



# COVID-19 Impacts on Air Quality

Predicting NO<sub>2</sub> Air Concentrations using Meteorological Data

Adam Seybert

LTC, US Army Student Detachment

# Introduction and Background

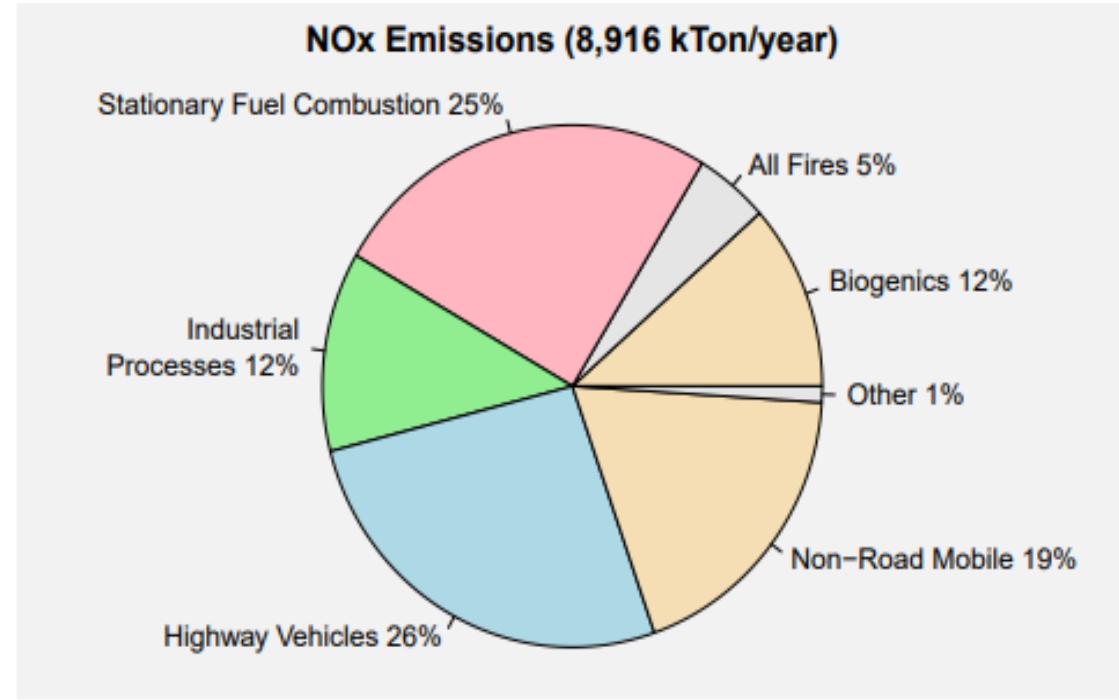
- COVID-19 led to widespread lockdowns across the U.S.
  - Many workers transitioned to remote work, reducing daily commuting.
  - This shift significantly impacted transportation-related emissions and air quality, especially in major cities.
- Over 200 studies have analyzed pandemic-related air quality changes using key pollutants ( $\text{NO}_2$ ,  $\text{O}_3$ , CO, PM2.5).
  - Initial studies often relied on preliminary data; more robust data is now available.
- The U.S. EPA Air Data repositories offer high-quality, comprehensive pollutant data.

# Purpose

- Investigate the impact of COVID-19 lockdowns on air quality in major U.S. metropolitan areas.
- Analyze NO<sub>2</sub> daily summary data from EPA in-situ sensors.
- Study five cities:
  - Chicago, IL
  - Los Angeles, CA
  - New York City, NY
  - Denver, CO
  - Washington, DC
- Incorporate meteorological data to account for weather-related influences on air quality.
- Use EPA data on hourly wind, temperature, barometric pressure, relative humidity, and dew point.

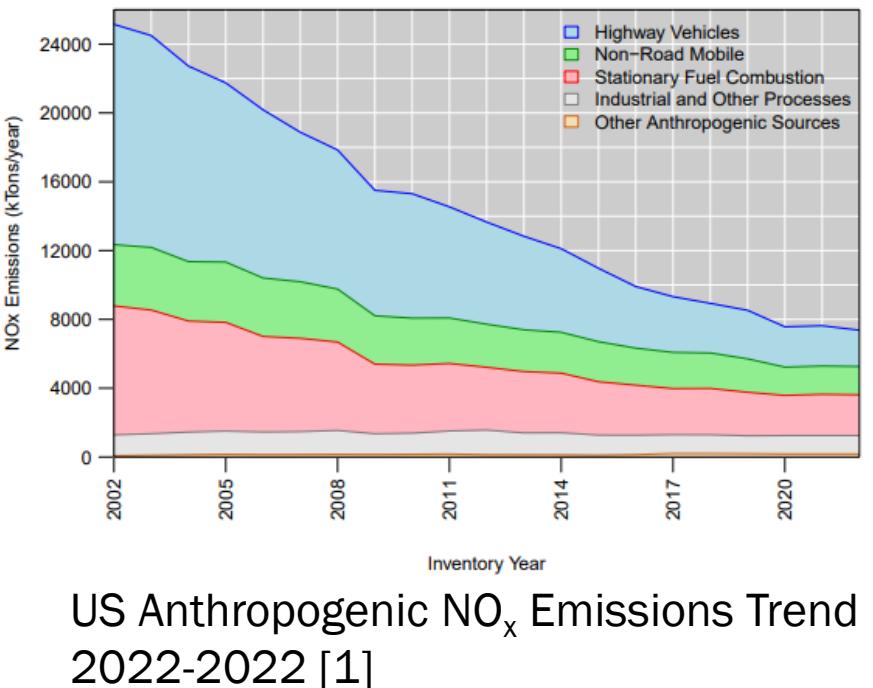
# Research Questions

- What factors impact NO<sub>2</sub> air concentration as measured from EPA ground-based sensors?
  - How well do these factors predict NO<sub>2</sub> air concentrations?
- Did COVID-19 lockdowns impact NO<sub>2</sub> air concentration?
  - To what extent did the lockdowns impact NO<sub>2</sub> air concentration



NO<sub>x</sub> Emissions by sector [1]

# Background (NO<sub>2</sub> Trends)

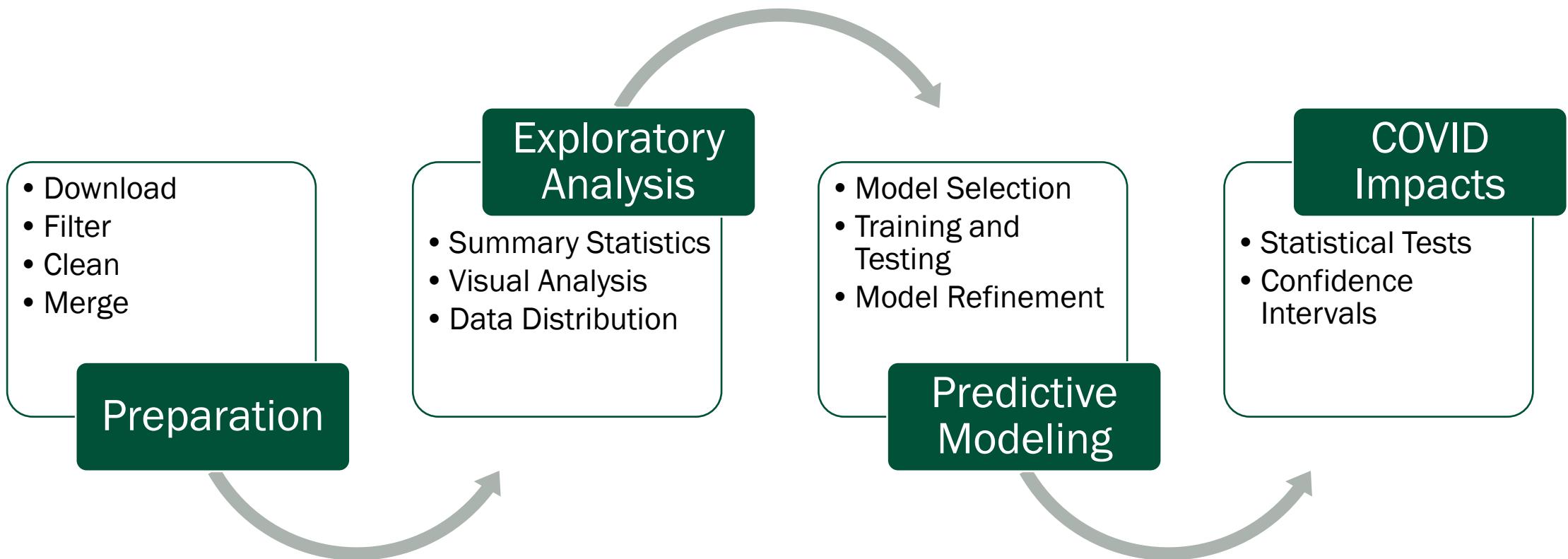


- EPA NO<sub>2</sub> Overview [1]
  - NO<sub>2</sub> concentrations are generally higher during fall and winter months, and lower during spring and summer.
  - Most monitoring sites show decreasing trends in NO<sub>2</sub> concentrations.
  - Annual mean and 98th percentile daily maximum 1-hour NO<sub>2</sub> concentrations have remained relatively constant over the past decade, well below the National Ambient Air Quality Standards (NAAQS)

# Data Sources

- US EPA Monitoring Network for Nitrogen Dioxide [1] [2]
  - 481 monitoring sites reported hourly NO<sub>2</sub> data to the EPA.
  - Monitoring networks include:
    - State and Local Air Monitoring Stations (SLAMS): Over 80% of sites
    - National Core (NCore) multi-pollutant monitoring network
    - Photochemical Assessment Monitoring Stations (PAMS)
    - Near-road monitoring network: 71 monitors
  - Hourly data is aggregated to daily and yearly data sets
- EPA Hourly Meteorological Data [2]
  - Hourly wind, temperature, barometric pressor, relative humidity, and dew-point measurements.

# Computational Workflow

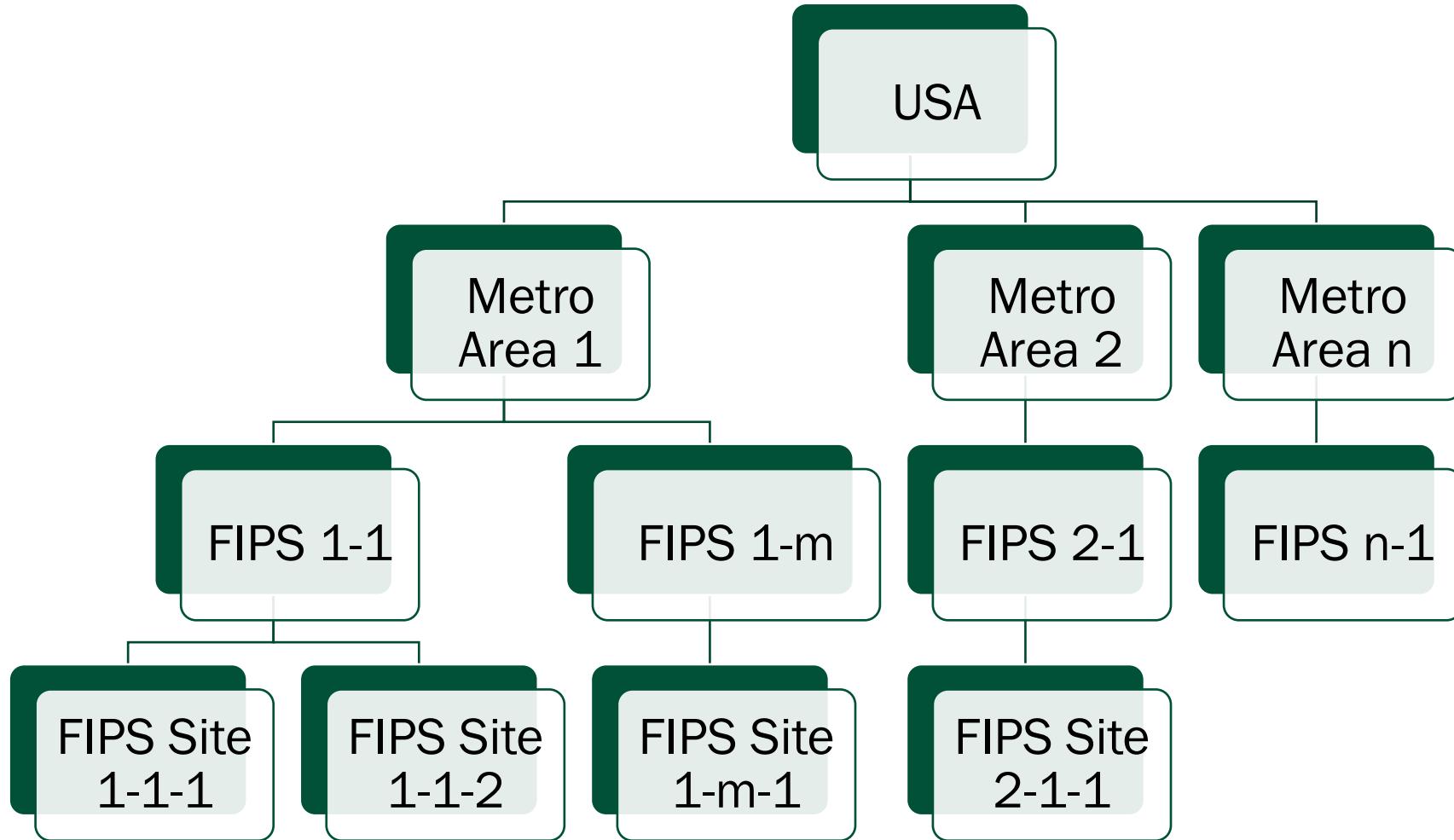


# Data Preparation (Metro Statistical Area Data)

- Defining a region of interest
  - Metropolitan Statistical Area (MSA) Geography Reference File [3]
    - MSA Area
    - County Name
    - Federal Information Processing Standards (FIPS) codes
      - State and County (example: Los Angeles is 06037 and 06059)

Metro Areas	MSA name	FIPS Codes
Chicago	Chicago-Naperville-Elgin, IL-IN-WI Metro Area	14
Denver	Denver-Aurora-Lakewood, CO Metro Area	10
Los Angeles	Los Angeles-Long Beach-Anaheim, CA Metro Area	2
New York	New York-Newark-Jersey City, NY-NJ-PA Metro Area	25
Washington DC	Washington-Arlington-Alexandria, DC-VA-MD-WV Metro Area	24
	Total	75

# Data Preparation (Geographic Level of Detail)



# Data Preparation (NO<sub>2</sub> Download)

- NO<sub>2</sub> Data Daily [2]
  - Average NO<sub>2</sub> measurement (ppb)
  - Max NO<sub>2</sub> measurement (ppb)
  - Hour of max measurement
  - Indexed by FIPS (state and county), and Site Number (non-fips)
  - Site details (name, address, latitude, longitude)
- Reducing to Area of Interest
  - Filter by FIPS codes
  - Index by FIPS codes and site number (example: 06037\_1103)
- Clean
  - 2 FIPS-Sites have multiple same-day measurements
    - ~1500 observations (mean difference of about 1.2ppb (15%))
  - Outliers
    - ~1,000 observations less than 1ppb
- Downloaded Data Summary
  - 24 FIPS Codes (54 FIPS-Sites)
  - 5 Years of daily data (1826 days)
  - 86,165 Total Observations
    - Of a possible 98,604 (~90%)

# Data Preparation (NO<sub>2</sub> Download)

---

Year	Chicago	Denver	Los Angeles	New York	Washington
2017	1,802	1,904	6,054	3,780	3,275
2018	1,309	1,739	5,947	3,799	3,246
2019	1,818	2,092	6,155	4,120	3,232
2020	2,196	2,098	6,483	3,880	3,283
2021	2,295	2,036	6,390	3,967	3,265

---

# Data Preparation (Weather Download)

- Weather Data Hourly [2]
  - WIND: Wind direction and speed
  - TEMP: Temperature
  - PRESS: Pressure
  - RH\_DP: Humidity and Dew Point
- Reducing to Area of Interest
  - Filter by FIPS codes
  - Index by FIPS codes and site number  
(example: 06037\_1103)
- Clean
  - Outliers
    - 103 observations with wind speeds greater than 47 knots (Gale force winds @ 4 FIPS-Sites)
  - Summarized to Daily
- Downloaded Data Summary
  - 31 FIPS Codes (70 FIPS-Sites)
  - 5 Years of hourly data (1826 days)
  - 95,069 Total Observations
    - Of a possible 127,838 (~75%)

# Data Preparation (Weather Download)

- Weather data is not \*complete
  - 1667 observations with full data
- Dropping barometric pressure and dew point increases observations
  - ~47,00 (still less than half)
- Dew point only measured in Chicago

Measure	Missing	Unique	Observed
bar_pressure	50,013	9,637	45,056
dew_point	91,818	1,789	3,251
out_temperature	10,046	13,159	85,023
rel_humidity	37,813	8,549	57,256
wind_direction	20,846	24,414	74,223
wind_speed	20,402	6,521	74,667

# Exploratory Analysis Questions

## Factor Analysis

- What weather factors are important?
- Are there time-series trends to account for?
- At what geographics scale should we model?

## Model Selection

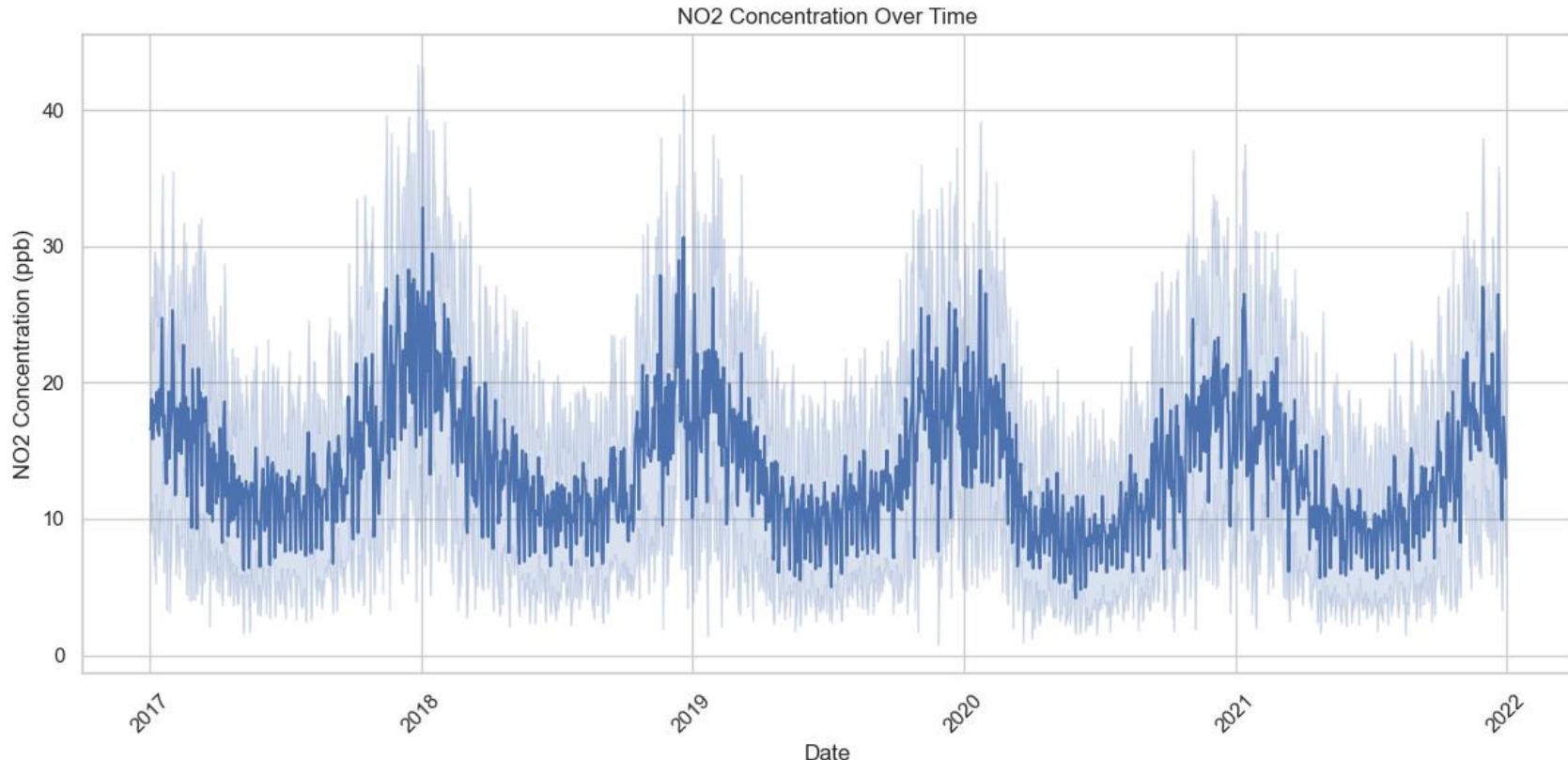
- How is the data distributed?
- What models are reasonable based on the distributions?

