Air Quality & Pollution Final Report

DS3000

Jayden Ye, Esha Kanakapura, Jason Balayev, Ian Chung

Abstract

This report develops ML models to forecast the air quality of five air pollutants (PM10, PM2.5, CO, CO2, and ozone) based on key factors such as time-of-day and previous trends. Of all the regression models we built, Polynomial Regression with Interactions performed best; however, it showed poor performance and explained only 1% of the ozone concentration. In contrast, the predictions for the remaining pollutants, Polynomial Regression with Interactions, performed worse than a naive baseline that always predicts the mean concentration. Linear and Polynomial Regression yielded worse results, with prediction errors 5% to 15% larger than the mean-based baseline, highlighting that different tactics might be necessary, such as more advanced models that can account for more variables. Critical assumptions, including linearity, independence, and constant variance, were consistently proved wrong by the analysis, indicating that factors not included in our dataset may have an impact on pollutants. In all, our results highlight the difficulties in forecasting actual air pollution with the limited resources and models that we had, and called for a more advanced model that could keep up with the different factors that affect air pollution.

Introduction

Air pollution is on the rise in developed and urban areas such as Boston. Exposure to these air pollutants over time can cause long-term health issues and worsen the conditions of vulnerable communities. Understanding these trends can offer valuable information as to how air pollutant patterns vary across different times of day within the Boston area. Visualizations of the data can also make the reality of air pollution more apparent to the public. This analysis will focus on identifying trends in air pollution across an extended period of time for certain air pollutants (PM10, PM2.5, CO, CO2, and ozone) in the Boston area. Additionally, we aim to forecast pollutant levels through machine learning models, which can educate the Boston community and improve quality of life for vulnerable populations. These were our key questions:

1. How does time of day affect air pollution?
2. Can we use previous data to predict when a spike will occur in air pollution?

Data Description

Summary of the Data Processing Pipeline:

1. Obtain cleaned data from the API, including time of day, pollutant concentrations, and location
2. Visualize the data using Seaborn, Plotly, and Matplotlib

3. Apply machine learning models and evaluate performance using regression analysis

To start, we extracted data from the Open-Meteo Weather API by setting up and requesting information from various categories in a Jupyter Notebook. This involved deciding which data points were applicable to answering the question at hand. In order to compare air pollution levels, we needed to choose categories that best represented harmful but common pollutants. These were: particulate matter 10, particulate matter 2.5, carbon monoxide, carbon dioxide, and ozone. The API had information on the levels of these pollutants throughout hours spanning several days, this being the time of day data that allowed us to forecast. Additionally, we wanted to look into urban areas where air pollution tends to be high, for which we chose the city of Boston which the API contained information. We observed that the data was already clean, without any NaN or 0.0 values, so we didn't need to prepare it further on that front. However, we did need to convert Unix timestamps from the API into pandas datetime objects.

After deciding what to extract, we called our requests from the API. Calling the location of Boston required the latitude and longitude (42.3398°N, -71.0892) of the city. Next, we requested every required pollutant and its hourly data within the area. In order to predict the concentration of a pollutant at some given hour, we created a pandas data frame with a date column containing hourly timestamps from start to end as well as columns for each air quality measurement. This was to help answer our second question so the spikes at certain hours could be more apparent. To further analyze our data, we created visualizations. This included two heatmap diagrams to show PM2.5 and ozone concentrations between the different hours and days in the Boston area. Darker colors showed higher pollution levels and lighter colors showed lower concentrations. This also included a line plot that converted the 6 hour timeframes of PM2.5 and ozone into averages to help simplify the data. This made it possible to observe trends throughout the day to determine if there are spikes for pollution at certain phases of the day. Finally, we prepared our data for machine learning by selecting the relevant features of time of day as well as each pollutant. We had multiple pollutants to predict, so we created 5 predictive models for each regression model. These regression models included linear regression, polynomial regression, and polynomial regression with interactions. Each prediction model had its own mean square error and coefficient of determination, which helped us see what our models predicted best and worst.

