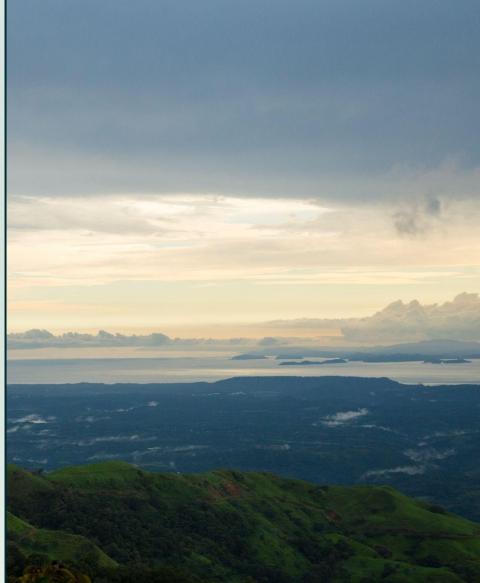


Environmental ML Analytics: NC Urban Areas



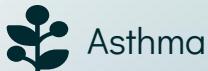
Ariana, Natalie, Mary, Jolciel

TABLE OF CONTENTS

| | | | |
|--------------------------------------|---|--|---------------------|
| Data Clean Up | 1 | | |
| Database Sets | 2 | | |
| Visualizations & Analysis | 3 | | |
| | |  | |
| | | | The ML Model |
| | 4 | | |
| | | | Conclusions |
| | 5 | | |
| | | | Questions |
| | 6 | | |

WHY AIR?

Want to predict Air Quality in North Carolina.



Asthma



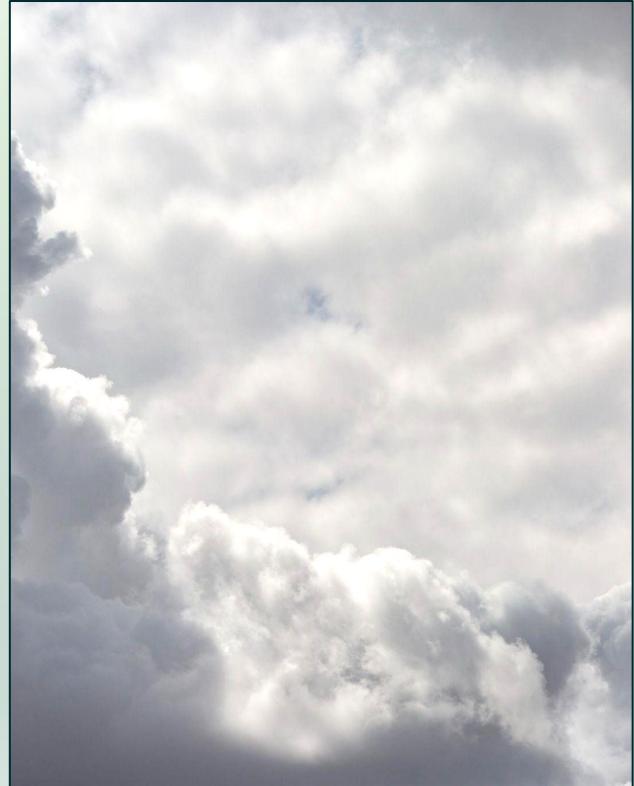
Allergies



Respiratory health



Environmental Health



Our Goal:

**Predict the number of unhealthy days in the 2023 year,
using data that goes as far back as 1980.**

Our Focus:

**Primarily focus on Mecklenburg County and Wake
County, home to the largest city and capital of North
Carolina.**

What is AQI?