# Exploratory Data Analysis (NO<sub>2</sub> Data Summary)

NO<sub>2</sub> average concentrations vary across metro areas  
NO<sub>2</sub> average concentrations vary from year to year

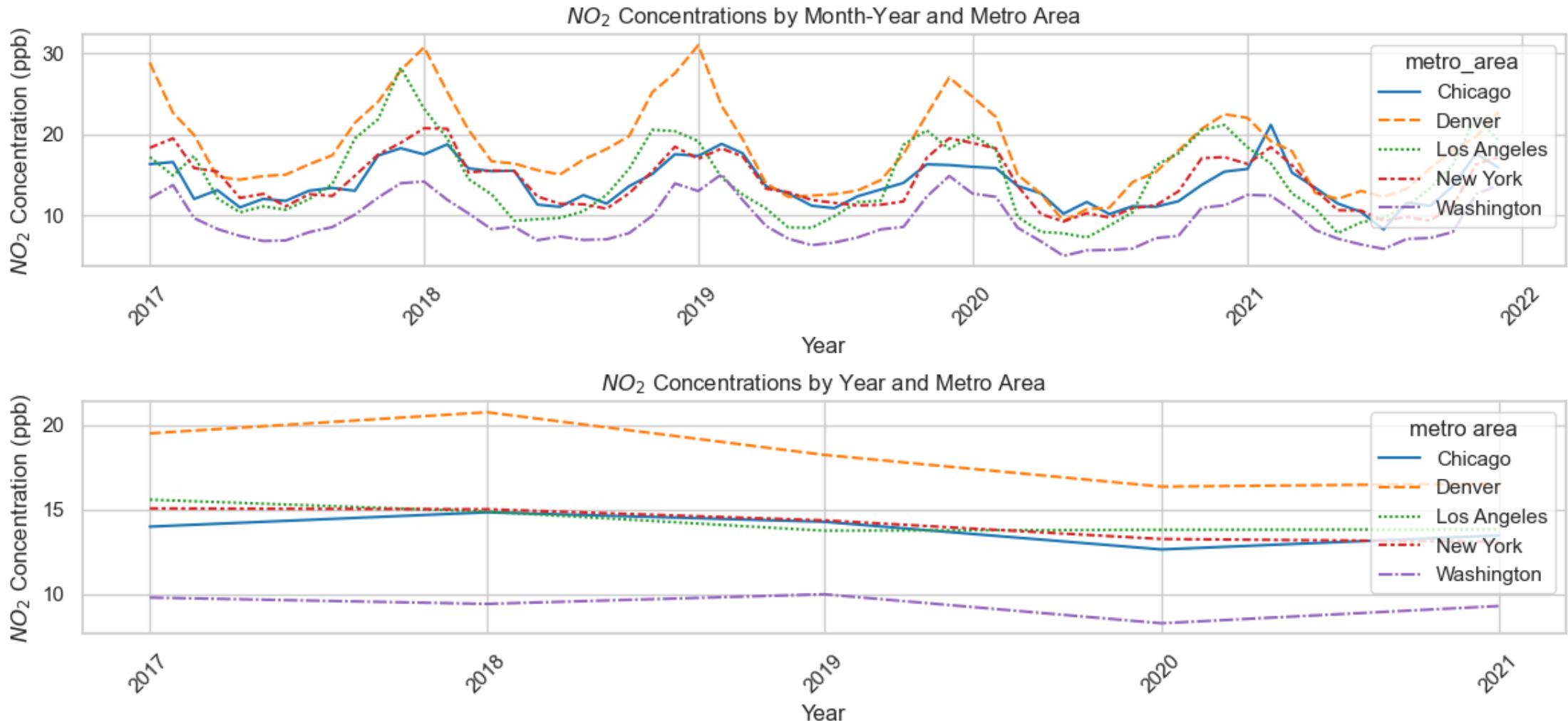
count	86,165							
mean	13.74	Year	Chicago	Denver	Los Angeles	New York	Wash DC.	AVG
std	8.79	2017	14.00	19.51	15.59	15.07	9.81	14.62
min	1.00	2018	14.84	20.75	14.89	15.02	9.44	14.45
25%	7.00	2019	14.28	18.23	13.76	14.37	10.00	13.81
50%	12.03	2020	12.65	16.36	13.82	13.27	8.29	12.81
75%	18.72	2021	13.47	16.54	13.84	13.11	9.31	13.11
		AVG	13.73	18.18	14.36	14.16	9.37	13.74
max	70.62							

# Exploratory Analysis (NO<sub>2</sub> Concentration Over Time)

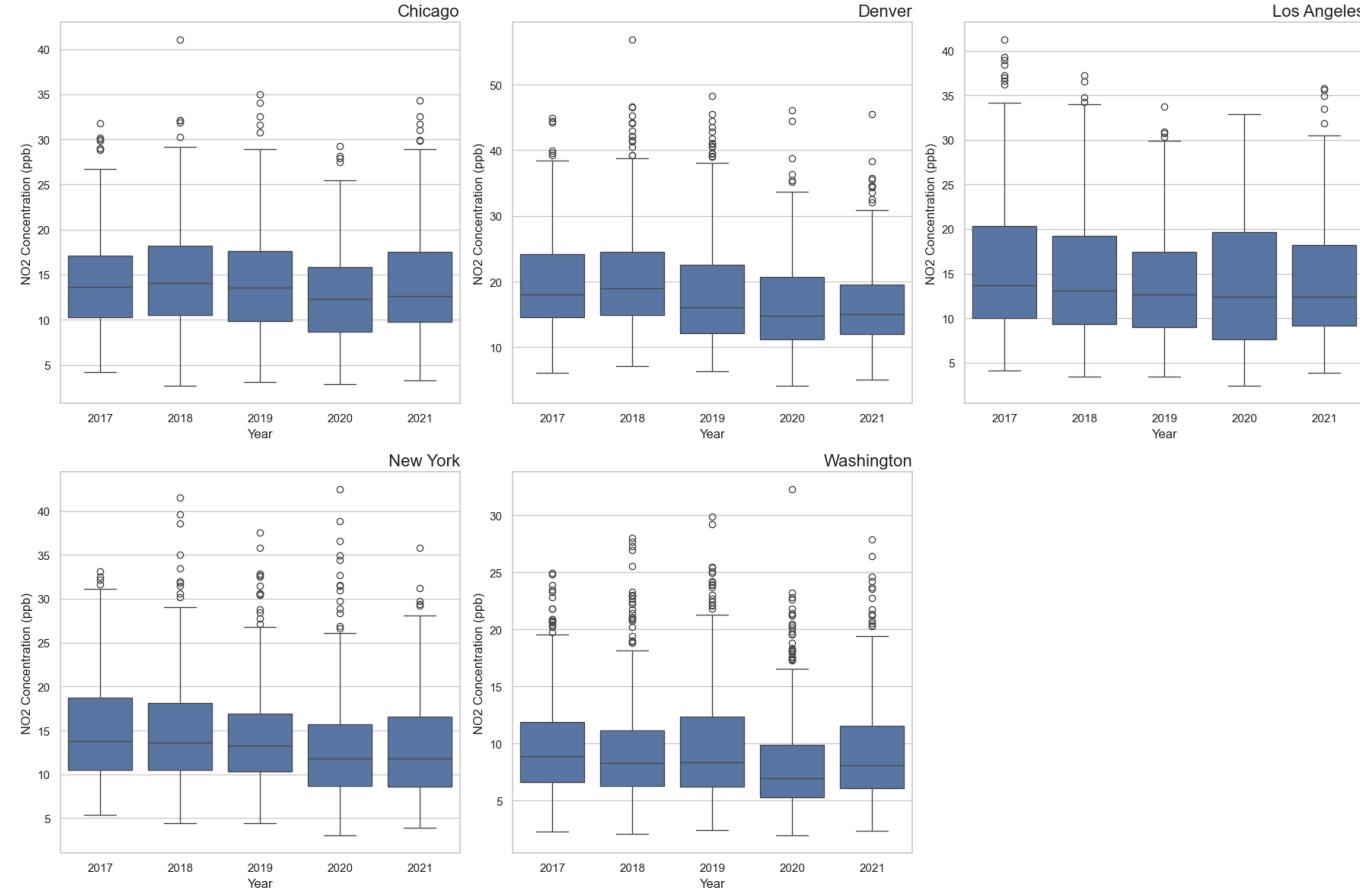


NO<sub>2</sub> average concentrations have seasonality

# Exploratory Analysis (NO<sub>2</sub> Concentration Over Time)



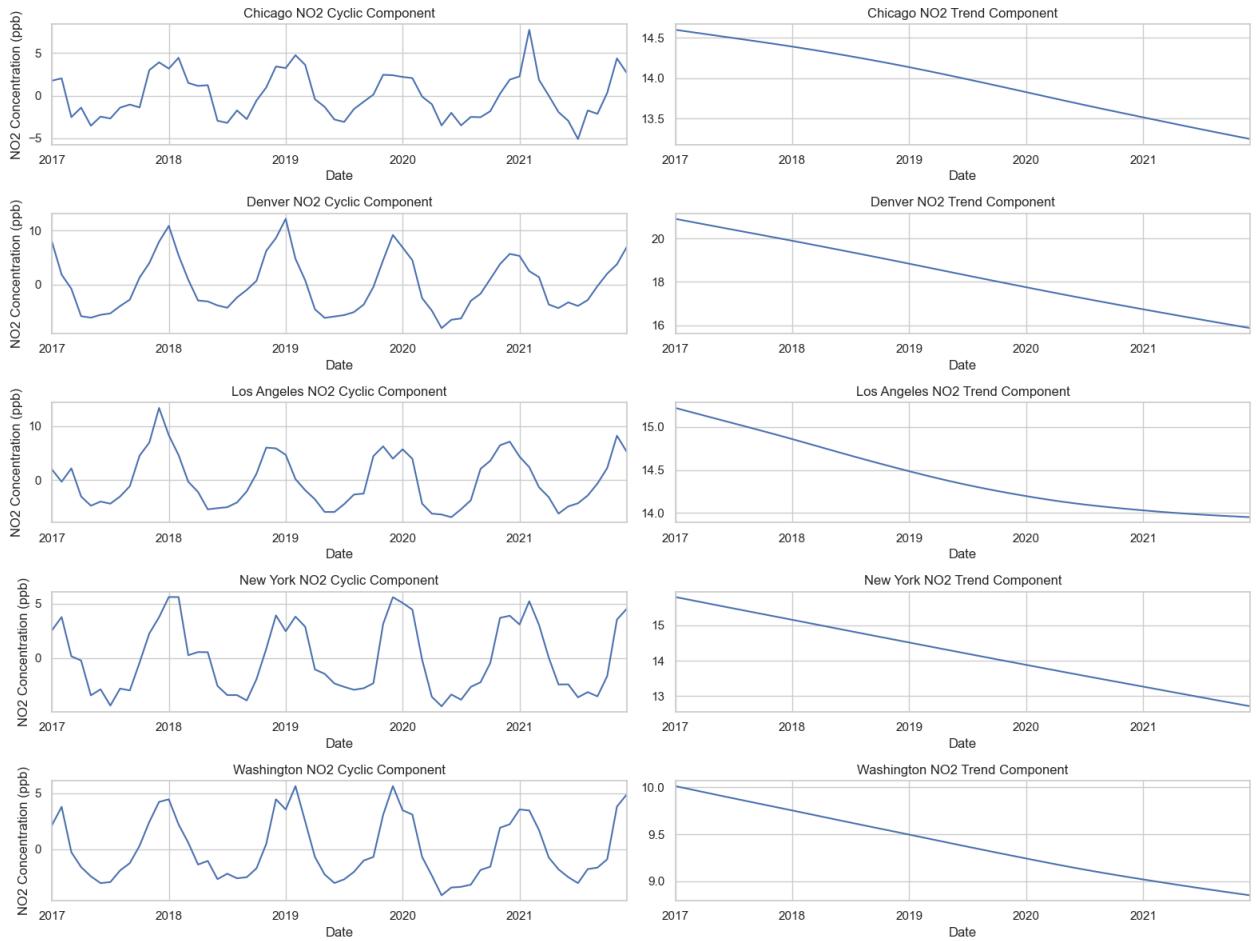
# Exploratory Analysis (NO<sub>2</sub> Concentration Over Time)



Year is likely an important factor

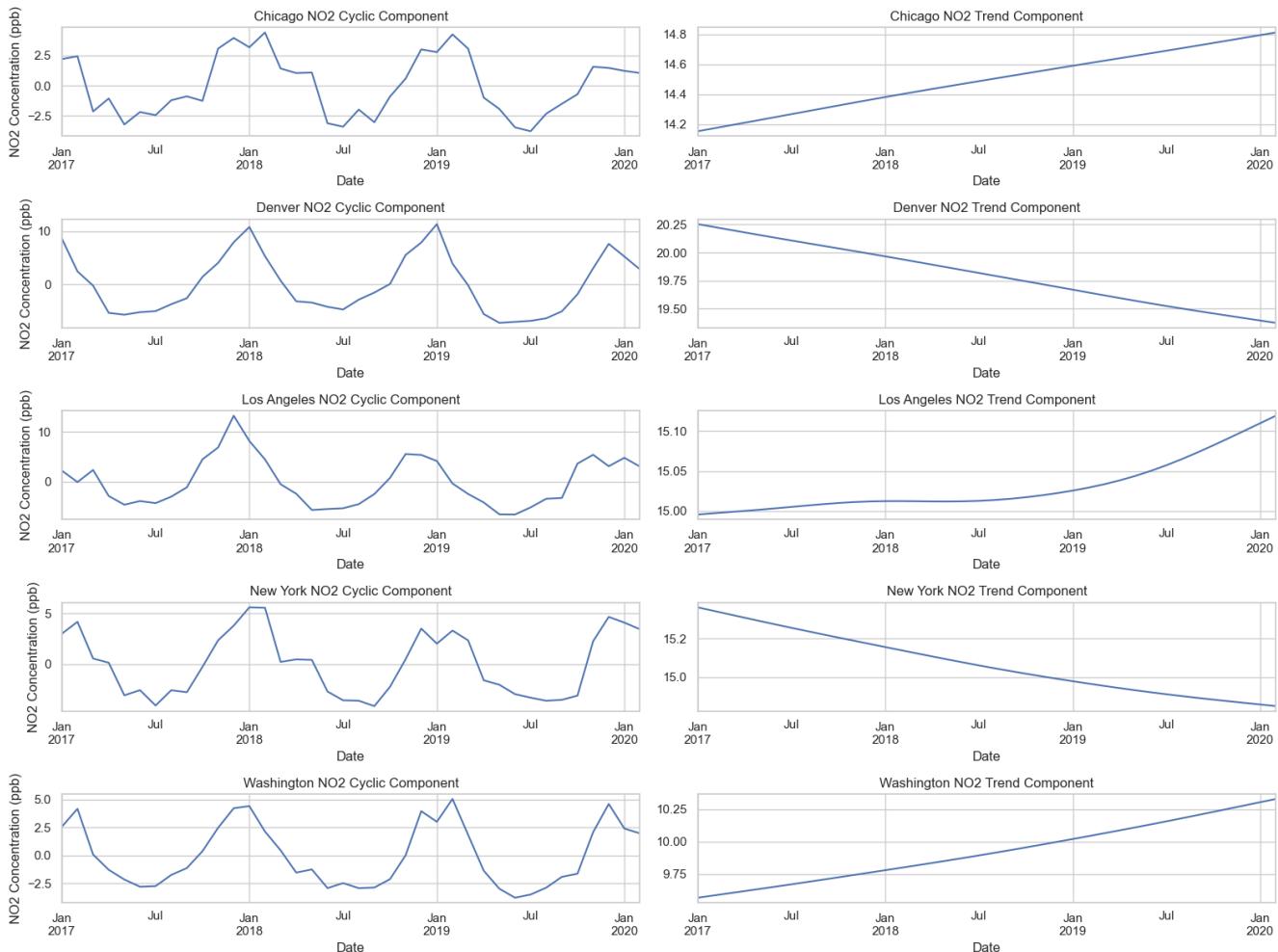
# Exploratory Analysis (NO<sub>2</sub> Concentration Over Time)

- Hodrick-Prescott filter
  - $\lambda = 129600$
- NO<sub>2</sub> concentration has a cyclic component (yearly period)
- NO<sub>2</sub> concentration has a trend component (decreasing)

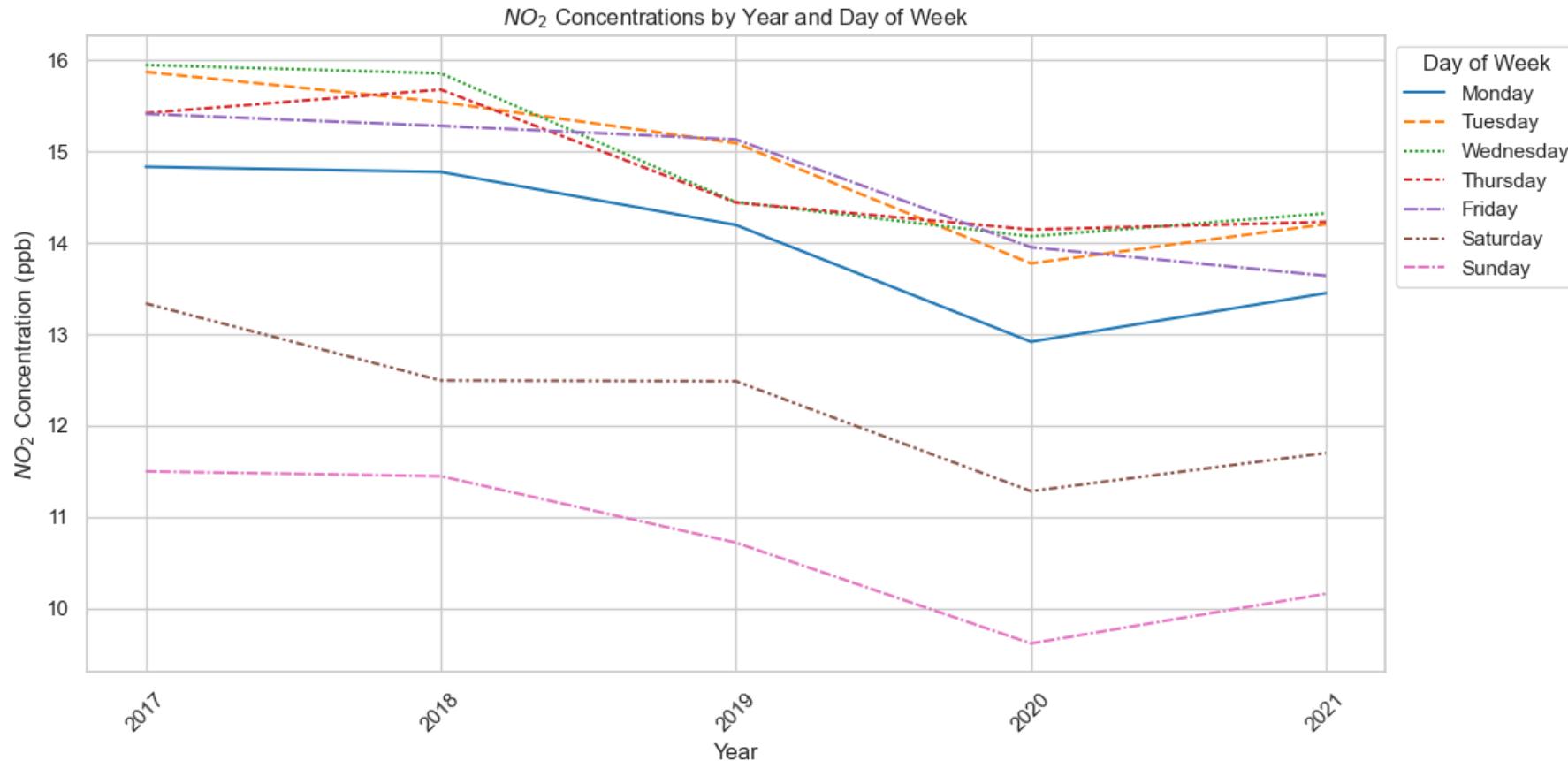


# Exploratory Analysis (NO<sub>2</sub> Concentration Over Time)

- HP Filter is sensitive to shocks. One-time shifts can cause trend shifts that do not occur
- Remove post COVID data results in same cyclical pattern but different trends

