AQI DETECTIVES 🕵️ 🕵️ 🕵️ 🕵️ 🕵️ 💨

Air Quality: A Critical Global Concern

Collaborators: Shephali Dubey/Girish Hosalli/Christine Chung/Chearine Pringle/Xiwu Dai

Project Overview & Goals

Air quality is a vital aspect of our environment and human health. The quality of the air we breathe directly impacts our well-being and the sustainability of our planet. Our project aims to discover trends in air pollution in New York City and Long Island and do a deep dive into the air quality index data from the past (12) years, using Machine Learning models, to predict the AQI for the next year. This will help individuals , especially those who fall under the category of sensitive groups, know when it is okay to go outside, and help governments manage traffic patterns and industrial emissions, in an effort to reduce Climate Change.

Project Techniques & Models

1

Data Collection & EDA

We pulled AQI data from EPA.com, using their API , cleaned and preprocessed the data, and performed EDA. Following key metrics were used to finalize the dataset (.csv) for observations, analysis and modeling:

- by State = NY
- by Pollutant = AQI (PM2.5, PM10, Ozone, SO2,NO2,Carbon Monoxide)
- by County = 5 Boroughs of NY & LI
- by Date = Jan 1, 2013-Apr1, 2024

2

Model Development

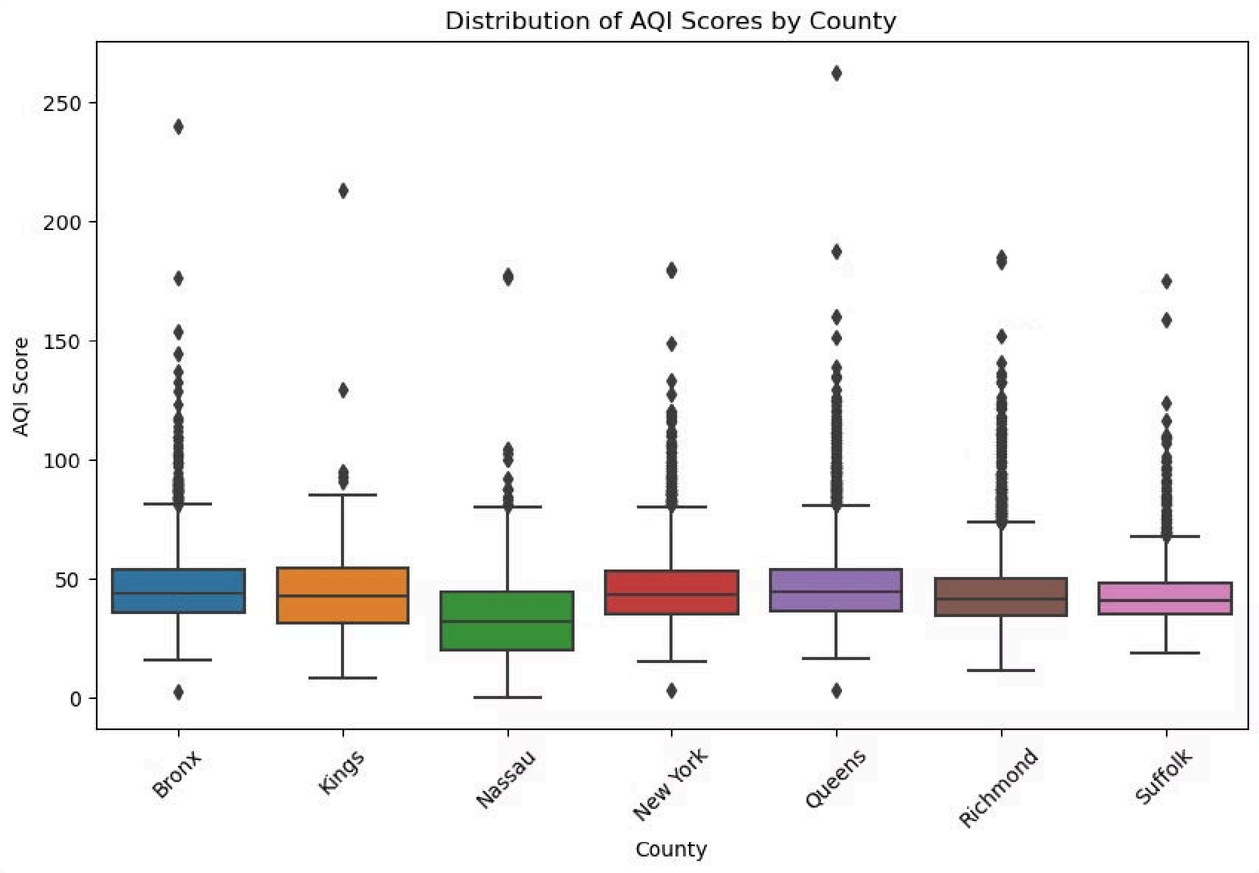
We used Supervised Learning ML Models such as Logistic Regression, Random Forest Regressor , Linear Regression, and Prophet , to model the data, checked for data leakage , performed accuracy, precision, r2, mse, cross validation checks to optimize the models.

