Pollutants and Climate: Predictive Modeling in NYC

Victor Gavrilenko Ilay Cohen Shay Harush
Lidor Mashiach
Student IDs: [209406255, 206515744 , 314804287 , 209280098]

June 2025

Abstract

Climate change has been a major concern in recent decades, and in our project, we aimed to explore one specific aspect of this global issue: the relationship between pollutants and climate. Our hypothesis was that changes in pollution levels directly influence climate variables, in particularly temperature. To investigate this, we focused on a defined geographic and temporal scope - New York City during the years 2016, 2018, 2020, and 2021. We specifically selected these years to examine potential environmental impacts during the COVID-19 pandemic, a time when urban activity, and presumably pollution, declined significantly.

We collected air pollutant data from the U.S. Environmental Protection Agency (EPA) and weather data from a public open-sorce weather API named open-mateo . After extensive preprocessing-filtering for relevant boroughs near NYC (Bronx, Queens), removing distant counties, merging datasets by location and date, and handling missing values - we obtained a clean dataset of 2,896 records. Key features included pollutant concentrations (e.g., Ozone, NO), Air Quality Index (AQI), and various climate variables such as temperature and precipitation.

Exploratory data analysis revealed significant insights, particularly a strong positive correlation between ozone levels and temperature. We also observed that pollutant levels dropped in 2020, supporting our hypothesis about reduced human activity during the pandemic.

To predict temperature from pollutant data, we developed a neural network model (MLP) named VISL and benchmarked it against traditional models like Linear Regression and Random Forest. While the VISL model performed reasonably well, its results were not yet satisfactory. This suggests that improvements are needed. In future work, we plan to enhance the model by engineering additional relevant features (e.g., urban activity indicators) and improving the architecture to better capture complex relationships. Code is available in VISL github repository[1]

1 Related Work

Air pollution interacts in complex ways with weather conditions, a topic that has drawn increasing attention in recent literature. We review key contributions that inform our analysis and distinguish our study in scope, scale, and methodology.

He et al. (2024): *Air Pollution Interactions with Weather and Climate Extremes.* This work provides a global synthesis of how aerosols and gaseous pollutants influence extreme weather events such as heatwaves and heavy rainfall, and highlights feedback loops between pollution and meteorological dynamics. Their review supports our hypothesis that air pollution can both impact and be influenced by local weather conditions. However, unlike their broad, literature-focused perspective, we conduct a detailed, city-specific empirical analysis using multi-year data from New York City. Moreover, we integrate an AI-based forecasting framework to quantify and predict pollutant–weather interactions.

[1] https://github.com/Gavision97/VISL_Project.git

Berman and Ebisu (2020): *Changes in U.S. Air Pollution During the COVID-19 Pandemic.* This study analyzes NO_2 and $PM_{2.5}$ reductions across U.S. counties during early 2020 lockdowns, comparing them statistically to 2017–2019 baselines. We adopt their statistical framework to assess significant declines in pollution during NYC's lockdown period, while extending the analysis to include 2021 rebound effects. Our approach also differs by explicitly linking pollution trends to concurrent changes in weather conditions, in particular temperature.

Pitiranggon et al. (2022): *Effects of the COVID-19 Shutdown on Spatial and Temporal Patterns of Air Pollution in New York City.* Employing difference-in-difference and land-use regression methods, this study maps $PM_{2.5}$ and NO_2 concentrations across 93 monitoring sites in NYC during March–June 2020. Their spatial-temporal methodology informs our own neighborhood-level mapping of pollutant–weather correlations. We expand on their work by including four pollutants (adding CO and O_3), extending the time frame to four years, and introducing machine learning models to forecast meteorological outcomes based on spatial pollution trends.

Liu et al. (2020): *Exploring the Relationship Between Air Pollution and Meteorological Conditions in China.* Using data from 896 Chinese monitoring stations (2014–2019), this paper examines how temperature, wind speed, and humidity affect pollutant concentrations ($PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3) under various environmental policy conditions. Their spatial interpolation and seasonal analysis inform our method for coupling daily weather and pollution data in NYC. In contrast, we focus on daily city-wide time series, include COVID-era comparison periods, and incorporate AI models for predictive insights.