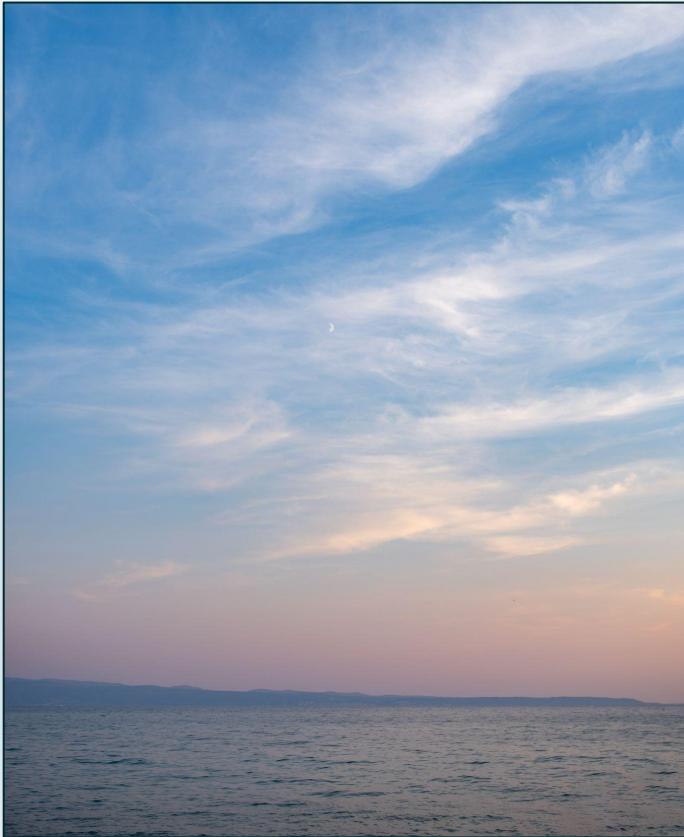
AQI is a numerical scale that communicates the quality of the air.

-  Calculated based on the concentrations of specific air pollutants
-  AQI scale is divided into different categories
-  Categories often include ranges such as "Good," "Moderate," "Unhealthy for Sensitive Groups," "Unhealthy," etc

| Air Quality Index (AQI) Values | Levels of Health Concern | Colors |
|---------------------------------------|--------------------------------------|--|
| <i>When the AQI is in this range:</i> | <i>..air quality conditions are:</i> | <i>...as symbolized by this color:</i> |
| 0-50 | Good | Green |
| 51-100 | Moderate | Yellow |
| 101-150 | Unhealthy for Sensitive Groups | Orange |
| 151 to 200 | Unhealthy | Red |
| 201 to 300 | Very Unhealthy | Purple |
| 301 to 500 | Hazardous | Maroon |



1



Data Clean Up

Data Sets

- Obtained CSV files for US counties (1980-2022)
- Merged files
- Selected NC counties
- Dropped null/duplicates
- Created single merged file

Source: *United States Environmental Protection Agency*

https://aqs.epa.gov/aqswb/airdata/download_files.html#Annual

| | County | Year | Days with AQI | Good Days | Moderate Days | Unhealthy for Sensitive Groups Days | Unhealthy Days | Very Unhealthy Days | Hazardous Days | Max AQI | 90th Percentile AQI | Median AQI | Days CO | Days NO2 | Days Ozone | Days PM2.5 | Days PM10 |
|----------------|-----------|------|---------------|-----------|---------------|-------------------------------------|----------------|---------------------|----------------|---------|---------------------|------------|---------|----------|------------|------------|-----------|
| State | | | | | | | | | | | | | | | | | |
| North Carolina | Alamance | 2000 | 104 | 37 | 66 | | 1 | 0 | 0 | 104 | 76 | 56 | 0 | 0 | 0 | 104 | 0 |
| North Carolina | Alexander | 2000 | 211 | 94 | 64 | | 48 | 5 | 0 | 182 | 126 | 58 | 0 | 0 | 211 | 0 | 0 |
| North Carolina | Avery | 2000 | 365 | 211 | 117 | | 34 | 3 | 0 | 172 | 101 | 47 | 0 | 0 | 353 | 12 | 0 |
| North Carolina | Buncombe | 2000 | 260 | 129 | 110 | | 15 | 6 | 0 | 179 | 93 | 51 | 0 | 0 | 176 | 84 | 0 |
| North Carolina | Cabarrus | 2000 | 115 | 36 | 78 | | 1 | 0 | 0 | 118 | 86 | 56 | 0 | 0 | 0 | 114 | 1 |

