New York City Air Quality Analysis

Zeb Moffat

Yida Fang

New York City Air Quality Surveillance Data

- https://catalog.data.gov/dataset/air-quality
- 15,499 records of air quality surveillance data
- Each record contains location information, time (year, month), pollutant type and amount
- Key air-pollution indicators such as particulate matter (PM2.5) , nitrogen oxides (NOx), sulfur dioxide (SO2), ozone (O3), and other boiler emissions

These indicators provide a view into air conditions across NYC. Air pollution is a major environmental risk. Exposures vary greatly by neighborhood due to differences in traffic density, industrial activities, and housing characteristics. Understanding where and when pollutant levels rise can inform policies aimed at reducing health and environmental issues associated with poor air quality.

Association

Apriori

FP-Growth

Apriori Association Rule Analysis

Analysis purpose:

Search for the combination patterns of HIGH values of pollutants and asthma to determine if there are potential health risk signals.

Analysis method:

- Classify NO_2, PM2.5, O_3 and Asthma as HIGH or LOW.
- Apriori with minsupport=0.05 and minconfidence=0.5
- Sort and select the most meaningful rules by lift.

Apriori Results

Result explanation:

Support (6.38%): There are 6.38% of the records in the dataset that simultaneously have NO2_HIGH and ASTHMA_HIGH.

Confidence (52.17%): When NO_2 is at a high level, there is a 52.17% probability of experiencing a high asthma emergency event.

Lift (1.043): Slightly greater than 1, indicating that an increase in NO_2 slightly raises the risk of asthma, but the correlation is not strong.

Although the correlation is not strong, it provides a possible public health indicator.

```
=== Association Rules (Sorted by Lift) ===
  antecedents    consequents  support  confidence      lift
2  (NO2_HIGH)  (ASTHMA_HIGH)  0.06383    0.521739  1.043478
```

The only rule that meets the conditions in the data is: NO2_HIGH → ASTHMA_HIGH

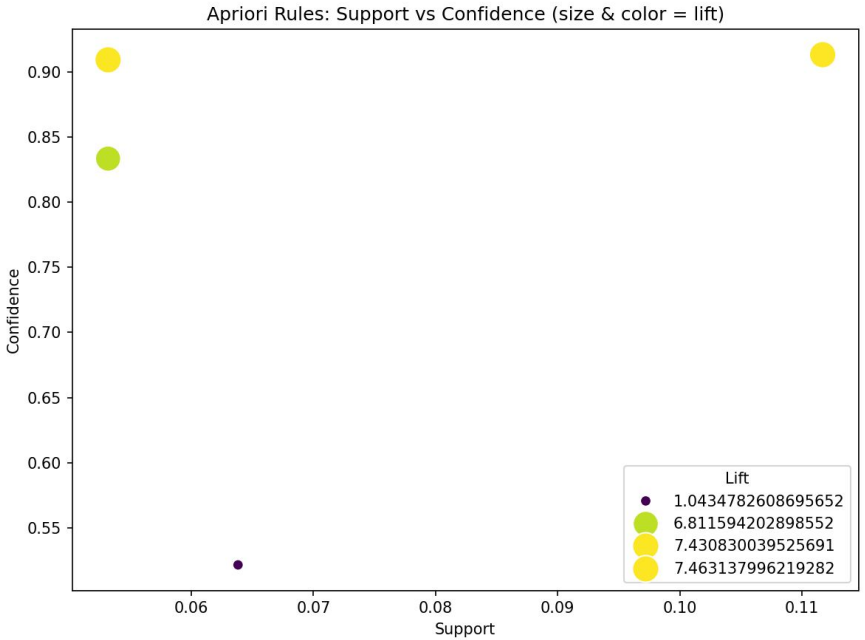
PM2.5: Although the data volume is similar to that of NO_2, the distribution of PM2.5 is relatively low. The number of times HIGH was determined was very few → support < 0.05 → Automatically filtered by Apriori.