Ngarambe et al. (2021): *Air Pollution and Urban Heat Island Dynamics in Seoul, Korea.* This work analyzes nine years of pollutant and temperature data from 13 stations in Seoul, employing regression and $ANOVA$ to explore how the urban heat island (UHI) effect correlates with pollutant concentrations. Their focus on pollutant–temperature interactions parallels our study, though we broaden the scope to include humidity and precipitation. Additionally, we examine a critical four-year period that includes the COVID-19 lockdown and recovery, applying machine learning to yield both descriptive and predictive analyses of NYC's pollutant–weather dynamics.

2 Data

2.1 Data Collection

Data were collected from two different sources: air pollutant data from the U.S. Environmental Protection Agency (EPA) [2] and weather data from a public open-source weather API named open-mateo[3]. We first concatenated individual pollutant datasets (four years for each pollutant, resulting in 16 datasets) to create one dataset per pollutant - O_3, CO, NO_2, and $PM_{2.5}$.

2.2 Data Preprocessing

We had a total of 16 pollutant datasets. This is because we considered four pollutants (O_3, CO, NO_2, and $PM_{2.5}$), and for each pollutant, we had separate datasets for the years 2016, 2018, 2020, and 2021. First, we concatenated the four yearly datasets for each pollutant, resulting in one combined dataset per pollutant - leaving us with four datasets in total. Next, The four pollutant datasets were merged based on common attributes: date, county, latitude, and longitude. We retained data only for Bronx and Queens counties, which are geographically close enough to NYC to align with our weather dataset. The merged pollutant dataset was then joined with the weather dataset using the date feature to obtain the final dataset.

Next, we handled the missing values, which were present only in the weather dataset (173 in total). These were removed to ensure data consistency and model readiness.

[2] https://www.epa.gov/outdoor-air-quality-data/download-daily-data

[3] https://open-meteo.com/en/docs/historical-weather-api

2.2.1 Dataset Normalization

Min-Max normalization was applied to all numerical features to rescale values to the [0, 1] range, which improves model convergence and performance. The formula used for Min-Max scaling is:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where X is the original feature value, $X_{\min}$ and $X_{\max}$ are the minimum and maximum values of the feature, and X' is the normalized value.

3 Methodology

3.1 Data Visualization

3.1.1 Correlation Heatmap

From the figure below, we can see a clear correlation between ozone (i.e., O_3) and temperature, suggesting that ozone levels may influence temperature values. Moreover, there is noticeable correlation among the pollutants. For example, a correlation of nearly 0.7 between CO and NO_2. We also observe expected correlations among various weather factors, which aligns with our assumptions. Moreover, there is no correlation between any pollutant and rain precipitation - we will not include those label in our future ML model.