3

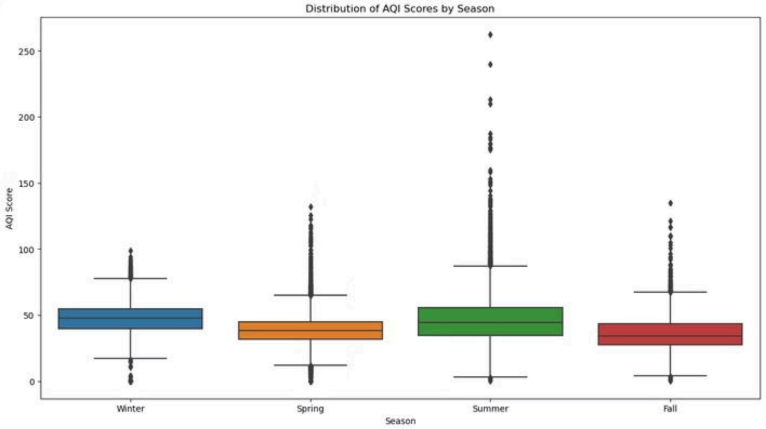
Forecasting

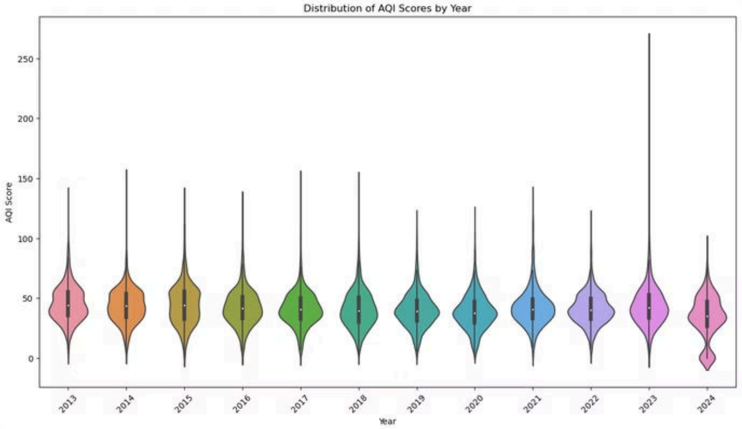
We decided on Random Forest and Prophet Models to try to make predictions for next 1 year.

EDA



Temporal Trends: - The data spans multiple years, allowing us to observe long-term trends in air quality.
- There are clear seasonal patterns in air quality, with generally higher AQI scores in summer months.
- Over the years, there's a slight improvement in overall air quality, but this trend isn't consistent across all pollutants





Seasonal Variation: Compare the median (middle line of each box) across seasons. Higher medians indicate worse air quality.

Variability: The size of each box represents the interquartile range. Larger boxes suggest more variable air quality within that season.

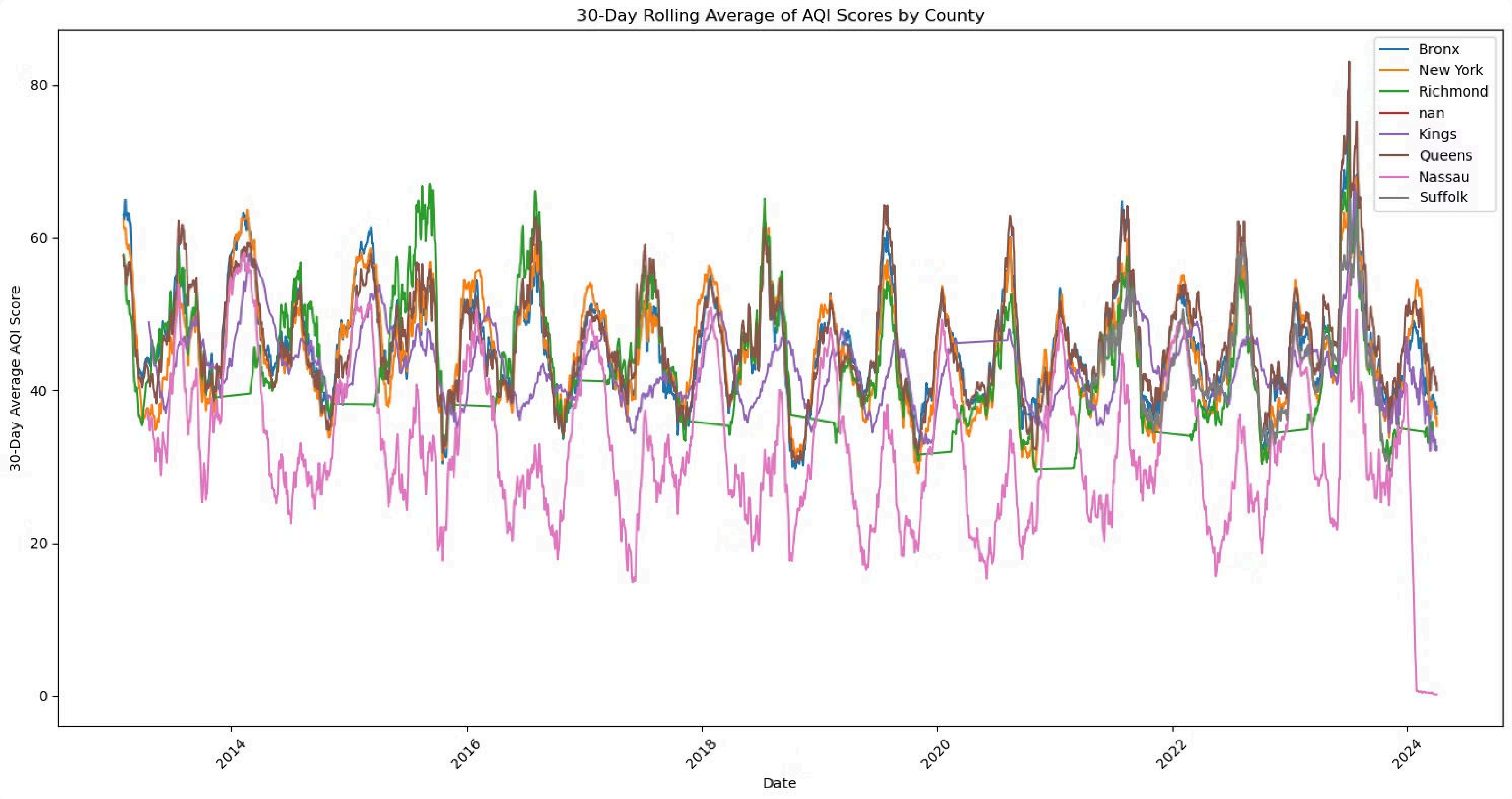
Outliers: Points beyond the whiskers are outliers, representing unusually high AQI scores. More outliers in a season suggest more frequent pollution events.

