Miguel Angel Perez Porras

12/04/2024

ENV3040C

Course Project

Air quality is a critical factor for public and environmental health, reflecting the impacts of urbanization, industrial activity, and population density. It has been overlooked by the general population, and most of the time they take the air quality for granted. I will be studying the distribution of the Air Quality Index (AQI) in urban and rural areas of the United States, focusing on how population density affects air quality. I will look at the examined number of days with recorded AQI, the presence of "good days," and population density metrics for each state. The analysis goal is to identify patterns and singularities between urban and rural areas. The insights from this research contribute to understanding the interaction between human activities and environmental conditions, providing an insight foundation for targeted environmental interventions and policies.

For my hypotheses, I suspect the urban states, characterized by higher population density and industrial activity, will exhibit worse air quality than rural states, as indicated by fewer "good days" with AQI and a greater number of days with recorded AQI. Population density will have a significant negative correlation with air quality, suggesting that more densely populated states are likely to have higher pollution levels.

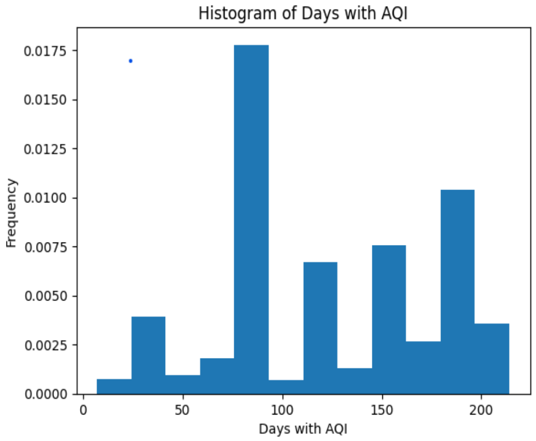
Several studies have explored the relationship between urbanization, population density, and air quality, providing insights to the hypothesis I am working on. There is a study that analyzed air quality disparities between urban and rural counties in the U.S. According to the CDC, they found that urban counties exhibited higher levels of pollutants, such as ozone and particulate matter, compared to rural counties. These findings suggest that urbanization contributes to increased air pollution, potentially impacting public health.

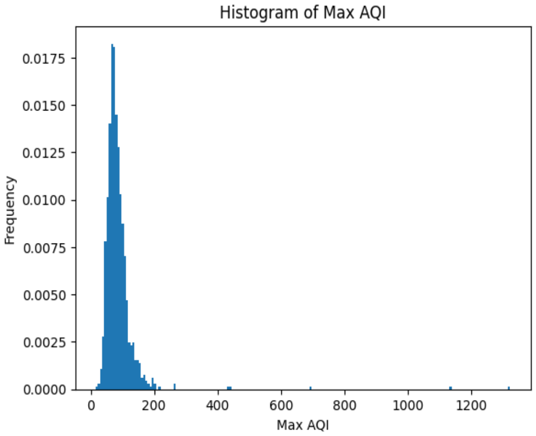
The data was sourced from the U.S. Environmental Protection Agency (EPA) and provides comprehensive air quality measurements across various states. It includes key variables such as the number of days with a recorded Air Quality Index (AQI), the number of "good days," maximum air quality index, the median AQI, etc. Each data point corresponds to a specific county value, allowing for an insightful analysis of air quality distribution among all counties in the US.

The data was measured using standard EPA protocols for monitoring air quality, which involve collecting daily pollutant concentrations for ozone (O_3), particulate matter (PM_2.5 and PM_{10}), carbon monoxide (CO), and nitrogen dioxide (NO_2). These measurements are used to calculate the AQI and classify the air quality on a scale from "Good" to "Hazardous."

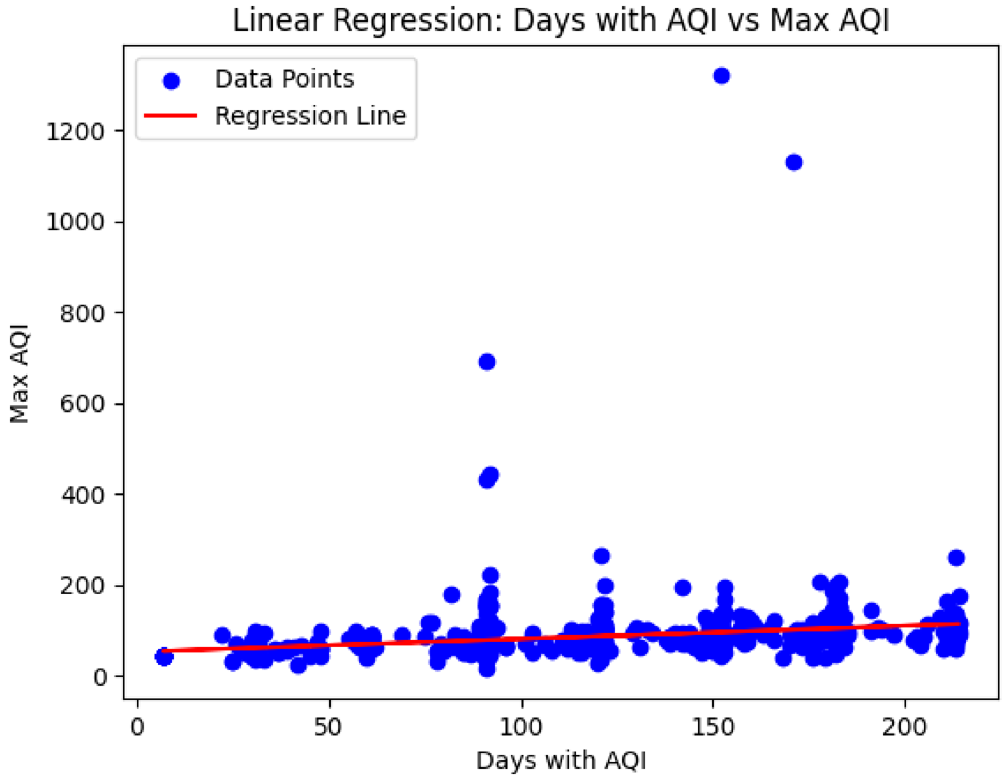
For the purposes of this study the data points from Mexico and the District of Columbia were excluded to focus mainly on U.S. states. It is really useful that they provided international values to the measurements, but for the purpose of this analysis they are not necessary. There are no direct replicates, as each data point represents unique annual metrics for each specific county in the state.