Gather 2023 Data

1. Gather individual csv files for each pollutant
 - Contain Daily records & AQI value for pollutant
2. Merge all csv files
3. Create DataFrames
 - Focus on 2 counties with ALL available data
4. Sort By Date & Gather Required Data



| | DATE | COUNTY | CO_DAILY_AQI_VALUE | NO2_DAILY_AQI_VALUE | OZONE_DAILY_AQI_VALUE | PM2_DAILY_AQI_VALUE | PM10_DAILY_AQI_VALUE |
|------|------------|--------|--------------------|---------------------|-----------------------|---------------------|----------------------|
| 0 | 2023-01-01 | Wake | 5.0 | NaN | NaN | NaN | NaN |
| 544 | 2023-01-01 | Wake | NaN | 16.0 | NaN | NaN | NaN |
| 1082 | 2023-01-01 | Wake | NaN | NaN | 23.0 | NaN | NaN |
| 1459 | 2023-01-01 | Wake | NaN | NaN | NaN | 35.0 | NaN |
| 273 | 2023-01-01 | Wake | 5.0 | NaN | NaN | NaN | NaN |

Breakdown of Required Data

Days of { Pollutant }

- Total count of { Pollutant } with the max AQI for the day

90th Percentile AQI

- From all max Daily AQI Values

| Year | County | 90th Percentile AQI | Days CO | Days NO2 | Days OZONE | Days PM2.5 | Days PM10 |
|------|-------------|---------------------|---------|----------|------------|------------|-----------|
| 2023 | Mecklenburg | 71.0 | 0 | 1 | 218 | 108 | 0 |
| 2023 | Wake | 63.0 | 0 | 2 | 175 | 148 | 0 |

Columns Required for Model:

- 90th Percentile AQI
- Days CO
- Days NO2
- Days OZONE
- Days PM2.5
- Days PM10



Breakdown of Required Data

To Gather Data:

- Change null values to zero
- Group by Date, keep maximums
- Create new columns for required data
- Use for loop to find:
 - Max AQI values for each pollutant
 - Pollutant with Max AQI value
 - Keep count of pollutants
 - Overall Max AQI for the date
- Find 90th Percentile of all Daily Max AQI Values
- Find Sum of count of pollutants
- New Data Frame with required columns
- Export to CSV to Use in Model



Resources

→ [*nc_2023_data.csv*](#)





2

Database Sets

Data Retrieval

1. Loading the Data

```
# load the data from your CSV file into a pandas DataFrame.
```

```
import pandas as pd

# Load the data
file_path = '/Users/nataliabondarenko/Desktop/Github/project-4/Resources/nc_aqi_1988-2022.csv'
data = pd.read_csv(file_path)
```

```
# Preview the data
print(data.head())
```

```
... County Year Days with AQI Good Days Moderate Days \
0 Bunccombe 2000 260.0 120.0 118.0
1 Bunccombe 2001 253.0 141.0 100.0
2 Bunccombe 2002 260.0 144.0 83.0
3 Bunccombe 2003 303.0 178.0 119.0
4 Bunccombe 2004 357.0 187.0 166.0

Unhealthy for Sensitive Groups Days Unhealthy Days Very Unhealthy Days \
0 15.0 6.0 0.0
1 11.0 1.0 0.0
2 27.0 6.0 0.0
3 6.0 0.0 0.0
4 4.0 0.0 0.0

Hazardous Days Max AQI 98th Percentile AQI Median AQI Days CO \
0 0.0 179.0 93.0 51.0 0.0
1 0.0 173.0 87.0 48.0 0.0
2 0.0 179.0 108.0 42.0 0.0
3 0.0 137.0 72.0 46.0 0.0
4 0.0 133.0 74.0 49.0 0.0

Days NO2 Days Ozone Days PM2.5 Days PM10
0 0.0 176.0 80.0 0.0
1 0.0 174.0 78.0 0.0
2 0.0 172.0 88.0 0.0
```

```
# Checking for missing values
print(data.isnull().sum())
```

```
[4]

... County 0
Year 0
Days with AQI 0
Good Days 0
Moderate Days 0
Unhealthy for Sensitive Groups Days 0
Unhealthy Days 0
Very Unhealthy Days 0
Hazardous Days 0
Max AQI 0
98th Percentile AQI 0
Median AQI 0
Days CO 0
Days NO2 0
Days Ozone 0
Days PM2.5 0
Days PM10 0
dtype: int64
```

3. Data Cleaning

Based on the inspection, perform necessary cleaning steps

```
# Removing duplicates
data = data.drop_duplicates()

# Confirming if any duplicates were removed
print(f'Data shape after removing duplicates: {data.shape}')
```