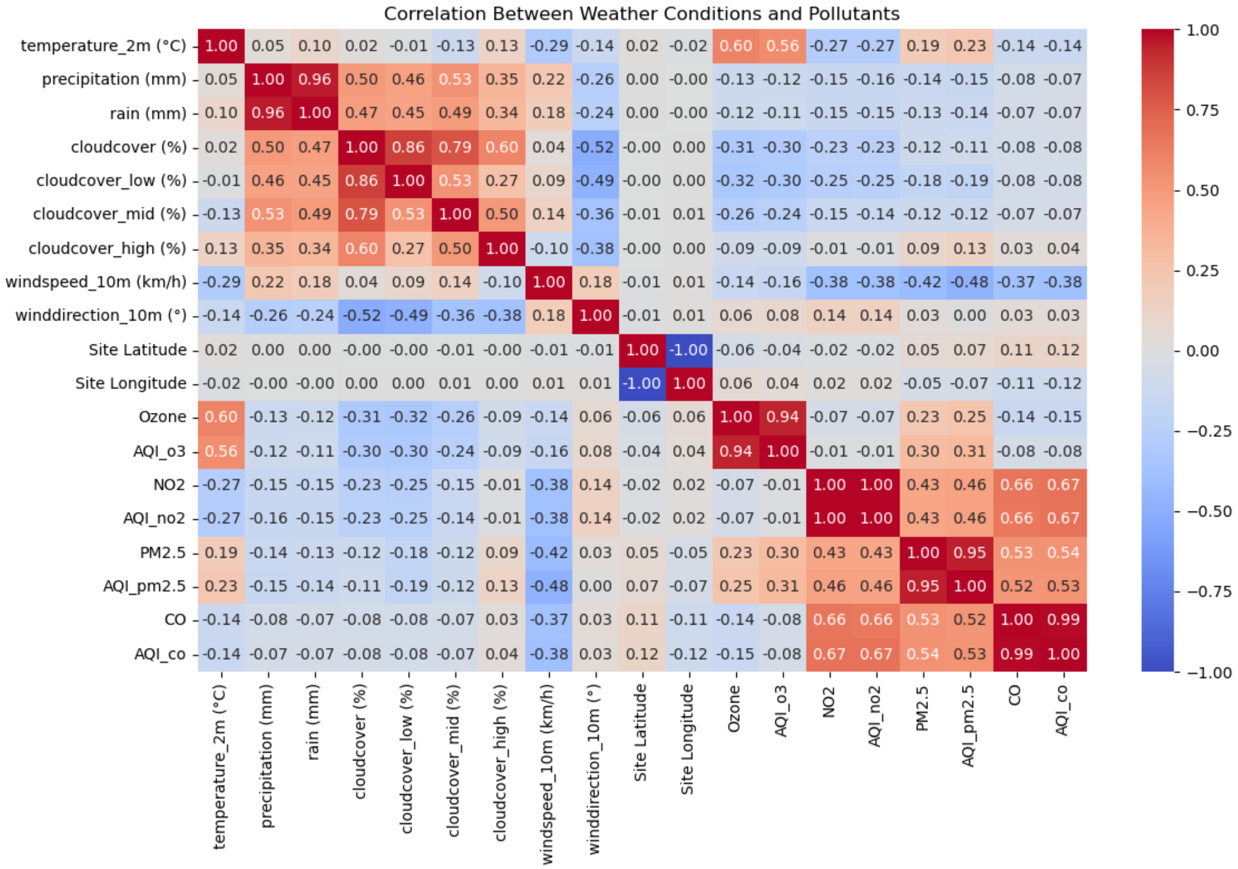

Figure 1: Correlation heatmap

3.1.2 Correlation Between Ozone and Temperature

In the figure below, we can observe a slight correlation between ozone and temperature. Specifically, when the temperature is low, ozone levels are also low, and as the temperature increases, ozone levels tend to rise as well

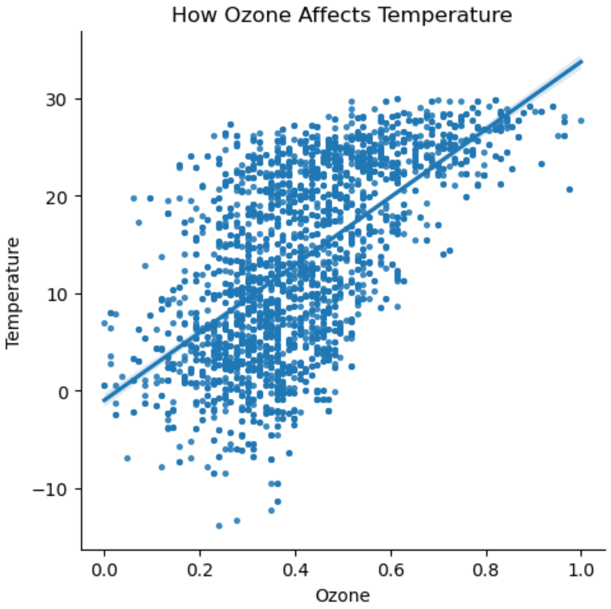

Figure 2: Correlation between ozone and temperature

3.1.3 Yearly Averages of Pollutants

The figure below clearly shows that the average value of each individual pollutant was at its lowest in 2020 (except O_3). This supports our earlier discussion and further suggests that pollution levels significantly dropped during that year (pick COVID year).