Seasonal Patterns: Patterns such as higher AQI scores in summer (potentially due to increased ozone) or winter (possibly due to increased particulate matter from heating).

Distribution of AQI scores for each year: Distribution Shape, The width of each "violin" shows how common different AQI scores are. Wider sections indicate more frequent occurrences of those scores.

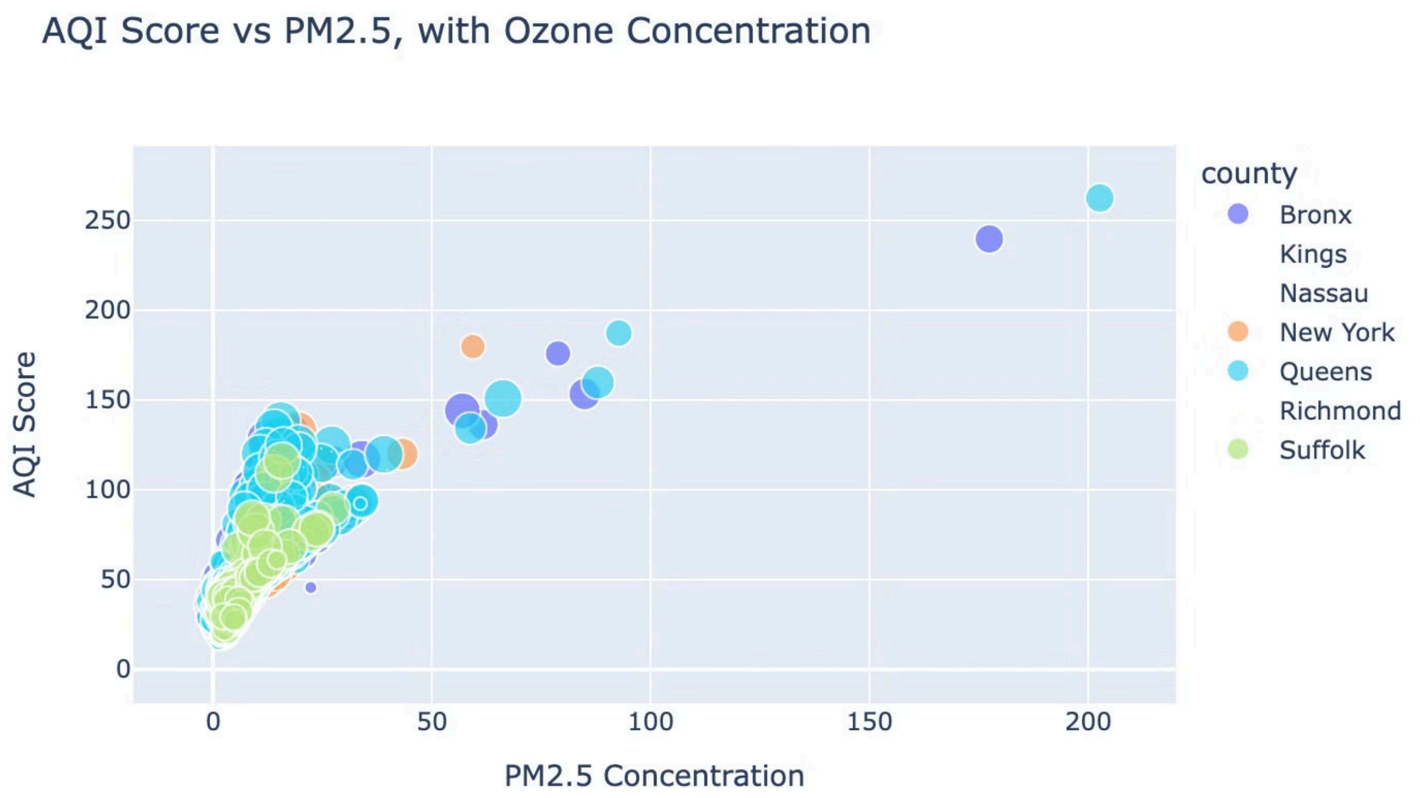
Central Tendency: The thickest part of each violin represents the most common AQI scores for that year. Shifts up or down indicate worsening or improving air quality.

Extremes: The extent of the thin ends of the violins shows the range of extreme values experienced each year.



Spatial Variations:

- Different counties show varying levels of air quality, likely due to factors such as population density, industrial activity, and local geography.
- Urban areas like Bronx consistently show higher levels of certain pollutants, particularly NO2 and PM2.5, compared to less densely populated areas.



Pollutant-specific Observations:

- Ozone (44201): Tends to be higher in summer months, likely due to increased sunlight and heat.

- PM2.5 (88101 and 88502): Shows significant variability and is often the primary driver of high AQI scores.