Hourly data

	date	pm10	pm2_5	carbon_monoxide	carbon_dioxide	ozone
0	2025-11-10 00:00:00+00:00	11.5	9.1	202.0	449.0	74.0
1	2025-11-10 01:00:00+00:00	6.5	5.3	198.0	449.0	77.0
2	2025-11-10 02:00:00+00:00	4.1	3.5	201.0	449.0	81.0
3	2025-11-10 03:00:00+00:00	4.1	3.6	205.0	451.0	82.0
4	2025-11-10 04:00:00+00:00	4.8	4.2	209.0	454.0	77.0
..	...	...	...	...	...	...
139	2025-11-15 19:00:00+00:00	2.3	2.3	221.0	450.0	57.0
140	2025-11-15 20:00:00+00:00	2.4	2.4	250.0	452.0	50.0
141	2025-11-15 21:00:00+00:00	4.7	4.6	281.0	454.0	42.0
142	2025-11-15 22:00:00+00:00	9.7	9.6	317.0	458.0	32.0
143	2025-11-15 23:00:00+00:00	10.3	10.1	355.0	463.0	22.0

Example of processed data for November 2025 in Boston

Methods

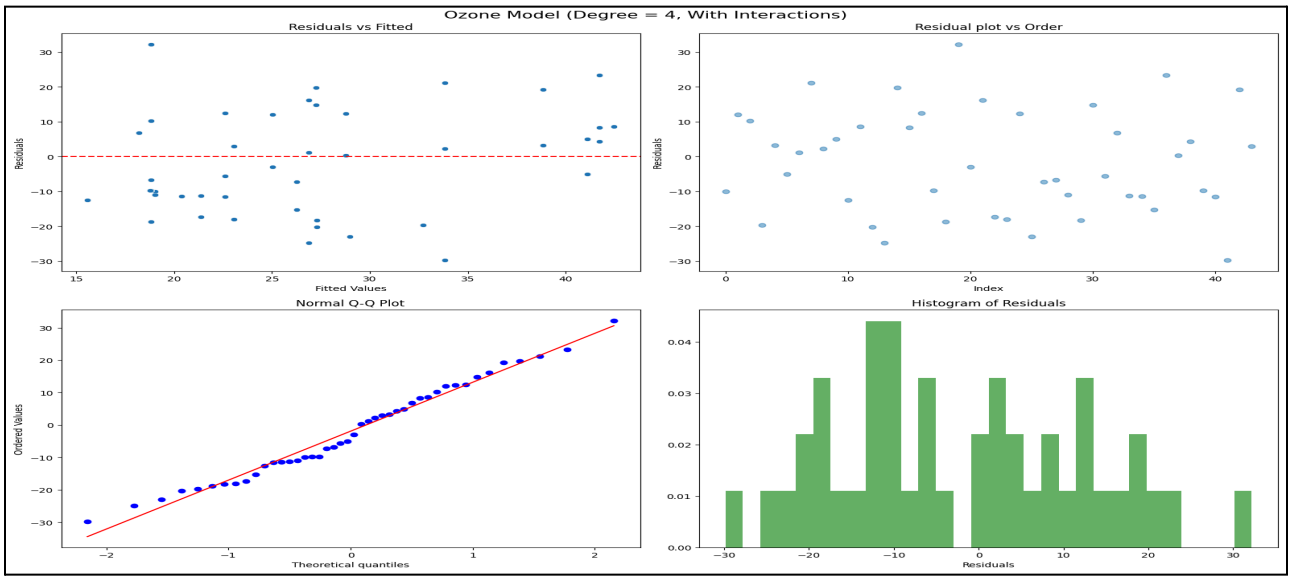
In order to determine if hourly air pollution could be predicted by machine learning, we developed three different regression models. Each model was evaluated and tested based on multiple air pollutants (Ozone, PM, CO, and CO2). Our main research question was “Can we predict air quality indicators (ozone, particulate matter, and carbon emissions) based on the hour of the day in Boston?” Because air pollution varies across different times and atmospheric conditions, each model that we developed was tested for: Residuals vs. Fitted Plots, Residuals vs. Order Plots, Q-Q Plots, and Histograms of Residuals.

For the first model, we used hour-of-day as the predictor for Ozone levels, and used polynomial regression for Ozone prediction. Polynomial regression allows predictions to be easy to interpret, but it is sensitive to violated independent assumptions. Since Ozone formation depends on sunlight and other photochemical processes, we expected non-linearity and potential hourly dependence through this model. For our second model, we used polynomial regression to predict Particulate Matter (PM10 & PM2.5). Particulate Matter often rises sharply at certain times of the day (e.g., early morning), so we expected a curved model, but potentially results more suitable than a linear one. For our third model, we used polynomial regression to determine and predict hourly interactions of Carbon Dioxide and Carbon Monoxide. CO and CO2 typically change based on traffic cycles, human activity, and mixing height, so we expected this to be the most difficult to predict.