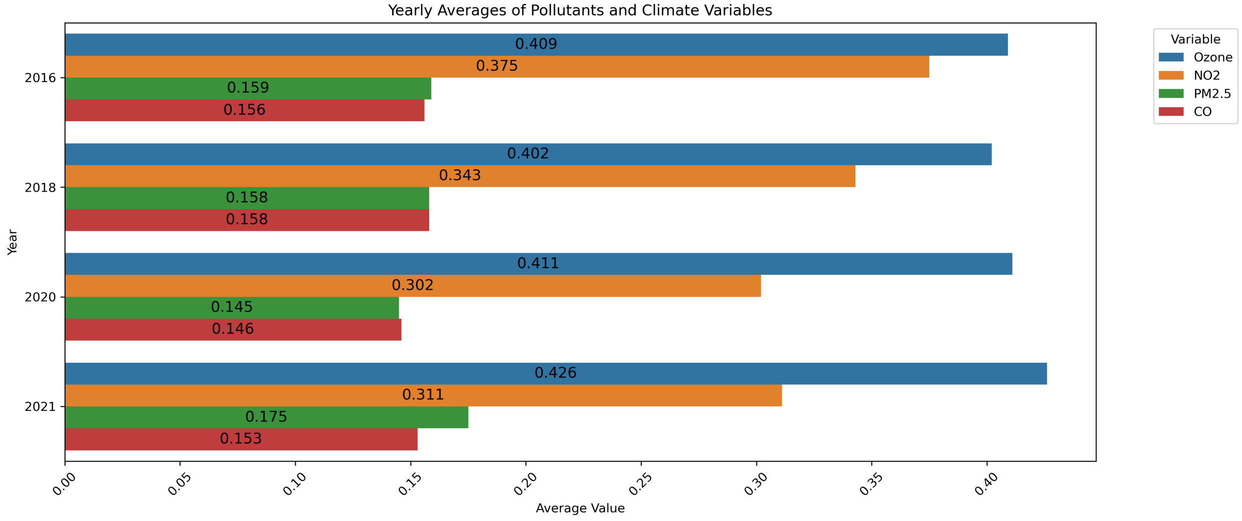

Figure 3: Yearly average concentrations of pollutants (O_3, CO, NO_2, and $PM_{2.5}$) in 2016, 2018, 2020 & 2021

3.2 Feature Engineering & Selection

3.2.1 One-Hot Encoding

One of the remaining issues in our dataset was the `Date` column, which was in a standard date format. To enable the model to learn from temporal patterns, we needed to convert this information into a numerical form.

We addressed this by extracting three temporal features - `Month`, `Day`, and `Year` - from the date. Each of these was then transformed using one-hot encoding, allowing the model to treat them as categorical variables with no inherent ordinal relationship.

For example:

- January (Month 1) → [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
- July (Month 7) → [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0]

This approach ensures that the model treats each month, day, and year as distinct categories rather than continuous values, thereby preventing misleading numerical relationships between them.

Next, we applied one-hot encoding to the `County` feature. As previously noted, our final dataset includes only two counties: Bronx and Queens. One-hot encoding converts this feature into two binary columns, ensuring that the model does not assume any ordinal relationship between the two locations.

3.2.2 KMeans Clustering for Latitude and Longitude

To transform the latitude and longitude features into more meaningful categorical variables, we applied KMeans clustering to the location data. We chose to cluster the data into two groups, which aligns with our dataset containing only two counties - Bronx and Queens. As expected, the clustering results reflected this separation, effectively encoding the geographic regions in a way that can be better understood by machine learning models.

3.2.3 Feature Selection

After applying one-hot encoding to the date and county features and clustering location data using KMeans, we were left with a total of 39 features and 2,896 rows.

We decided to exclude rain and precipitation from the set of target labels. As observed in the heatmap (see Figure 1), these variables showed high correlation with several other weather-related features, as well as with each other. Such multicollinearity can negatively impact model performance, particularly in regression tasks. Therefore, to ensure more robust and reliable predictions, we selected temperature as the sole label for our regression models.