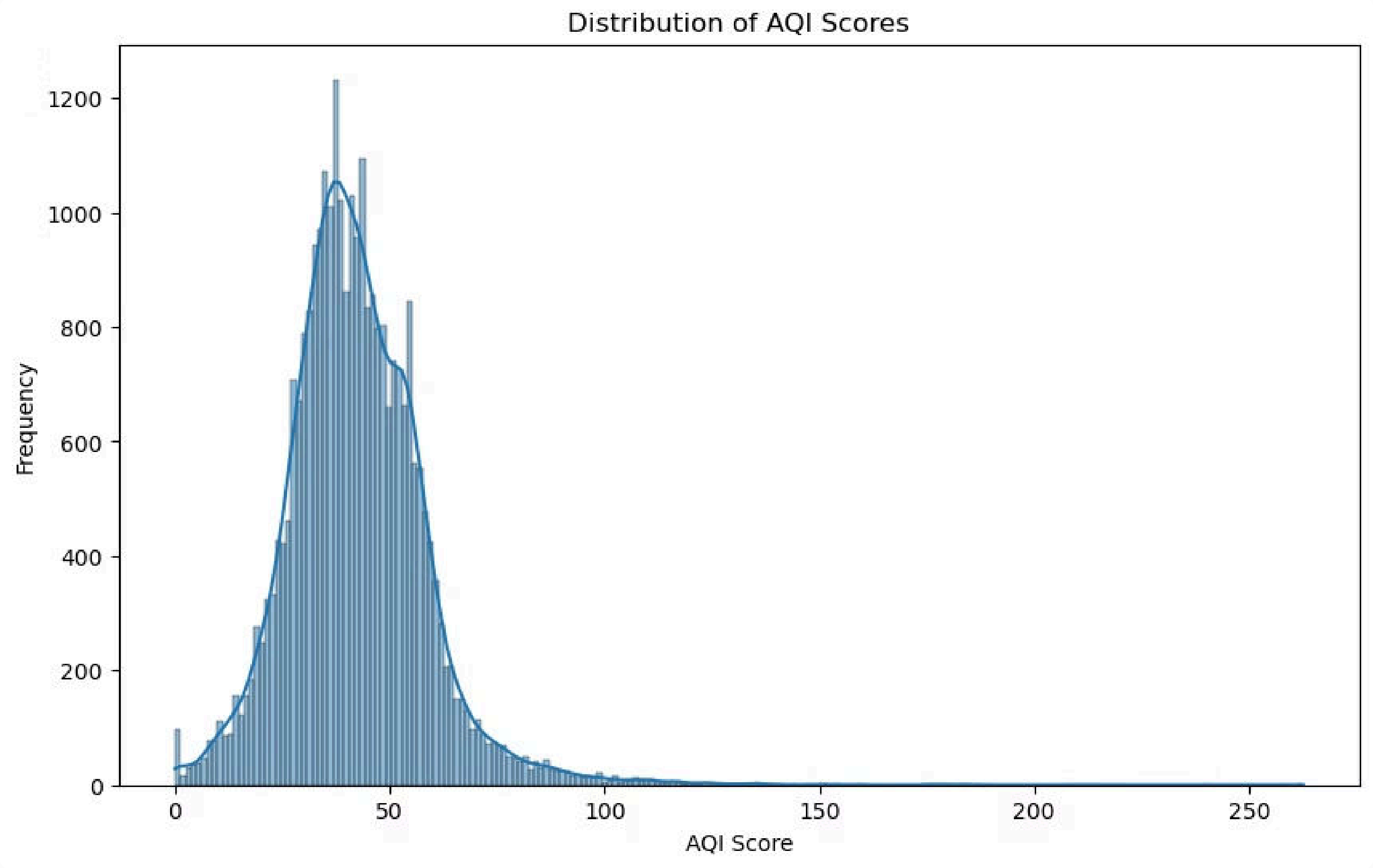
- Nitrogen Dioxide (42602): Generally higher in urban areas and during winter months.

- Sulfur Dioxide (42401): Levels have decreased over the years, possibly due to stricter emissions controls.

- Carbon Monoxide (42101): Generally low levels, with occasional spikes in urban areas.

AQI Score Distribution:

- Most days fall in the "Good" (0-50) to "Moderate" (51-100) AQI range.
- There are occasional spikes into "Unhealthy for Sensitive Groups" (101-150) or "Unhealthy" (151-200) categories.



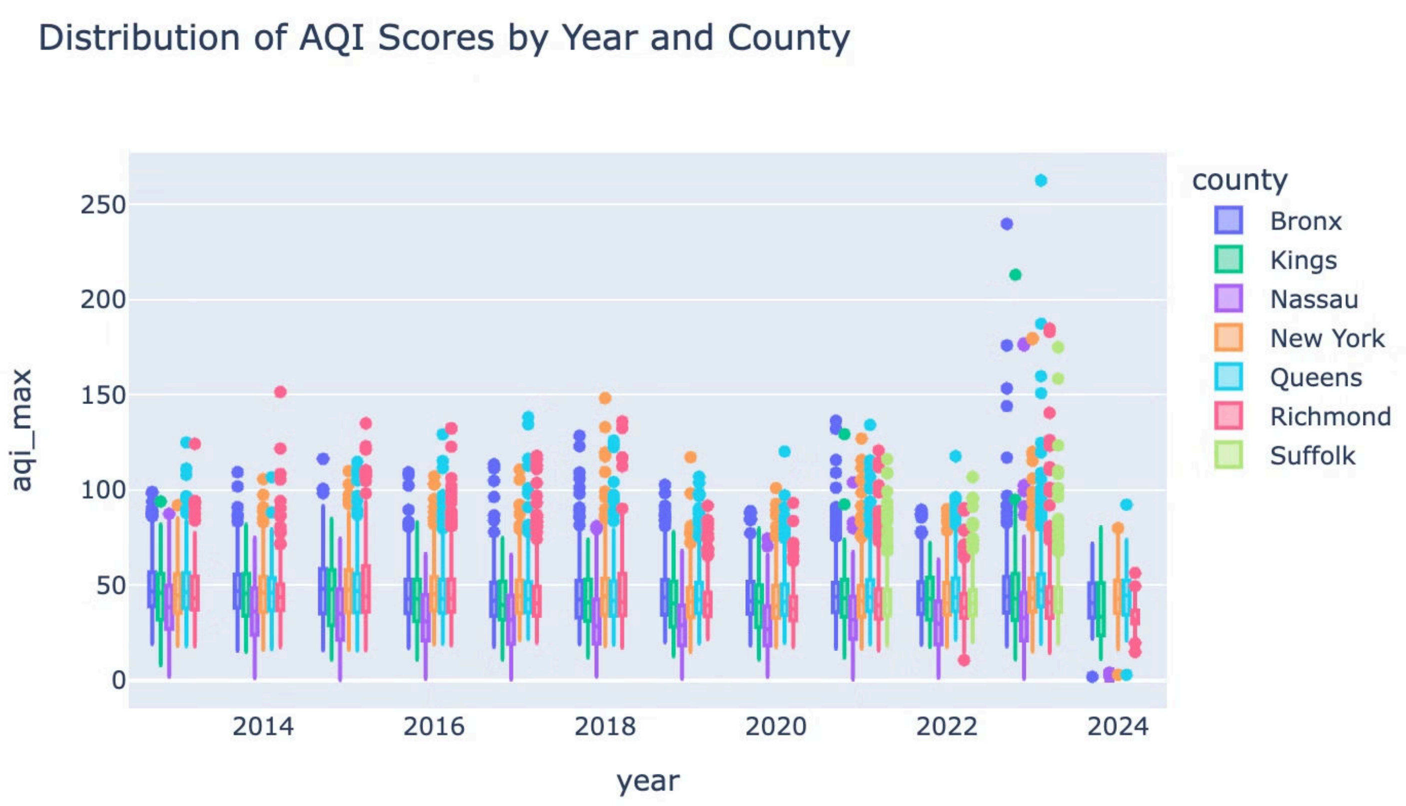
Histogram of AQI Scores: This histogram shows the distribution of AQI scores across all counties and time periods.