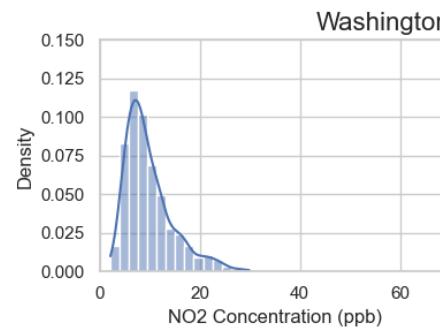
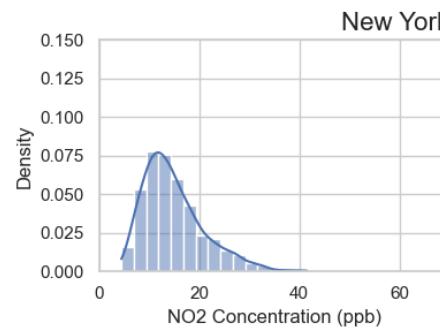
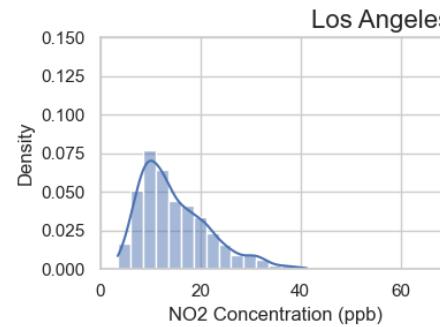
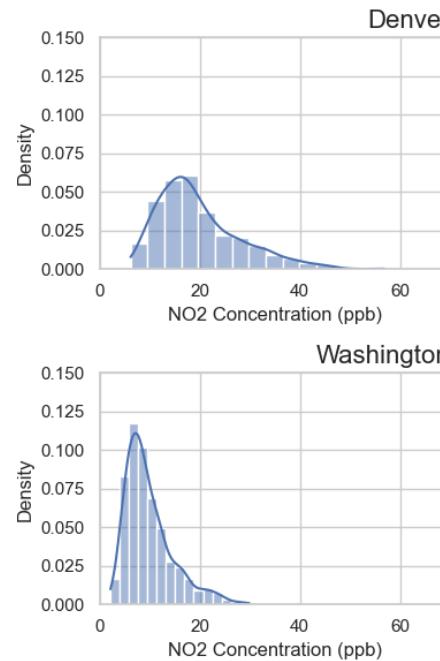
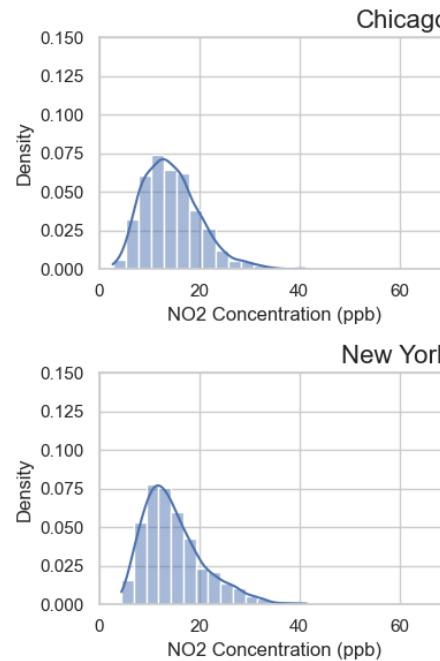
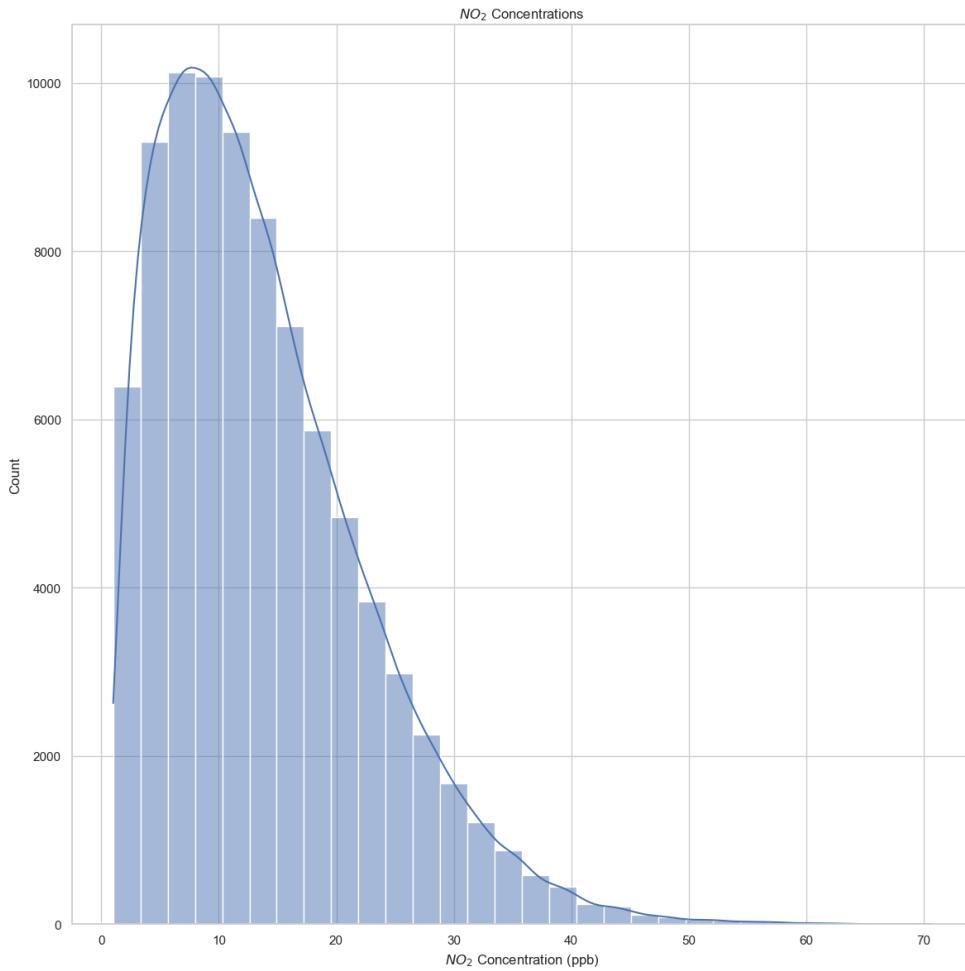


# Exploratory Analysis (NO<sub>2</sub> Concentration by Day of Week)



Day of the weeks is likely an important factor

# Exploratory Analysis (NO<sub>2</sub> Concentration Distribution)

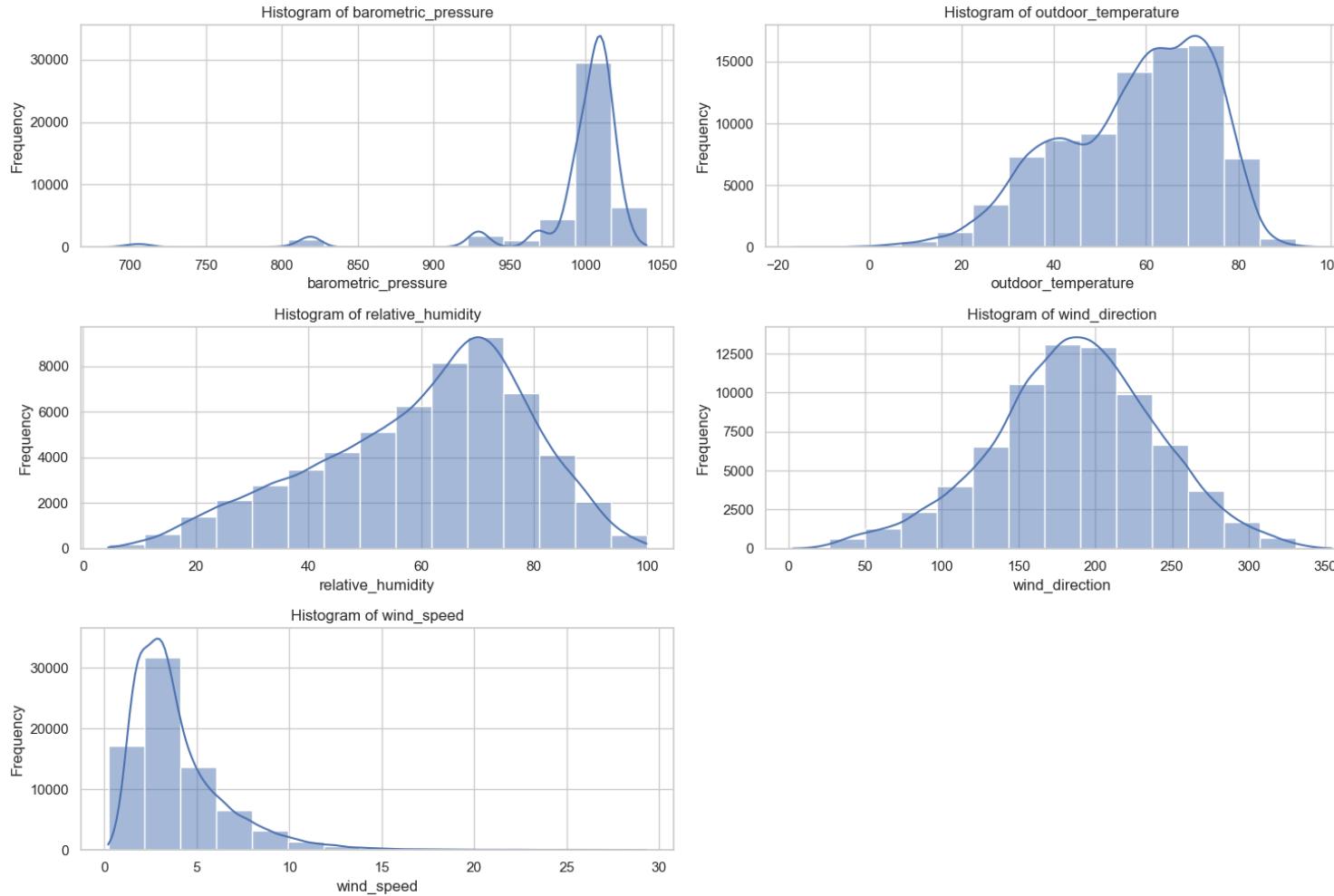


Data appears log-normal distributed

# Exploratory Analysis (Weather Summary Data)

Metro Area	Barometric Pressure (mb)		Dew Point (F)		Outdoor Temperature (F)		Relative Humidity (%)		Wind Direction (card degree)		Wind Speed (knots)	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Chicago	993	6.8	40.23	19.54	51.8	18.6	72.1	13.3	183.0	68.2	6.3	2.7
Denver	790	48.9			52.1	17.8	45.3	18.7	183.9	44.4	4.5	2.5
Los Angeles	992	25.6			65.1	8.8	59.9	18.8	186.8	47.0	2.7	1.1
New York	1,001	13.1			54.3	17.5	62.4	14.7	198.8	68.3	4.6	2.6
Washington	1,010	8.5			57.7	16.5	64.0	15.1	189.7	67.7	3.4	2.6

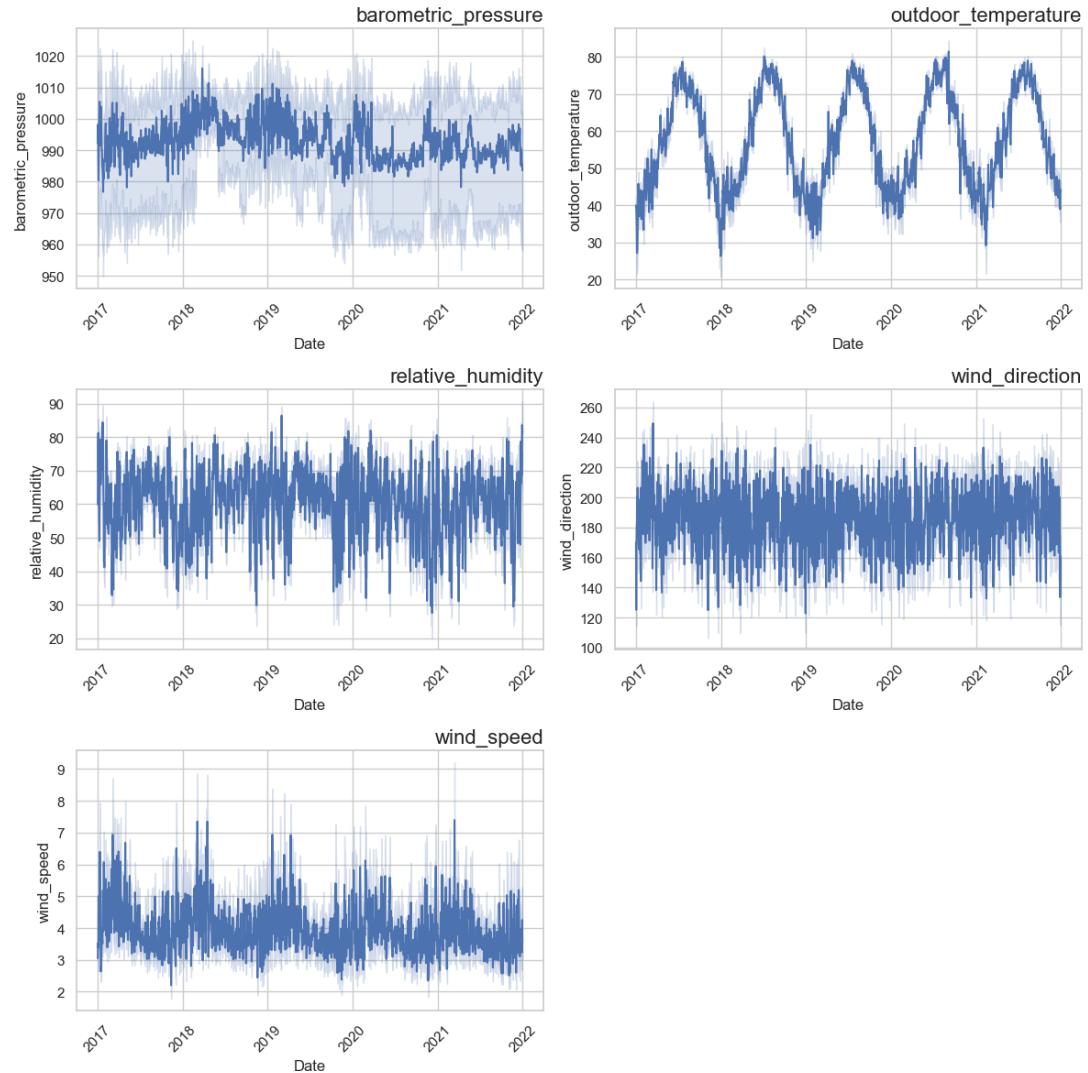
# Exploratory Analysis (Weather Distributions)



- Barometric pressure is dependent on metro area
- Wind direction is normal with average direction around  $180^\circ$
- Windspeed appears log-norm

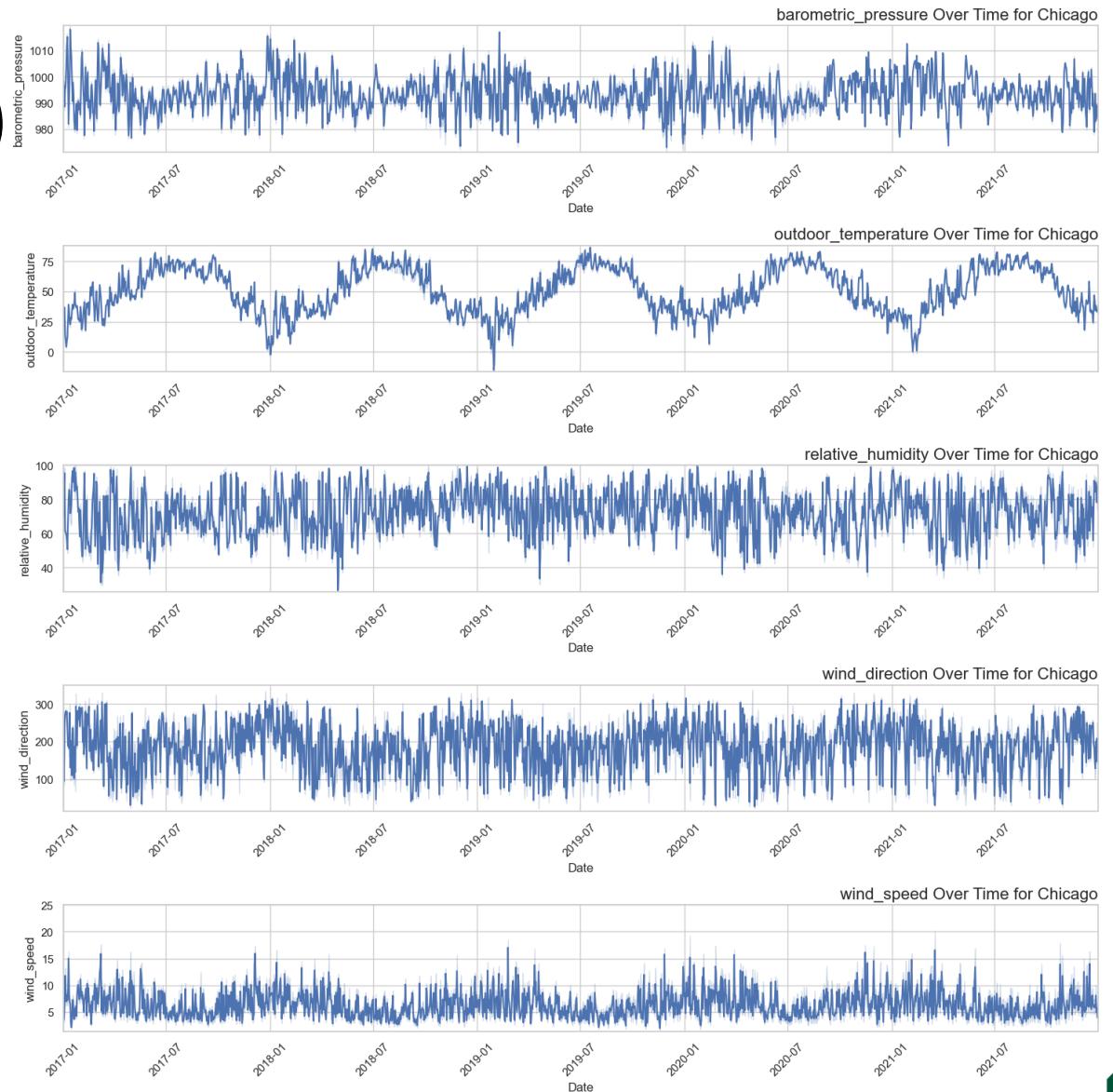
# Exploratory Analysis (Weather Time Series)

- Outdoor temperature shows cyclical component
- Relative humidity shows increased variability in the winter
- Windspeed shows increased variability in the winter

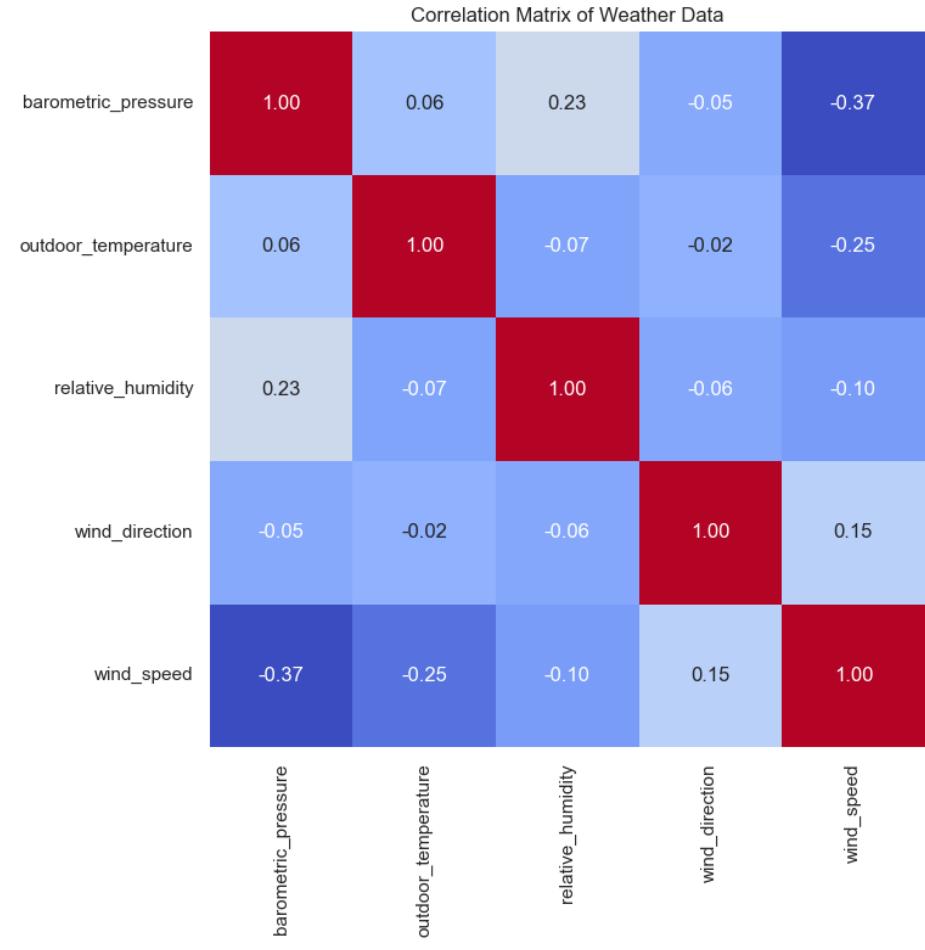
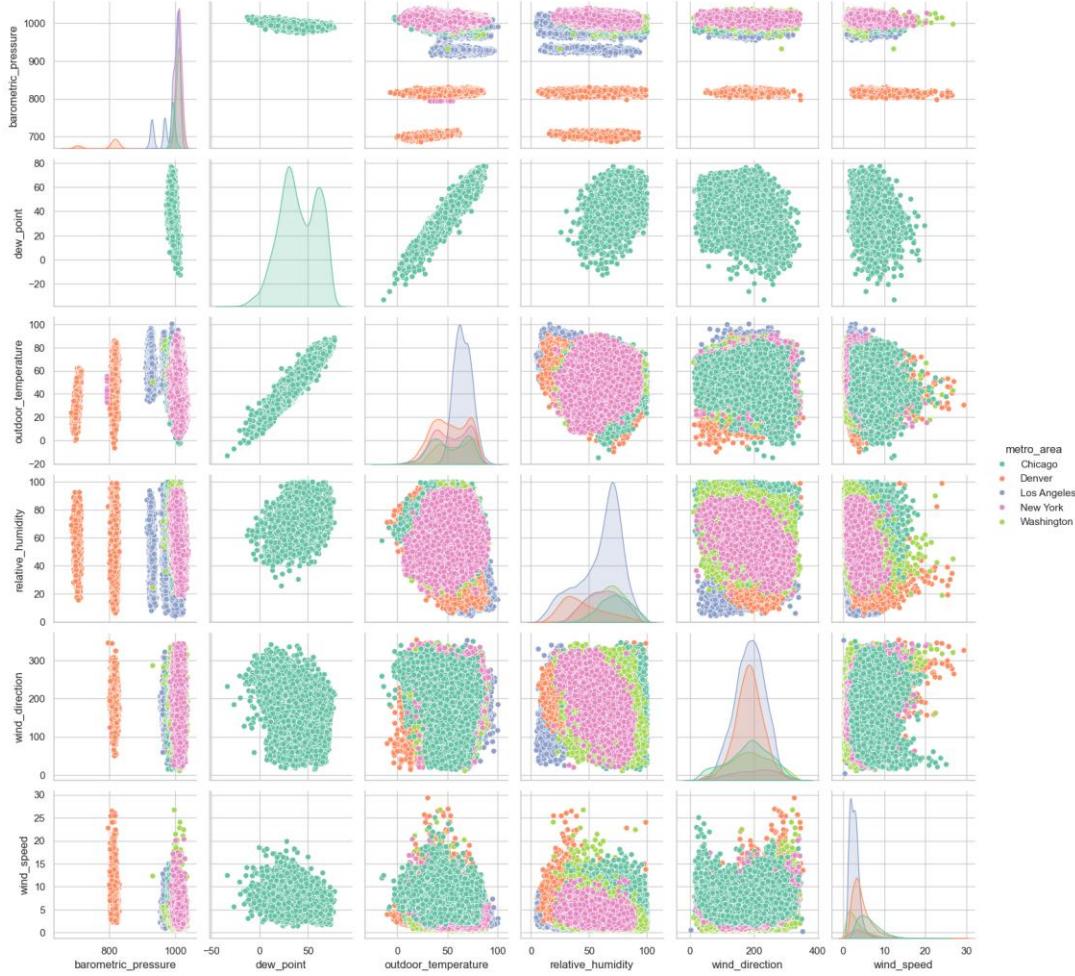