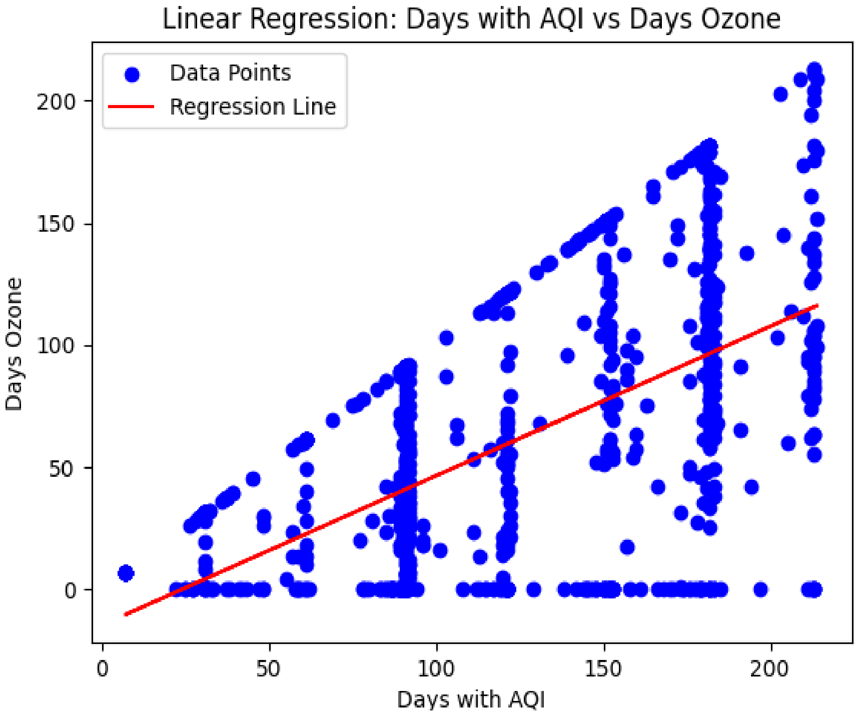
It is important to note that while the data provides aggregate annual values, potential biases could arise from variations due to geographical factors across states and differences in seasonal pollution patterns. These factors should be considered when interpreting results, particularly when comparing urban and rural states.