Analysis: Central Tendency: The peak of the distribution shows the most common AQI scores.

Spread: The width of the distribution indicates the range of typical AQI scores.

Skewness: If the distribution is skewed right (tail to the right), it indicates occasional high pollution events.

Multiple Peaks: Multiple peaks could suggest different typical conditions, perhaps corresponding to different seasons or locations.

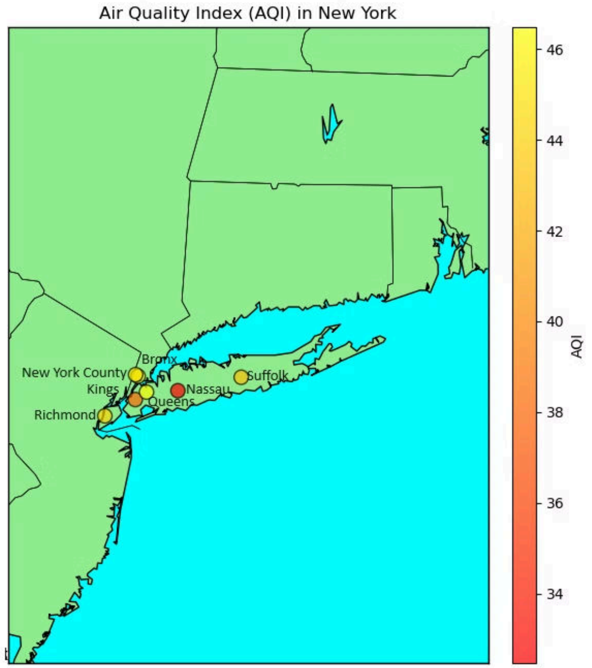


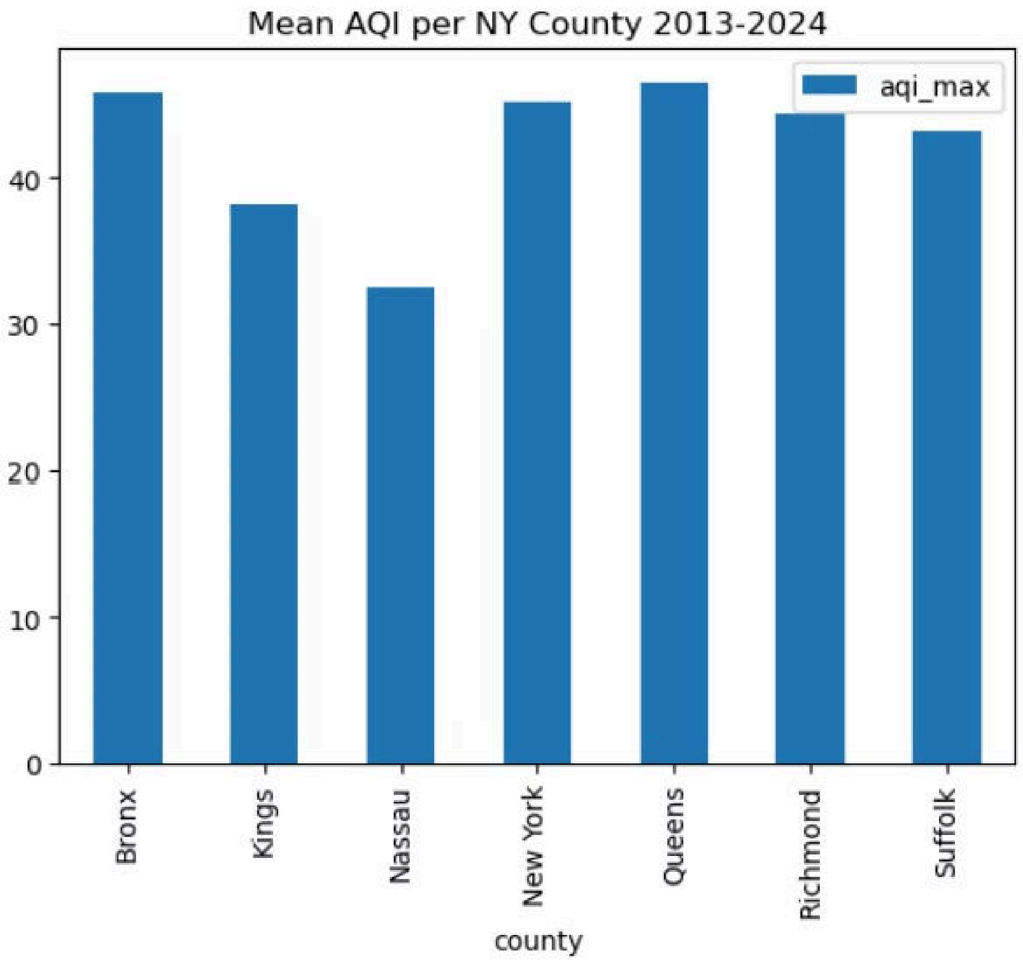
Observations and Outliers:
Canadian Wildfires: Most recently, in June 2023, severe wildfires in Quebec led to significant air quality issues in New York and other parts of the northeastern United States. This event caused some of the worst air quality readings in recent history for the region, with AQI levels reaching "Hazardous" levels in some areas. While the provided dataset doesn't explicitly mention Canadian wildfires, we can infer their impact by looking for unusual spikes in PM2.5 levels, especially during summer months.