# Exploratory Analysis (Weather Time Series)

- By Region
- Outdoor temperature shows cyclical component
- Relative humidity shows increased variability in the winter
- Windspeed shows increased variability in the winter
- Plots for all metro areas available in [back-up slides](#).



# Exploratory Analysis (Weather Correlations)



# Exploratory Analysis (Conclusions)

## Factor Analysis

- What weather factors are important?
  - Wind speed, wind direction, outdoor temperature, and relative humidity should be considered
- Are there time-series trends to account for?
  - Weekly and Yearly cyclical trends
- At what geographics scale should we model?
  - Metro Area scale is appropriate but should include FIPS or FIPS-Site ([backup slides available](#))

# Exploratory Analysis (Conclusions)

## Model Selection

- How is the data distributed?
  - $\text{NO}_2$  is log-normal while weather data is a mix of normal and log-normal
- What models are reasonable based on the distributions?
  - Generalize linear models are appropriate for the data.
  - Other regression models should also be explored
  - Time-series models like ARIMA/SARIMA or rolling averages will likely have good prediction accuracy of short time scales because of their ability to “adjust” to shocks

# Predictive Modeling (Linear Regression)

- **Purpose:**

Predict a continuous outcome ( $\hat{y}$ ) based on one or more input features( $x_n$ ).

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

- **Training Objective:**

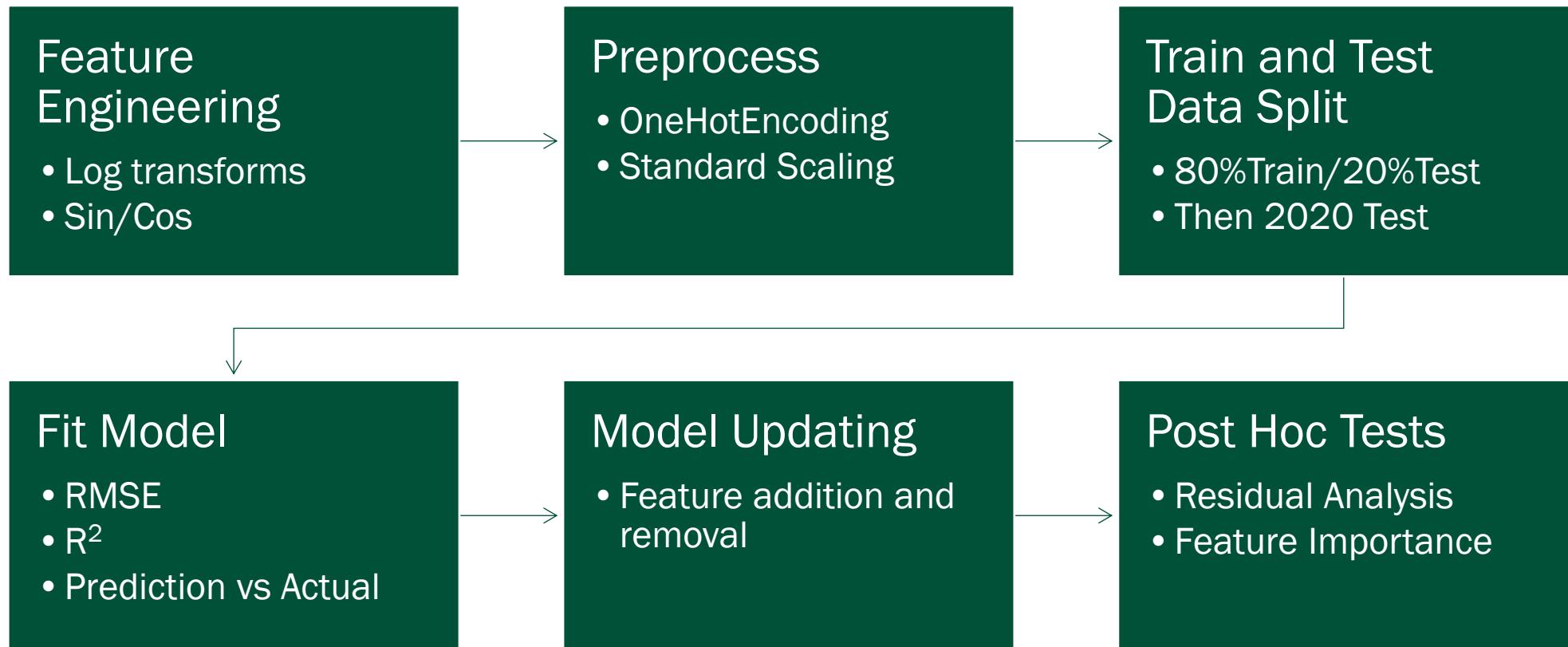
Minimize the **sum of squared errors (SSE)** between predicted and actual values:

$$SSE = \sum(y_i - \hat{y}_i)^2$$

- **Interpretation:**

Each coefficient  $\beta_i$  shows the expected change in  $y$  for a one-unit increase in  $x_i$ , holding others constant.

# Predictive Modeling (Method)



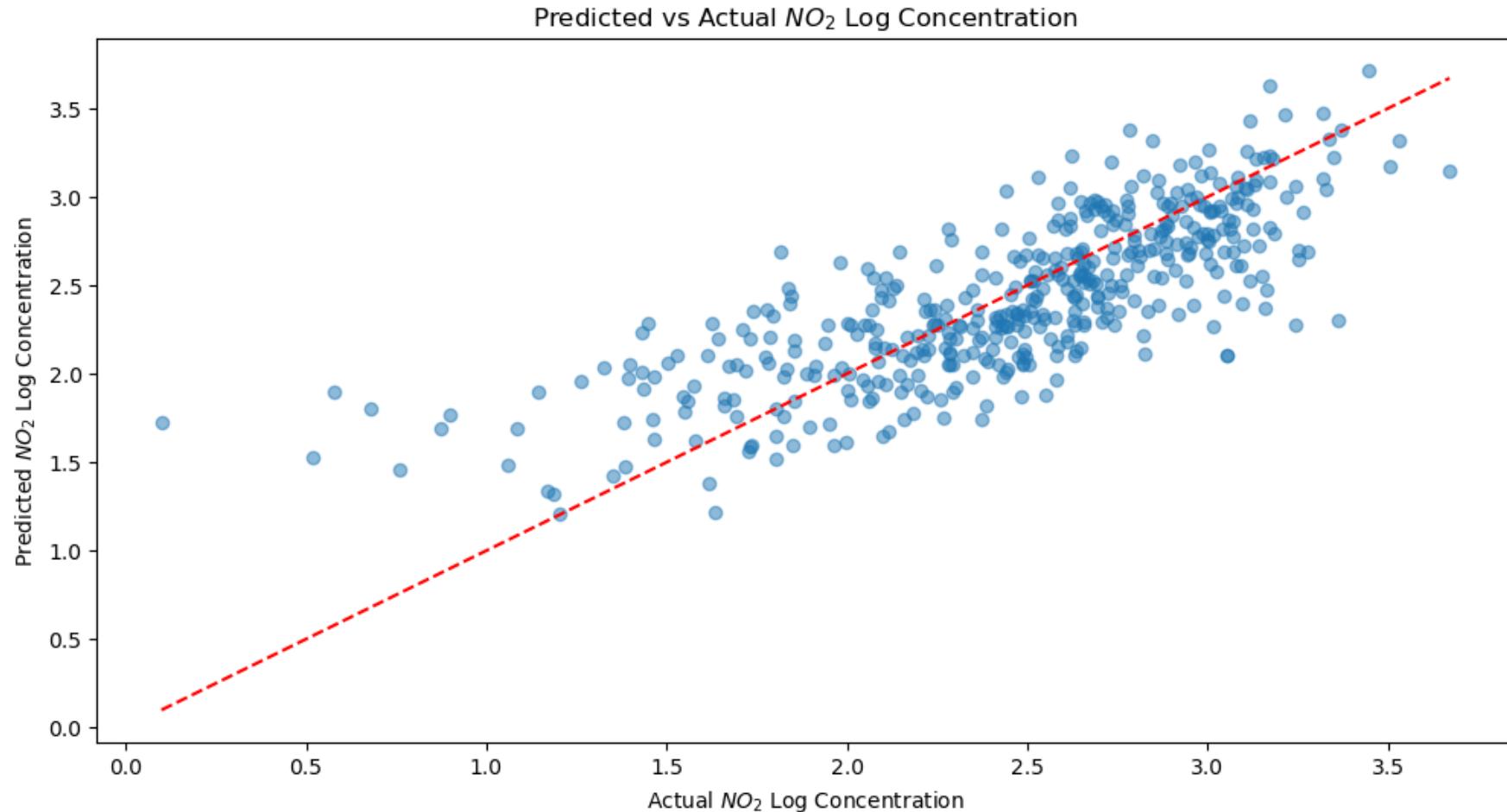
# Predictive Modeling (Factor Engineering)

- Time Series Factors
  - Yearly trends by metro area
  - Yearly cyclical components by metro area
    - $day_{yearSin} = \sin\left(\frac{2\pi day_{year}}{365}\right)$
    - $day_{yearCos} = \cos\left(\frac{2\pi day_{year}}{365}\right)$
  - Weekly component by metro area
    - Day of the week (oneHotEncoder)
- Weather Factors
  - Wind Speed
    - $wind_{speedLog} = \log(wind_{speed})$
- Location Factor
  - FIPS-Site (oneHotEncoder)

# Predictive Modeling (Model Fit – Chicago Example)

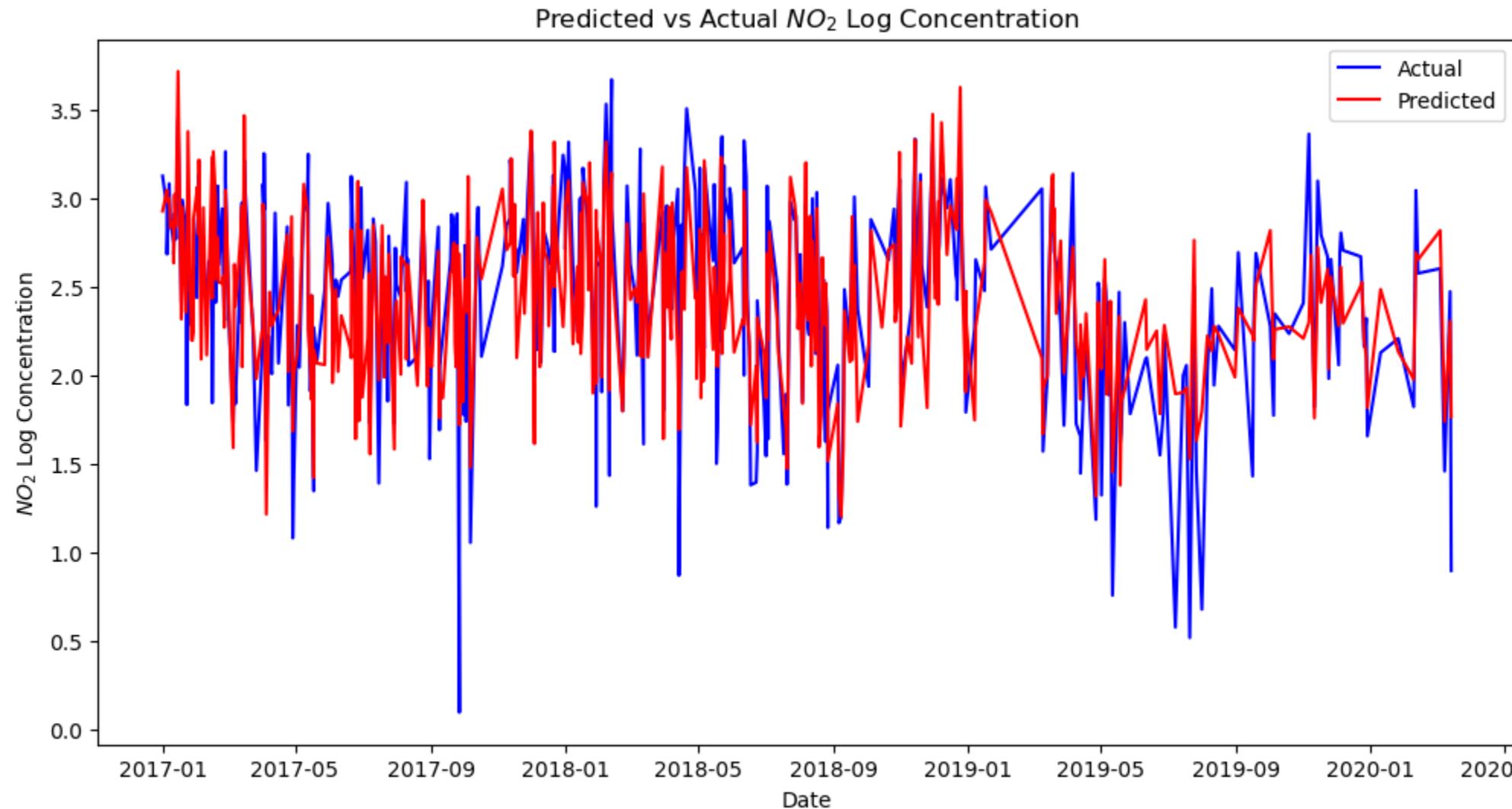
- Using only 2017 to Pre-COVID (14 March 2020)
- 80% train vs 20% test
- Log( $\text{NO}_2$ ) vs
  - fip\_site,
  - day\_of\_week, day\_of\_year\_sin, day\_of\_year\_cos,
  - outdoor\_temp, relative\_humidity, wind\_direction, wind\_speed\_log
- Results
  - Mean Squared Error: 0.123
  - Root Mean Squared Error: 0.350
  - R-squared: 0.593

# Predictive Modeling (Model Fit – Chicago Example)

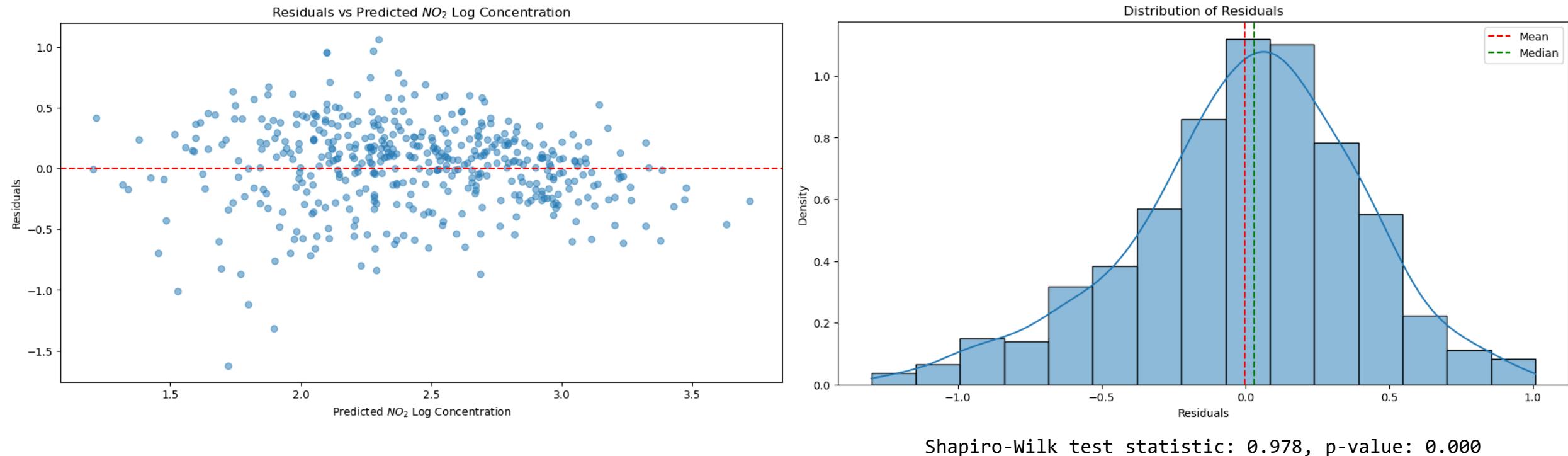


Low log( $NO_2$ ) concentrations are predicted to be higher

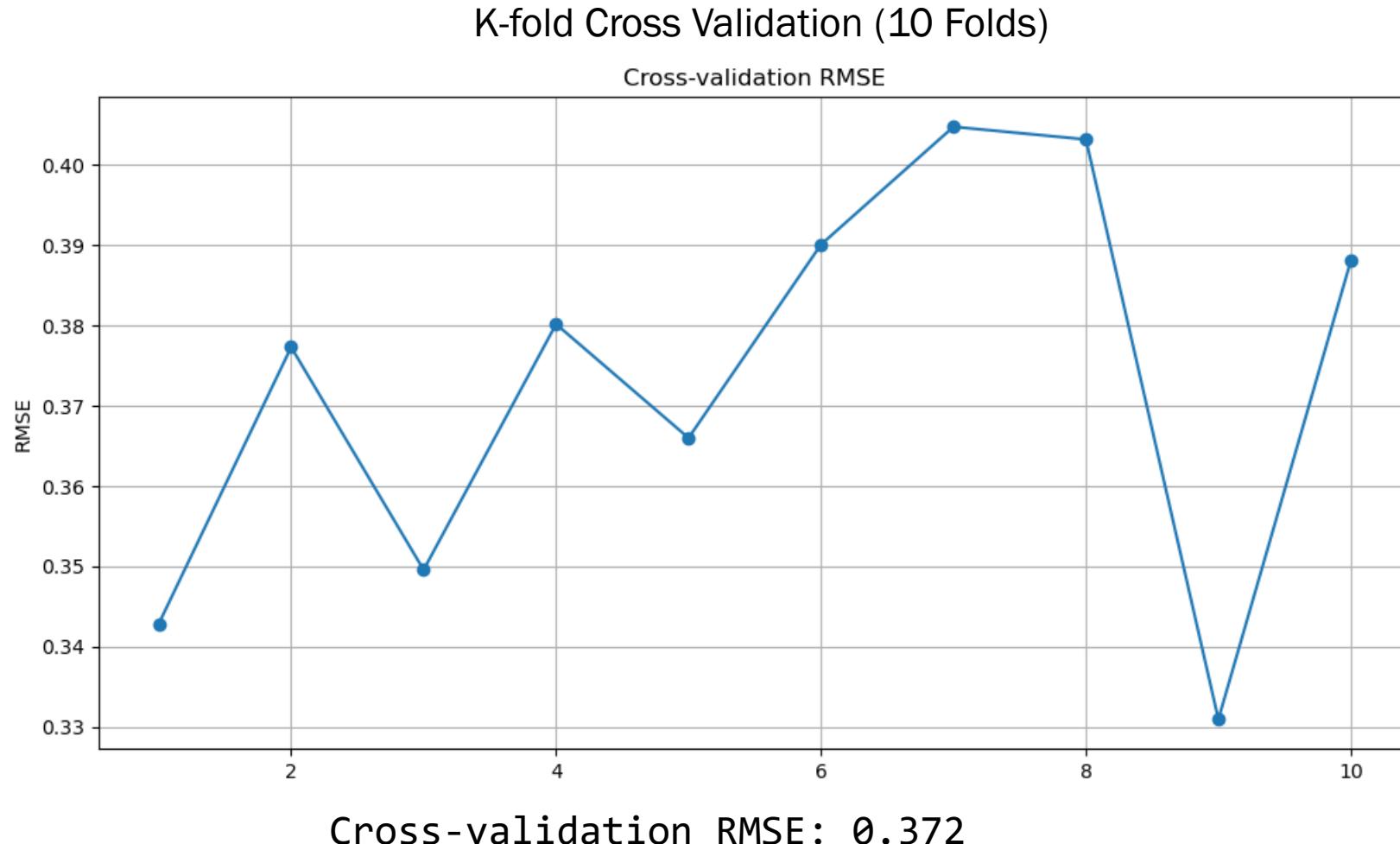
# Predictive Modeling (Model Fit – Chicago Example)