O_3: O_3 shows almost no variation throughout the entire dataset. Unable to generate O3_HIGH → No HIGH/LOW differentiation

Apriori scatter

Scatter Plot Interpretation

- Each point represents one candidate association rule generated by the Apriori algorithm.
- X-axis (Support): How frequently the rule appears in the dataset.
- Y-axis (Confidence): How reliable the rule is when the antecedent occurs.
- Point size & color: Represent the lift, showing the strength of the rule (higher = stronger).



Key Insight

Although the scatter plot shows multiple potential rules, only one rule eventually meets both the support (≥0.05) and confidence (≥0.5) thresholds:

NO2_HIGH → ASTHMA_HIGH

- Other rules have:
- Very low support (rare events)
- Or low confidence (weak reliability)
- Or high lift but insufficient support

Therefore, NO_2 is the only pollutant that forms a statistically valid rule.

FP-Growth

What specific combinations of Location, Season, and Pollutant Type most frequently occur together with 'High Pollution' events?

Analyzed only high pollution (Top 20% z-score) events to discover which factors consistently appear together:

Features Analyzed:

- **Geographic Type** (e.g., Borough, Community District)
- **Geographic Location** (e.g., Bronx, Manhattan)
- **Pollutant Type** (e.g., PM 2.5, Ozone, NO2)
- **Month** (temporal patterns)

Key Technical Decisions:

- Min Support = 5% – Pattern must appear in at least 5% of high pollution events to be considered significant
- Lift ≥ 1.0 – Only keep associations where items occur together more often than random chance
- Used TransactionEncoder to convert each pollution event into a basket of attributes
- Sorted results by confidence to identify the most reliable patterns

FP-Growth Results 1

Rank	Antecedents	Consequents	Support	Confidence	Lift
1	Ozone (O3), CD	Month_6 (June)	5.5%	100%	3.04
2	Ozone (O3)	Month_6 (June)	13.6%	100%	3.04
3	UHF34, Nitrogen dioxide (NO2)	Month_12 (December)	8.3%	90.4%	1.48
4	Nitrogen dioxide (NO2)	Month_12 (December)	36.2%	88.3%	1.44
5	Nitrogen dioxide (NO2), CD	Month_12 (December)	16%	88.2%	1.44
6	UHF42, Nitrogen dioxide (NO2)	Month_12 (December)	11%	86.9%	1.42
7	CD (Community District)	Month_12 (December)	26.6%	64%	1.04
8	UHF34	Month_12 (December)	13.9%	63.1%	1.03
9	UHF42, Fine particles (PM 2.5)	Month_12 (December)	7.8%	62.3%	1.02
10	Month_12 (December), CD	Nitrogen dioxide (NO2)	16%	60%	1.46

FP-Growth Results 2

Top Discovery: Ozone + Community Districts -> June (100% Confidence) When high ozone pollution occurs in community districts, it happens in June 100% of the time (Lift = 3.04)

Key Patterns Identified:

1. Seasonal Nitrogen Dioxide (NO2) Pattern
    - NO2 pollution strongly associated with December (Month 12)
    - 88.3% confidence across all locations
    - Appears in 36% of high pollution events (highest support)
2. Geographic-Pollutant Combinations:
    - UHF34 + NO2 -> December (90.4% confidence)
    - UHF42 + NO2 -> December (86.9% confidence)
    - UHF42 + PM 2.5 -> December (62.3% confidence)
3. Community District Pattern:
    - Community districts see high pollution in December 64% of the time

Key Insight: High pollution events show seasonal patterns. Ozone peaks in summer (June) while Nitrogen Dioxide and particulate matter peak in winter (December). Some areas (UHF regions 34 and 42) are more susceptible to winter NO2 pollution.

Actionable Finding: Public health alerts should target June for ozone warnings and December for NO2/PM 2.5 warnings in high-risk neighborhoods.

Classification

Decision Tree

Random Forest

Decision tree classification model

Analysis objective:

Evaluate whether the status of pollutants (HIGH/LOW) can be used to predict whether asthma is in a HIGH state.