For example:
There are several instances where PM2.5 levels (88101 and 88502) spike dramatically, sometimes reaching into the "Unhealthy" or even "Very Unhealthy" AQI categories. These spikes often occur across multiple counties simultaneously, suggesting a large-scale event like wildfire smoke. The timing of these spikes (typically in summer months) aligns with the usual wildfire season in Canada.

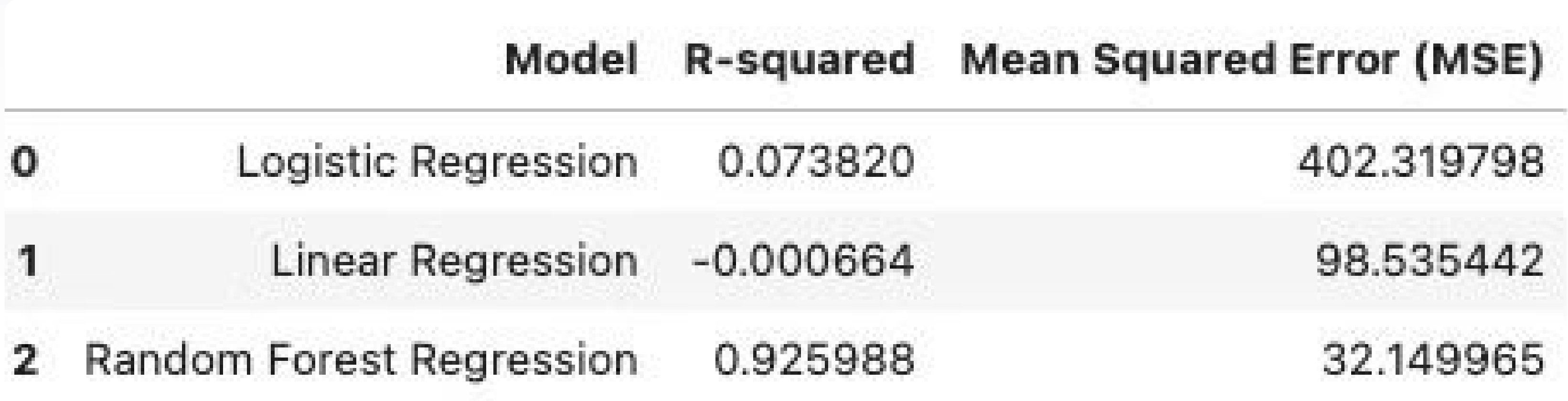
Other Notable Observations:
- Winter Inversions: There are periods in winter months where pollutant levels (particularly PM2.5 and NO2) remain elevated for several days, possibly due to temperature inversions trapping pollutants near the ground.
- Summer Ozone Events: Occasionally, there are days with very high ozone levels, typically during hot, sunny days in summer. These could be considered "ozone events" and may trigger air quality alerts.
- Urban Heat Island Effect: Urban areas like Bronx consistently show higher temperatures and pollutant levels compared to surrounding areas, demonstrating the urban heat island effe