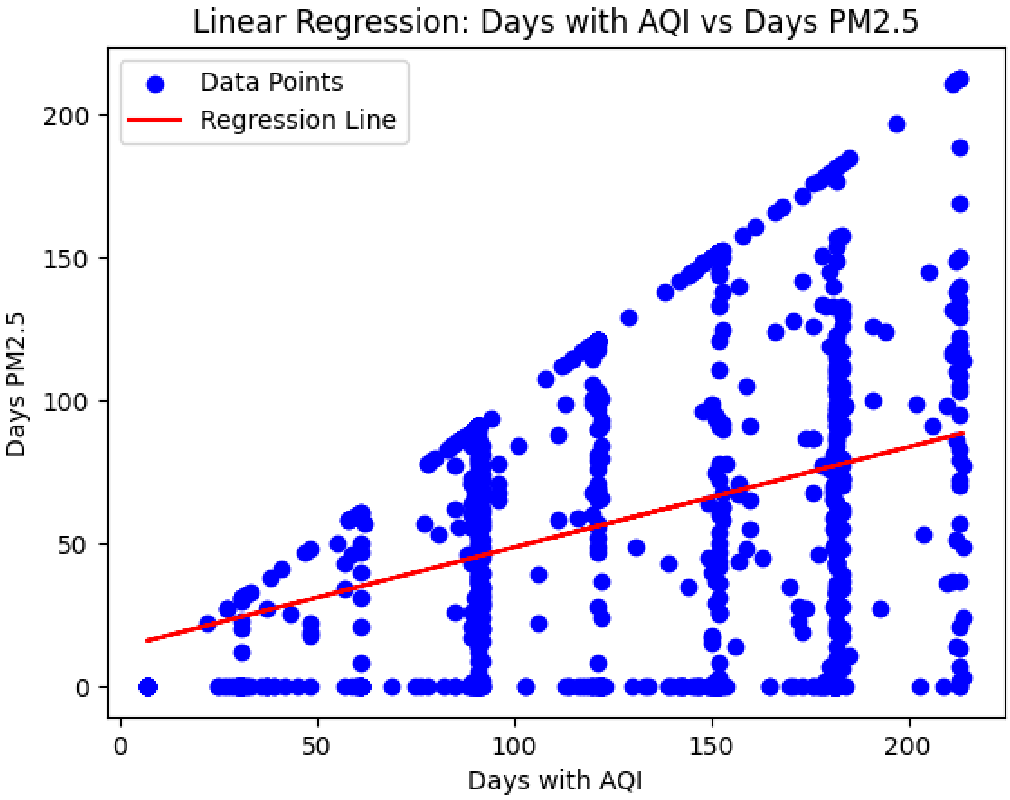


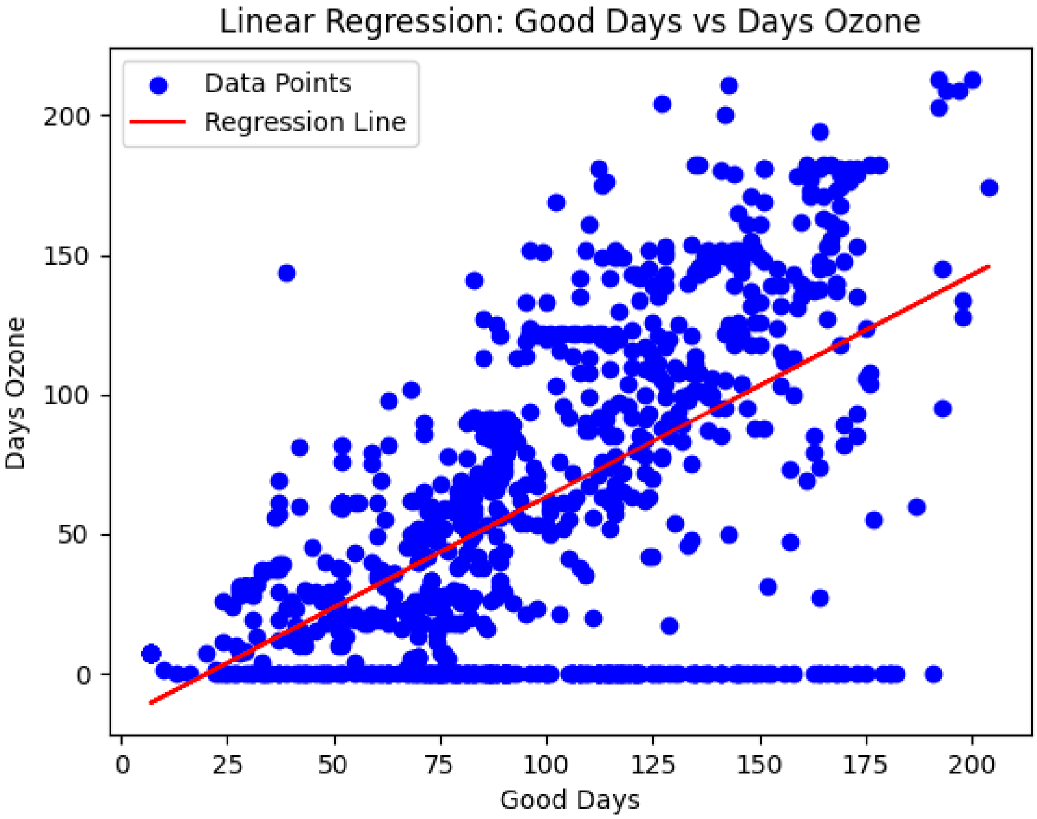


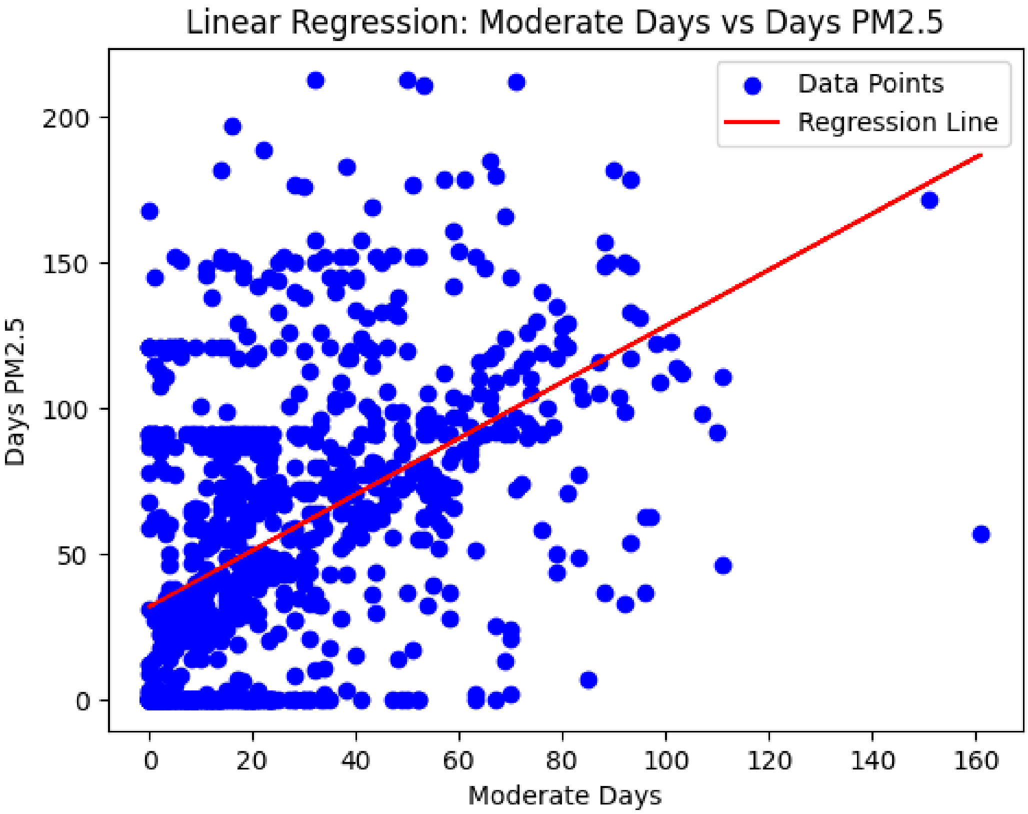
From the calculated values from the Pearson and Spearman correlation, we can see both correlations are strong and positive for days with AQ and good days, indicating that as the number of days with AQI increases, the number of good days also increases. This could suggest regions with consistent AQI monitoring tend to experience better air quality overall. The comparison between Days with AQI and the Moderate days give a positive correlation suggesting that areas with many AQI days also experience a significant number of moderate-quality air days. Furthermore, between the maximum AQI and the Days with AQI the Spearman correlation is much stronger, suggesting a flat and not necessarily linear relationship. It's worth noting that higher AQI days may signigy to more extreme pollution events. When we compare it with the days with Ozone, we get a moderate positive correlation suggests that more AQI days are associated with a higher number of days dominated by ozone pollution.