4. Data Transformation

Transform the data as needed for analysis.

```
# Converting 'State' and 'County' to category data types
data['State'] = data['State'].astype('category')
data['County'] = data['County'].astype('category')
```

```
# Checking the data types again to confirm the changes
data.dtypes
```

```
[5]

... State category
County category
Year int64
Days with AQI int64
Good days int64
Moderate Days int64
Unhealthy for Sensitive Groups Days int64
Unhealthy Days int64
Very Unhealthy Days int64
Hazardous Days int64
Max AQI int64
98th Percentile AQI int64
Median AQI int64
Days CO int64
Days NO2 int64
Days Ozone int64
Days PM2.5 int64
Days PM10 int64
dtype: object
```

```
# Drop the 'State' column
data = data.drop(columns=['State'])
```

```
# Confirming if the column was dropped
print(f'Data shape after dropping "State" column: {data.shape}')
```

Process:

- Data cleaning and data inspection
- Data transformation
- Data export
- DB storage



(Python Pandas)

SQLite

SQLite format 3

```
A%  
A&a))Û}tablepredicted_2023predicted_2023CREATE TABLE predicted_2023 (  
    "County" TEXT,  
    "Year" BIGINT,  
    "Days with AQI" FLOAT,  
    "Good Days" FLOAT,  
    "Moderate Days" FLOAT,  
    "Unhealthy for Sensitive Groups Days" FLOAT,  
    "Unhealthy Days" FLOAT,  
    "Very Unhealthy Days" FLOAT,  
    "Hazardous Days" FLOAT,  
    "Max AQI" FLOAT,  
    "90th Percentile AQI" FLOAT,  
    "Median AQI" FLOAT,  
    "Days CO" FLOAT,  
    "Days NO2" FLOAT,  
    "Days Ozone" FLOAT,  
    "Days PM2.5" FLOAT,  
    "Days PM10" FLOAT  
)EX##Û}tableair_qualityair_qualityCREATE TABLE air_quality (  
    "County" TEXT,  
    "Year" BIGINT,  
    "Days with AQI" FLOAT,  
    "Good Days" FLOAT,  
    "Moderate Days" FLOAT,  
    "Unhealthy for Sensitive Groups Days" FLOAT,  
    "Unhealthy Days" FLOAT,  
    "Very Unhealthy Days" FLOAT,  
    "Hazardous Days" FLOAT,  
    "Max AQI" FLOAT,  
    "90th Percentile AQI" FLOAT,  
    "Median AQI" FLOAT,  
    "Days CO" FLOAT,  
    "Days NO2" FLOAT,  
    "Days Ozone" FLOAT,  
    "Days PM2.5" FLOAT,  
    "Days PM10" FLOAT
```

5. Data Export

After cleaning export it to a new CSV file or directly use it for further analysis or machine learning.

```
[8] # Exporting to a new CSV file  
cleaned_file_path = '/Users/nataliabondarenko/Desktop/GitHub/project-4/Resources/cleaned_data.csv'  
data.to_csv(cleaned_file_path, index=False)
```

6. Database Storage

```
[8] from sqlalchemy import create_engine  
  
# Create an engine that stores data in the local directory's file  
engine = create_engine('sqlite:///nc_aqi_data.db')  
  
[9] # Store the data in a table named 'air_quality'  
data.to_sql('air_quality', con=engine, if_exists='replace', index=False)  
  
# Store the data in a table named 'predicted_2023'  
data.to_sql('predicted_2023', con=engine, if_exists='replace', index=False)
```

[9]
... 1596

(SQL Database)

Data Preprocessing

Feature engineering is a step in the data preprocessing phase which is an essential part of preparing data for machine learning models

```
# Adding new features based on the existing data

# Ratio of Good to Moderate Days
data['Good_to_Moderate_Ratio'] = data['Good Days'] / data['Moderate Days']

# Percentage of Unhealthy Days (of all types)
data['Unhealthy_Days_Percentage'] = (
    data['Unhealthy for Sensitive Groups Days'] + data['Unhealthy Days'] + data['Very Unhealthy Days'] + data['Hazardous Days']
) / data['Days with AQI'] * 100

# Display the first few rows with the new features
data[['County', 'Year', 'Good_to_Moderate_Ratio', 'Unhealthy_Days_Percentage']].head()
```

| | County | Year | Good_to_Moderate_Ratio | Unhealthy_Days_Percentage |
|---|----------|------|------------------------|---------------------------|
| 0 | Buncombe | 2000 | 1.172727 | 8.076923 |
| 1 | Buncombe | 2001 | 1.410000 | 4.743083 |
| 2 | Buncombe | 2002 | 1.734940 | 12.692308 |
| 3 | Buncombe | 2003 | 1.495798 | 1.980198 |
| 4 | Buncombe | 2004 | 1.126506 | 1.120448 |