Overall AQI Per County





Comparison of Regression Models

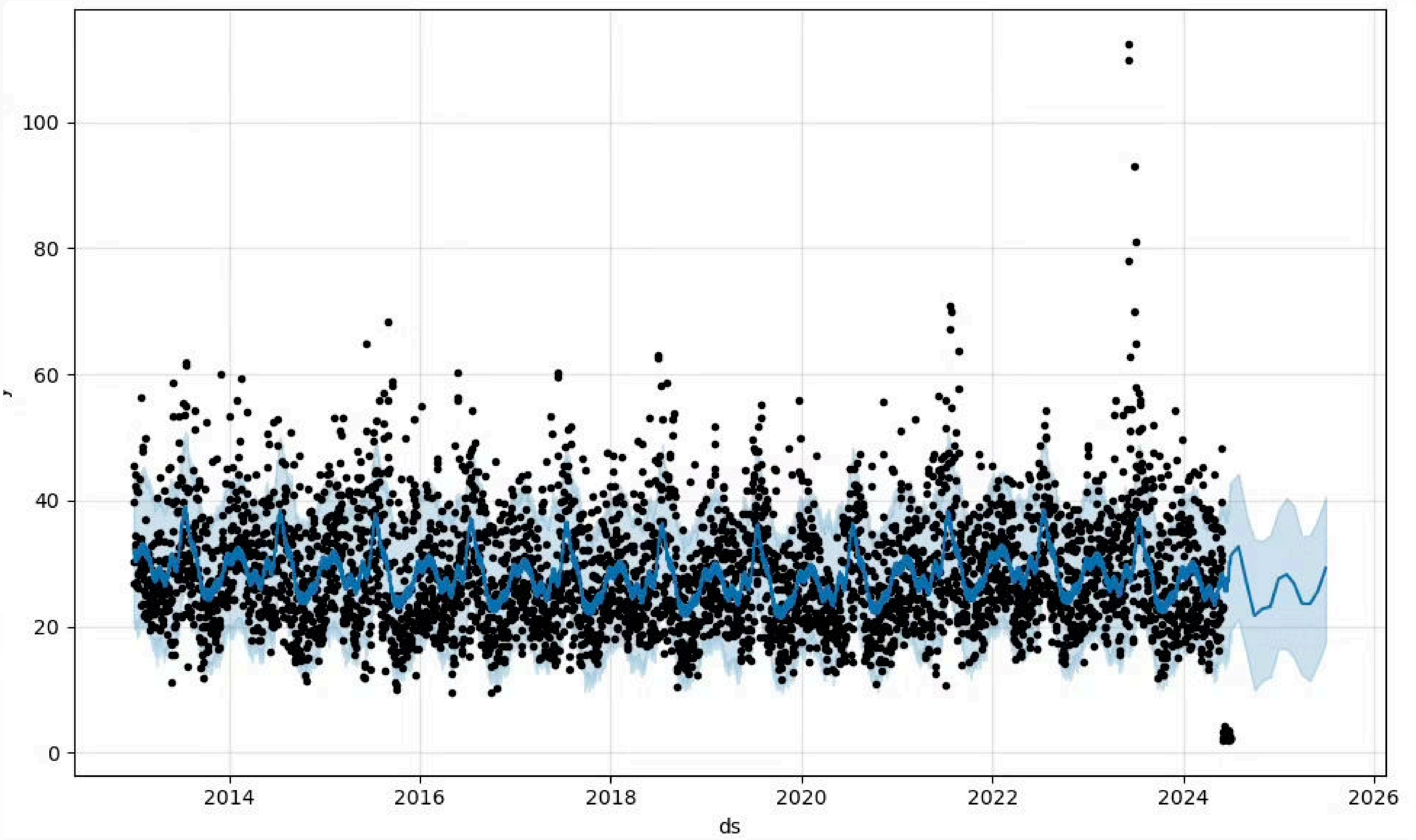


	Model	R-squared	Mean Squared Error (MSE)
0	Logistic Regression	0.073820	402.319798
1	Linear Regression	-0.000664	98.535442
2	Random Forest Regression	0.925988	32.149965

Random Forest Regressor Model

- Train Accuracy = 98.9%
- Test Accuracy = 92.6%
- It fits!!!

