Air Quality Index in Houston

Nicole McCarthy

June 24, 2025

Abstract

This project builds upon previous work on the Houston metroplex climate investigations by attempting to correlate AQI and refinery activity in the surrounding regions. First, a linear regression model was built using crude oil prices as a proxy for refinery processing rates. This did not explain the variation in AQI. Second, principal component analysis was conducted and only 60% of the data's variation could be explained by the first two components. Additionally, AQI was not strongly correlated with the other climate measurements in determining the principal components.

Contents

Introduction

This project builds on previous projects for this specialization by continuing to work on climate data from Houston but expanding to identifying and predicting extreme climate behavior in context. In previous work, a potential correlation between air quality and the price of oil was speculated. This paper identifies longer trends in this lane and attempts to strengthen claims of correlation between air quality and trackable activities in Houston.

With over 2.3 million residents in Houston itself (not counting the sprawl of suburbs surrounding the downtown area), the city is home to a significant portion of Texans (United States Census Bureau, 2024). Houston is known for its vast array of highways and its energy industry. Prior to 1990, energy made up over 80% of jobs, leading to the popular nickname, “Energy Capital of the World.” Today, that percentage is closer to 50% as other industries have grown (City of Houston, 2025).

This paper attempts to identify the effect that the offshore drilling and refinery activities nearby have, if any, on the air quality. The rates of this activity are linked to the price of crude oil, which is well documented and easily accessible data. Higher prices for crude oil incentivize higher rates of drilling, though a drop in prices typically takes around two months before a drop in drilling is seen (United States Energy Information Administration, 2023).

More importantly, refinery activity releases pollutants into the air, including carbon dioxide, methane, nitrogen oxides, and sulfur dioxide (Kunak, 2025). Of these, nitrogen oxides and sulfur dioxides are among those compounds measured as pollutants when determining air quality (United Nations, 2022). Of the top ten most productive refineries in the United States, four are found in the surrounding regions of Houston (Zehl & Associates, 2025). These are

located in Port Arthur, Beaumont, Baytown, and Texas City (Figure 1). These refineries and others in the Houston metro area can process 2.6 million barrels of crude oil a day (Zehl & Associates, 2025).

