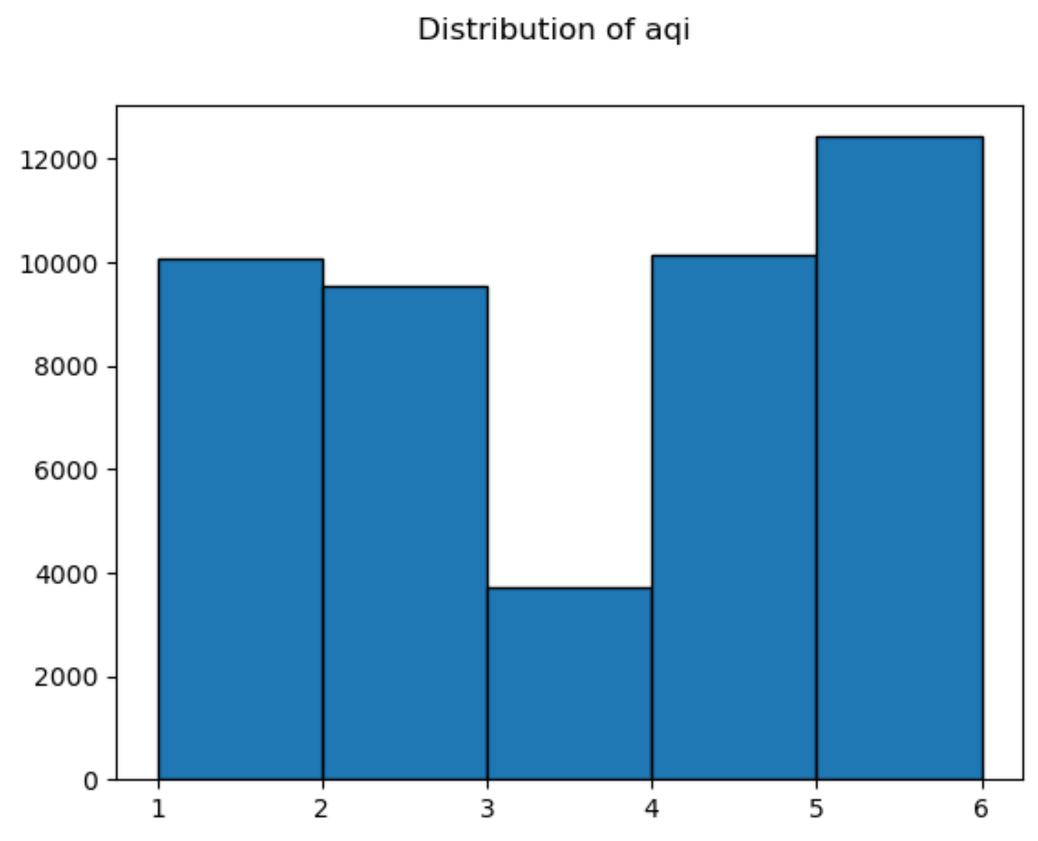
Problem Statements:

- We will build some models that can classify the Air Quality Index of Vietnam (1-5) using the given features.
- Solving this problem offers diverse benefits, informing health interventions, policies, and fostering global collaboration.
- Moreover, accurate forecasting serves as a crucial tool for addressing health, environmental, and societal challenges tied to air pollution, promoting a comprehensive and sustainable approach.

Data Modeling

Step 1: Data Preparation

1.1: Investigating the distribution of the label values to determine whether there is a bias in `aqi`.