3.3 Model Building

We built a Multi-Layer Perceptron (MLP) regression model, named **VISL** (Very Interpretable Structured Learner). This model is designed to predict temperature values based on the 39 input features described in the previous sections.

The VISL model architecture consists of several fully connected (dense) layers, with Batch Normalization and LeakyReLU activations applied between them to introduce non-linearity and stabilize learning. The model pipeline is as follows:

- A linear layer that maps the 39-dimensional input to 128 neurons (without bias).
- Batch Normalization on 128 features.

- LeakyReLU activation:

$$\text{LeakyReLU}(x) = \begin{cases} x, & x \geq 0 \\ \alpha x, & x < 0 \end{cases}$$

  where α is a small positive slope (by default 0.01).

- A linear layer from 128 to 256 neurons, followed by Batch Normalization and LeakyReLU.
- A linear layer from 256 to 512 neurons, followed by Batch Normalization and LeakyReLU.
- A linear layer from 512 back to 256 neurons, followed by Batch Normalization and LeakyReLU.
- A linear layer from 256 to 128 neurons, followed by Batch Normalization and LeakyReLU.
- A linear layer from 128 to 32 neurons, followed by Batch Normalization and LeakyReLU.
- A final linear layer that maps the 32 features to a single output neuron for regression (temperature prediction).

Overall, the VISL model uses the following sequence of layers (implemented using PyTorch):

```
self.mlp = nn.Sequential(
    nn.Linear(in_features=39, out_features=128, bias=False),
    nn.BatchNorm1d(128),
    nn.LeakyReLU(),
    nn.Linear(in_features=128, out_features=256, bias=False),
    nn.BatchNorm1d(256),
    nn.LeakyReLU(),
    nn.Linear(in_features=256, out_features=512, bias=False),
    nn.BatchNorm1d(512),
    nn.LeakyReLU(),
    nn.Linear(in_features=512, out_features=256, bias=False),
    nn.BatchNorm1d(256),
    nn.LeakyReLU(),
    nn.Linear(in_features=256, out_features=128, bias=False),
    nn.BatchNorm1d(128),
    nn.LeakyReLU(),
    nn.Linear(in_features=128, out_features=32, bias=False),
    nn.BatchNorm1d(32),
    nn.LeakyReLU(),
    nn.Linear(in_features=32, out_features=1)
)
```

The final layer outputs a continuous value suitable for regression tasks. We trained this model using the **Mean Absolute Error (MAE)** loss function, which is defined as:

$$\text{MAE} = \frac{1}{n} \sum_{i=0}^{n-1} |y_i - \hat{y}_i|$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, and n is the number of data points. MAE measures the average magnitude of the errors without considering their direction, making it robust to outliers and easy to interpret.

The model was optimized using the **AdamW** optimizer, which is a variant of the Adam optimizer that decouples weight decay from the gradient update. After hyperparameter tuning, we selected a learning rate of 1×10^{-3} and a weight decay of 1×10^{-5}. The model was trained for 150 epochs.

4 Results

4.0.1 Evaluation of VISL Model

To evaluate the performance of the VISL model, we rely on three commonly used regression metrics: **Mean Absolute Error (MAE)**, **Coefficient of Determination (R^2)**, and **Mean Absolute Percentage Error (MAPE)**. These metrics offer different perspectives on prediction accuracy and error spread. Their mathematical definitions and explanations are as follows:

- **Mean Absolute Error (MAE)**:

  $$\text{MAE} = \frac{1}{n}\sum_{i=0}^{n-1} |y_i - \hat{y}_i|$$

  MAE measures the average magnitude of errors between predicted and true values without considering their direction. It provides an easily interpretable error measure in the same units as the target variable.

- **Coefficient of Determination (R^2)**:

  $$R^2 = 1 - \frac{\sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1} (y_i - \bar{y})^2}$$

  R^2 indicates the proportion of the variance in the target variable that is predictable from the input features. A score of 1 indicates perfect prediction, while 0 means the model does no better than the mean of the target.

- **Mean Absolute Percentage Error (MAPE)**:

  $$\text{MAPE}(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1} \frac{|y_i - \hat{y}_i|}{\max(\epsilon, |y_i|)}$$

  MAPE expresses the prediction error as a percentage of the actual values, making it easy to interpret. The use of ϵ prevents division by zero when true values are very small or zero.