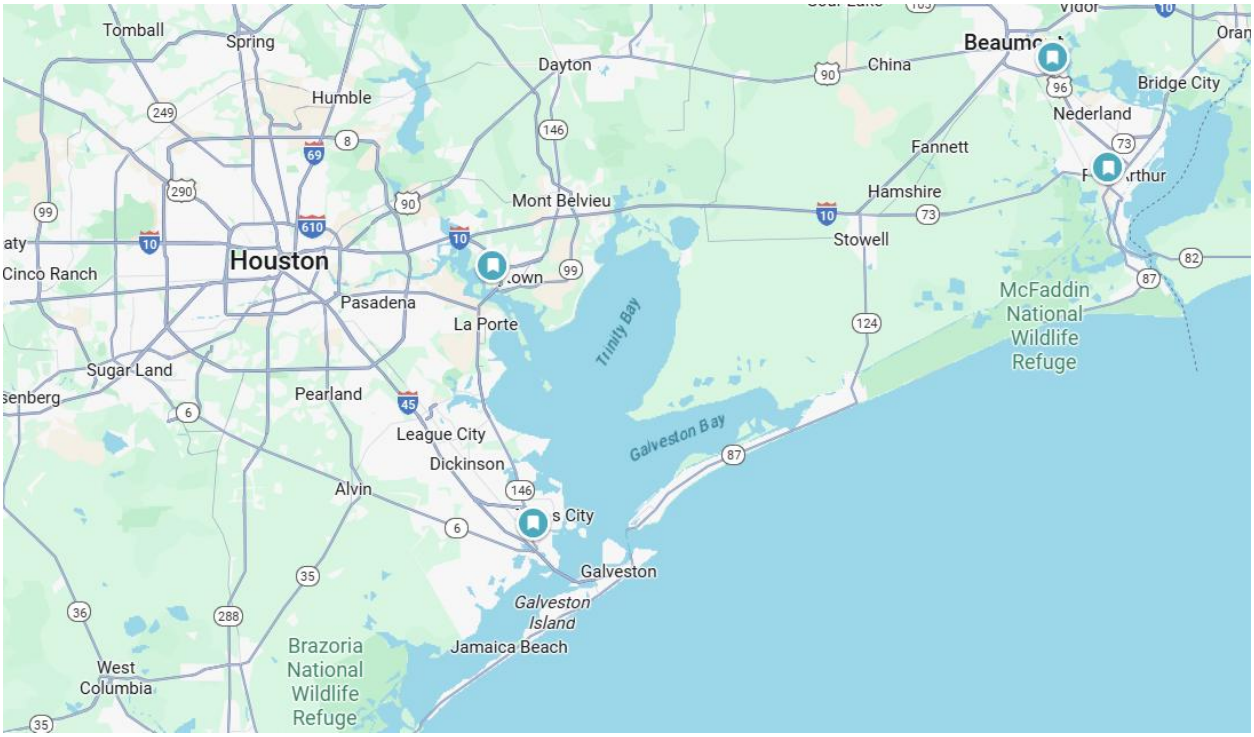

Figure 1. Locations of four of the top ten petroleum refineries in the United States

Other environmental factors are considered as well including hurricanes and wind speed and direction. Furthermore, a special look at 2020 and the following two years is investigated to determine whether car traffic is a major player in recent years of air quality. This time is useful for this part of the investigation due to the lockdown during the COVID-19 pandemic significantly decreasing the amount of driving Americans at large were doing (McCahill, 2023). Originally, gasoline prices at the pump were considered for analysis, but historically there is not a strong correlation between gas prices and driving behavior (Brand, 2009).

Methodology

The data for weather, including precipitation, temperature, and wind, were downloaded from NOAA. The data for crude oil price reflects dollars per barrel in Cushing, OK from the Energy Information Administration. Finally, data about air pollution came from the Environmental Protection Agency.

The supervised method, linear regression, was used to determine the relationship between air quality index (AQI) and refinery processing (estimated by the price of crude oil). The unsupervised method, principal component analysis (PCA), was used to quantify the relationship between air quality, crude oil prices, wind speed and direction, precipitation, and temperature.

Results

After loading in the AQI and crude oil price, the linear regression model had an r^2 score of 0.032 and a mean squared error of 178.184. A small positive correlation is seen in Figure 2. Though visually, this positive trend is promising, only 3% of the variation in the AQI can be explained by the crude oil price. This negligible effect nullifies the initial hypothesis that crude oil prices, as a measure of refinery processing rates, could explain Houston's air quality. To understand AQI better, we must turn to our unsupervised methods.

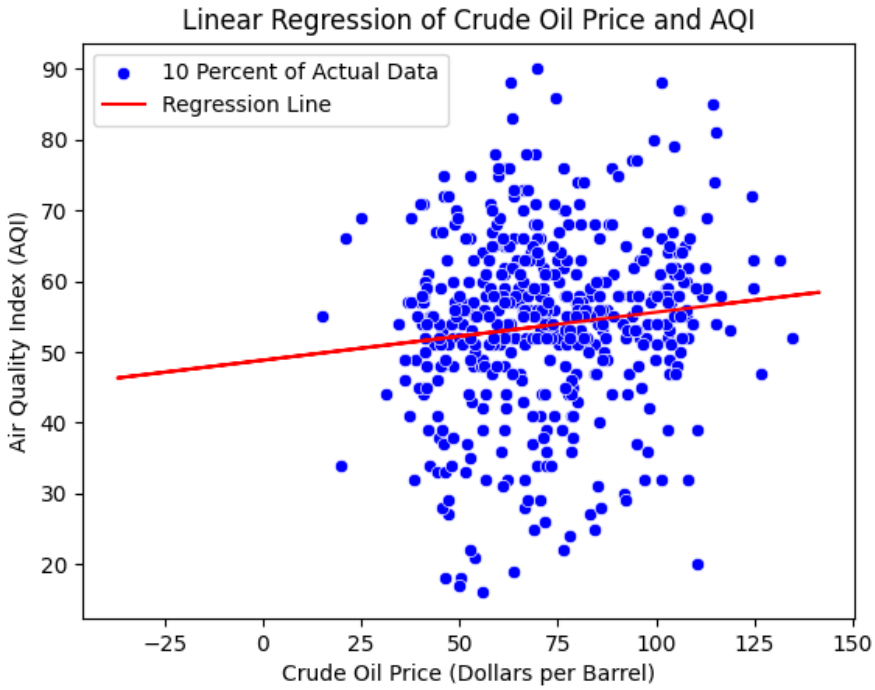

Figure 2. Linear Regression of Crude Oil Price and AQI for 2005-2024

The principal component analysis (PCA) on precipitation, wind speed and direction, temperature minimum and maximum, haze and dust, and AQI showed stronger correlations than the linear regression model but was still not completely explanatory. For instance, Figure 3 shows that only about 60% of the variation in the data could be explained by the first three components. This is rather unsatisfactory.