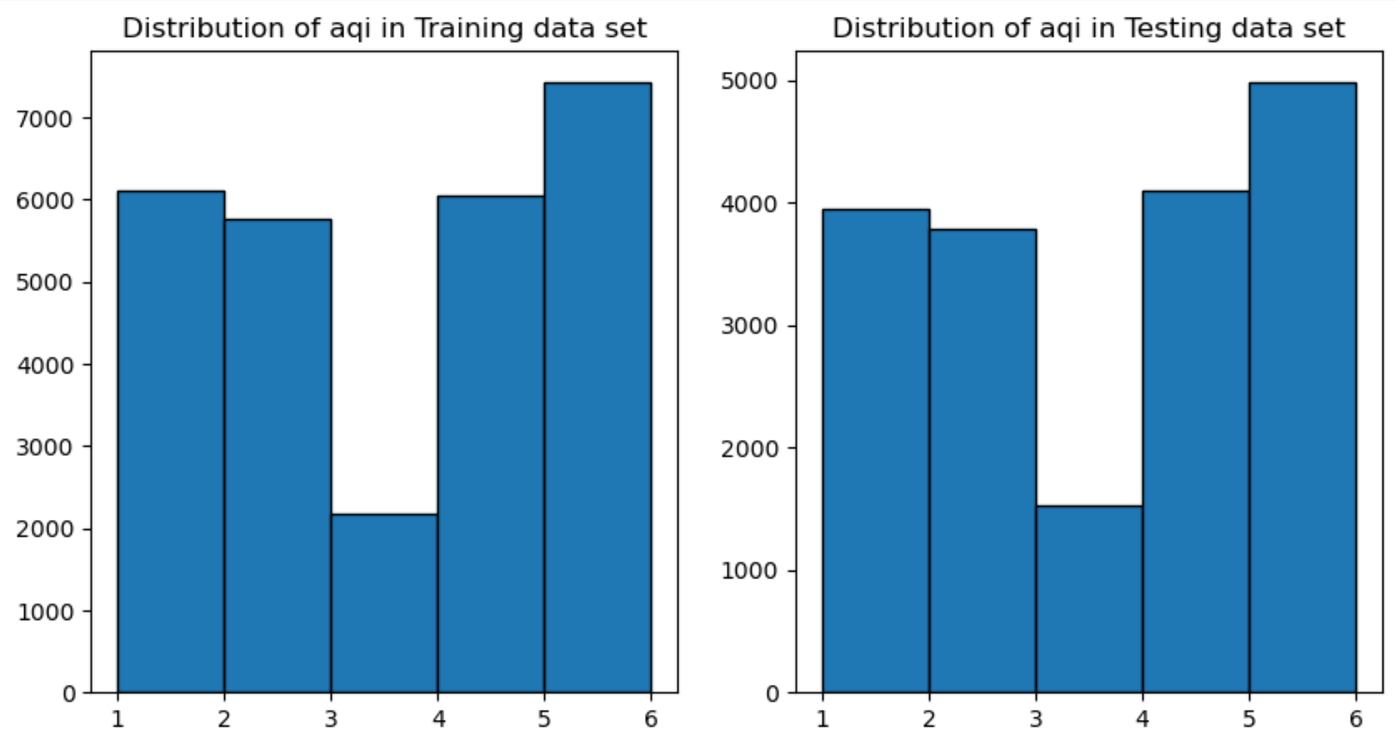


- In general, all the values in the label seems to be evenly distributed except the value 3.
- Moreover, most places throughout the country are characterized by poorer air quality.