Model Characteristics:

- Use the HIGH/LOW Boolean variables as input
- Predicting ASTHMA_HIGH
- The decision tree has a simple structure and strong interpretability.

Decision Tree Classification Results

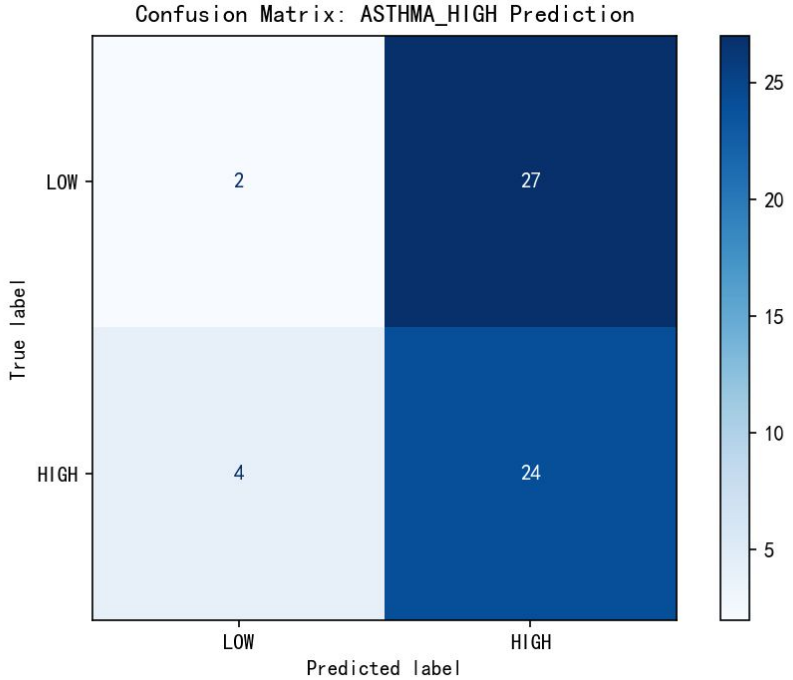
=== Decision Tree Classification Results ===

Accuracy: 0.456

Confusion Matrix:

[[2 27]

[4 24]]



Decision Tree Classification Results

Result explanation:

Accuracy: 0.456

- The model correctly predicts asthma levels only 45.6% of the time.
- This indicates that pollutant levels alone are not sufficient to produce a strong predictive model for asthma emergencies.

True LOW = 29 samples

- Only 2 were correctly predicted as LOW
- 27 were incorrectly predicted as HIGH
- The model over-predicts HIGH for almost all LOW cases.

True HIGH = 28 samples

- 24 were correctly predicted as HIGH
- 4 were incorrectly predicted as LOW
- The model predicts HIGH reasonably well, but fails on LOW cases.

The decision tree has a strong bias toward predicting HIGH asthma incidents, regardless of the input.

Random Forest

Can we predict if a location is currently experiencing 'High Pollution' based on the time of year, location, and the pollutant being measured?

We built a Random Forest model with 300 decision trees to predict pollution levels (High vs. Low) using three key features:

- **Indicator ID** (type of pollutant being measured)
- **Geo Join ID** (geographic location)
- **Month**

Key Technical Decisions:

- Applied **One-Hot Encoding** to convert categorical variables into binary indicators that the model can process
- Used **class_weight='balanced'** to address dataset imbalance (fewer high pollution instances)
- Split data 80/20 for training and testing to validate model performance
- Allowed trees to grow fully (max_depth=None) to capture complex patterns