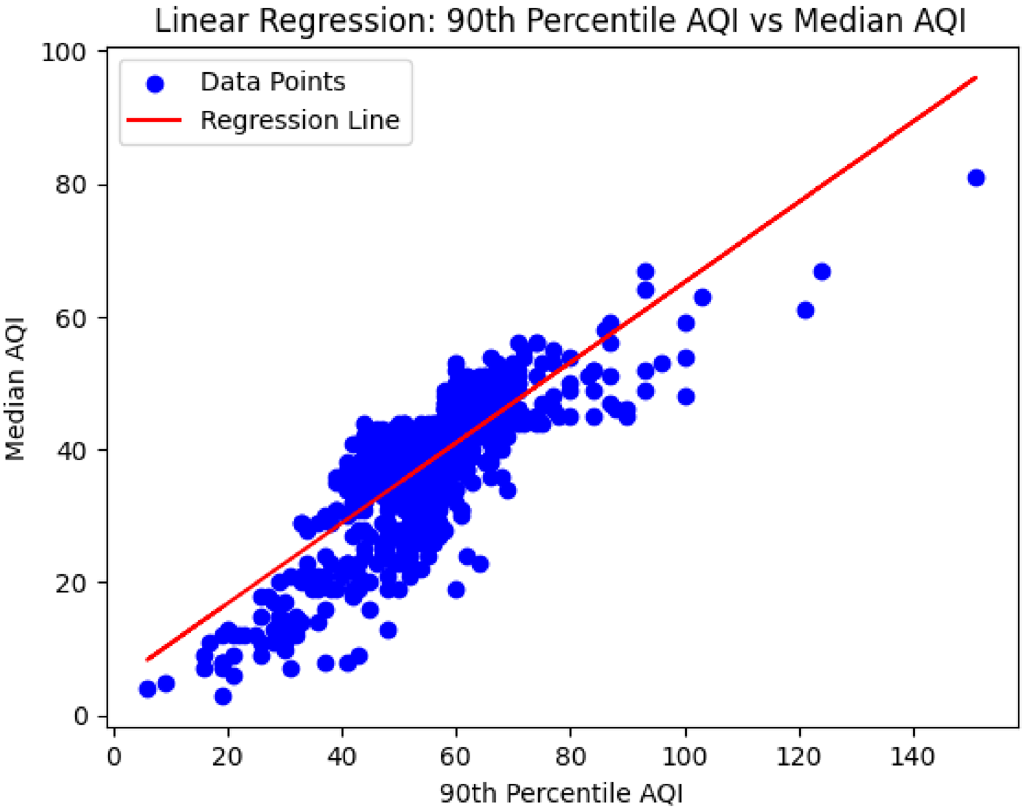





Linear Regression: Days with AQI vs Days PM2.5
Data Points
Regression Line
Days PM2.5
0
50
100
150
200
Days with AQI
0
50
100
150
200

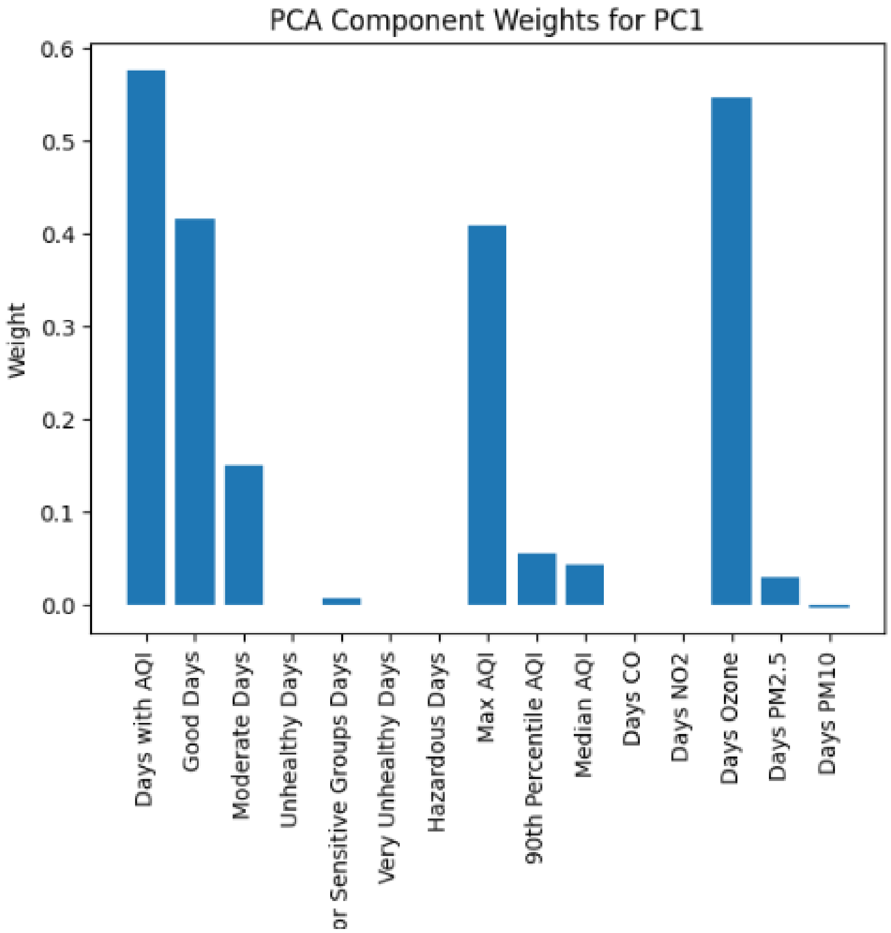

Linear Regression: Good Days vs Days Ozone
Data Points
Regression Line
Days Ozone
0
50
100
150
200
Good Days
0
25
50
75
100
125
150
175
200