Data Modeling

Step 1: Data Preparation

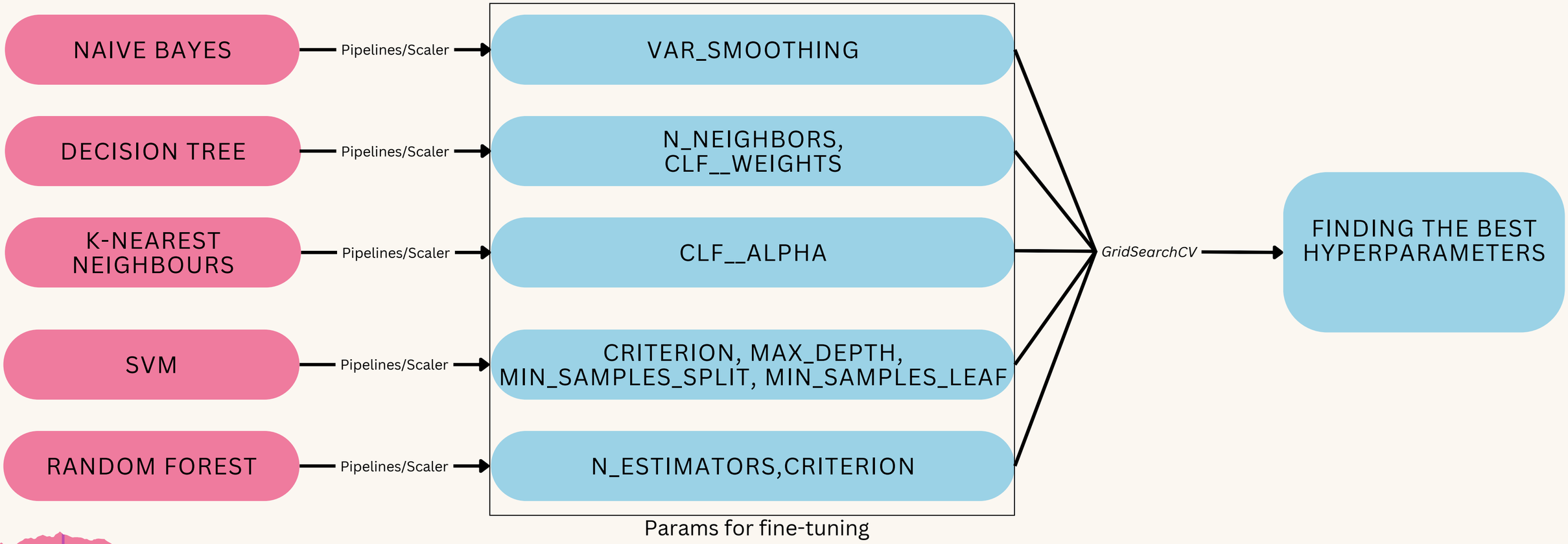
1.2: Dividing randomly the training and testing dataset into 6/4 ratio.



Looks like the distributions of both training and testing test are remain intact after being splitted, this may help the models generalize more effectively and reduce the bias when testing them.

Data Modeling

Step 2: Create, train & Test models



Data Modeling

Step 3: Evaluating models

3.1: Precision, recall, f1-score, support of each model after 5 folds.

	precision	recall	f1-score	support
1	0.56	0.89	0.69	3949
2	0.39	0.41	0.40	3790
3	0.00	0.00	0.00	1532
4	0.48	0.57	0.52	4090
5	0.93	0.61	0.73	4985
accuracy			0.57	18346
macro avg	0.47	0.50	0.47	18346
weighted avg	0.56	0.57	0.55	18346

NAIVE BAYES

	precision	recall	f1-score	support
1	0.89	0.89	0.89	3949
2	0.72	0.82	0.77	3790
3	0.52	0.32	0.39	1532
4	0.81	0.87	0.84	4090
5	0.96	0.93	0.95	4985
accuracy			0.83	18346
macro avg	0.78	0.76	0.77	18346
weighted avg	0.83	0.83	0.83	18346

K-NEAREST NEIGHBOURS

	precision	recall	f1-score	support
1	0.90	0.99	0.95	3949
2	0.45	0.69	0.54	3790
3	0.09	0.01	0.02	1532
4	0.58	0.39	0.47	4090
5	0.96	1.00	0.98	4985
accuracy			0.71	18346
macro avg	0.60	0.62	0.59	18346
weighted avg	0.68	0.71	0.69	18346

SVM

	precision	recall	f1-score	support
1	1.00	1.00	1.00	3949
2	1.00	0.99	1.00	3790
3	0.98	0.96	0.97	1532
4	0.99	0.99	0.99	4090
5	0.99	1.00	1.00	4985
accuracy			0.99	18346
macro avg	0.99	0.99	0.99	18346
weighted avg	0.99	0.99	0.99	18346

DECISION TREE

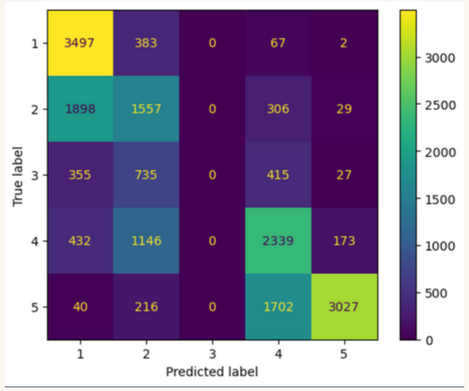
	precision	recall	f1-score	support
1	1.00	1.00	1.00	3949
2	1.00	0.99	0.99	3790
3	0.99	0.96	0.97	1532
4	0.99	0.99	0.99	4090
5	0.99	1.00	1.00	4985
accuracy			0.99	18346
macro avg	0.99	0.99	0.99	18346
weighted avg	0.99	0.99	0.99	18346

