Predicting Air Quality

STAT 451 Final Project

Lacey Dinh, Tan Bui, Minji Suh,
Maddie Young, and Diego Ugaz

Motivation

Why do we care about “fresh air”?

- Air pollution can lead to illness such as lung disease, asthma, etc.
- Poor air quality results in 100,000 premature deaths and costs $150 billion each year (National Weather Service)

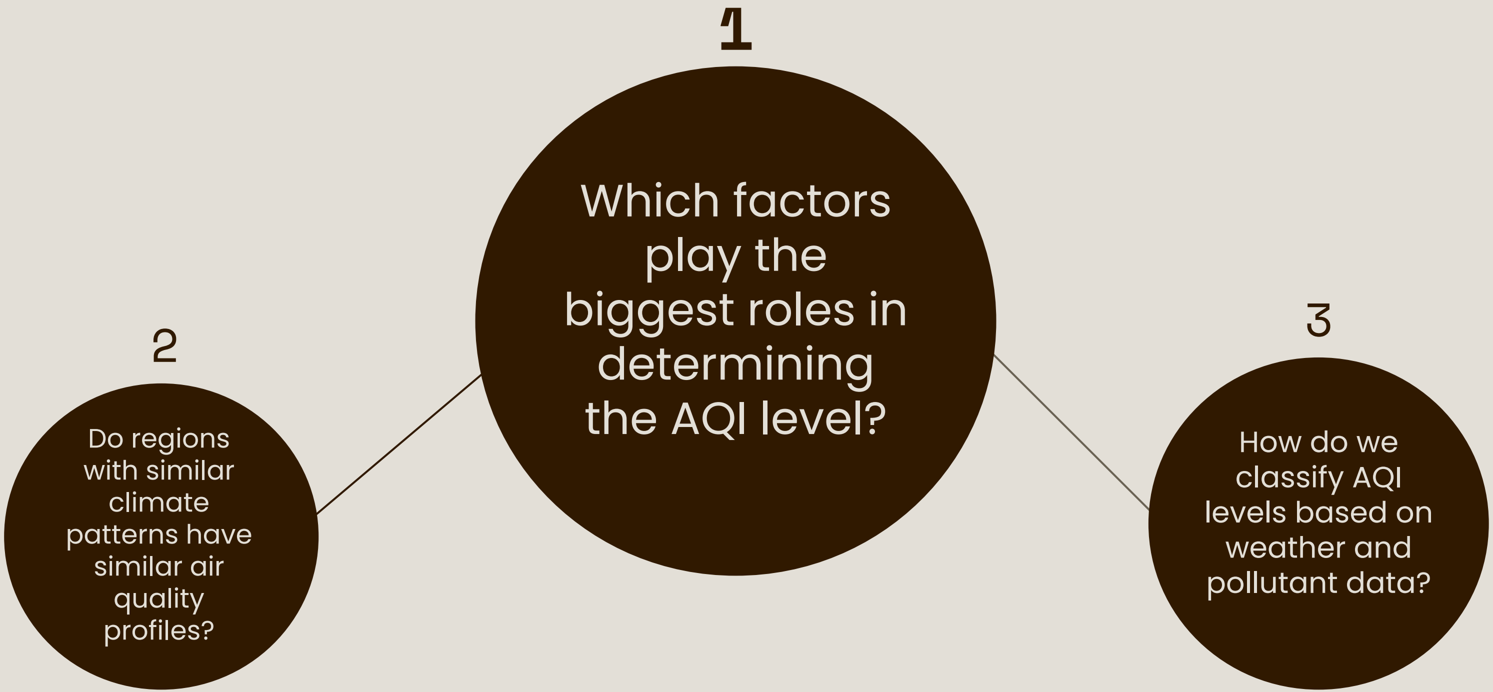
Goal: investigate how meteorological and environmental factors influence air quality

Dave Sanders (NYT)

Data Exploration

- Collected by N. Elgiriyewithana (Kaggle)
- Contains 40+ and over 37,000+ observations
- Key categories for features:
  - **Time and Location:** latitude, longitude, time of observation
  - **Weather:** temperature, precipitation, UV index
  - **Pollutant:** concentration of Nitrogen Oxide, Ozone
  - **Astronomy:** sunrise, moon phase

Research Questions



1

Which factors play the biggest roles in determining the AQI level?

2

Do regions with similar climate patterns have similar air quality profiles?

3

How do we classify AQI levels based on weather and pollutant data?

Preprocessing

- A subset of weather and pollutant features are used
  - Temperature, wind speed, humidity, cloud, UV index
  - Concentration of air pollutants
- No missing data!
- Variables are standardized to minimize the impact of outliers

Q1 Which factors play the biggest roles in determining the AQI level?