Within this 60%, wind speed (both average and peak) and minimum and maximum temperature explain the most variation within the top two principal components (Figure 4). This is perhaps unsurprising as these change the most throughout the year in the Houston climate. However, when looking at the biplot of the components in relation to the features, AQI is not strongly correlated with any other feature (Figure 5). Only dust and haze a barely correlated.

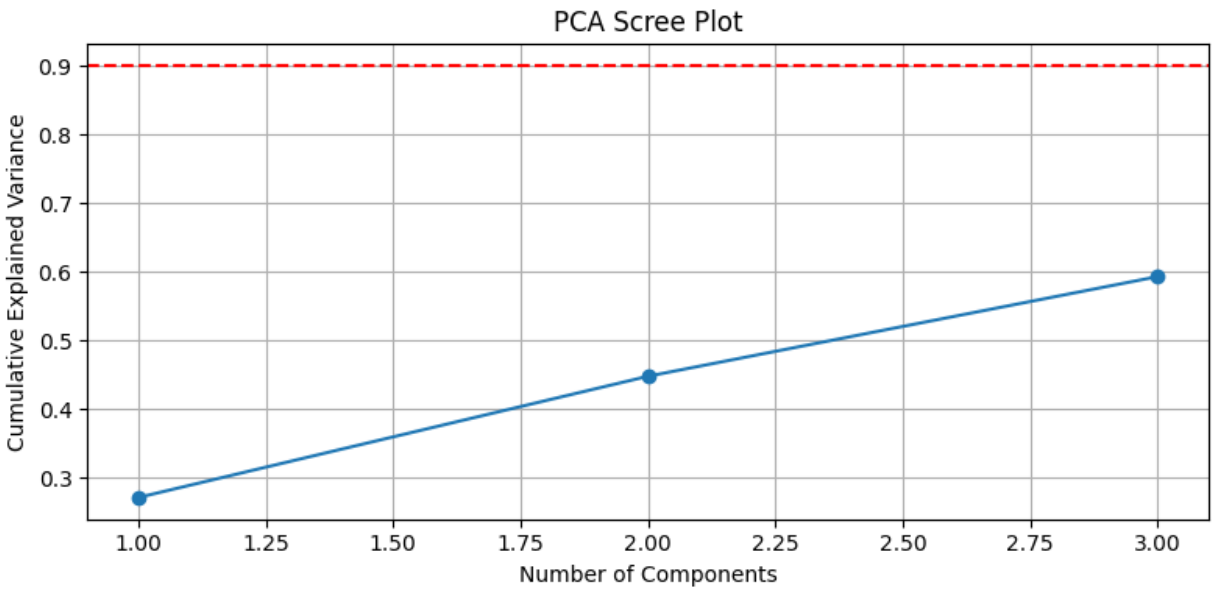

Figure 3. Scree Plot of PCA

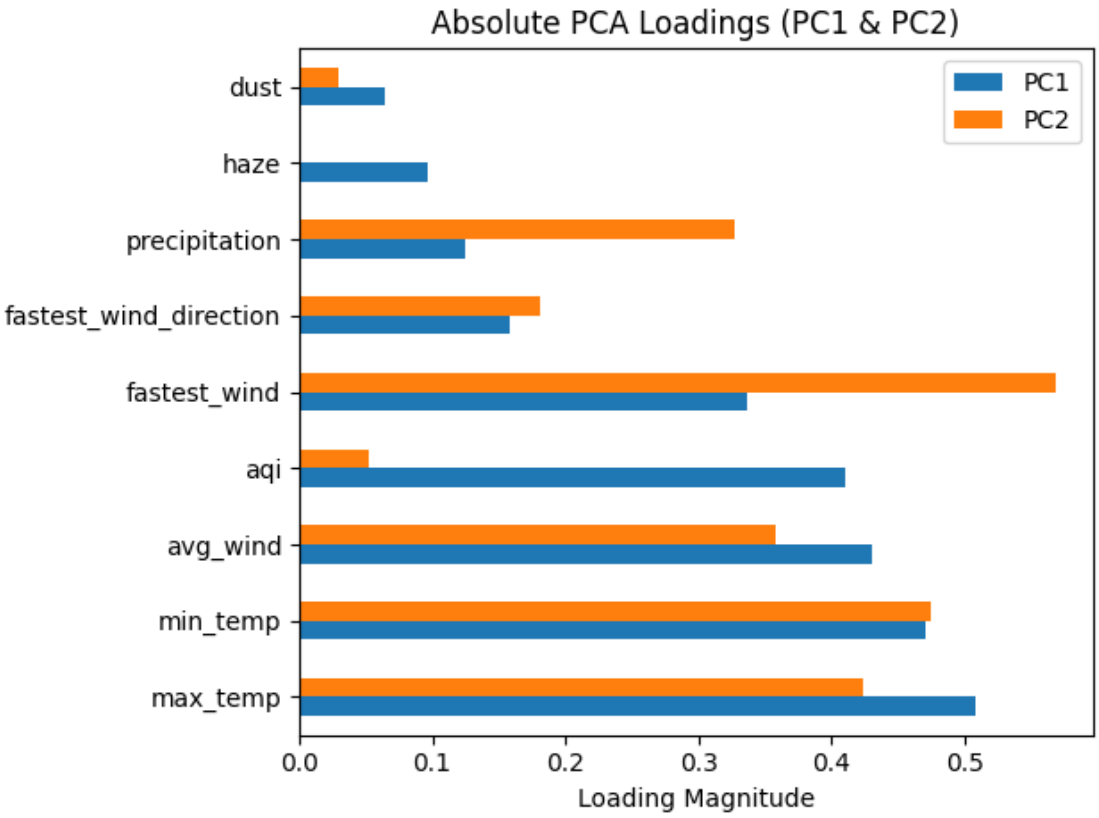

Figure 4. Feature Loadings of the Principal Components

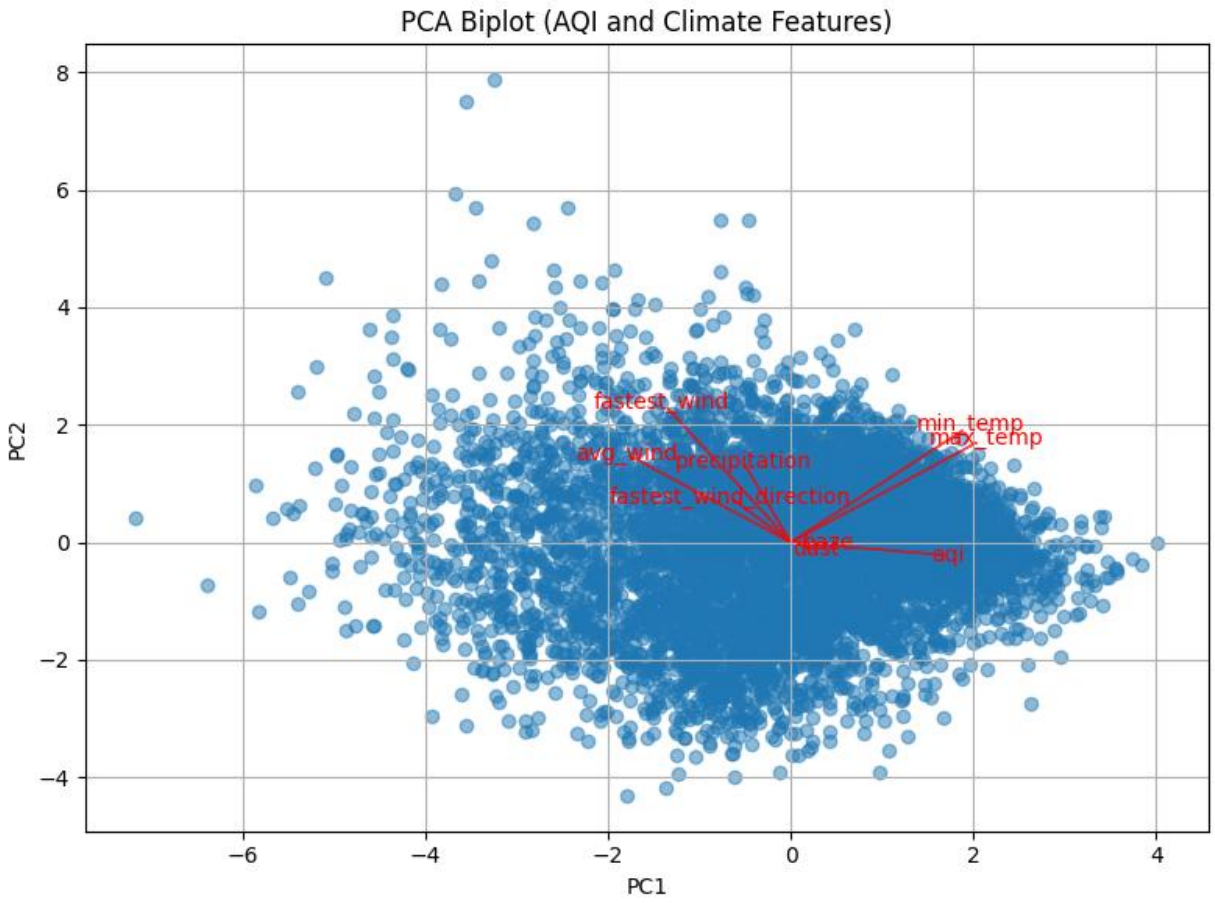

Figure 5. Biplot of principal components and feature influence

Limitations

This project was limited largely by the lack of data specific to refinery activity. Additionally, there were likely better explanations that could be found for AQI in Houston specifically, but the project was cut short due to time. Additionally, the primary component analysis was unsatisfactory in directly correlating AQI with the other climate features and a more direct clustering method would likely prove more beneficial. Finally, a multiple linear regression analysis for the climate features and AQI also holds potential.