RANDOM FOREST

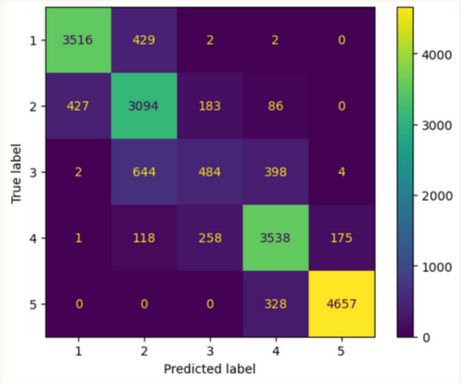
Data Modeling

Step 3: Evaluating models

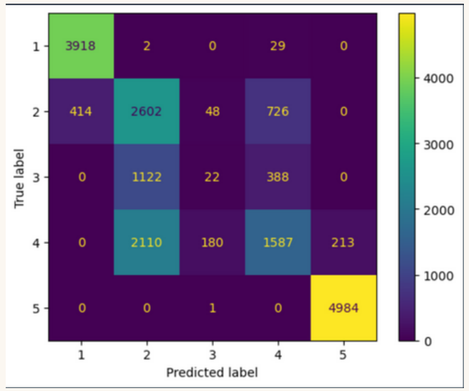
3.2: Confusion matrix of each model after 5 folds.