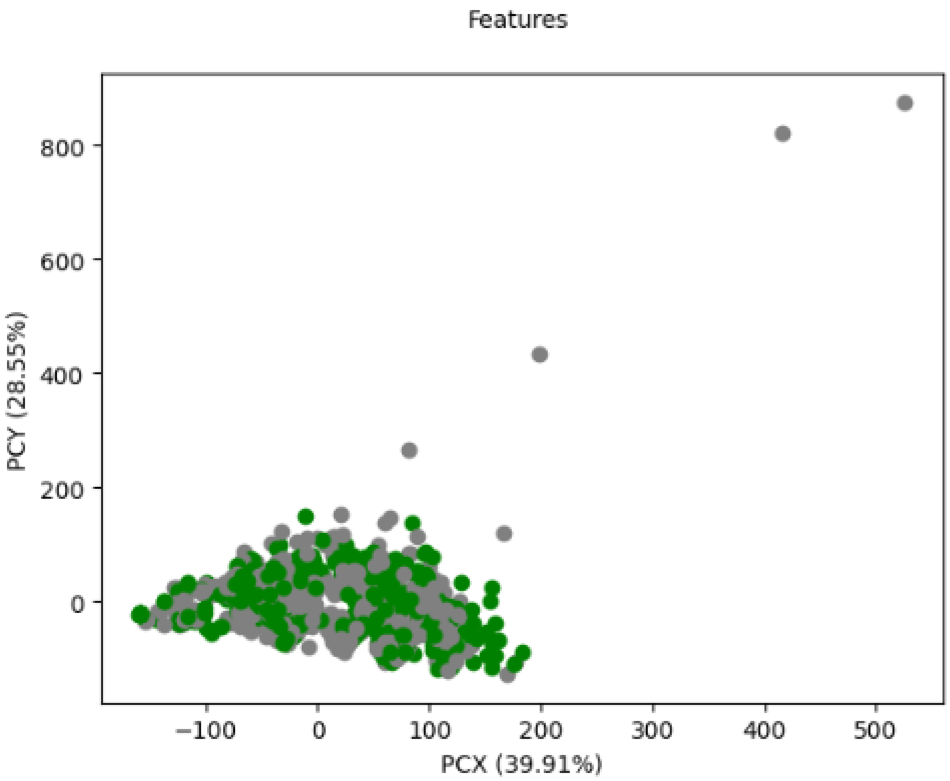

Linear Regression: Moderate Days vs Days PM2.5
Data Points
Regression Line
Days PM2.5
0
50
100
150
200
Moderate Days
0
20
40
60
80
100
120
140
160


Linear Regression: 90th Percentile AQI vs Median AQI
Data Points
Regression Line
Median AQI
0
20
40
60
80
100
90th Percentile AQI
0
20
40
60
80
100
120
140

Now, the Principal Component Analysis (PCA) is a dimensionality reduction technique that identifies patterns in large datasets by projecting them onto a smaller space. For this project, PCA was performed on quantitative air quality variables such as the number of AQI days, good days, and particulate matter days to identify patterns in air quality distribution across urban and rural states.

The grouping for PCA visualization was based on population density, using state populations as a reference. So, states with populations exceeding 6 million were denominated as urban, while those below this value were considered rural. This classification was chosen to explore how population density affects air quality metrics. Urban areas are expected to exhibit higher and more distinctive air quality behavior compared to rural areas.





The plot shows a highly scattered and overlapping distribution of states, with urban and rural categories failing to form clearly distinct clusters. This suggests that air quality variables may not differ sharply between urban and rural states or that other underlying factors, such as geographic or industrial variations, contribute significantly to the observed patterns. Interestingly, some urban states appear as outliers, isolated from the main cluster of rural states. This could indicate particular air quality challenges or monitoring practices in these areas. The overlap between categories may also reflect the influence of regional pollution transport or similar environmental regulations across states, which can confuse distinctions between urban and rural regions.

For this project, I have learned valuable insight into interpreting correlation coefficients, this made understand the strength and direction of relationships between variables. Additionally, I enhanced my coding skills by learning how to create, manipulate, and analyze datasets using Python. Finally, I developed a deeper understanding of the factors that influence the Air Quality Index (AQI), including population density, particulate matter levels, and the distribution of good and unhealthy air quality days, allowing me to get meaningful conclusions from the data.

The limitations of this data set and overall project were the classification of states as urban or rural based only on population may have oversimplified the complex factors that influence air quality. Factors such as industrial activity, transportation networks, and geographic features also play critical roles in shaping air quality. The analysis was conducted at the state level due to data limitations, which may have blurred variations in air quality. Also, the significant overlap between urban and rural states in the PCA scatter plot suggests that population density alone may not fully explain variations in air quality. This limits the strength of conclusions drawn about urban and rural distinctions.

For any future studies that will potentially analyze air quality, I suggest investing more time at the county level to capture finer graph readings and thus giving more spatial variation. This would allow for a more obvious understanding of how urban and rural areas within states contribute to overall air quality trends. Also, the inclusion of additional variables such as traffic density, industrial emissions, proximity to pollution sources, and meteorological data could provide a more comprehensive view of the factors influencing air quality.