Random Forest Results

Overall Performance: 70% accuracy in predicting pollution levels

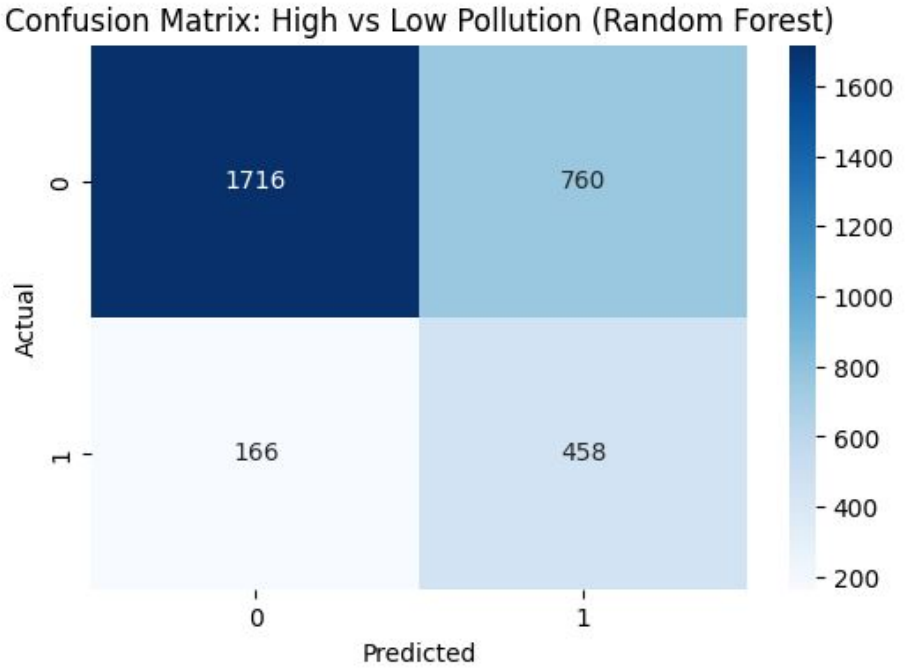
For Low Pollution:

- 91% precision - When the model predicts low pollution, it's correct 91% of the time
- 69% recall - The model identifies 69% of actual low pollution cases
- Strong at avoiding false alarms

For High Pollution:

- 38% precision - When predicting high pollution, it's only correct 38% of the time
- 73% recall - Catches 73% of actual high pollution events
- Model is sensitive to high pollution but produces many false positives

The model prioritizes catching high pollution events (important for public health alerts) but struggles with precision. The confusion matrix shows 760 false positives (predicted high when actually low) versus only 166 false negatives (missed high pollution events). This trade off may be acceptable if catching pollution events is more important than occasional false alarms.



Clustering

DBSCAN

K-Means

DBSCAN Clustering Analysis

Analysis objective:

Search for the natural stratification of air quality in NYC communities.

Model Settings:

- Use original continuous pollutant values (NO_2, PM2.5, O_3).
- Standardize features with z-score normalization so each variable has mean 0 and variance 1.

Algorithm:

- Apply DBSCAN
- Parameters: eps = 0.45, min_samples = 5

DBSCAN Result

DBSCAN successfully identified four distinct air quality patterns.

Cluster 0 represents the daily air quality in NYC.

- Low PM2.5, moderate O_3
- The most common and stable.

Cluster 1 & 2 represent severe pollution hotspots

- High NO_2 + High PM2.5
- Cluster 2 is even more extreme, possibly corresponding to pollution peaks during specific periods or locations.

Cluster 3 represents a unique high O_3 pattern.

- Unlike the previous two types of high NO_2/PM2.5 patterns,
- it may indicate a pollution situation dominated by photochemical reactions.

A small standard deviation indicates that the pollution patterns for each category are highly consistent.

```
DBSCAN Cluster Statistics:

Cluster 0:
  Size: 2317 observations
  NO₂:   19.92 ± 5.68
  PM2.5: 8.98 ± 1.80
  O₃:    30.40 ± 1.57

Cluster 1:
  Size: 13 observations
  NO₂:   28.64 ± 0.55
  PM2.5: 13.33 ± 0.37
  O₃:    27.68 ± 0.63

Cluster 2:
  Size: 5 observations
  NO₂:   31.28 ± 0.45
  PM2.5: 12.91 ± 0.27
  O₃:    21.98 ± 0.20

Cluster 3:
  Size: 6 observations
  NO₂:   14.57 ± 0.28
  PM2.5: 11.27 ± 0.19
  O₃:    36.07 ± 0.13
```

K-Means

Can we identify distinct 'Pollution Profiles' for neighborhoods?