Data Saving

*Saved as cleaned_data.csv
and nc_aqi_data.db*

| 1 | County | Year | Days with AQI | Good Days | Moderate Days | Unhealthy for Sensitive Groups Days | Unhealthy Days | Very Unhealthy Days | Hazardous Days |
|---|----------|------|---------------|-----------|---------------|-------------------------------------|----------------|---------------------|----------------|
| 2 | Buncombe | 2000 | 260.0 | 129.0 | 110.0 | 15.0 | 6.0 | 0.0 | 0.0 |
| 3 | Buncombe | 2001 | 253.0 | 141.0 | 100.0 | 11.0 | 1.0 | 0.0 | 0.0 |
| 4 | Buncombe | 2002 | 260.0 | 144.0 | 83.0 | 27.0 | 6.0 | 0.0 | 0.0 |
| 5 | Buncombe | 2003 | 303.0 | 178.0 | 119.0 | 6.0 | 0.0 | 0.0 | 0.0 |
| 6 | Buncombe | 2004 | 357.0 | 187.0 | 166.0 | 4.0 | 0.0 | 0.0 | 0.0 |
| 7 | Buncombe | 2005 | 362.0 | 192.0 | 159.0 | 10.0 | 1.0 | 0.0 | 0.0 |
| 8 | Buncombe | 2006 | 364.0 | 182.0 | 177.0 | 5.0 | 0.0 | 0.0 | 0.0 |
| 9 | Buncombe | 2007 | 344.0 | 161.0 | 173.0 | 10.0 | 0.0 | 0.0 | 0.0 |



AIR QUALITY INDEX ANALYSIS (AQI)

Here is where your presentation begins

Visualizations and Analysis

3



Environmental Data Analysis: Mecklenburg and Wake Counties

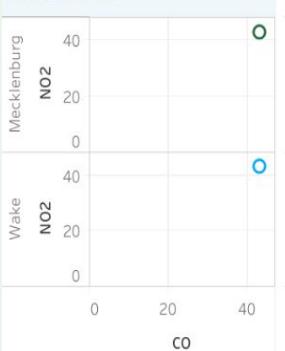
Past Patterns and Future Projections of Exposure Levels



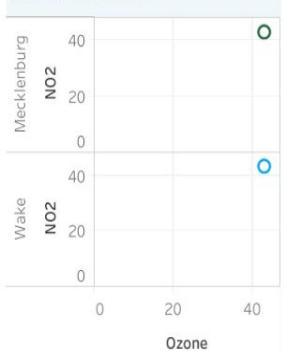
County-Wide Pollution Correlations: Mecklenburg and Wake Counties