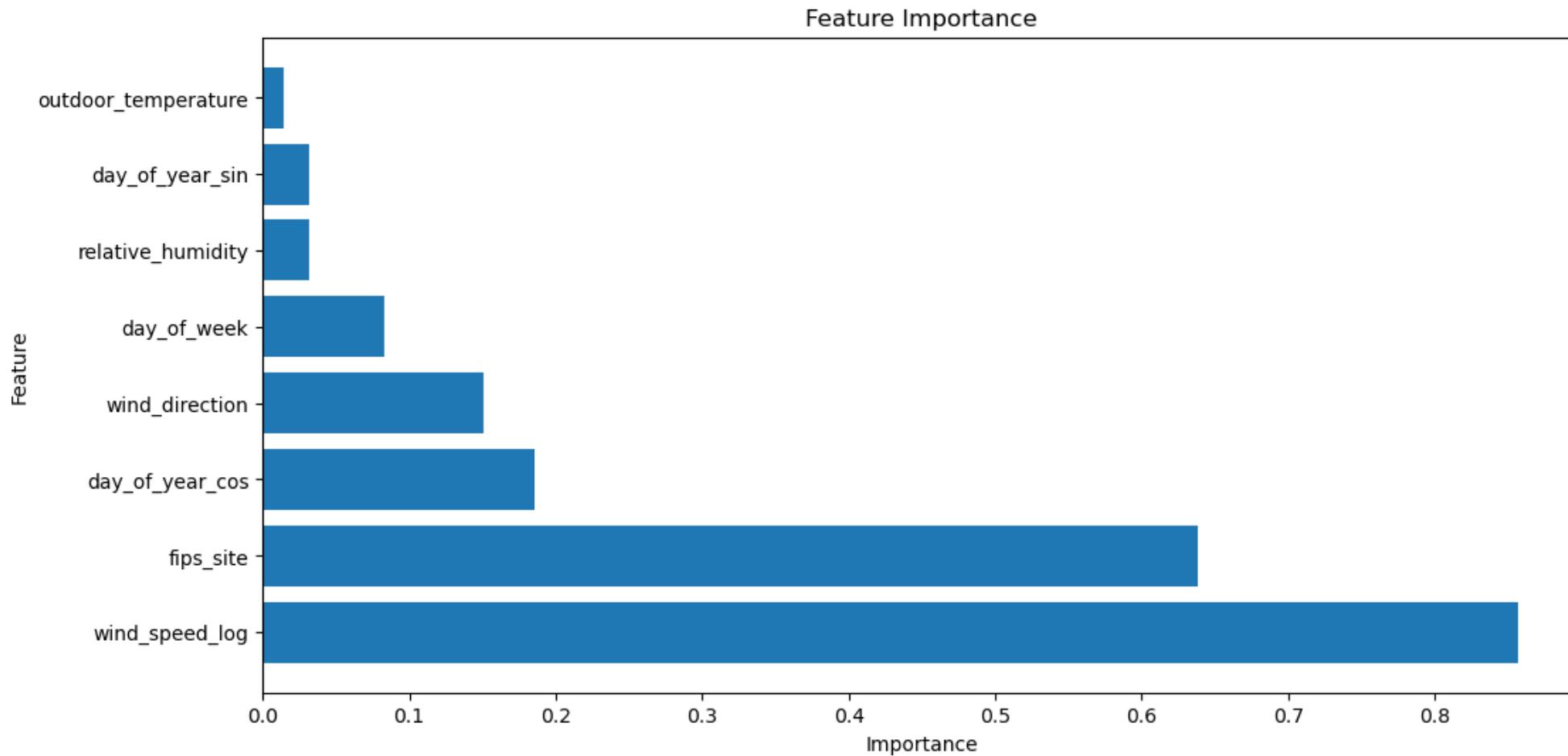
# Predictive Modeling (Residual Analysis – Chicago Example)



# Predictive Modeling (Validation – Chicago Example)



# Predictive Modeling (Feature Importance – Chicago Example)



# Predictive Modeling (All Metro Area Results Pre-COVID)

---

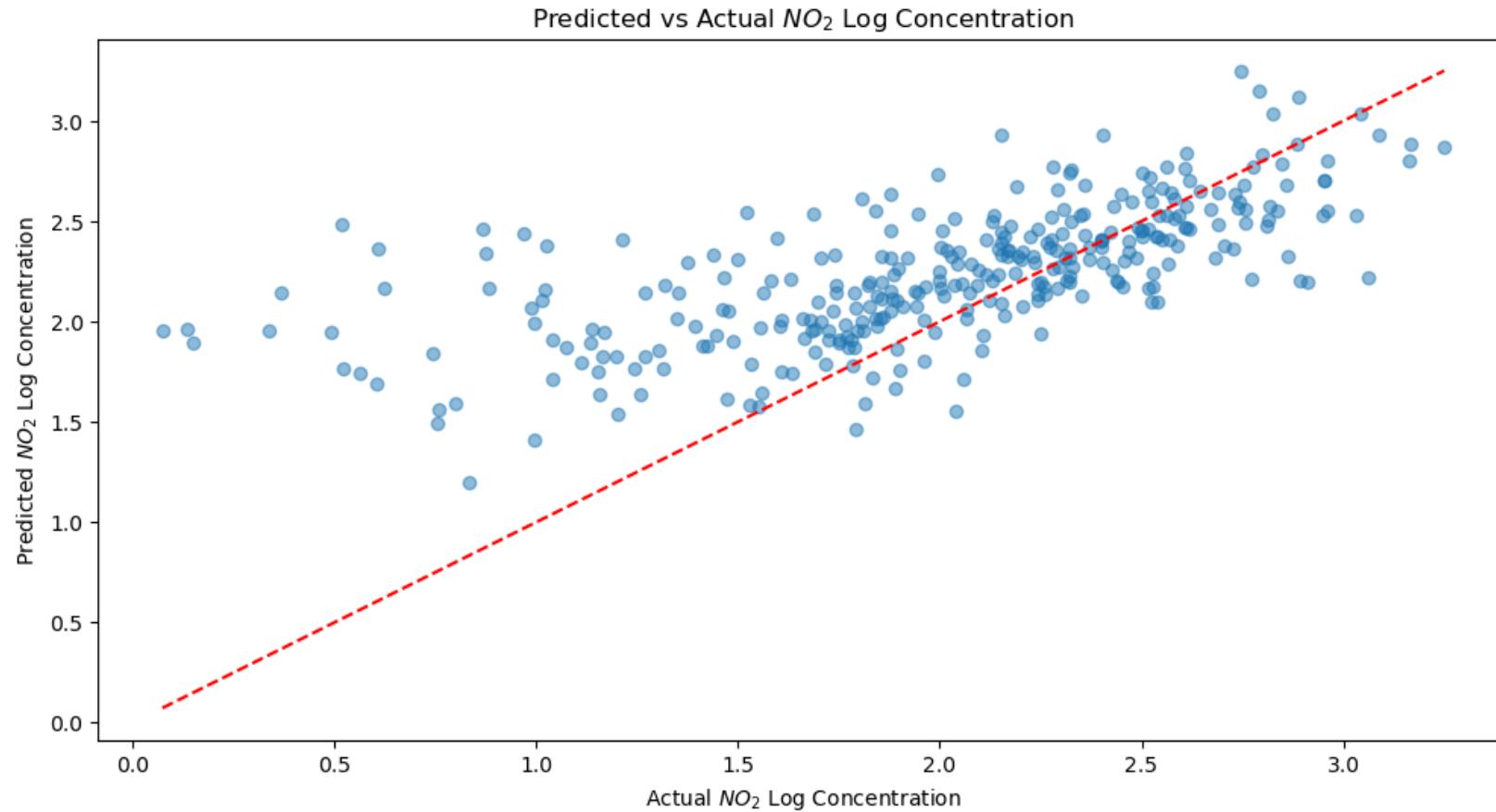
Metro Area	MSE	RMSE	R <sup>2</sup>
Chicago	0.12	0.35	0.60
Denver	0.09	0.3	0.85
Los Angeles	0.15	0.38	0.66
New York	0.14	0.37	0.45
Washington	0.16	0.4	0.60

---

# COVID-19 Impacts (Predictive Difference – Chicago Example)

- Using 100% Pre-COVID as training data
- Post COVID (2020-03-15 to 2021-03-15) as the testing data
- Log( $\text{NO}_2$ ) vs
  - fip\_site,
  - day\_of\_week, day\_of\_year\_sin, day\_of\_year\_cos,
  - outdoor\_temp, relative\_humidity, wind\_direction, wind\_speed\_log
- Results
  - Mean Squared Error: 0.271
  - Root Mean Squared Error: 0.521
  - R-squared: 0.274

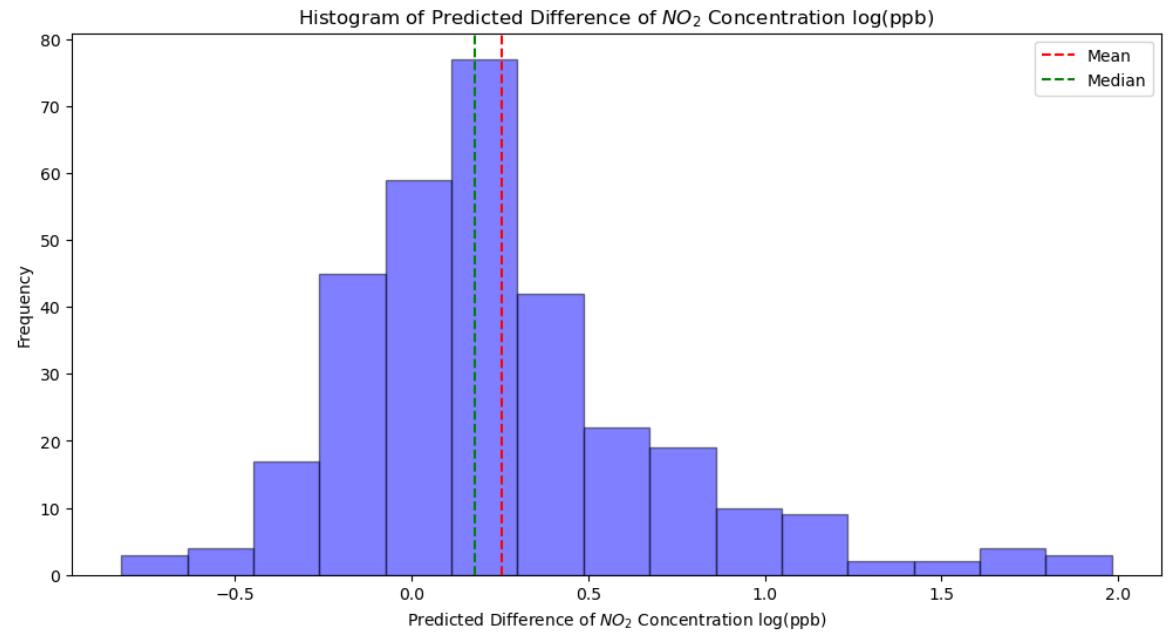
# COVID-19 Impacts (Predictive Difference – Chicago Example)



**Predictions are higher than actual observed values**

# COVID-19 Impacts (Predictive Difference – Chicago Example)

- Mean of predicted difference: 0.26
- Median of predicted difference: 0.18
- 95% confidence interval for the mean of the predicted difference:  
(0.19, 0.29)



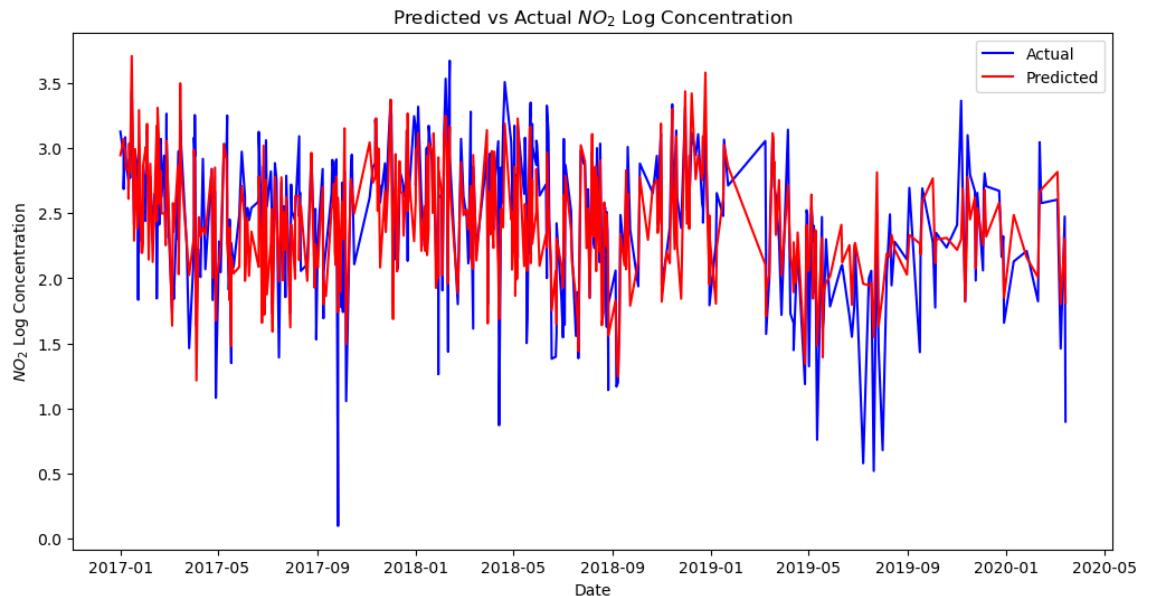
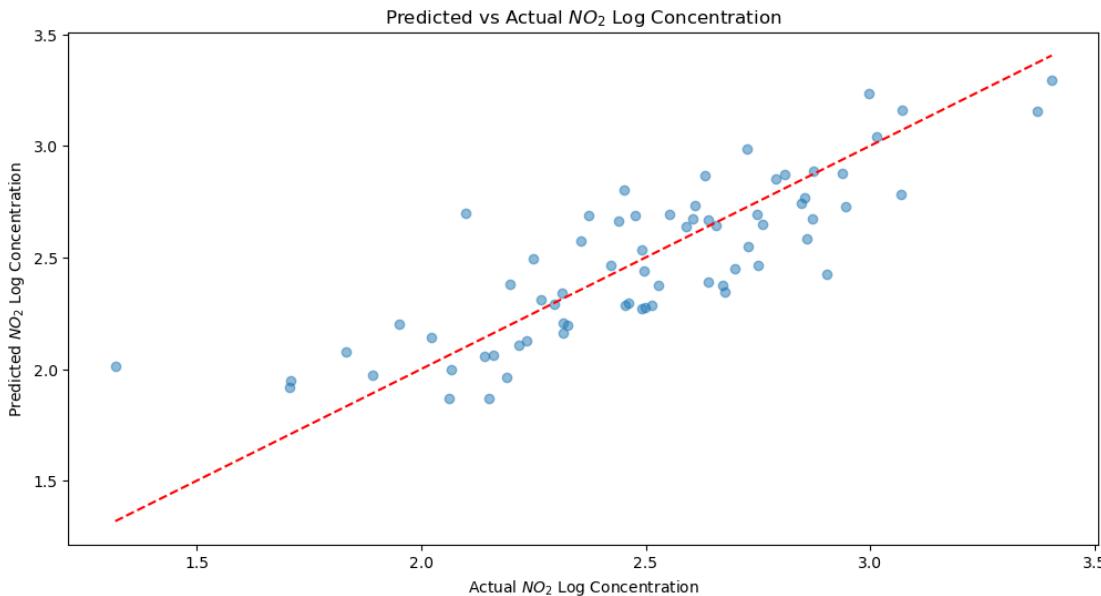
# COVID-19 Impacts (Predictive Difference – All Metro Areas)

Metro Area	R <sup>2</sup> Pre-COVID	R <sup>2</sup> Pos-COVID	Mean Difference	95% CI Lower	95% CI Higher
Chicago	0.60	0.26	0.25	0.20	0.30
Denver	0.85	0.8	0.18	0.17	0.20
Los Angeles	0.66	0.63	0.13	0.12	0.14
New York	0.45	0.17	0.21	0.17	0.26
Washington	0.60	0.06	0.38	0.36	0.41

# COVID-19 Impacts (Weekly Averaged Data–Chicago)

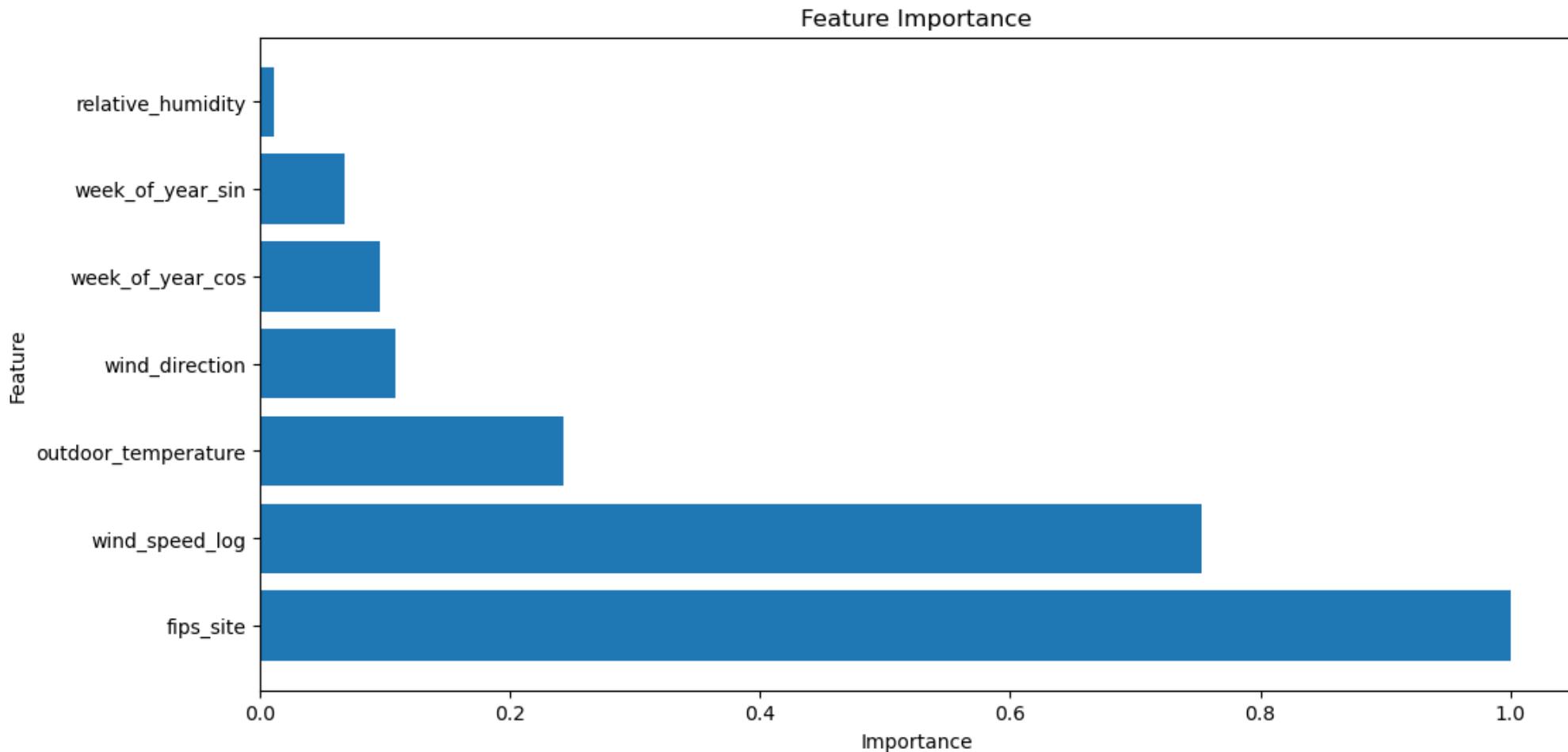
- Using only 2017 to Pre-COVID (14 March 2020)
- Averaged weekly values
- 80% train vs 20% test
- Log( $\text{NO}_2$ ) vs
  - fip\_site,
  - `week_of_year_sin`, `week_of_year_cos`,
  - outdoor\_temp, relative\_humidity, wind\_direction, wind\_speed\_log
- Results
  - Mean Squared Error: 0.047
  - Root Mean Squared Error: 0.217
  - R-squared: 0.678