Bibliography:

Strosnider H, Kennedy C, Monti M, Yip F. Rural and Urban Differences in Air Quality, 2008–2012, and Community Drinking Water Quality, 2010–2015 — United States. MMWR Surveill Summ 2017;66(No. SS-13):1–10. DOI: http://dx.doi.org/10.15585/mmwr.ss6613a1

Yaghjyan, L., Arao, R., Brokamp, C. *et al.* Association between air pollution and mammographic breast density in the Breast Cancer Surveilance Consortium. *Breast Cancer Res* **19**, 36 (2017). https://doi.org/10.1186/s13058-017-0828-3

Li, J., Han, L., Zhou, W. *et al.* Uncertainties in research between urban landscape and air quality: summary, demonstration, and expectation. *Landsc Ecol* **38**, 2475–2485 (2023). https://doi.org/10.1007/s10980-023-01744-5

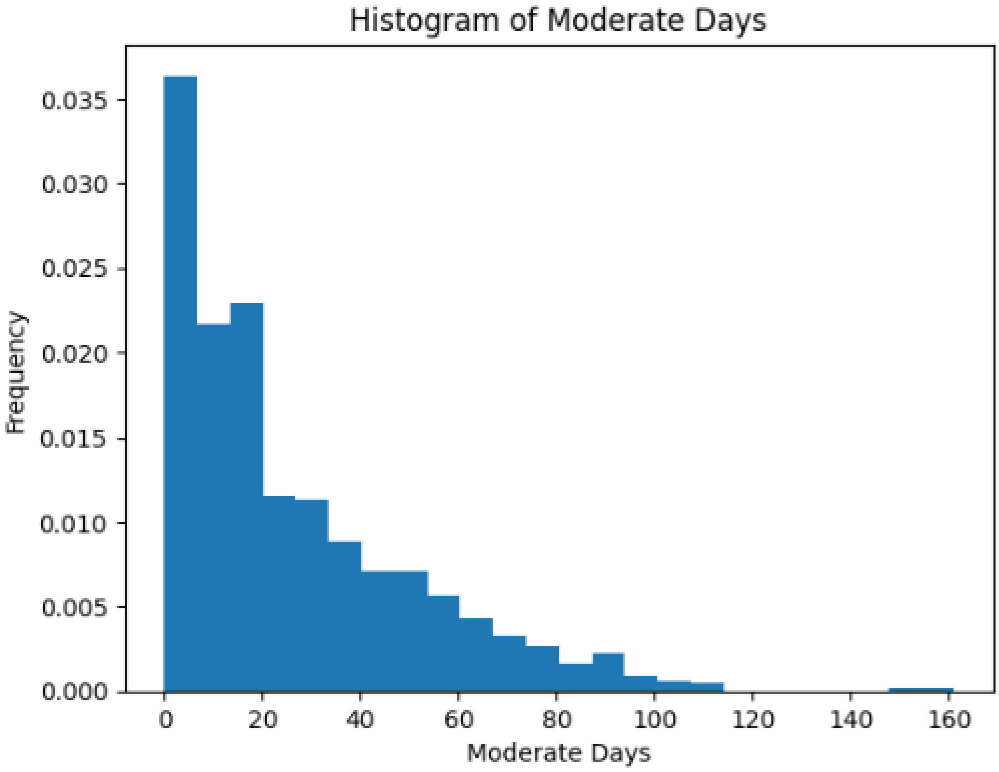
Escobedo, Francisco, Jennifer A. Seitz, and Wayne Zipperer. 2009. “Air Pollution Removal and Temperature Reduction by Gainesville’s Urban Forest: FOR216 FR278, 5 2009”. *EDIS* 2009 (5). Gainesville, FL. https://doi.org/10.32473/edis-fr278-2009.

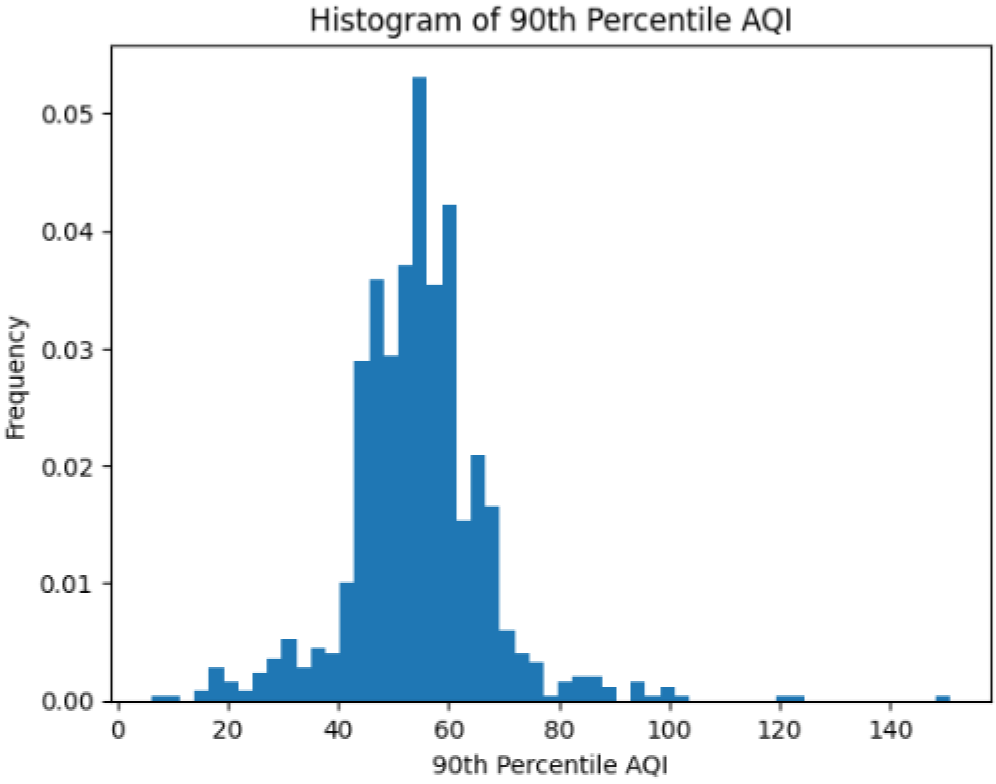
Borck, R., & Schrauth, P. (2019). *Population density and urban air quality*. CESifo Working Paper No. 7985. CESifo Group Munich. Retrieved from https://www.cesifo.org/en/publications/2019/working-paper/population-density-and-urban-air-quality