1. Create a pollution x location matrix
   - Pivot the dataset so each row is a neighborhood
   - Each column is a pollutant
   - Values are scaled pollution levels (z-scores), so different units can be compared
2. Ran K-Means Clustering (K=3)
   - Grouped neighborhoods based on similarities across all pollutants
3. Visualized clusters with PCA
   - Used PCA to compress multi-dimensional data into 2 principal components
   - They capture the two directions with the strongest variation for plotting

K-Means Results

Cluster 0

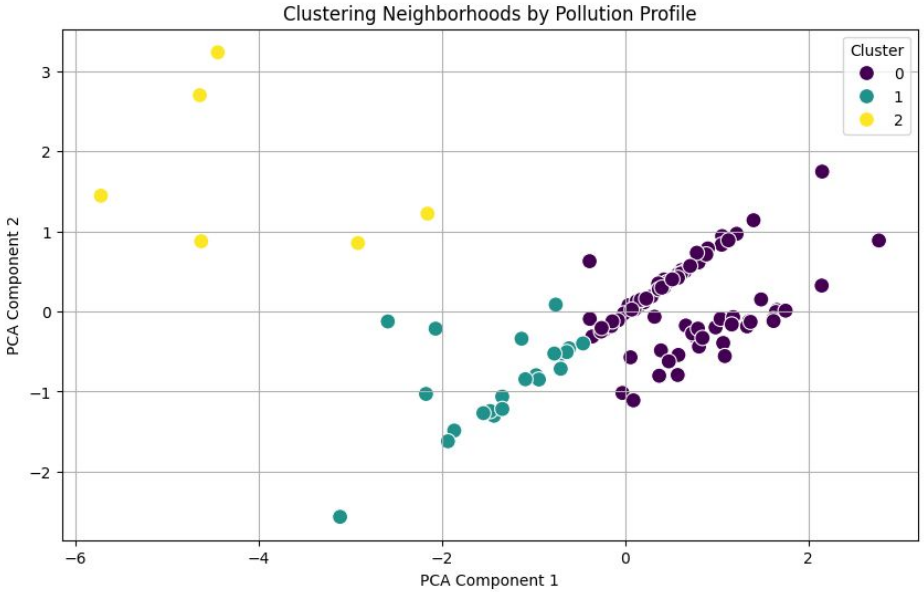
- Most pollutant values are slightly above average
- Cleaner, low emission neighborhoods

Cluster 1

- Above average for most pollutants
- Moderately polluted areas

Cluster 2

- Very high boiler emissions (NOx, PM2.5, S02)
- Heavy industrial or high-traffic pollution hotspots



Cluster 0 Top 5 Pollutants:
Outdoor Air Toxics - Benzene 0.284455
Ozone (O3) 0.269178
Outdoor Air Toxics - Formaldehyde 0.245921
Nitrogen dioxide (NO2) 0.244053
Boiler Emissions- Total NOx Emissions 0.198584

Cluster 1 Top 5 Pollutants:
Ozone (O3) 0.977650
Nitrogen dioxide (NO2) 0.834158
Outdoor Air Toxics - Benzene 0.618434
Fine particles (PM 2.5) 0.555617
Outdoor Air Toxics - Formaldehyde 0.497538

Cluster 2 Top 5 Pollutants:
Boiler Emissions- Total PM2.5 Emissions 2.291318
Boiler Emissions- Total SO2 Emissions 2.245354
Boiler Emissions- Total NOx Emissions 2.213942
Ozone (O3) 1.142782
Outdoor Air Toxics - Benzene 0.970350

END