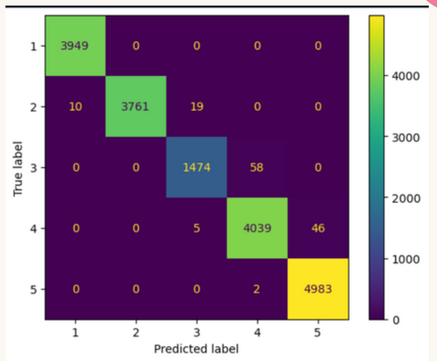
NAIVE BAYES



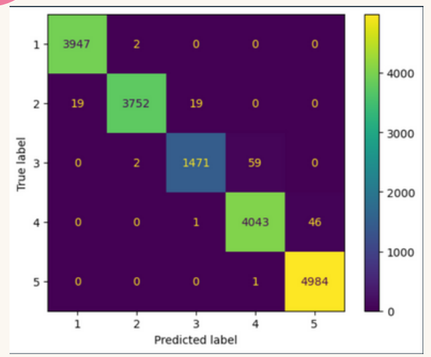
K-NEAREST NEIGHBOURS



SVM



DECISION TREE

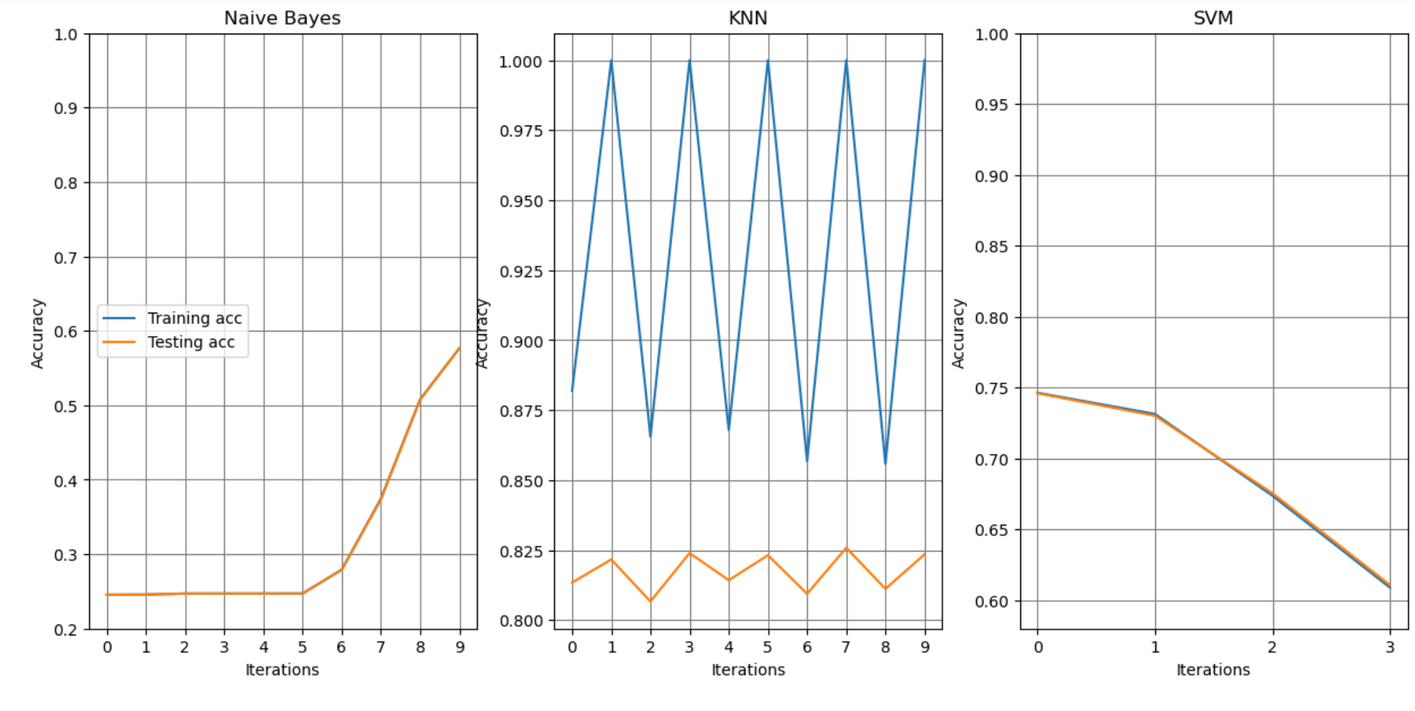


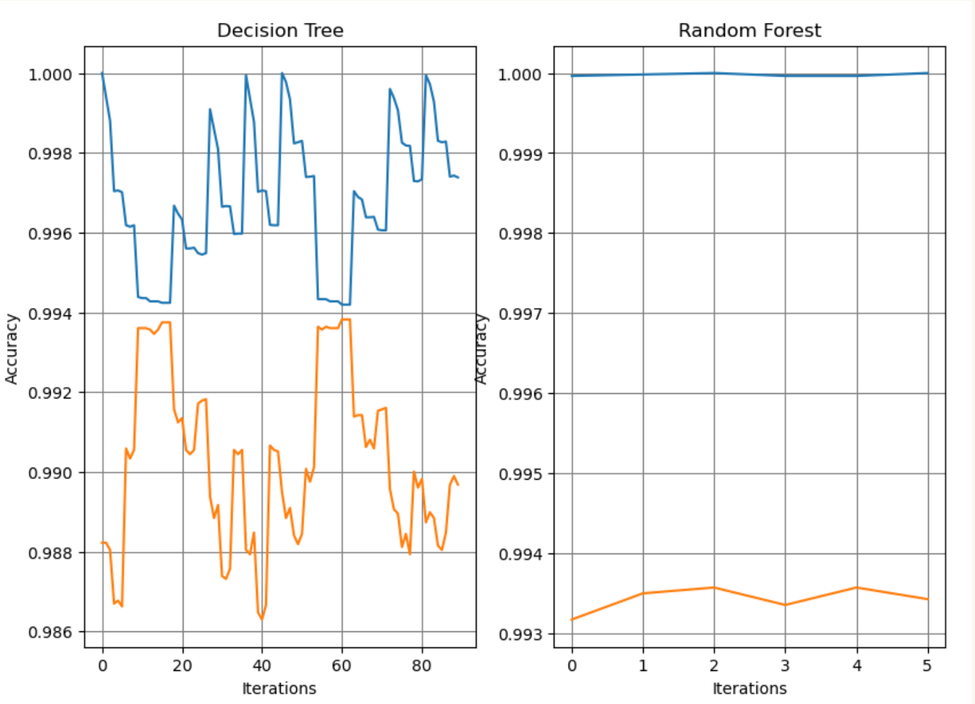
RANDOM FOREST

Data Modeling

Step 3: Evaluating models

3.3: Cross-validating visualization for all models.





THANK YOU!