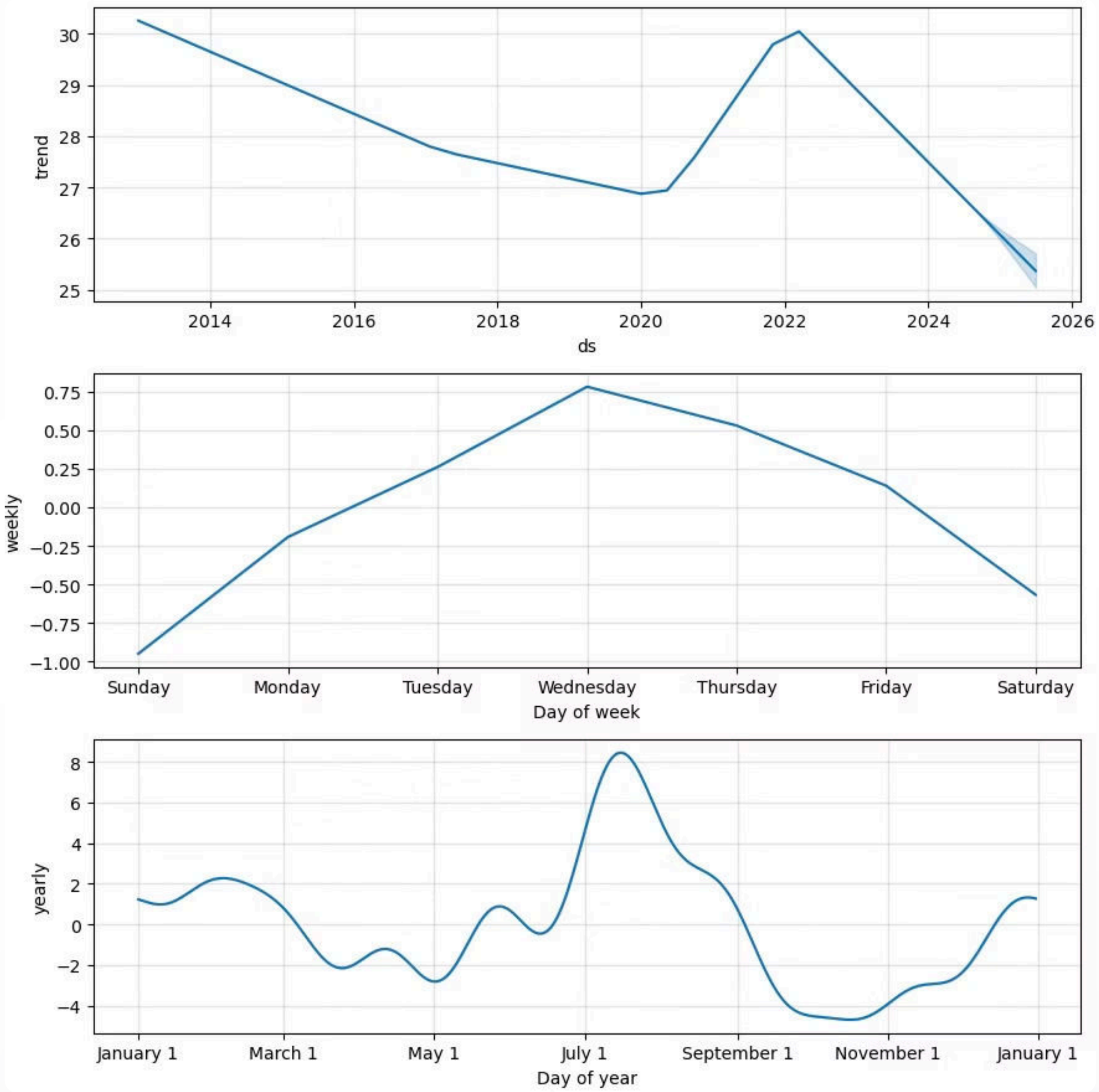
Prophet Model



Prophet Model

The plot above portrays that the Prophet Model is a good fit to predict future AQI trends!

Prophet Model Trends



Prophet Model

Yearly Trend - 2020 had the lowest AQI , possibly due to people being mostly home during Covid.

Weekly Trend - Wednesdays show peak AQI pollution

Monthly Trend - July shows the highest AQI pollution

Project Current & Future Approach

1

Project Leads

Each person was assigned as a lead person to manage each part of the project as follows:

- Christine = Git Hub
- Girish = Data extraction, cleaning & transformation
- Chearine = Models Evaluation, Optimization , Performance
- Xiwu = Readme
- Shephali = Presentation

2

Project Setbacks

- We lost a week in trying to get a good dataset due to the following reasons:
    - Data by County was not giving the needed dataset and it took time to figure out how to pull the needed Dataset i.e. Data by State.
    - Data could not be pulled for more than a year at a time.
- Preprocessing the dataset for further analysis took longer then anticipated due to complex data i.e. 1 hour, 8 hour, 24 hour data with different observation days frequency , missing observations in various months for various parameters and counties, etc.

3

Future Enhancements

If we had more time we would have liked to

- Connect weather , mortality and other factors
- Frontend Dashboard to pull data by Zip code
- Learn about ARIMA and implement in our model
- Create an ordinal classifier using the different AQI

Conclusion

Project Achievements

- Leveraging historical air quality data
- Using advanced machine learning models to predict future trends in New York City and Long Island

Key Findings

1. Temporal and Spatial Trends:

- Seasonal variations in air quality
- Pollutant-specific behaviors might contribute to higher AQI scores(e.g., higher ozone levels in summer, increased particulate matter in winter)

2. Predictive Insights:

- Employed Random Forest machine learning models and Prophet
- Provided predictive insights into the Air Quality Index (AQI) for the upcoming years

Improving air quality requires a collective effort, involving individuals, governments, and industries. By adopting sustainable practices and investing in clean technologies, we can create a healthier and more sustainable future for all.

Implications and Recommendations:

1. Public Health: The data underscores the importance of air quality alerts, especially for sensitive groups during high pollution events.
2. Policy: Targeted interventions may be needed in areas consistently showing higher pollutant levels.
3. Further Research: More detailed analysis of the correlation between wildfire events and local air quality could help improve predictive models and public health responses.
4. Monitoring: Continued comprehensive monitoring across different counties is crucial for understanding long-term trends and the effectiveness of air quality improvement measures.

Citations

https://aqs.epa.gov/aqsweb/documents/data_api.html#signup

https://aqs.epa.gov/aqsweb/documents/codetables/parameter_classes.html

Index: Air Quality, Explained in a nutshell

Key Air Pollutants :

- **Particulate Matter (PM2.5 & PM10) -** from combustion sources like vehicles and industrial processes.
- **Ozone (O_3)** - gas formed by chemical reactions involving volatile organic compounds (VOCs) and nitrogen oxides (NOx) emitted from vehicles and industrial activities.
- **Sulfur Dioxide (SO_2)** - primarily from burning fossil fuels
- **Nitrogen Dioxide (NO_2)** - gas formed from emissions from vehicles, power plants and off-road equipment.
- **Carbon Monoxide** - from combustion sources like vehicles , industrial processes and fossil fuels.

Health & Environmental Impacts:

- **Cardiovascular Problems** - due to fine particulate matter which can penetrate deep into the lungs, leading to inflammation and an increased risk of heart attacks and strokes.
- **Respiratory diseases** - such as asthma, bronchitis, and other respiratory illnesses, particularly in vulnerable populations like children and the elderly.
- **Environmental Degradation** - Acid rain, caused by sulfur dioxide and nitrogen oxides, damages forests, lakes, and aquatic ecosystems.
- **Climate Change** - air pollutants like carbon dioxide (CO_2) trap heat in the atmosphere, contributing to global warming and its associated effects

Factors Impacting Air Quality:

- **Wind Patterns** - Wind speed and direction can disperse or concentrate pollutants.
- **Temperature Inversions** - A layer of warm air above cooler air can trap pollutants near the ground.
- **Precipitation** - Rain and snow can cleanse the atmosphere of pollutants, improving air quality in the short term.
- **Human Activities** -Emissions from industries, vehicles, and other human activities are a major contributor to air pollution.

Air Quality Monitoring methods:

- **AQI -** A measure of air quality based on various pollutants, providing a standardized way to assess air quality.
- **Sensors -** that measure specific air pollutants, providing real-time data on air quality conditions.
- **Satellite Imagery -** Remote sensing techniques to monitor air pollution levels, especially over large areas/regions.