4 regression models:

- Linear Regression
- Decision Tree
- Random Forest
- Gradient Boosting

→ *Evaluated based on MSE and R^2*

Q1 Results

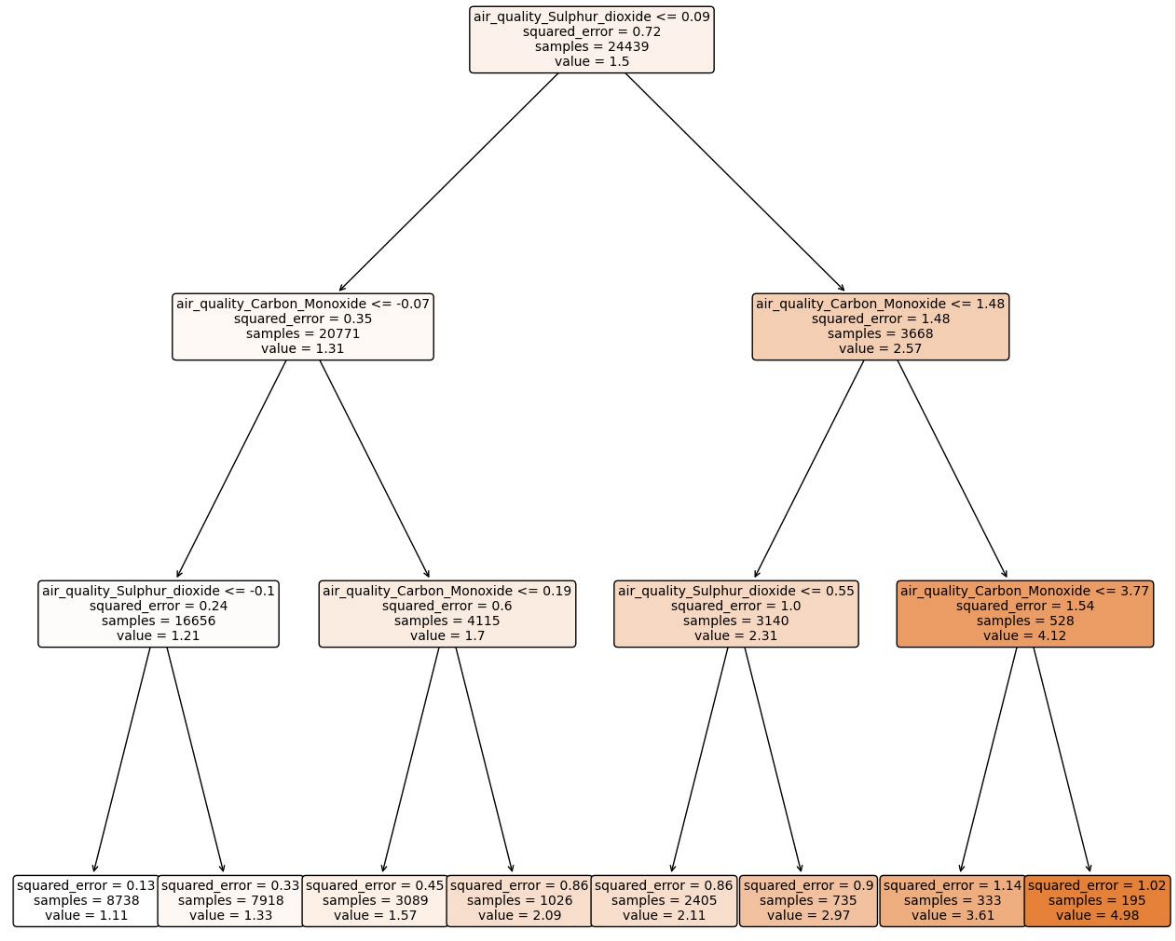
Scores on Validation Set

Model	***MSE***	***R^2***	***Optimal Hyperparameter***
Linear Regression	0.429	0.467	—
Decision Tree	0.405	0.498	—
Random Forest	0.201	0.750	200 decision trees with a max depth of 20
Gradient Boosting	0.284	0.648	200 decision trees with an alpha of 0.2

Q1 Results

Decision Tree:

Nodes near the root use concentrations of CO and SO_2 to branch.