Exploring Relationships in Exposure Levels

NO2 and CO



NO2 and Ozone



CO and NO2

0.07204 0.13535

Ozone and NO2

0.1253 0.2427

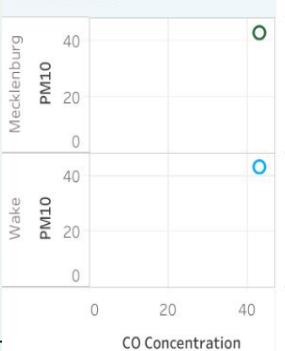
CO and PM2.5

0.1468 0.4293

PM10 and PM2.5

-0.49190 -0.46573

PM10 and CO

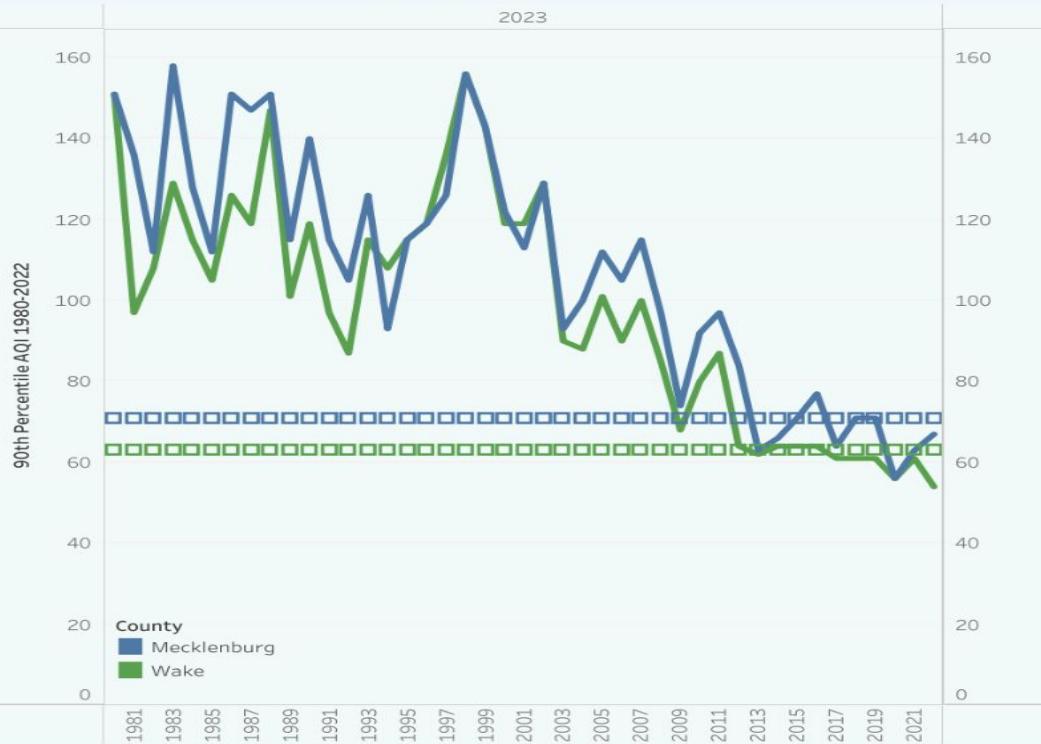


PM10 and PM2.5



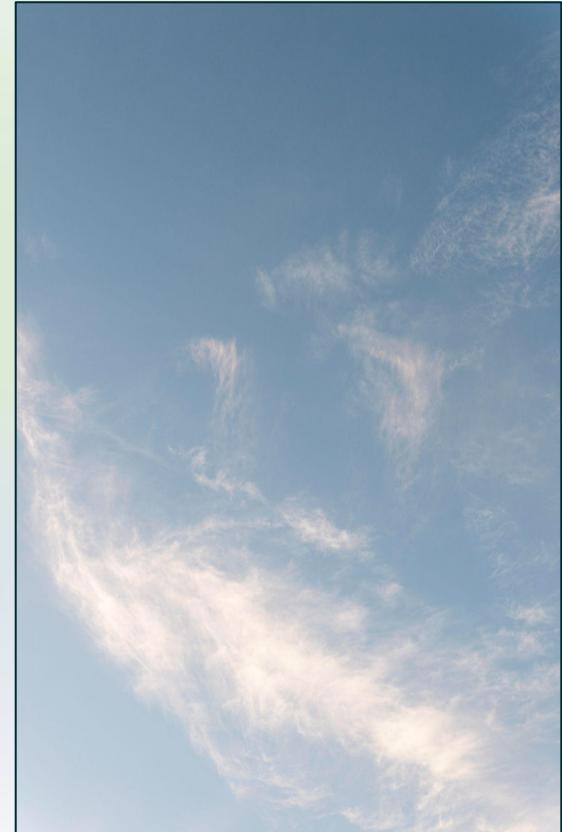
Temporal Analysis of 90th Percentile AQI: Mecklenburg & Wake