For all three of the models, the performance was measured by Mean Squared Error (MSE), Coefficient of determination (R^2), and visuals of residual patterns. This allowed us to assess whether the results from the models are effective and provide meaningful predictions. We chose these regression methods because they allow us to test whether hour-of-day has a measurable relationship with each air pollutant. The main assumptions of these models (linearity, independence, constant variance, and normality) can be violated in air pollution data, which may reduce accuracy. However, these methods were the most appropriate because they provide interpretable statistics and clear results that reveal whether time-based prediction is feasible.

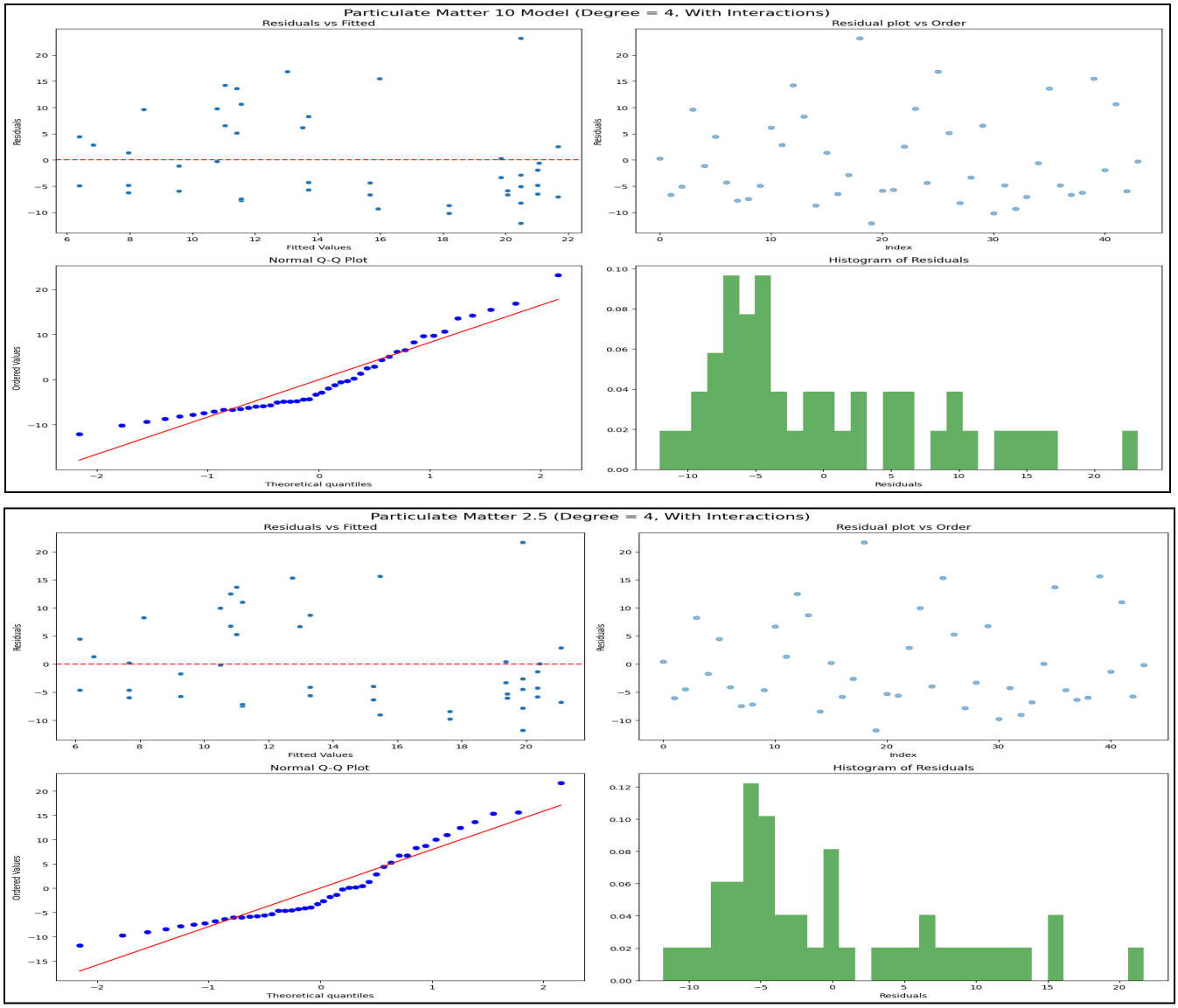
Results

For the first model (Ozone of Polynomial Regression with interaction terms), we were able to calculate MSE = 432.993 and R^2 = 0.0104, where the MSE value is lower than for the other polynomial regression models, and the R^2 value is slightly lower than the second polynomial regression model.