air_quality_Sulphur_dioxide <= 0.09
squared_error = 0.72
samples = 24439
value = 1.5

air_quality_Carbon_Monoxide <= -0.07
squared_error = 0.35
samples = 20771
value = 1.31

air_quality_Carbon_Monoxide <= 1.48
squared_error = 1.48
samples = 3668
value = 2.57

air_quality_Sulphur_dioxide <= -0.1
squared_error = 0.24
samples = 16656
value = 1.21

air_quality_Carbon_Monoxide <= 0.19
squared_error = 0.6
samples = 4115
value = 1.7

air_quality_Sulphur_dioxide <= 0.55
squared_error = 1.0
samples = 3140
value = 2.31

air_quality_Carbon_Monoxide <= 3.77
squared_error = 1.54
samples = 528
value = 4.12

squared_error = 0.13
samples = 8738
value = 1.11

squared_error = 0.33
samples = 7918
value = 1.33

squared_error = 0.45
samples = 3089
value = 1.57

squared_error = 0.86
samples = 1026
value = 2.09

squared_error = 0.86
samples = 2405
value = 2.11

squared_error = 0.9
samples = 735
value = 2.97

squared_error = 1.14
samples = 333
value = 3.61

squared_error = 1.02
samples = 195
value = 4.98

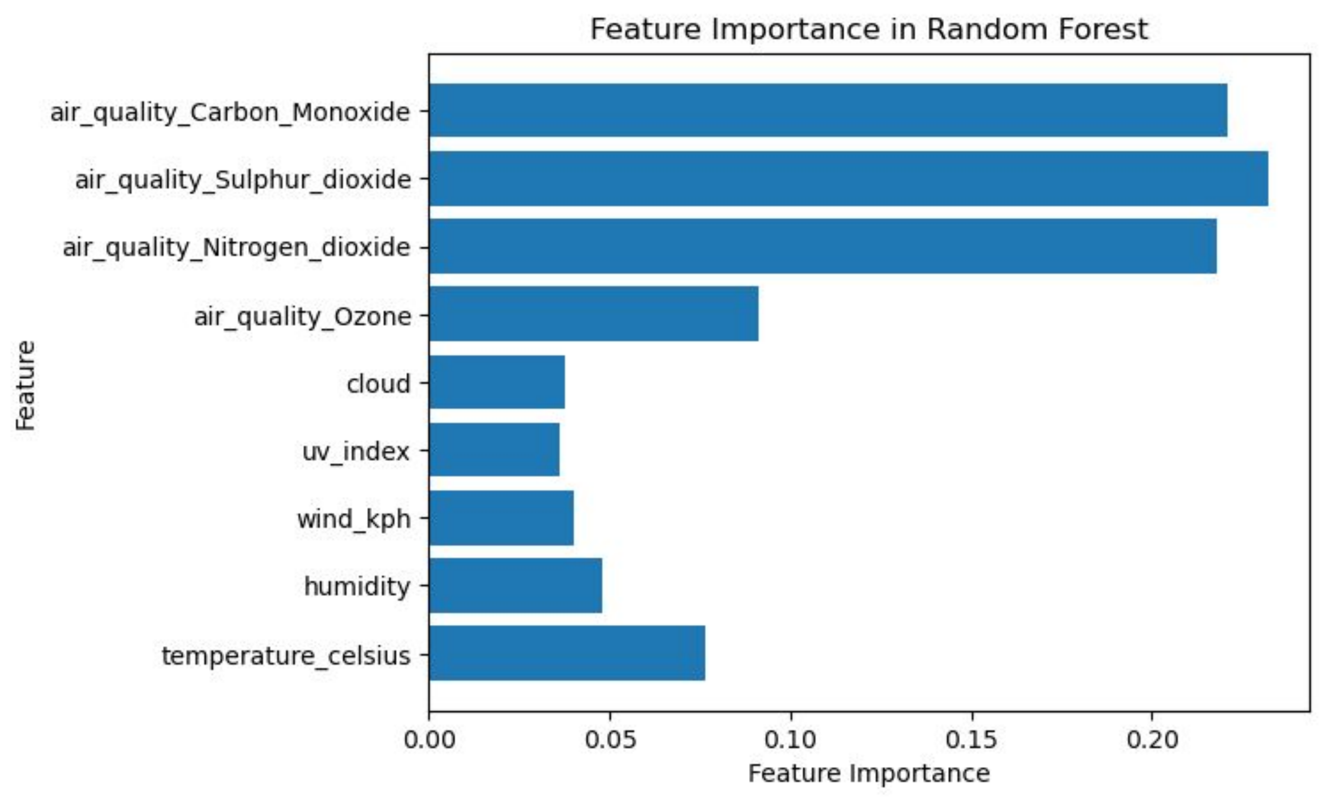
Q1 Results

Random Forest:

Most important features for AQI are CO, SO_2, NO_2

Scores on Test set:

- MSE: 0.201
- R^2: 0.759



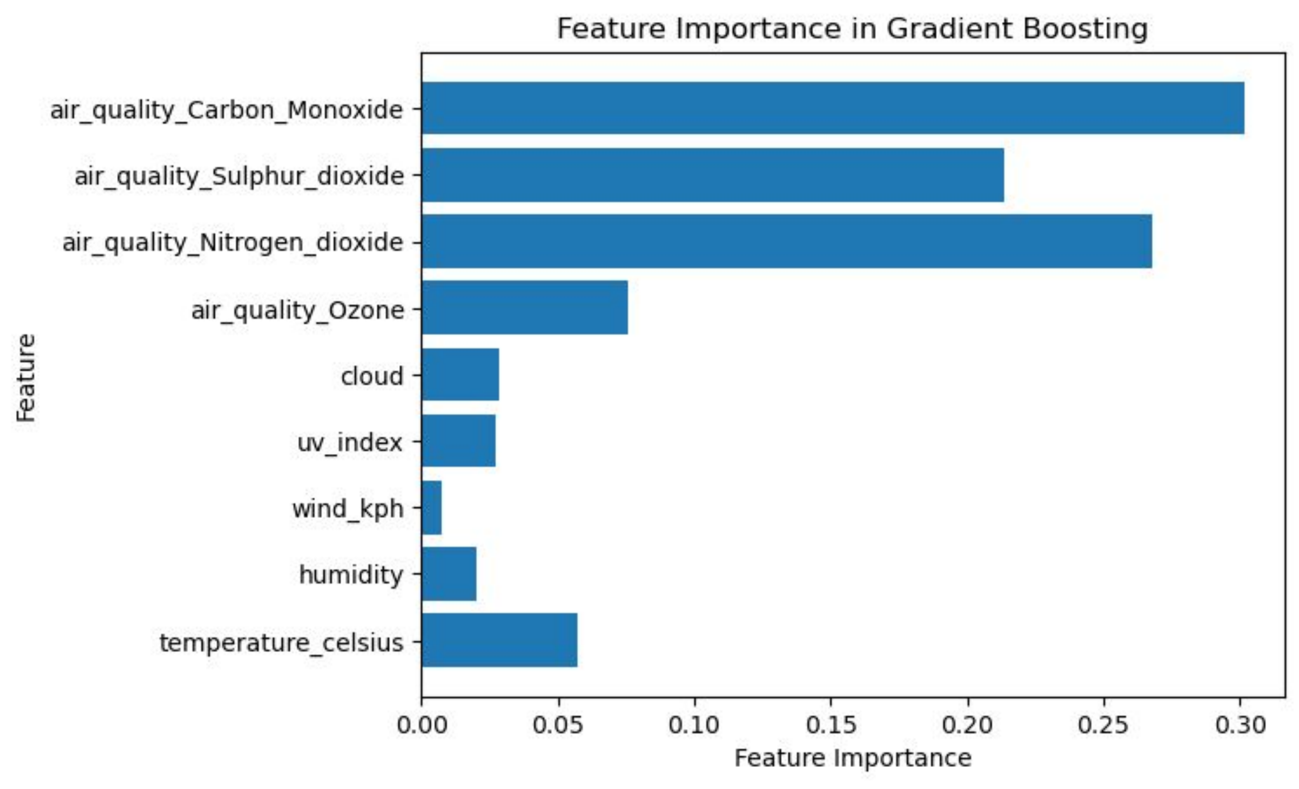
Q1 Results

Gradient Boosting:

Most important features for AQI are CO, SO_2, NO_2

Scores on Test set:

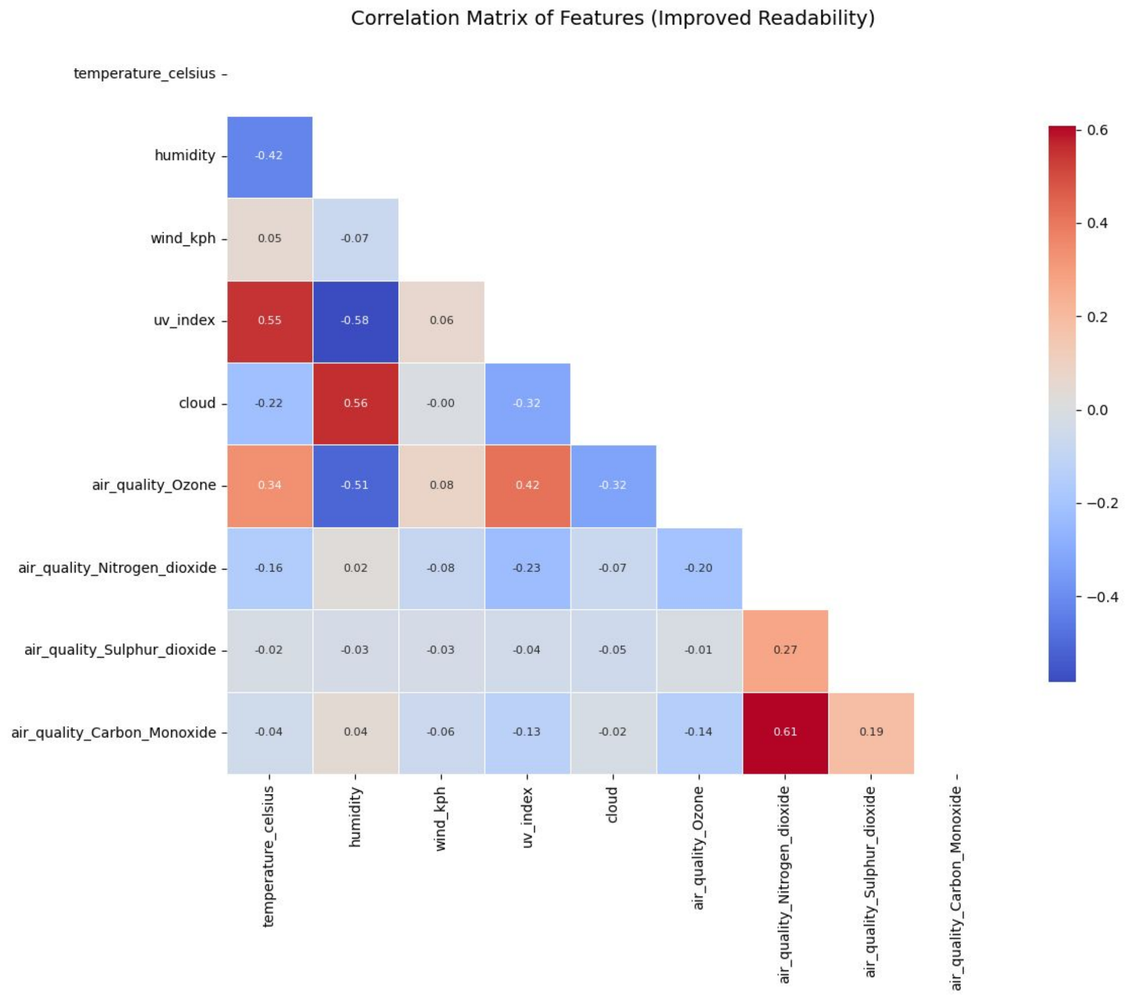
- MSE: 0.244
- R^2: 0.708



INTRO
DATA EXPLORATION
QUESTIONS
METHODS
RESULTS
DISCUSSION

Q1 Possible Complications

- Noticed similar importance of Ozone and Temperature in both ensemble models
- Ozone seems to be the most impactful pollutant on meteorological data
- Given these correlations, the meteorological features could be interpreted as repeat inclusions of Ozone as a feature



Correlation Matrix of Features (Improved Readability)