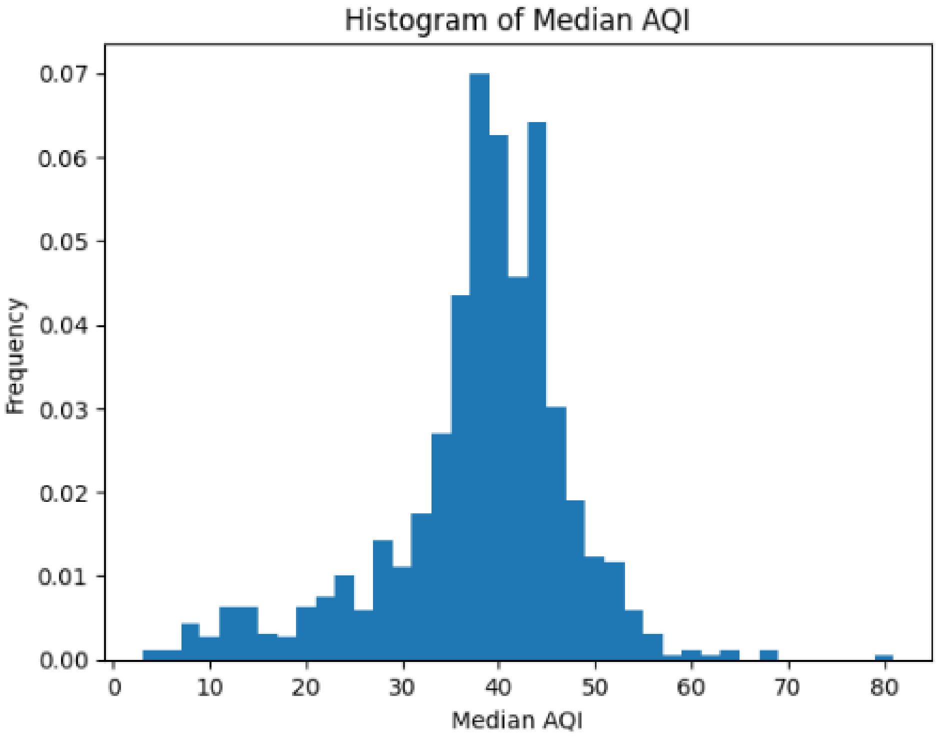
McKelvy, M., & Chapman, A. (2020). *Potential impacts of future urbanization and sea level rise on Florida's wildlife*. Journal of Fish and Wildlife Management, 11(1), 174 - 186. https://doi.org/10.3996/092019-JFWM-079

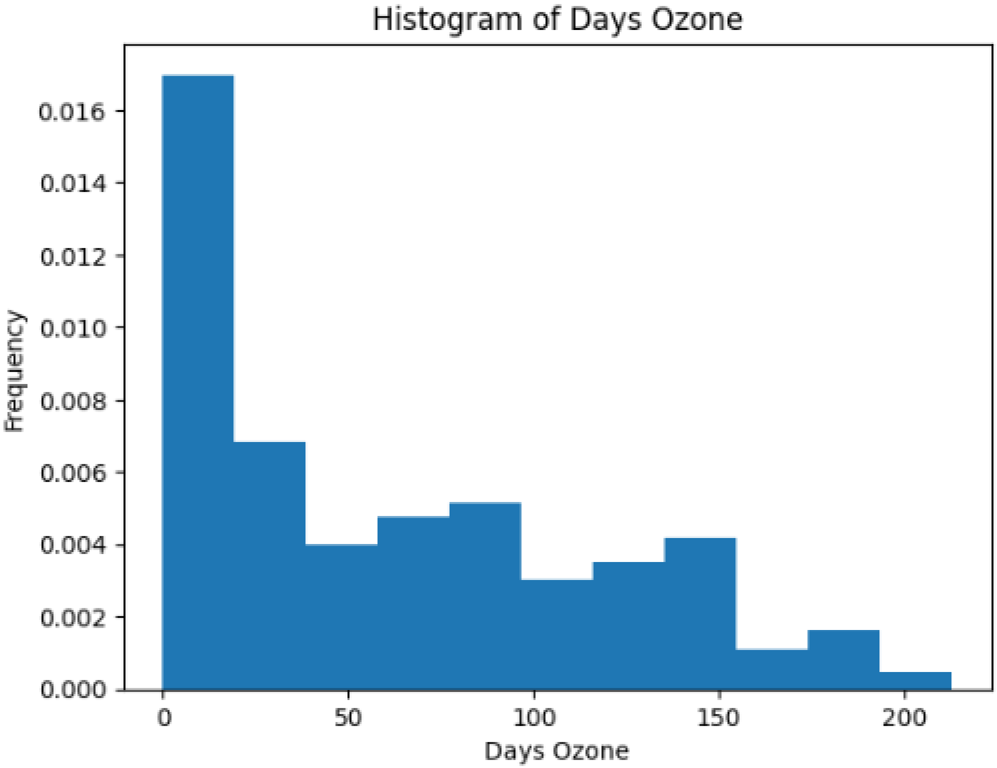
Nadarajah, M., Balakrishnan, P., & Ragavan, K. (2021). *Deep-AIR: A hybrid CNN-LSTM framework for air quality modeling in metropolitan cities*. arXiv. Retrieved from https://arxiv.org/abs/2103.14587