# COVID-19 Impacts (Weekly Averaged Data–Chicago)



Model still have issues with very low  $NO_2$  concentrations

# COVID-19 Impacts (Weekly Averaged Data–Chicago)



# COVID-19 Impacts (Weekly Averaged Data—All Metro Areas)

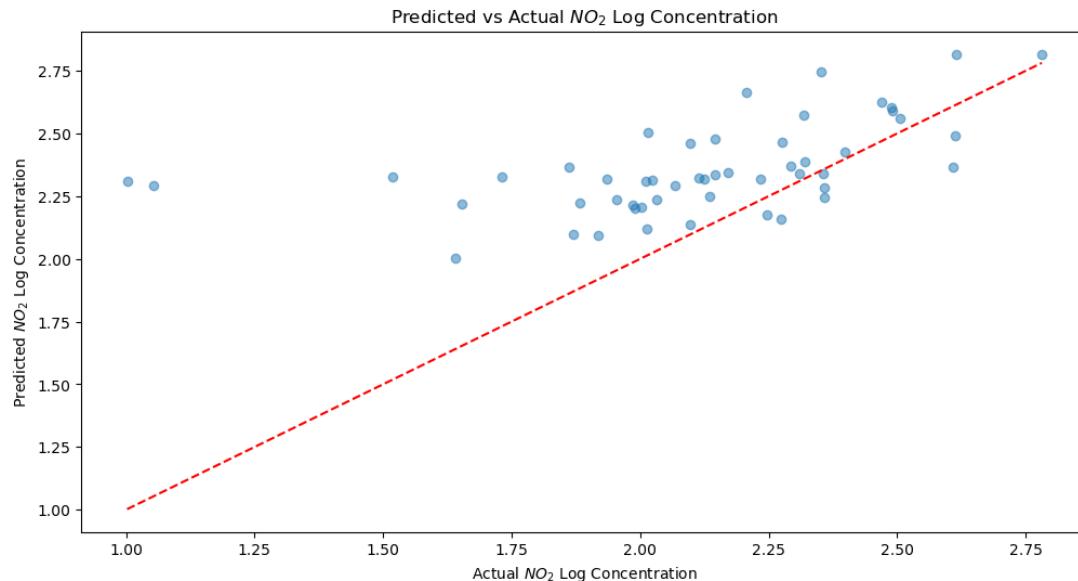
Metro Area	Weekly MSE	Weekly RMSE	Weekly R <sup>2</sup>	Daily R <sup>2</sup>
Chicago	0.05	0.22	0.68	0.60
Denver	0.03	0.17	0.96	0.85
Los Angeles	0.07	0.26	0.76	0.66
New York	0.04	0.19	0.60	0.45
Washington	0.08	0.28	0.67	0.60

Weekly averaged model has better fit than daily model

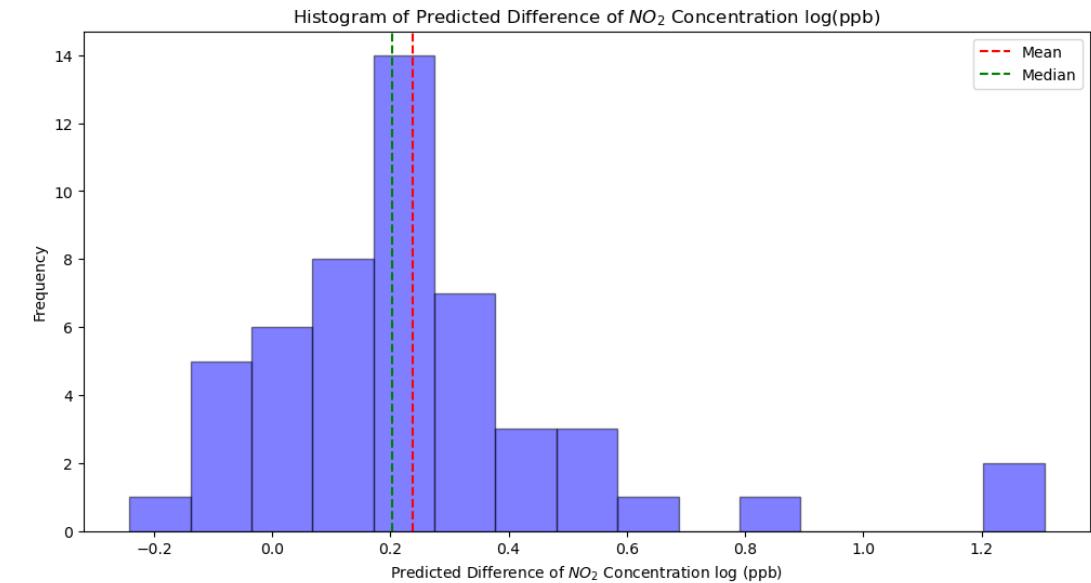
# COVID-19 Impacts (Predictive Difference – All Metro Areas)

- Using 100% Pre-COVID as training data
- Post COVID (2020-03-15 to 2021-03-15) as the testing data
- Log(NO<sub>2</sub>) vs
  - fip\_site,
  - week\_of\_year\_sin, week\_of\_year\_cos,
  - outdoor\_temp, relative\_humidity, wind\_direction, wind\_speed\_log

# COVID-19 Impacts (Predictive Difference—Chicago)



Mean Squared Error: 0.140  
Root Mean Squared Error: 0.374  
R-squared: -0.192

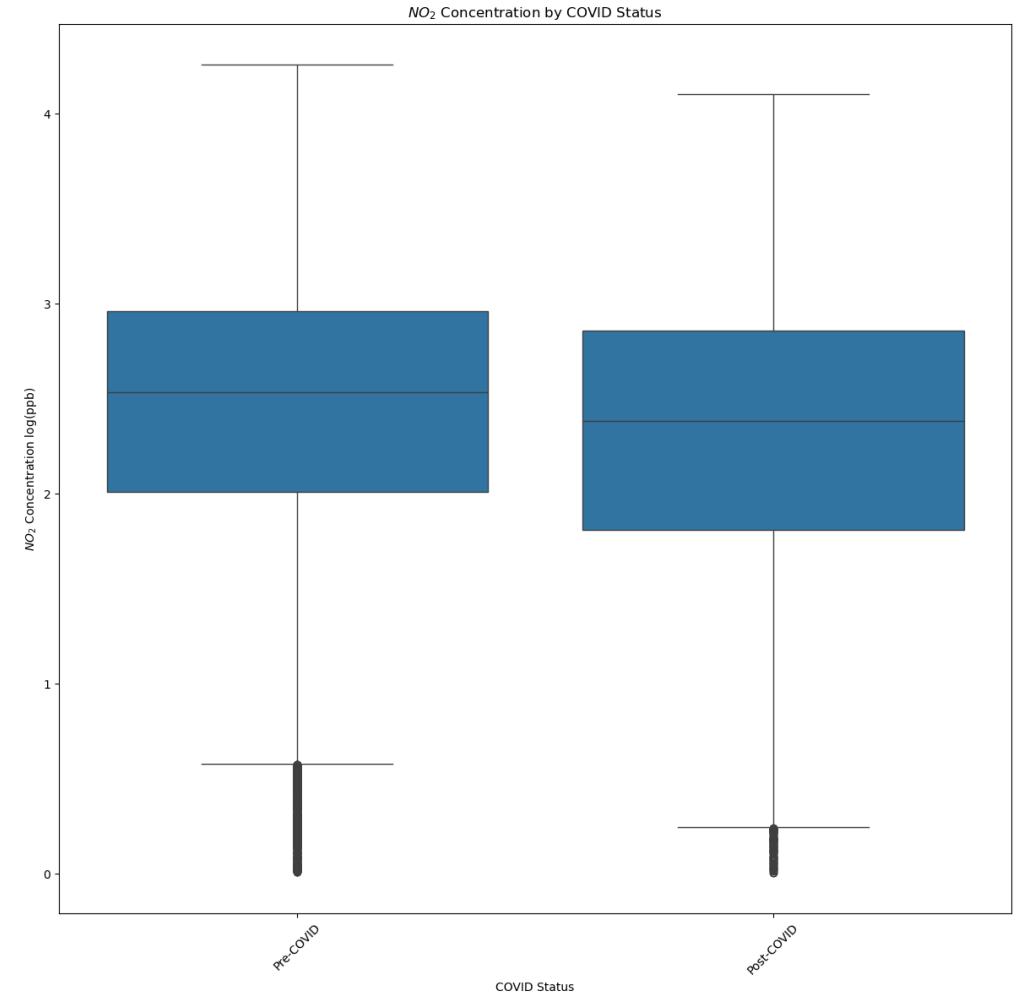
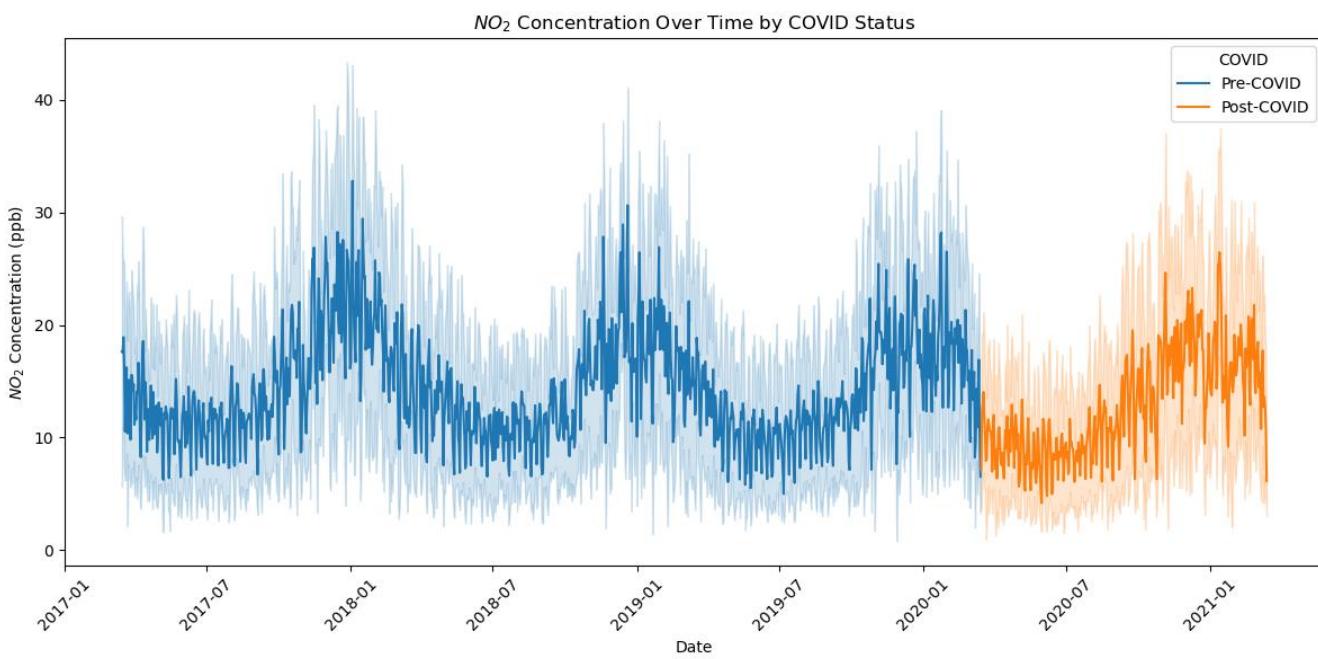


95% confidence interval for the mean of the predicted difference: (0.15, 0.32)

# COVID-19 Impacts (Predictive Difference – All Metro Areas)

Metro Area	R <sup>2</sup> Pre-COVID	R <sup>2</sup> Pos-COVID	Mean Difference	95% CI Lower	95% CI Higher
Chicago	0.68	-0.19	0.24	0.16	0.32
Denver	0.96	0.87	0.18	0.16	0.21
Los Angeles	0.76	0.72	0.12	0.10	0.14
New York	0.60	0.4	0.18	0.14	0.23
Washington	0.67	0.30	0.29	0.25	0.34

# COVID-19 Impacts (Grouped Analysis)



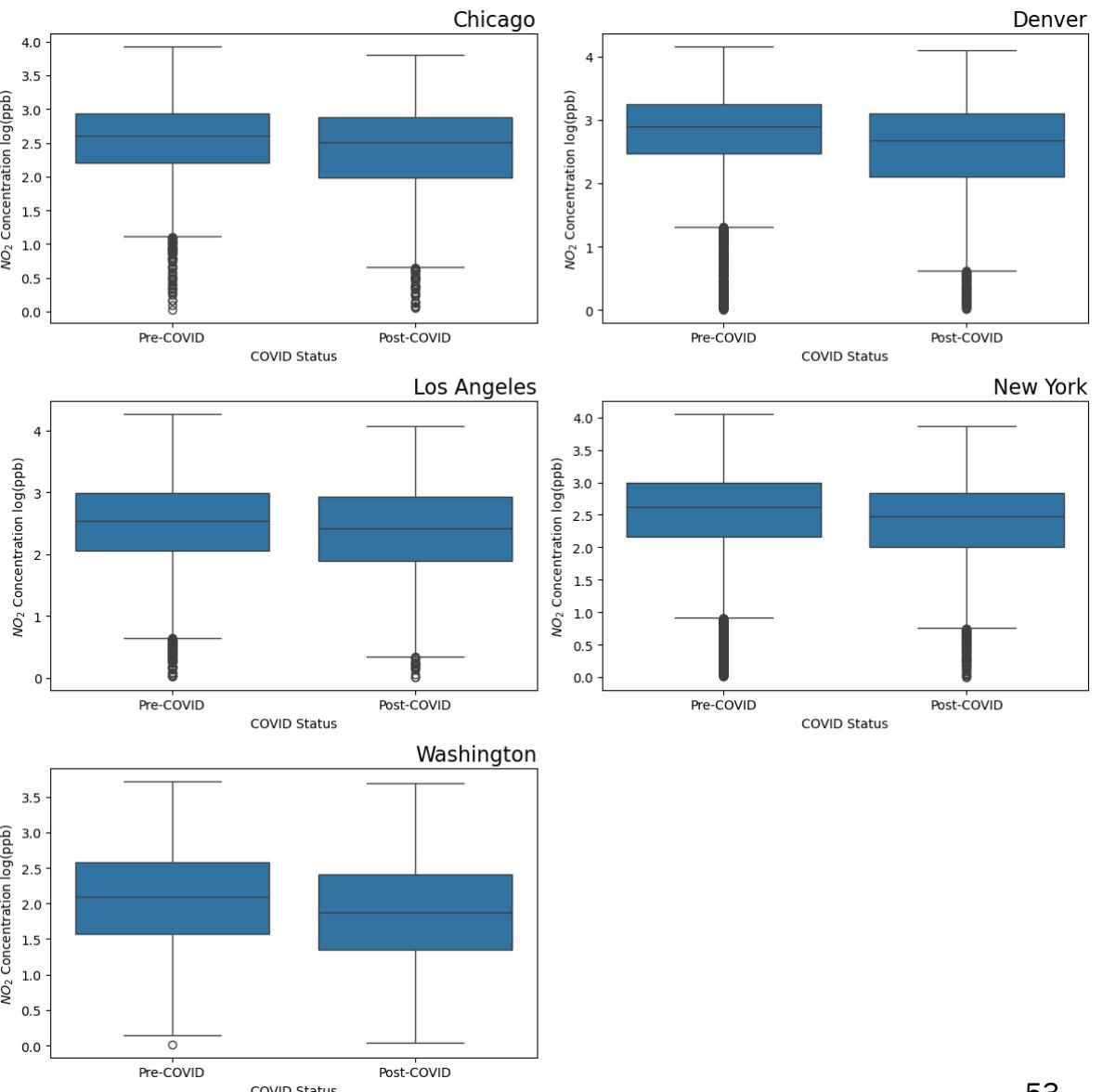
# COVID-19 Impacts (Generalized Linear Model)

- Python glm module with gaussian log-link

	pval	R2 (psedo)	Diff	95% CI Lower	95% CI Upper	Percent Change		
						Difference	95%CI Lower	95%CI Upper
Chicago	0	0.01	0.09	0.06	0.11	9%	6%	12%
Denver	0	0.02	0.19	0.16	0.22	21%	17%	25%
Los Angeles	0	0.00	0.08	0.06	0.1	9%	7%	11%
New York	0	0.01	0.14	0.12	0.16	15%	12%	17%
Washington	0	0.01	0.16	0.13	0.18	17%	14%	20%

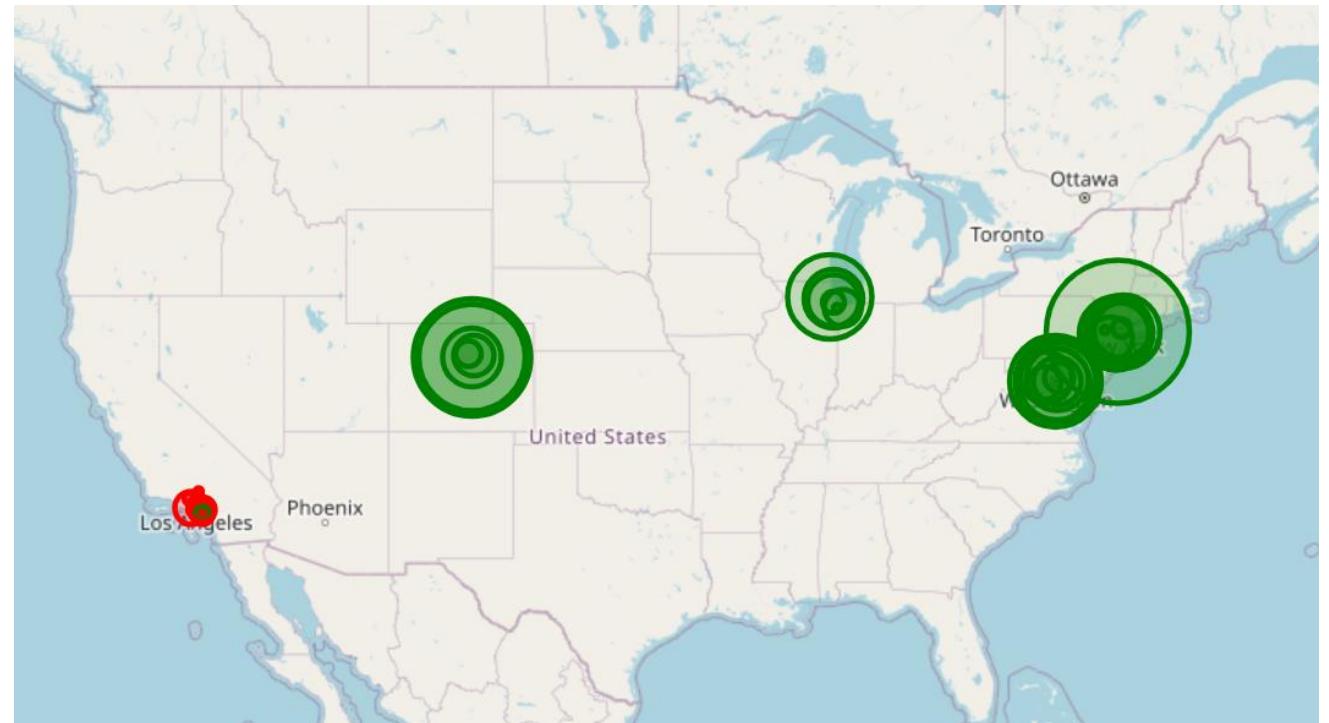
# COVID-19 Impacts (Grouped Analysis)

	Percent Change		
	Diff	95%CI Lower	95%CI Upper
Chicago	9%	6%	12%
Denver	21%	17%	25%
Los Angeles	9%	7%	11%
New York	15%	12%	17%
Washington	17%	14%	20%



# COVID-19 Impacts (Interactive Map of NO<sub>2</sub> Difference)

- Difference between 2020 and 2019 average yearly NO<sub>2</sub> concentration at each FIPS Site [4]
- Styled by
  - Color: decrease (green) or increase (red)
  - Size: magnitude of change (absolute difference)
- Shows 2020 increase in NO<sub>2</sub> concentrations in Los Angeles as compared to 2019



# Conclusions

- Predictions of daily NO<sub>2</sub> data with daily weather data was varied by metro area with Denver having the best predictability
- Daily fit-model has the worst performance for low NO<sub>2</sub> concentrations
- Predictive model fit with pre-COVID data generally predicted higher NO<sub>2</sub> concentrations than were observed during COVID
- COVID NO<sub>2</sub> concentrations were between 10-30% lower than predicted for all metro areas
  - Los Angeles saw the lowest reduction at about 10%
  - Washington DC saw the highest reduction at about 20-35%
- Weekly averaged model has better overall fit but still does poorly at predicting low NO<sub>2</sub> concentrations

# Future Work

- Explore the low NO<sub>2</sub> concentration ranges to improve model performance
  - (Box-Cox transform did not significantly improve model performance)
- Build models for FIPS Code
  - Use factor permutations for each FIPS code for best fit
- Explore more detailed time-series analysis options including yearly (365 days) and week of year (52 week) differenced data to remove cyclical components
- Explore deep learning methods

# Acknowledgements

Much of the code used in the project was initially used in class as part of homework sets or in-class projects.