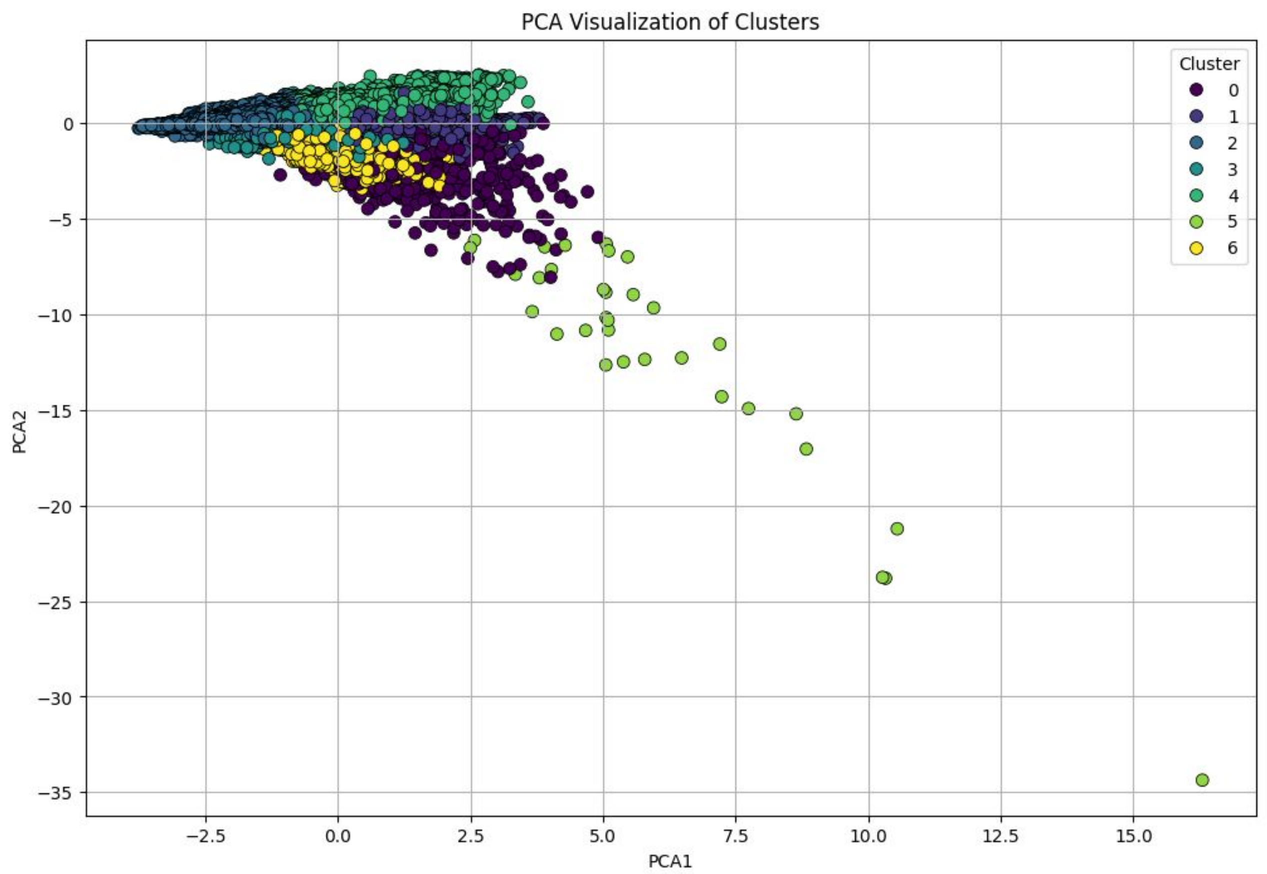
	temperature_celsius	humidity	wind_kph	uv_index	cloud	air_quality_Ozone	air_quality_Nitrogen_dioxide	air_quality_Sulphur_dioxide	air_quality_Carbon_Monoxide
temperature_celsius									
humidity	-0.42								
wind_kph	0.05	-0.07							
uv_index	0.55	-0.58	0.06						
cloud	-0.22	0.56	-0.00	-0.32					
air_quality_Ozone	0.34	-0.51	0.08	0.42	-0.32				
air_quality_Nitrogen_dioxide	-0.16	0.02	-0.08	-0.23	-0.07	-0.20			
air_quality_Sulphur_dioxide	-0.02	-0.03	-0.03	-0.04	-0.05	-0.01	0.27		
air_quality_Carbon_Monoxide	-0.04	0.04	-0.06	-0.13	-0.02	-0.14	0.61	0.19	

Q2 Do regions with similar climate patterns have similar air quality profiles?

Method:
K-means clustering

Features:

- Location & Weather variables
- Pollutants



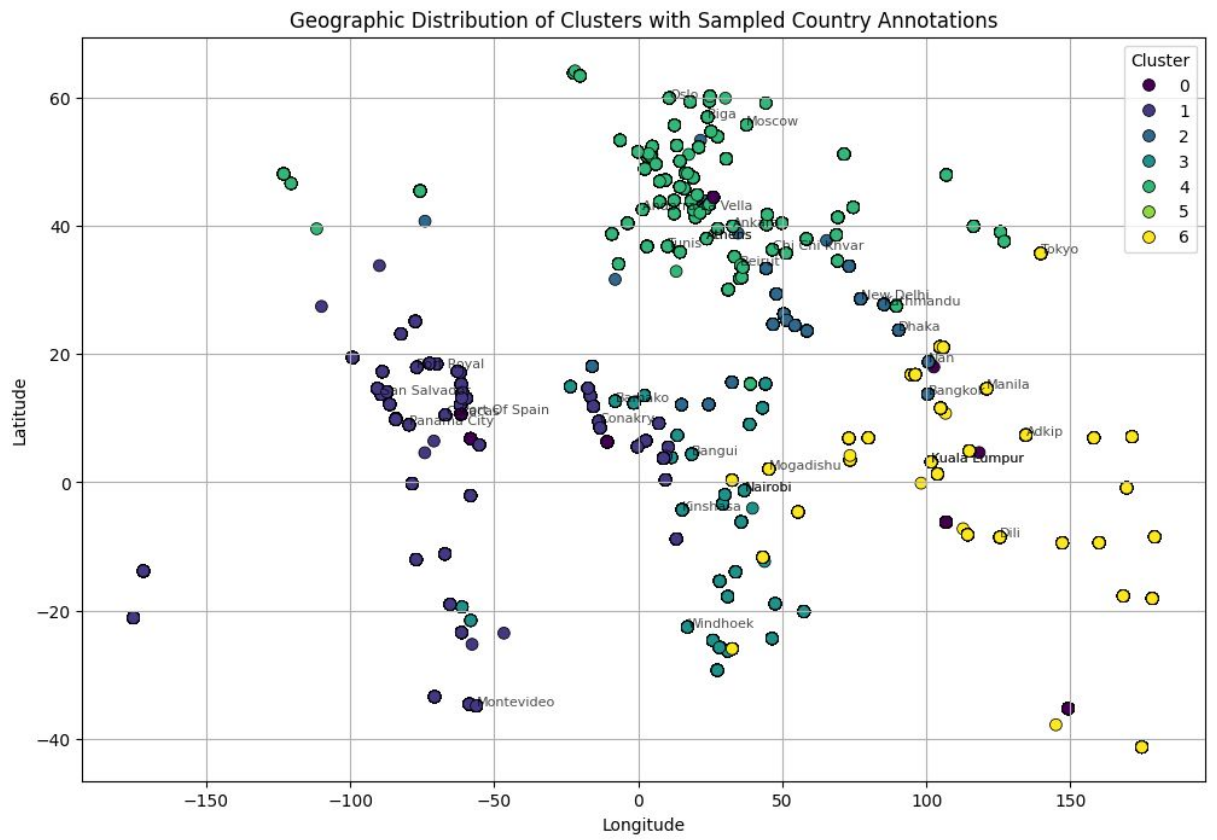
INTRO
DATA EXPLORATION
QUESTIONS
METHODS
RESULTS
DISCUSSION

Q2 Do regions with similar climate patterns have similar air quality profiles?

Cluster 2: Arid regions.

Cluster 4: Temperate regions.

Cluster 6: Tropical regions.



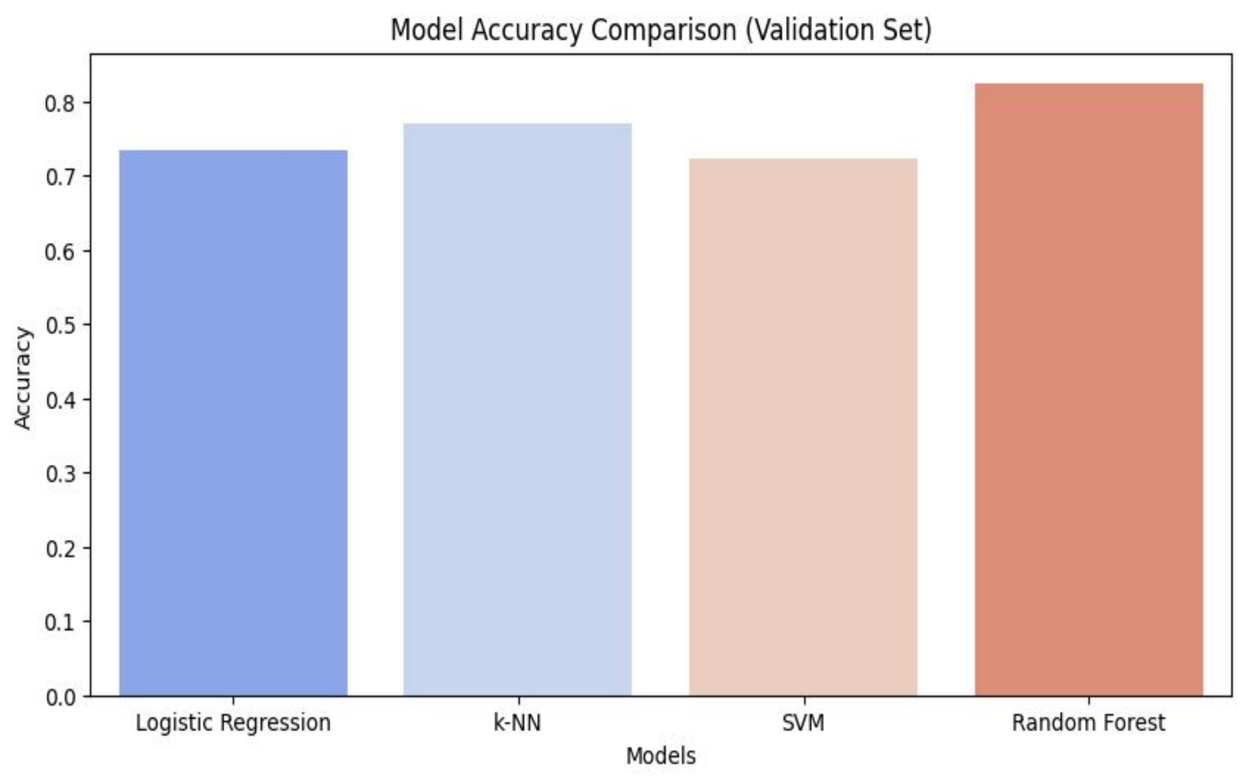
Q3 How do we classify AQI levels based on weather and pollutant data?