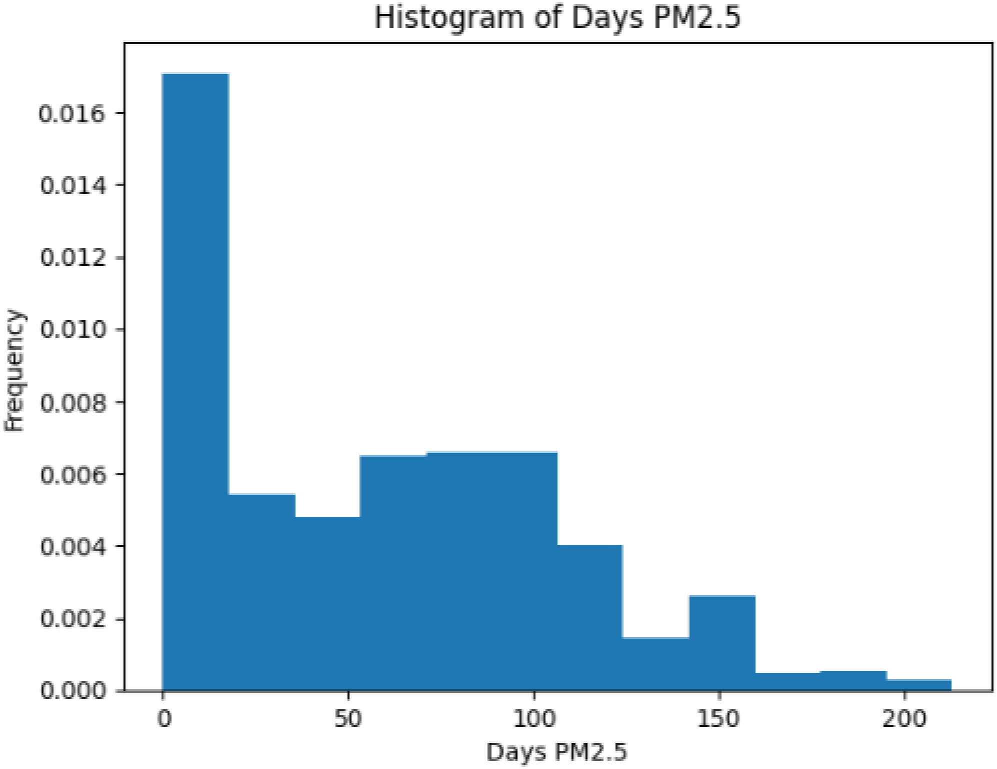
Supplementary Figures:

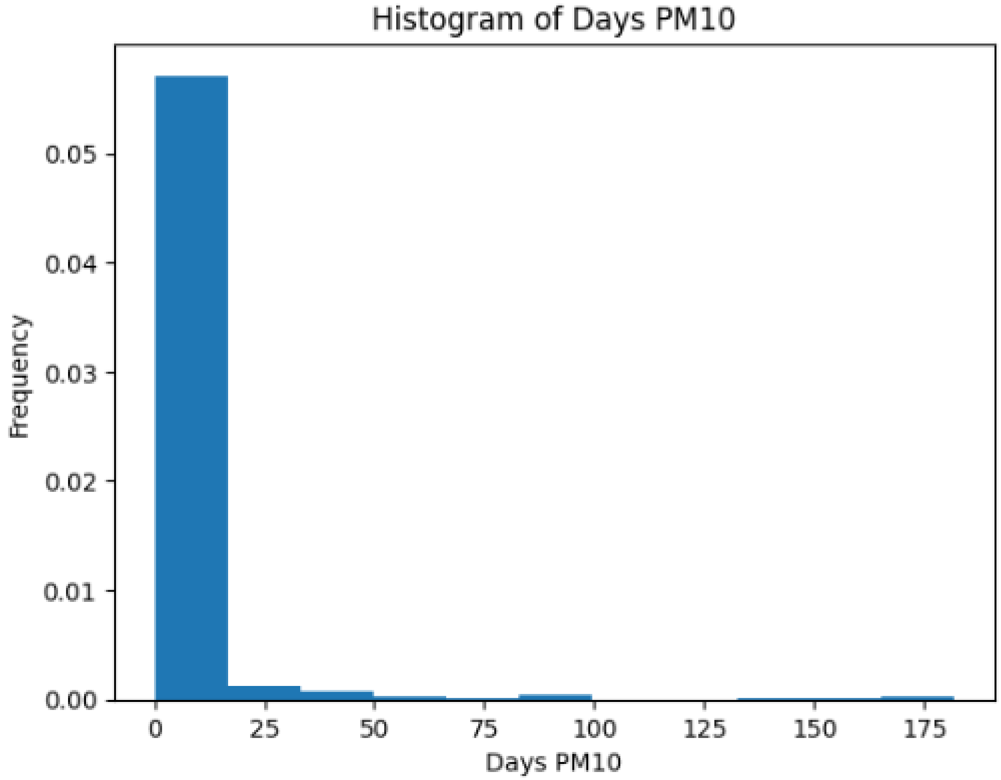



Histogram of 90th Percentile AQI
Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0
20
40
60
80
100
120
140
90th Percentile AQI


Histogram of Median AQI
Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
10
20
30
40
50
60
70
80
Median AQI


Histogram of Days Ozone
Frequency
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
50
100
150
200
Days Ozone


Histogram of Days PM2.5
Frequency
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
50
100
150
200
Days PM2.5


Histogram of Days PM10
Frequency
0.05
0.04
0.03
0.02
0.01
0.00
0
25
50
75
100
125
150
175
Days PM10