Conclusion

From both methods, we can conclude that Houston's AQI is not strongly correlated with refinery activity in the surrounding areas. The linear regression model correlating crude oil prices (a proxy for refinery activity) and AQI only explained 3% of the variation in the data. Furthermore, the primary component analysis only represented about 60% of the data's variation and did not reveal a strong correlation between AQI and any other climate measurement. Had the AQI been correlated with wind speed or direction, we might've been able to suspect that refineries did in fact play a part in the air quality. However, neither case proved helpful in unmasking the major factors in determining air quality in Houston.

References

Brand, D. (2009). Impacts of Higher Fuel Costs." *U.S. Department of Transportation: Federal Highway Administration*. https://www.fhwa.dot.gov/policy/otps/innovation/issue1/impacts.cfm#:~:text=The%20average%20gas%20price%20increase%20of%2028%25,as%20a%20whole%2C%20slightly%20below%202006%20levels.

City of Houston. (2025). "About Houston: Business Overview." *City of Houston, Texas*. https://www.houstontx.gov/abouthouston/business.html/.

Kunak. (2025). "Oil Refinery Emissions: Environmental Impact and Monitoring Solutions." *Kunak*. https://kunakair.com/oil-refinery-emissions/.

McCahill, C. (2023). "Americans are still driving less than before the pandemic." *State Smart Transportation Initiative*. https://ssti.us/2023/03/06/americans-are-still-driving-less-than-before-the-pandemic/.

United Nations. (2022). "How is air quality measured?" *UN Environment Programme*. https://www.unep.org/news-and-stories/story/how-air-quality-measured.

United States Census Bureau. (2024.) "Quick Facts: Houston city, Texas." *United States Census Bureau*. https://www.census.gov/quickfacts/fact/table/houstoncitytexas/PST045224.

United States Energy Information Administration. (2023). "STEO Perspectives: How do different crude oil prices affect U. S. crude oil and natural gas production?" *Short Term Energy Outlook.* https://www.eia.gov/outlooks/steo/report/perspectives/2023/11-WTIprice/article.php.

Zehl & Associates. (2025). “4 of Nation’s 10 Largest Oil Refineries Located Along Texas Gulf.” *Zehl & Associates*. https://www.zehllaw.com/4-of-10-largest-us-oil-chemical-refineries-located-in-houston-beaumont-port-arthur/.