- Dr. Kennedy: Foundational understanding of knowledge mining, model preparations, model tuning, and a reminder that data cleaning is 90% of the work.
- Dr. Miller: Applying statistical analysis to real world problems, namely ANOVA testing and linear regression.
- Dr Russo: Applying statistical visualization to real world problems, including the seaborn library.
- Dr Wolf (GGS): Applying time-series analysis to real world problems, including identifying cyclical and trend components in time-series data
- Dr Pfoser (GGS): Interactive map making, specifically deploying HTML in CodePen

# References

- [1] U.S. Environmental Protection Agency. (2023, June 29). Overview of nitrogen dioxide ( $\text{NO}_2$ ) air quality in the United States.url: [https://www.epa.gov/system/files/documents/2023-06/NO2\\_2022.pdf](https://www.epa.gov/system/files/documents/2023-06/NO2_2022.pdf)
- [2] U.S. Environmental Protection Agency. Air Data: Air Quality Data Collected at Outdoor Monitors Across the US. 2025. url: <https://www.epa.gov/outdoor-air-quality-data>.
- [3] U.S. Census Bureau. (n.d.). Metro area geography reference. U.S. Department of Commerce. Retrieved April 13, 2025, from <https://www.census.gov/programs-surveys/cbp/technical-documentation/reference/metro-area-geography-reference.html>
- [4] Folium Developers. (n.d.). Folium documentation (latest). Python Visualization. Retrieved April 14, 2025, from <https://python-visualization.github.io/folium/latest/index.html>



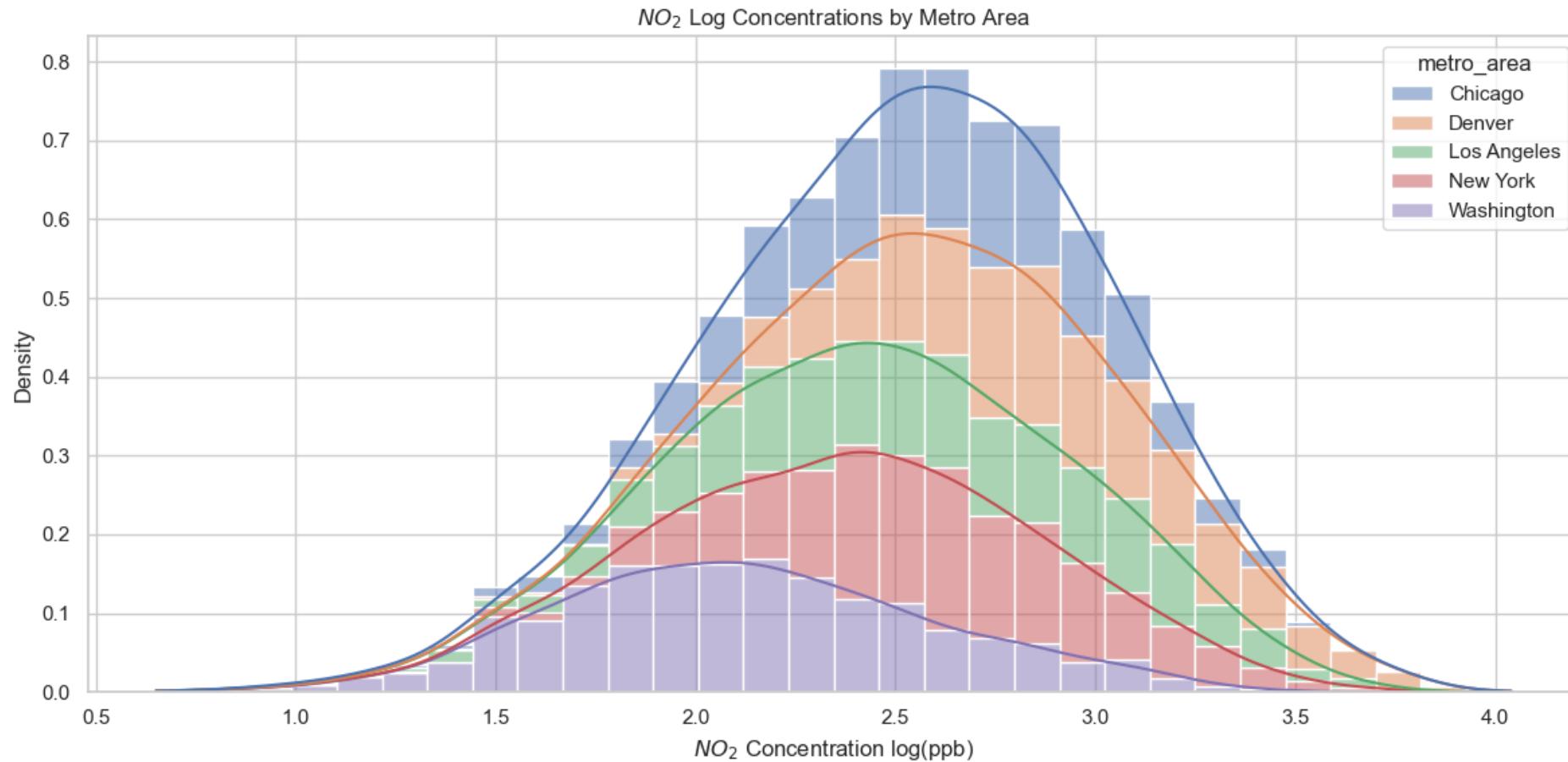
# Questions



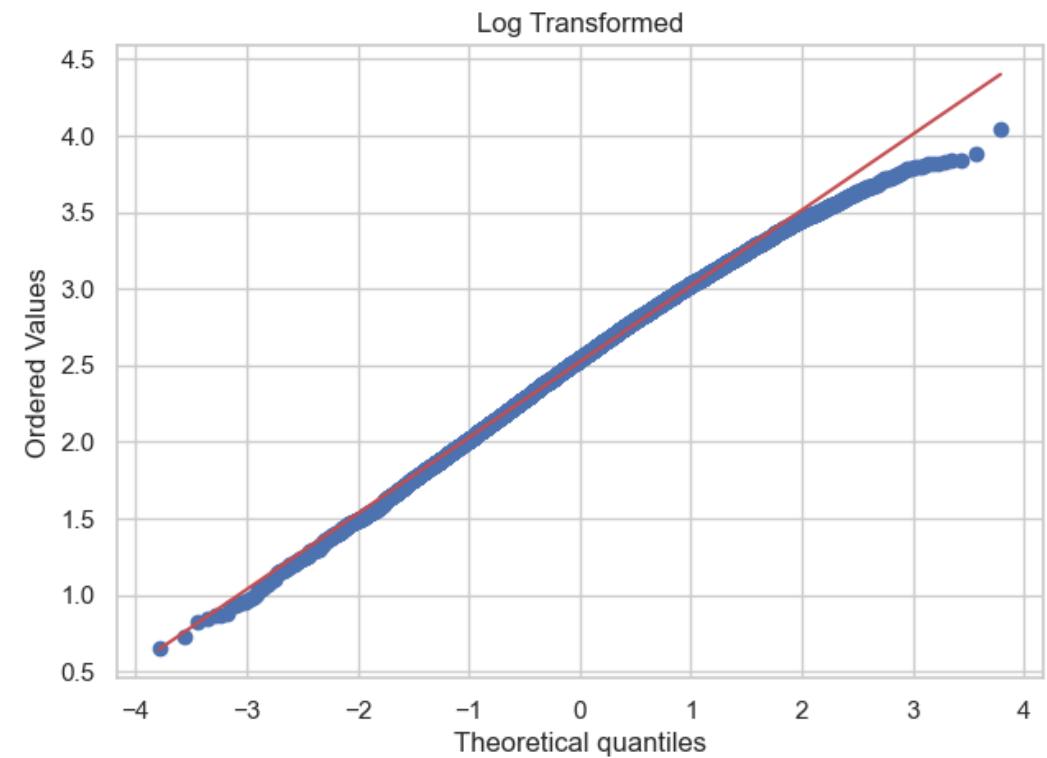
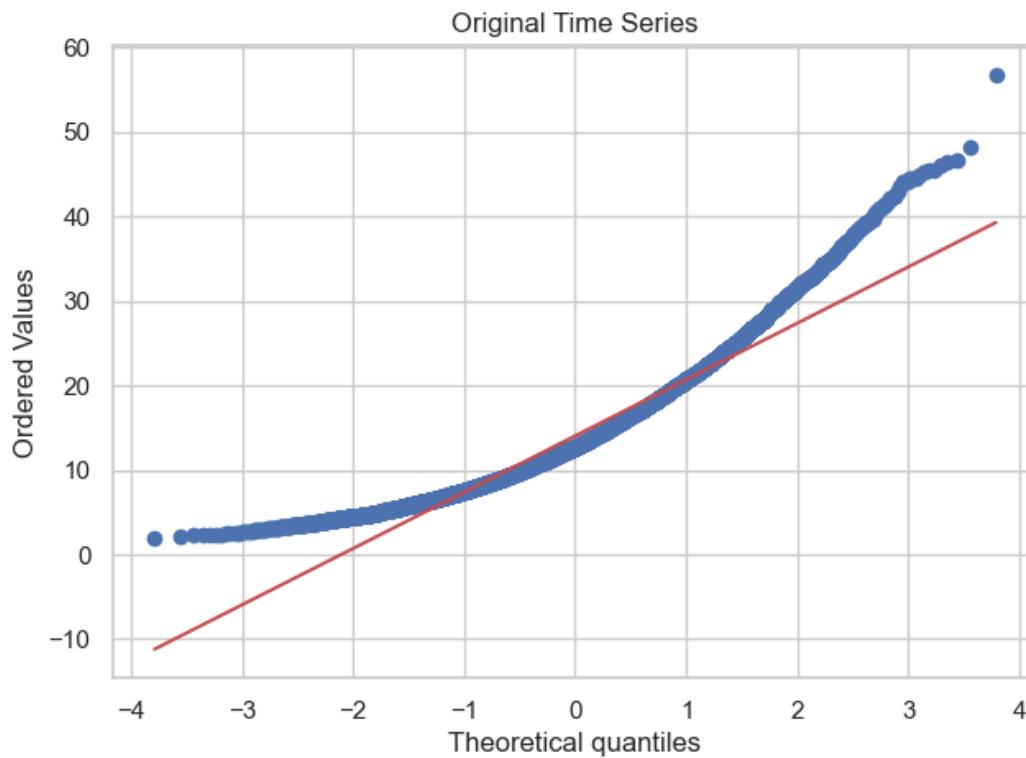
# Backup Material

Supporting Graphics and Model  
Assumptions

# Modeling Assumptions (Normality)



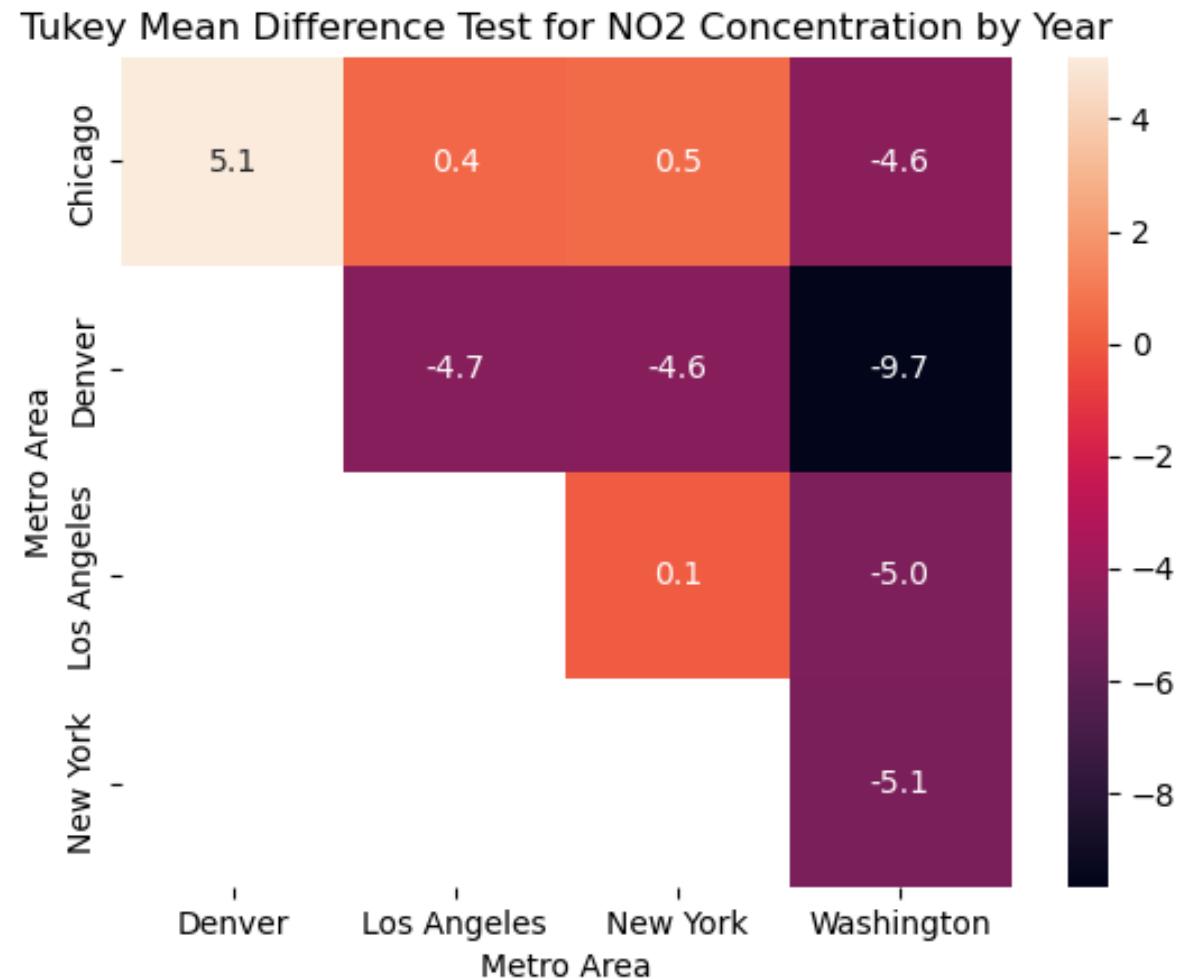
# Modeling Assumption (Normality)



Log transformed data appears more normal

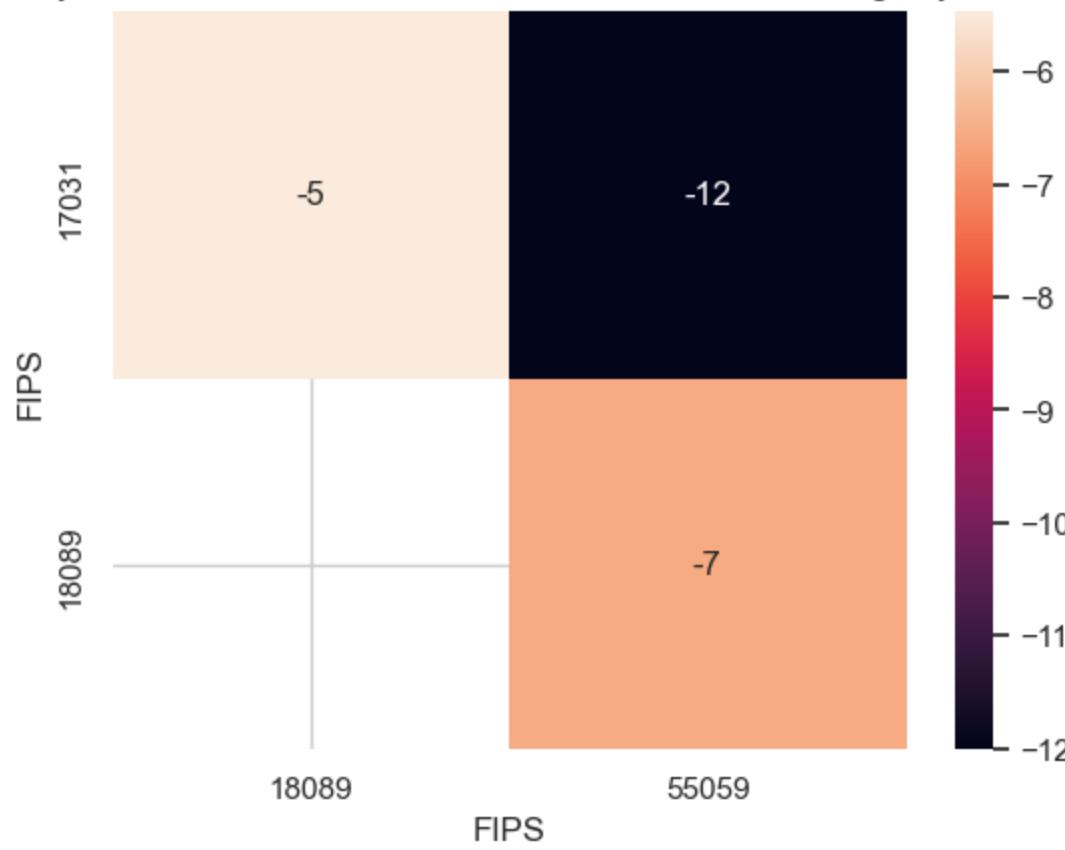
# $\text{NO}_2$ Concentration Difference Metro Area

- ANOVA test of  $\text{NO}_2$  concentration by metro area for 2017-2019
- The difference in  $\text{NO}_2$  concentrations for the metro areas is statistically significant
- The only exception is New York and Los Angeles