Tracking Air Quality Trends Over Time



The ML Model

4



Model Overview: Linear Regression

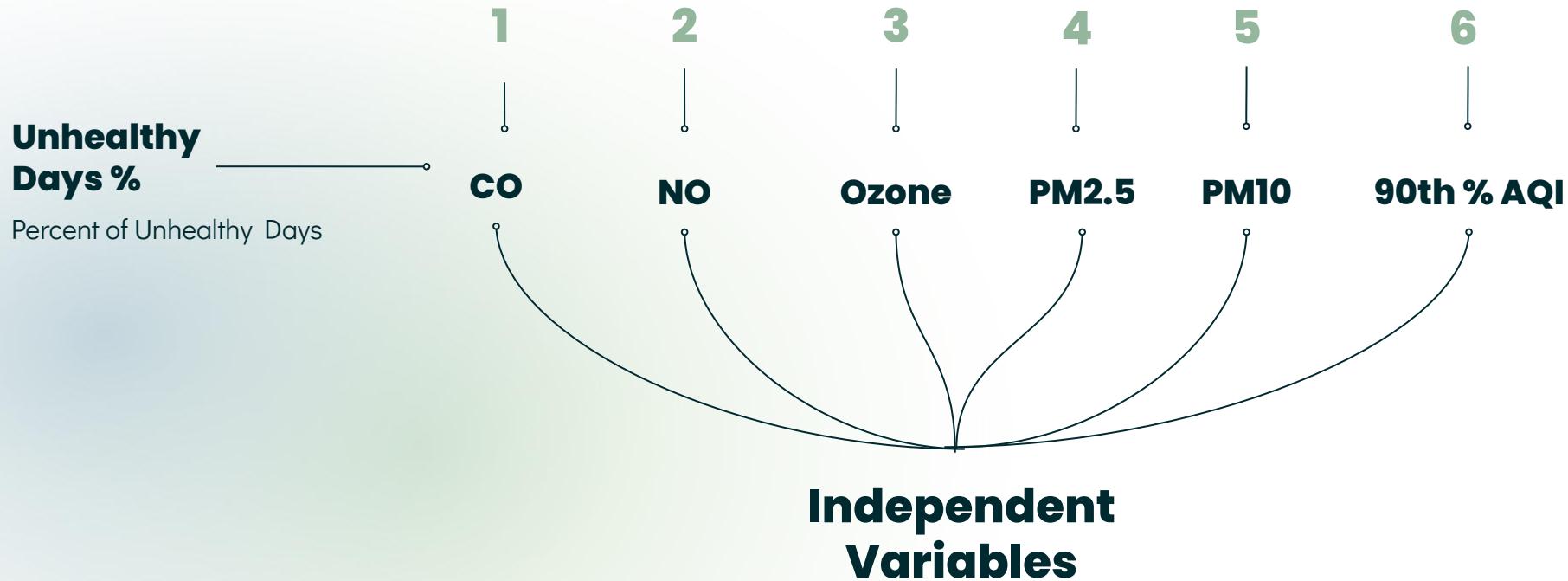
Target Variable - Percent of Unhealthy Days in 2023
for Charlotte and Raleigh

| | | Days CO | Days NO2 | Days Ozone | Days PM2.5 | Days PM10 | 90th Percentile AQI | Predicted_Unhealthy_Days_Percentage |
|-------------|------|---------|----------|------------|------------|-----------|---------------------|-------------------------------------|
| County | Year | | | | | | | |
| Mecklenburg | 2023 | 0 | 1 | 218 | 108 | 0 | 71.0 | 3.021587 |
| Wake | 2023 | 0 | 2 | 175 | 148 | 0 | 63.0 | 0.926351 |

1. Ability to decipher complex patterns
2. Identifies relationships within large datasets



Features and Training



Model Results

87%

R- Squared Value

| County | Year | Days CO | Days NO2 | Days Ozone | Days PM2.5 | Days PM10 | 90th Percentile AQI | Predicted_Unhealthy_Days_Percentage |
|-------------|------|---------|----------|------------|------------|-----------|---------------------|-------------------------------------|
| Mecklenburg | 2023 | 0 | 1 | 218 | 108 | 0 | 71.0 | 3.021587 |
| Wake | 2023 | 0 | 2 | 175 | 148 | 0 | 63.0 | 0.926351 |



Conclusions

5



Final Conclusion



Final Analysis: Our air is getting healthier!

Predicted Unhealthy/ Healthy Days Result:

- 97% Healthy and 3% Unhealthy for Charlotte
- 99% Healthy and 1% Unhealthy for Raleigh

Reasons as to Why?

- Traffic and Transportation
- Population Density



THANKS

DO YOU HAVE ANY QUESTIONS?



CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#) and infographics & images by [Freepik](#)



RESOURCES

All information was gathered from the United States Environmental Protection Agency. Here are a few pages that were the most useful:

- [Air Data Basic Information](#)
- [About Air Data Reports](#)
- [Download Daily Data](#)
- [U.S. Census Bureau](#)