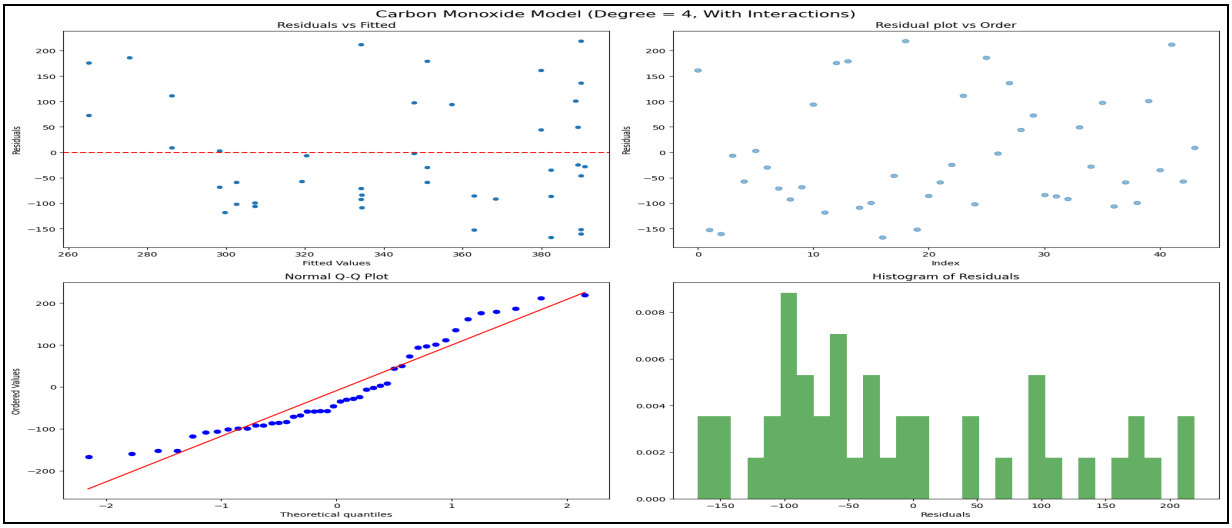
Based on the first plot of Residuals vs Fitted, we can see that the assumption of Linearity and Constant Variance is contradicted as the residuals do not follow a linear trend about 0 and have multiple outliers. The second plot of Residuals vs Order shows that there is some density of the data points near the y = 0 with multiple outliers as well, indicating there might be a violation of the independence assumption as well. Looking at the Histogram of Residuals and Q-q plot, we can witness there is strong right-skewedness, showing that there is no normal distribution, meaning a violation of normality assumption. These findings reflect the Ozone's dependence on factors other than just the hour alone, especially sunlight, weather conditions, etc.

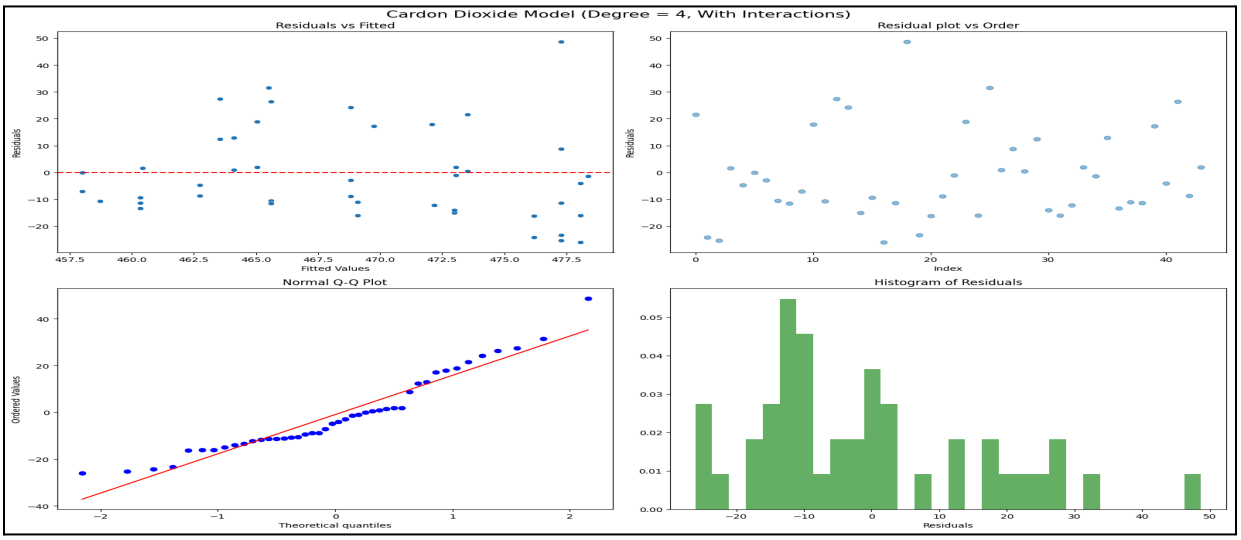
For the second model (Particulate Matter of Polynomial Regression with interaction terms), we calculated MSE = 122.559 and R^2 = -0.1708. For PM2.5, we calcuated MSE = 109.98 and R^2 = -0.1452. Both values showed negative R^2 values even after expanding the model to polynomial terms. These calculations show and confirm that hour-of-day alone does not predict accurate PM levels.