To evaluate the robustness and consistency of the VISL model, we trained and tested it ten times using randomized data splits. In each run, the dataset was randomly divided into training, validation, and test sets, and a new instance of the model was initialized and trained from scratch.

This repeated training process allowed us to observe the stability of the model's performance under different data partitions. After each run, we recorded three key evaluation metrics - MAE, R^2, and $MAPE$. By computing the mean, variance, and standard deviation of these metrics across all ten runs, we were able to assess the model's robustness and reliability more comprehensively than a single training instance would allow.

After training and evaluating the VISL model, we will compare its performance with two baseline models: **Linear Regression** and **Random Forest Regression**, to assess its effectiveness.

4.0.2 Comparison of VISL Performance with Existing Models

To evaluate the performance of the VISL model, we compared it against two well-known regression algorithms: ***Linear Regression*** and ***Random Forest Regression***. Each model was trained and evaluated 10 times using different randomized train, validation and test splits - consistent with the process used for ***VISL***. We computed three evaluation metrics for each model - MAE, R^2, and $MAPE$ - and recorded their **mean**, **variance**, and **standard deviation** across runs.

Linear Regression Statistics: {`'mae': [2.5144, 0.0, 0.0], 'r2': [0.8642, 0.0, 0.0], 'mape': [3.8326, 0.0, 0.0]`}

Random Forest Regression Statistics: {'mae': [1.22545, 0.00051, 0.02265], 'r2': [0.95557, 0.0, 0.00174], 'mape': [1.31997, 0.01619, 0.12724]}

VISL Model Statistics: {'mae': [1.09117, 0.00501, 0.07075], 'r2': [0.95183, 1e-05, 0.00366], 'mape': [0.53603, 0.00714, 0.08449]}

4.0.3 Performance Summary

- The ***Linear Regression*** model performed the weakest, with a high MAE of 2.5144 and no variance across runs, suggesting consistent but suboptimal predictions. Its R^2 of 0.8642 and MAPE of 3.8326 further highlight its limited accuracy compared to the other models.

- The ***Random Forest Regression*** model significantly outperformed Linear Regression, achieving a lower MAE of 1.2255 ± 0.0227, a higher R^2 of 0.9556 ± 0.0017, and a MAPE of 1.3200 ± 0.1272. These results indicate strong and stable predictive power with minimal variability across runs.

- The ***VISL*** model demonstrated the best overall performance. It achieved the lowest MAE of 1.0912 ± 0.0708, an R^2 of 0.9518 ± 0.0037, and the lowest MAPE of 0.5360 ± 0.0845. The VISL model outperformed both baselines on MAE and MAPE while maintaining competitive R^2, showing a good balance of accuracy, generalization, and stability across different data splits.