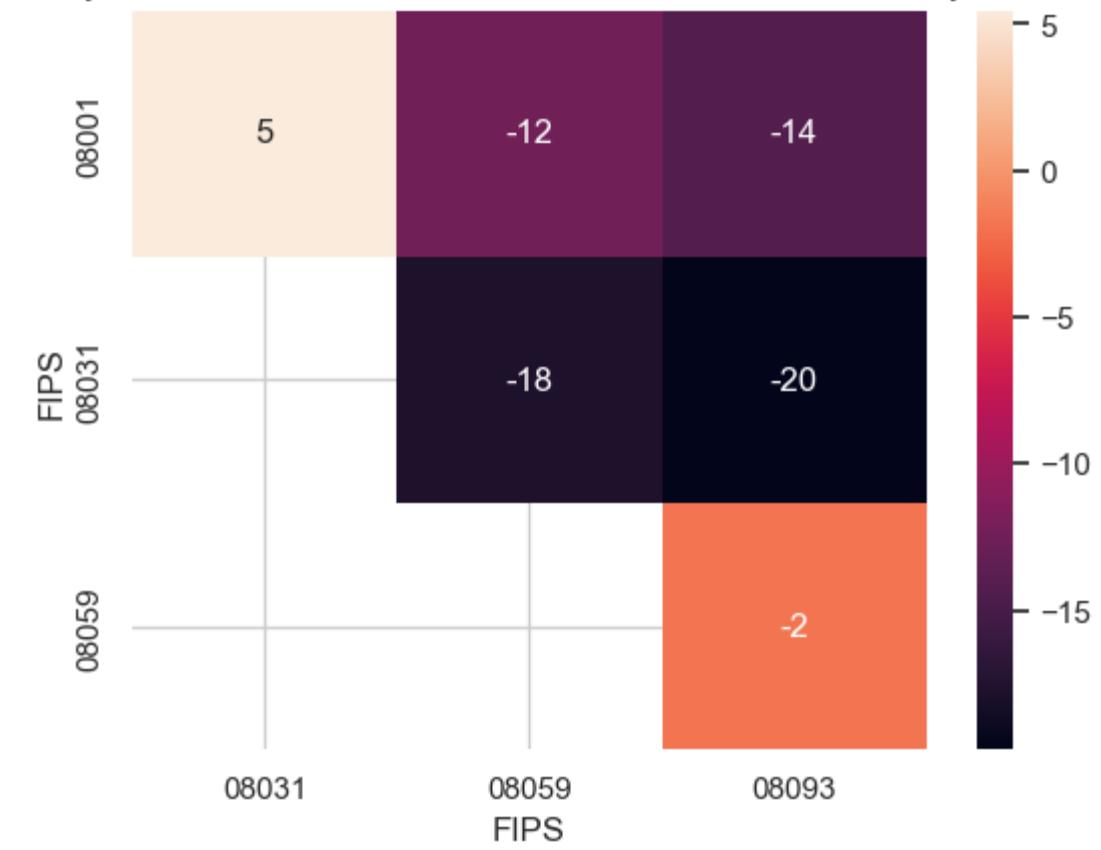


# $\text{NO}_2$ Concentration Difference FIPS Code

Tukey Mean Difference Test for  $\text{NO}_2$  Concentration in Chicago by FIPS

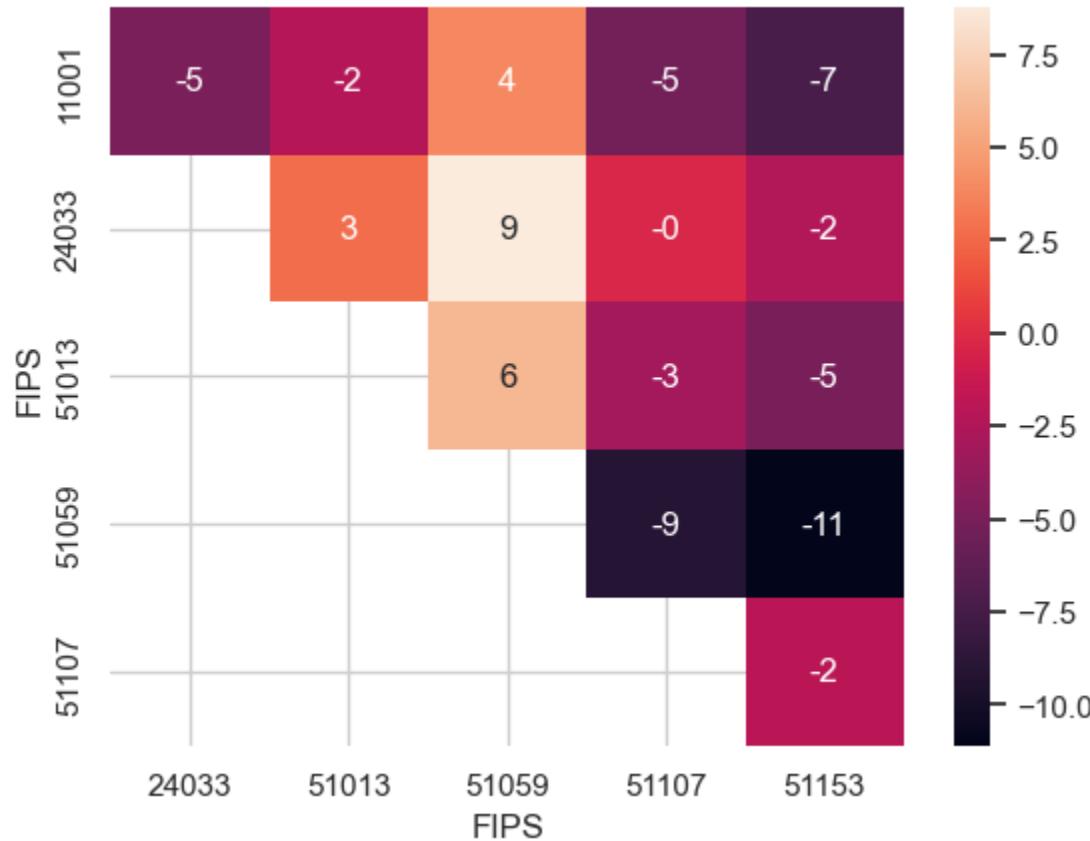


Tukey Mean Difference Test for  $\text{NO}_2$  Concentration in Denver by FIPS

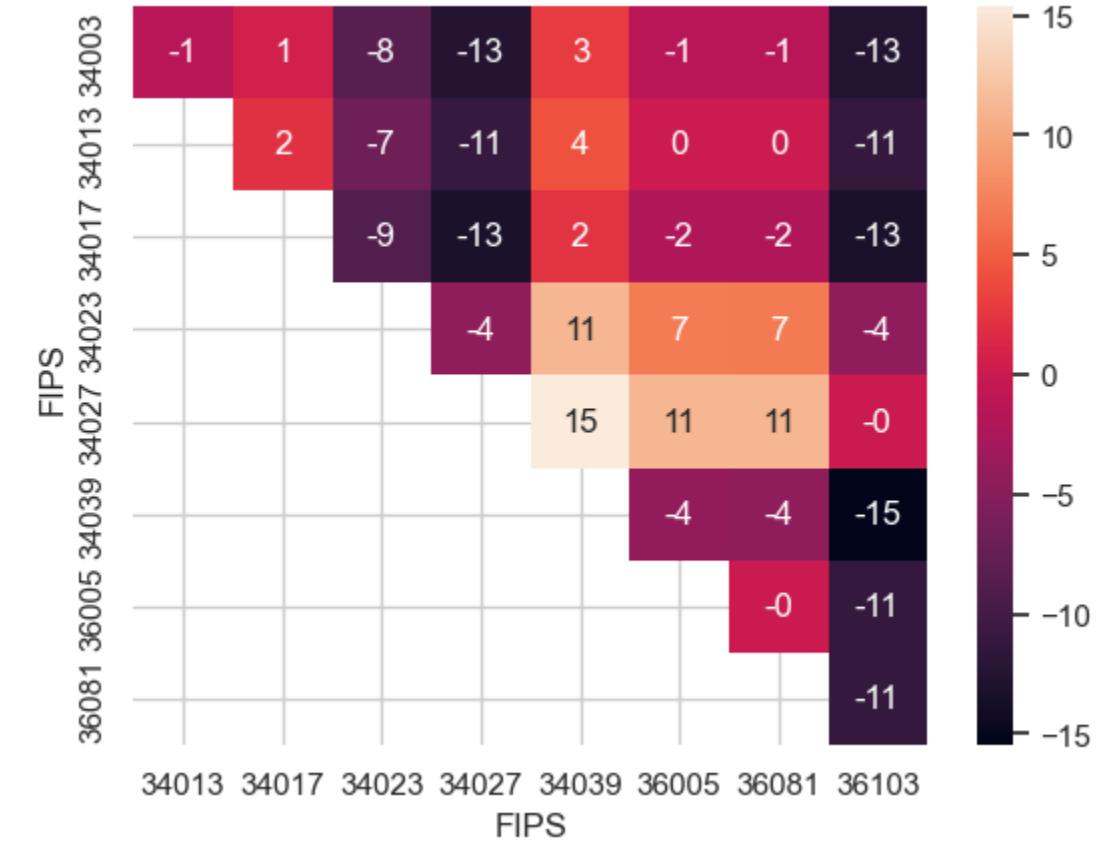


# $\text{NO}_2$ Concentration Difference FIPS Code

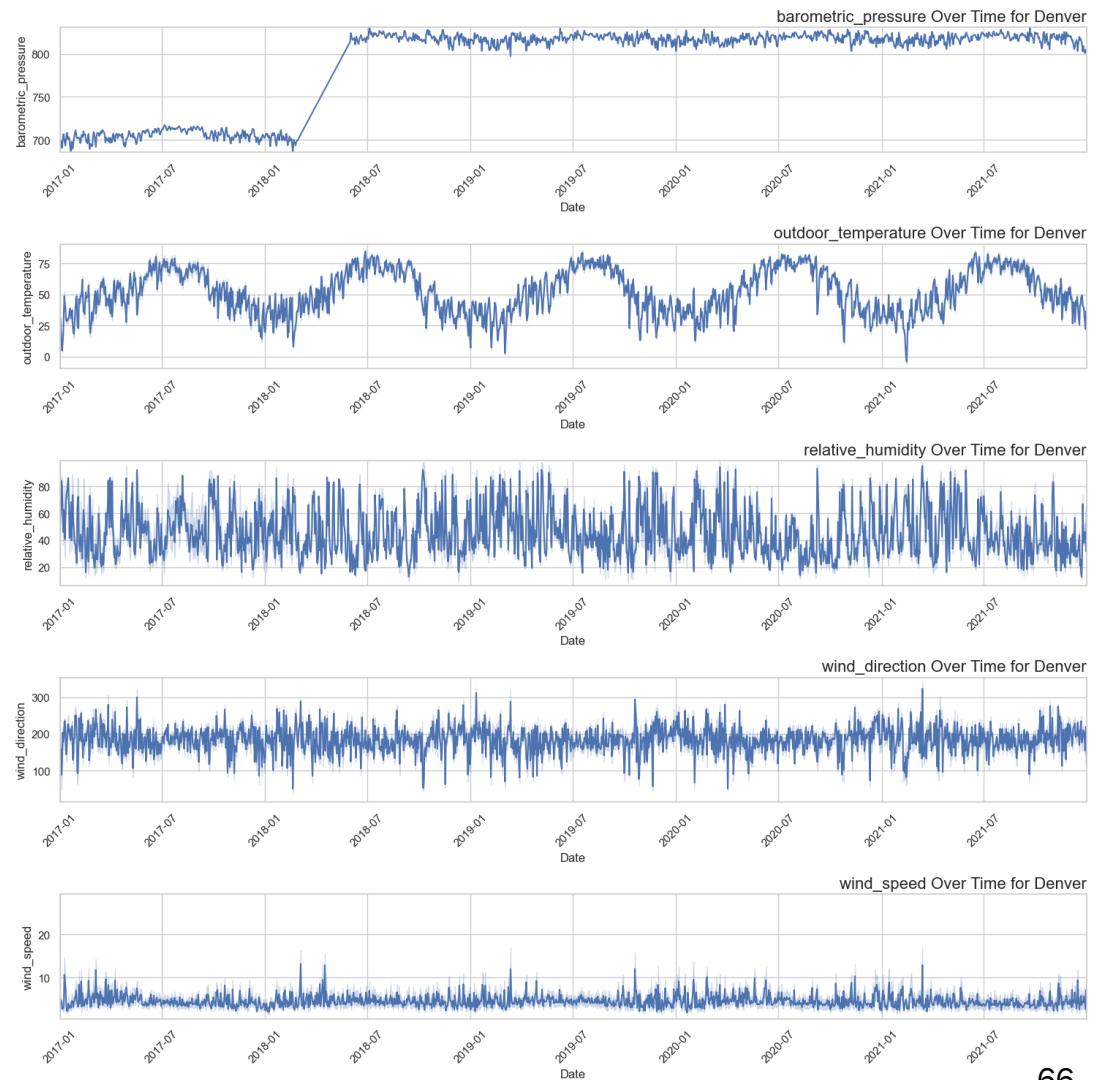
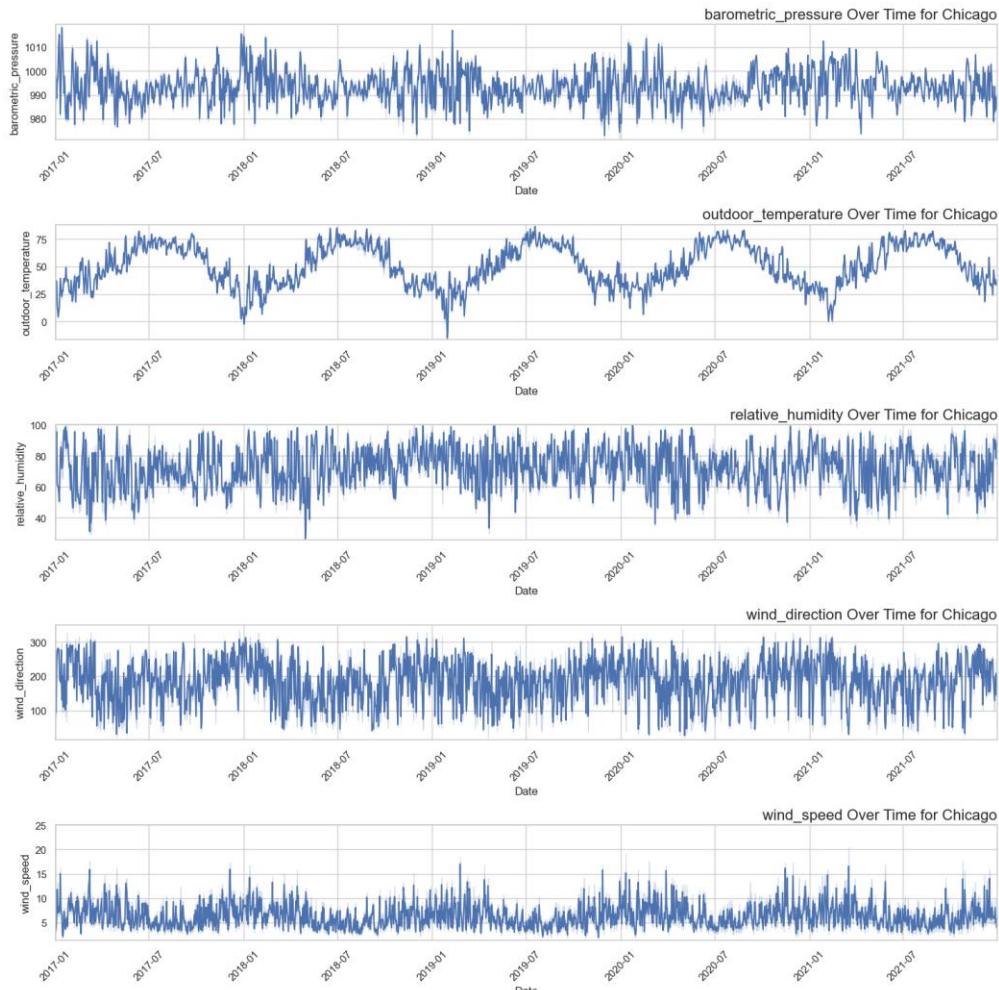
Tukey Mean Difference Test for  $\text{NO}_2$  Concentration in Washington by FIPS



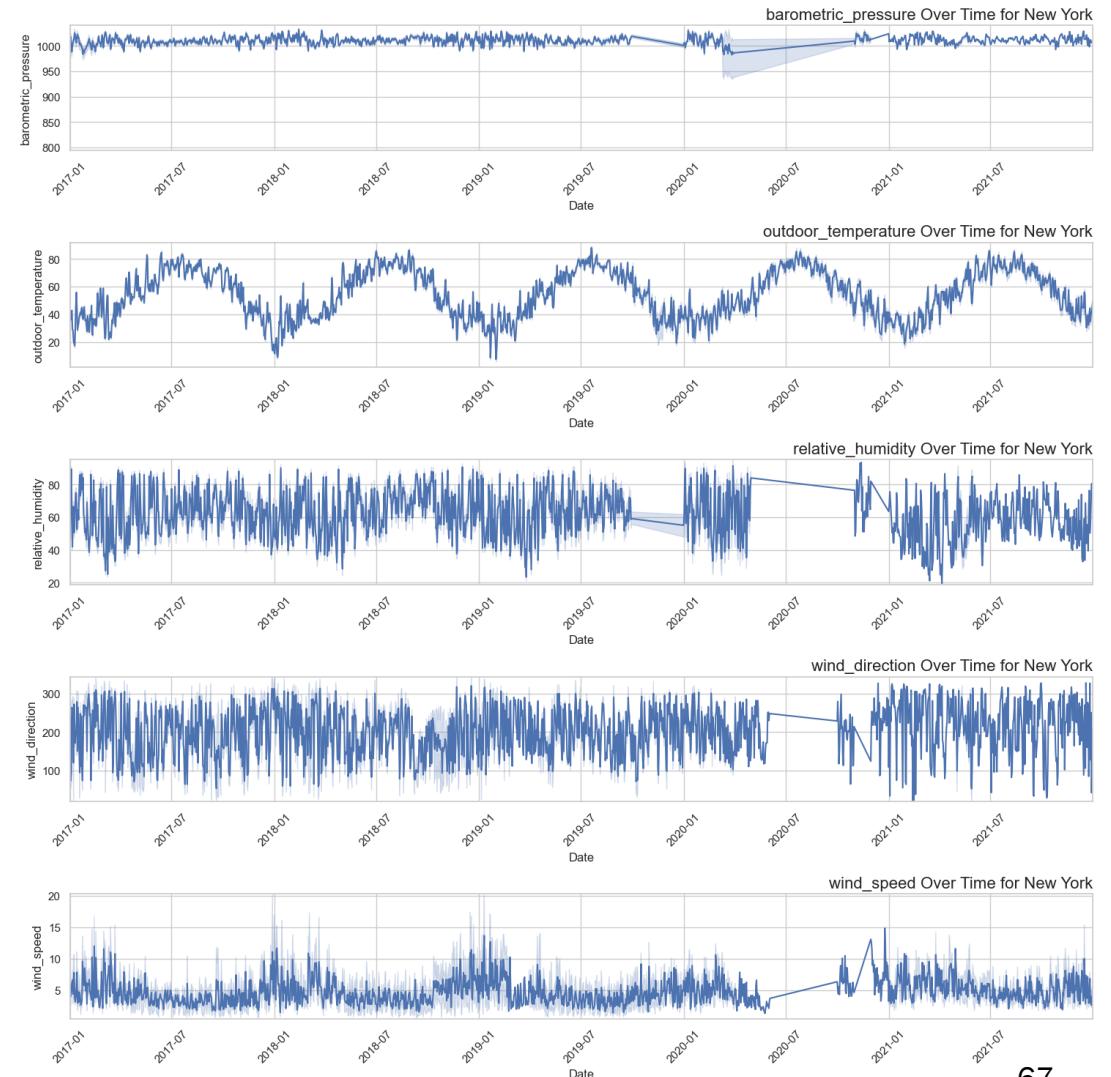
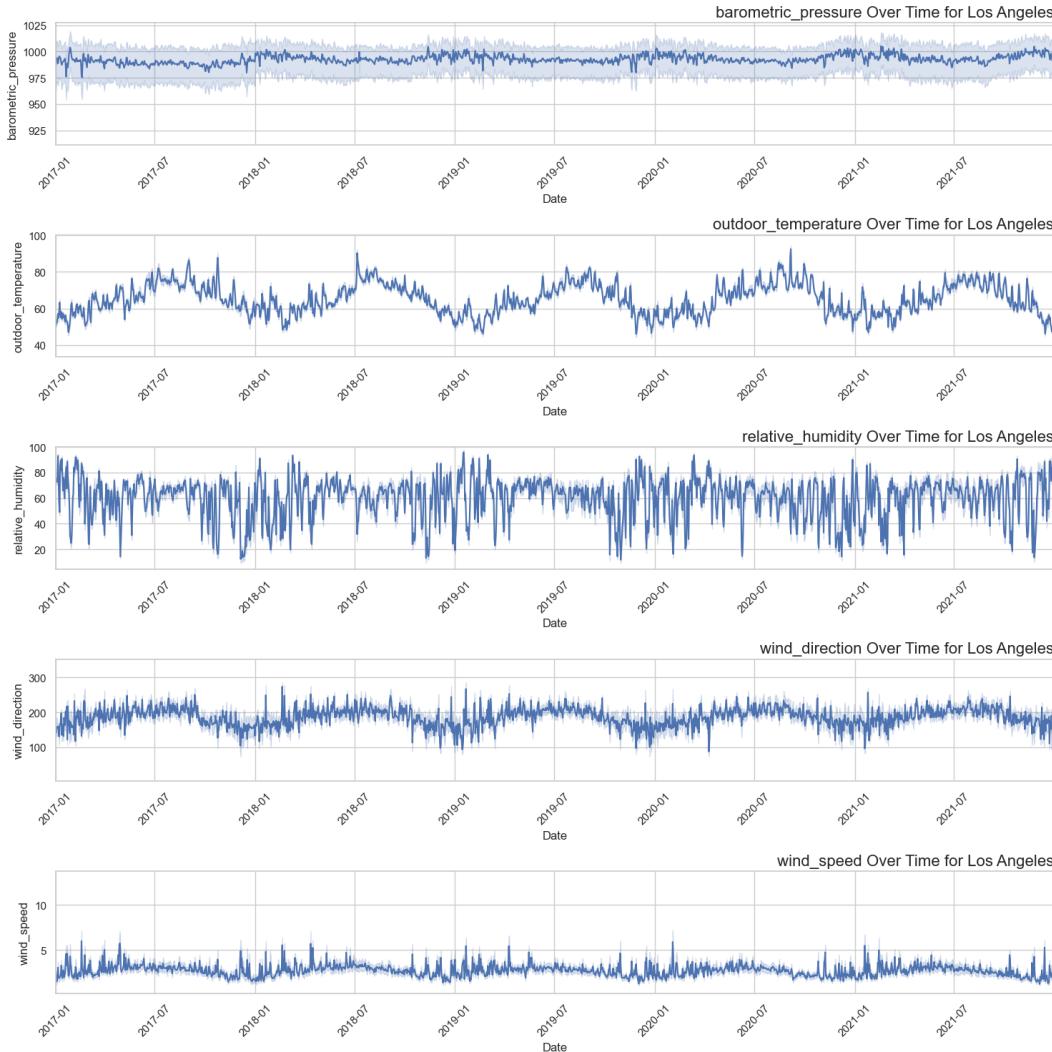
Tukey Mean Difference Test for  $\text{NO}_2$  Concentration in New York by FIPS



# Weather Data Over Time



# Weather Data Over Time



# Weather Data Over Time