The Residuals vs Fitted plot shows scatter without a trend but with inconsistency. The Residuals vs Order suggests weak independence because PM levels depend on external environment factors such as wind, humidity, etc. The Q-Q plots and histograms show deviations from normality where both models have skewed distributions. These findings emphasize that PM levels change due to complex human factors/activities that cannot be approximated or predicted just by temporal change alone.

For our third model (Carbon Monoxide / Carbon Dioxide Polynomial Regression model), we calculated: For Carbon Monoxide, we calculated MSE = 15802.879 and R^2 = -0.0738. For Carbon Dioxide, we calculated MSE = 369.704 and R^2 = -0.0647. Both carbon measurements produced negative R^2 values and showed large variance due to the large scale of measurements.





For both models, the Residual patterns were showing significant spread and several outliers. The Q-Q plots showed right-skewing, and the Histograms were irregular, which shows that both carbon calculations are not suited for hour-based regression. These pollutants change mainly due to human activities (traffic, industrial activities, etc), and rather than just hourly emissions, this explains the weak model performance.

Discussion

The findings that we've gotten help highlight how complex a problem we currently have at hand. Polynomial regression with interaction terms yielded our best results; however violates several critical assumptions. These violations of assumptions reveal the issues with the model we used.

The various assumption violations in our model reveal issues in our approach to air quality prediction. The patterns and violations observed in the Histogram of Residuals for Particulate

Matter 2.5 show skewed distributions and no clear normal curve. This can help suggest that our model isn't equipped or complex enough to handle the various factors that can affect PM 2.5 and other indicators of air pollution. For instance, just based on the data given alone, the model has no way of accounting for human activities. If construction is being done in the area, leading to an increase in air pollution, the model has no way of knowing that and adjusting data accordingly.

Additionally, the heteroscedasticity observed in our model suggests that the accuracy of our model can vary under different atmospheric conditions. Under stable and constant weather conditions, our model may be able to accurately predict the concentrations of pollutants in the air; however, with other complex factors in account may not. Human activities and outlier atmospheric conditions may throw off the model since it isn't capable of taking such things into account. These little factors that can all contribute to differences in air pollution are very complex, and all help highlight the complexities in trying to model such a vast system.

Our low R^2 values were a little alarming. For ozone, our best model only explained 1% of the variance in concentrations, while the other models for PM 2.5, PM 10, Carbon Dioxide, and Carbon Monoxide all produced negative R^2 values ranging from -0.0647 to -0.1708. This meant that the model was better off just predicting the mean pollution level every time.

Based on our low R^2 values along with multiple violations of critical assumptions, the information given from our model shouldn't be taken at face value. Essentially, every critical regression assumption was violated, and in combination with the low R^2 values, this undermines the validity of the results given by the model.

Ethical concerns

- False alarms: If our model were to falsely predict a major spike in air pollution that caused vulnerable groups (such as the elderly, children, or those with health conditions) to evacuate, it could cause panic within the area and could potentially affect work/school for those.
- On the other hand, if our model wasn't accurate enough to predict a major air pollution spike, vulnerable groups could unknowingly be negatively affected. This could cause them to experience increased levels of air pollution and damage their health, especially if they are more sensitive to those pollutants.

We can address these ethical concerns by overall improving our models to account for more complex factors such as human activity. By improving our models, we can better predict spikes in air pollution and be able to better help the public.

https://github.com/JasonBalayev/DS-3000-Final-Project