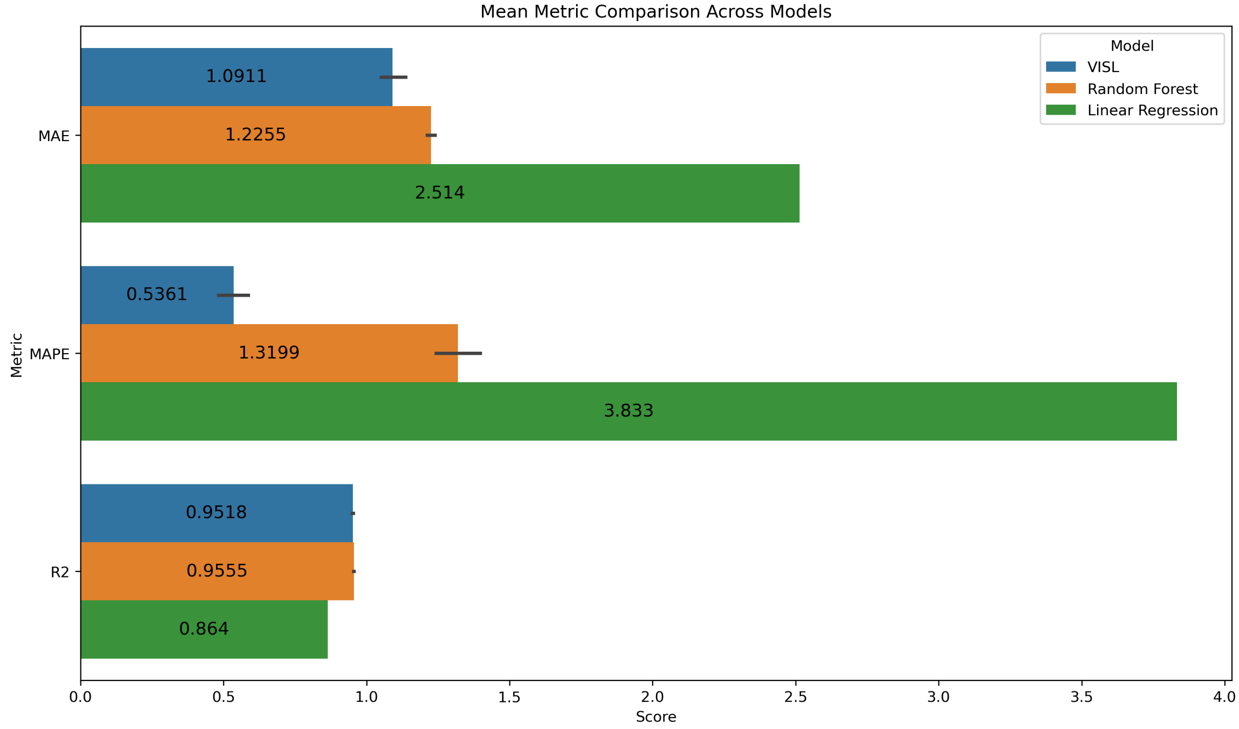

Figure 4: Bar plot comparing the average evaluation metric scores - Mean Absolute Error (MAE), Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE) - for the three models: Linear Regression, Random Forest Regression, and the VISL model. This visual highlights the superior performance of the VISL model on MAE and MAPE, while Random Forest achieves the highest R^2. Both VISL and Random Forest outperform Linear Regression across all metrics.

5 Summary

This study investigates the relationship between air pollution and temperature in New York City during four key years - 2016, 2018, 2020, and 2021 - with a special focus on the COVID-19 pandemic's impact. Our hypothesis proposed that fluctuations in pollution levels are reflected in local weather conditions, such as temperature. The dataset used in this analysis was assembled from publicly available sources, including the EPA for air quality data and the Open-Meteo API for weather information. After an intensive cleaning and preprocessing phase, the final dataset consisted of 2,896 clean and normalized records representing pollutant concentrations and meteorological measurements across Bronx and Queens counties.

Exploratory data analysis revealed notable patterns: a strong positive correlation between ozone (O_3) levels and temperature, and significant declines in pollutants such as CO and NO_2 during 2020 - the peak pandemic year when human activity dropped. These findings align with similar work in the literature, including studies on air pollution reduction during lockdowns and pollutant–weather interactions.

To test our hypothesis quantitatively, we built several predictive models aimed at forecasting temperature from pollutant levels. These included traditional machine learning approaches like Linear Regression and Random Forest, as well as a custom-built neural network model, VISL. While the neural network offered some predictive improvements, performance remained modest, indicating room for future enhancement. Potential improvements include the integration of urban activity indicators (such as traffic or mobility data) and the exploration of more complex model architectures.

Compared to related literature, our study is unique in its localized, multi-year, and AI-driven approach. Most prior work either focused on broader geographic scales or emphasized descriptive statistics rather than prediction. By contrast, our study combined data integration, correlation analysis, and predictive modeling to provide a comprehensive view of how air quality and weather interact over time in NYC.

In summary, our results affirm the hypothesis that air pollution levels influence temperature and that machine learning can offer predictive insights into these relationships. However, the modest performance of our initial models suggests that capturing such environmental dynamics remains a complex challenge. Future work should focus on enriching datasets, refining model architectures, and possibly incorporating causal inference techniques to deepen our understanding of these interactions.