Method:

- Random Forest
- k-NN
- SVM
- Logistic Regression

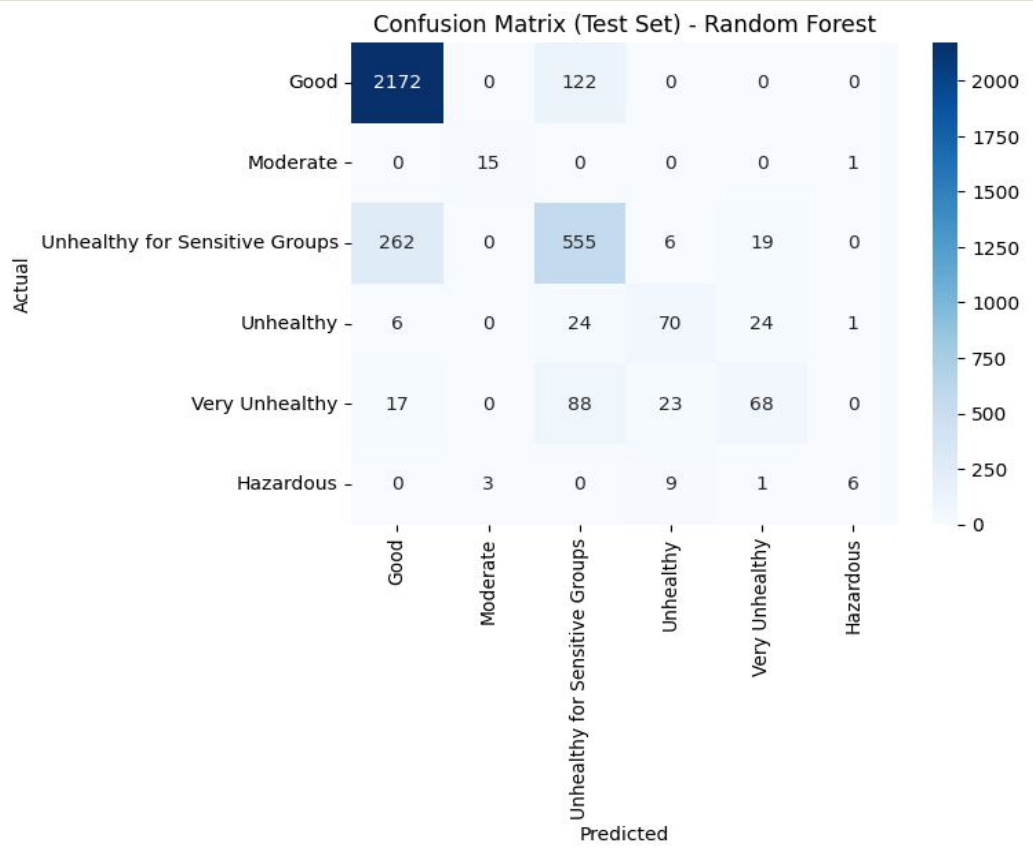
Features:

- Weather variables
- Pollutants



Q3 How do we classify AQI levels based on weather and pollutant data?

Confusion Matrix:



Confusion Matrix (Test Set) - Random Forest

Actual \ Predicted	Good	Moderate	Unhealthy for Sensitive Groups	Unhealthy	Very Unhealthy	Hazardous
Good	2172	0	122	0	0	0
Moderate	0	15	0	0	0	1
Unhealthy for Sensitive Groups	262	0	555	6	19	0
Unhealthy	6	0	24	70	24	1
Very Unhealthy	17	0	88	23	68	0
Hazardous	0	3	0	9	1	6

AQI Documentation:

https://document.airnow.gov/technical-assistance-document-for-the-reporting-of-daily-air-quailty.pdf

INTRO
DATA EXPLORATION
QUESTIONS
METHODS
RESULTS
Conclusions

Conclusions and Limitations

Q1:

Most important features: carbon monoxide, sulfur dioxide, nitrogen dioxide.

Q2:

- Climate shapes air quality profiles across regions.
- Lack of temporal data limits insights into seasonal variations.

Q3:

- Skewed distribution toward "Good" AQI, reducing performance for rare categories.
- Model can aid in early warnings for air quality, improving public health response.

Thank You!

Let us know if you have any suggestions or questions