# Breathe Easy: Unmasking the Elements of Indian Air Quality Data

Yicen Ye
Data Science Institute
Brown University
https://github.com/YicenYe/Data1030Project

## Introduction

### Problem, Purpose, and Importance

Air pollution is a pressing global issue, and India faces significant challenges in monitoring and controlling air quality. Poor air quality poses serious risks to public health, environmental sustainability, and economic development. To address these concerns, this project focuses on using **regression** models to predict the Air Quality Index (AQI) using time series data on pollutant levels. Predicting AQI values can help raise awareness, guide public policies, and inform citizens about potential health risks.

### Data Source

The dataset used for this project, "Air Quality Data in India (2015-2020)," is sourced from Kaggle and compiled by the Central Pollution Control Board (CPCB), the official environmental monitoring agency under the Government of India. The dataset includes daily pollutant measurements across Indian cities, capturing key variables such as PM2.5, PM10, NO2, CO, and SO2, among others. For this analysis, we focus on the city_day subset of the dataset.

| | City | Date | PM2.5 | PM10 | NO | NO2 | NOx | NH3 | CO | SO2 | O3 | Benzene | Toluene | Xylene | AQI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2304 | Amaravati | 2018-05-25 | 40.32 | 66.25 | 4.36 | 11.30 | 9.65 | 10.61 | 0.50 | 7.57 | 24.35 | 39.22 | 4.47 | 8.46 | 62.0 |
| 2305 | Amaravati | 2018-05-26 | 13.07 | 37.35 | 3.80 | 8.59 | 7.75 | 9.78 | 0.46 | 6.09 | 29.05 | 28.37 | 5.06 | 4.02 | 78.0 |
| 2306 | Amaravati | 2018-05-27 | 16.56 | 41.96 | 3.03 | 7.25 | 6.31 | 9.48 | 0.41 | 6.48 | 62.01 | 35.84 | 0.44 | 1.52 | 114.0 |
| 2307 | Amaravati | 2018-05-28 | 19.57 | 44.72 | 3.72 | 7.72 | 7.22 | 10.79 | 0.34 | 5.99 | 64.58 | 53.89 | 26.79 | 125.18 | 103.0 |
| 2308 | Amaravati | 2018-05-29 | 30.96 | 72.18 | 3.63 | 10.82 | 8.80 | 11.18 | 0.36 | 11.72 | 82.13 | 48.64 | 6.44 | NaN | 134.0 |

# EDA

Indian Air Quality Dataset is a time series dataset, which means that the data is non-iid. We have numerical columns being all the pollutants, and then one categorical column being cities.

**Preprocess**:
We start dealing with the data by dropping the AQI bucket column, which is simply the ranked bucket for the target value for the regression problem, AQI index. This categorical column is almost the same as the target value but not our target in the regression problem, so it will be redundant and unnecessary for later training and calculation. We also need to make sure that the date column is in Datetime form to use any time series-related methods. The dataset contains a significant amount of missing values. Since these features are continuous, handling them requires careful consideration to preserve the dataset's validity and avoid introducing biases. To prepare the data for models that do not inherently handle missing values, I used feature-reduced approach. This method ensures that the filling process maintains temporal consistency and does not introduce artifacts that could lead to overfitting.

Figure 1 shows the overall trend of key pollutants—PM10, PM2.5, CO, and NO2—alongside AQI over time (2015–2020). The solid lines represent pollutant levels, while the dashed blue line tracks AQI. The graph highlights a rising trend in pollution levels, with clear seasonal spikes. PM10 and PM2.5 show the strongest correlation with AQI, while CO and NO2 remain lower but follow similar patterns, which we can also confirm from the feature importance graph and the heatmap. These trends emphasize worsening air quality over time and recurring seasonal effects, underscoring the need for improved pollution control measures.

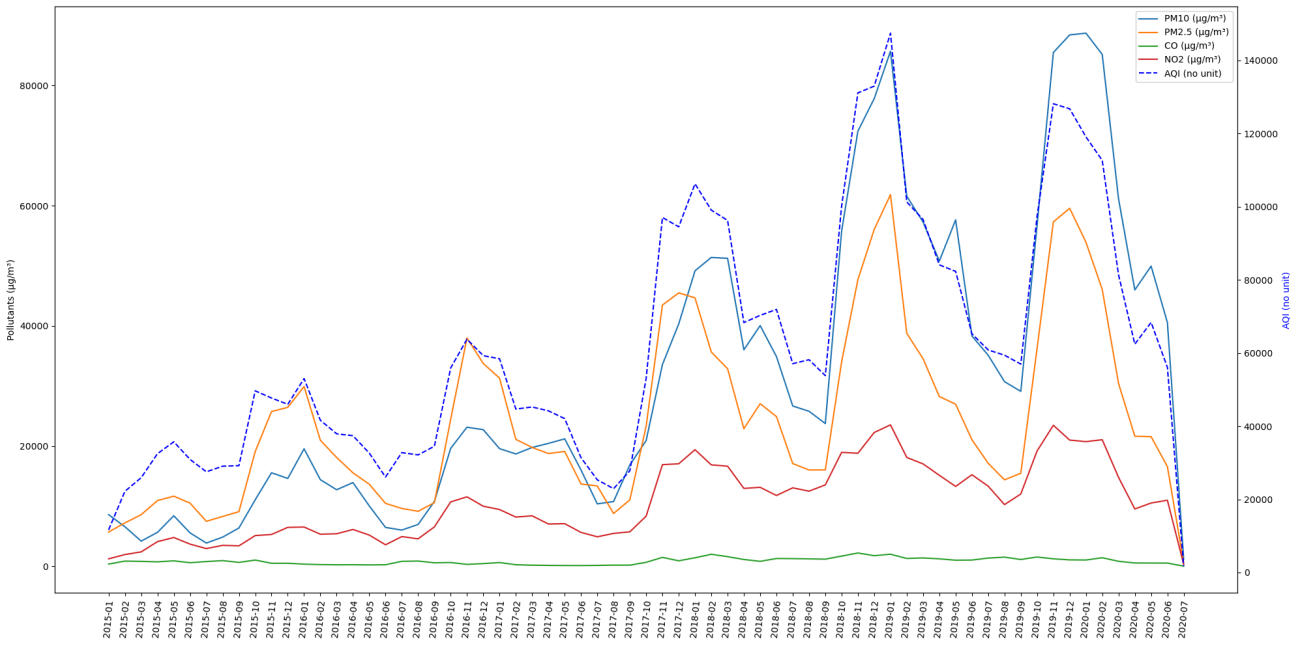


Figure 1: Trends of key pollutants (PM10, PM2.5, CO, NO2) and AQI over time reveal seasonal patterns and simultaneous peaks, indicating a strong correlation between pollutants and air quality.

Figure 2 illustrates the top 10 city-wise rankings for AQI, PM10 levels, and CO concentrations. Ahmedabad and Delhi lead with the highest AQI and CO values, respectively, indicating severe air pollution in these cities. Delhi also records the highest PM10 levels, followed by Gurugram and Talcher, emphasizing particulate matter as a major concern. The charts reveal significant variation across cities, with certain regions consistently ranking high across all three metrics.

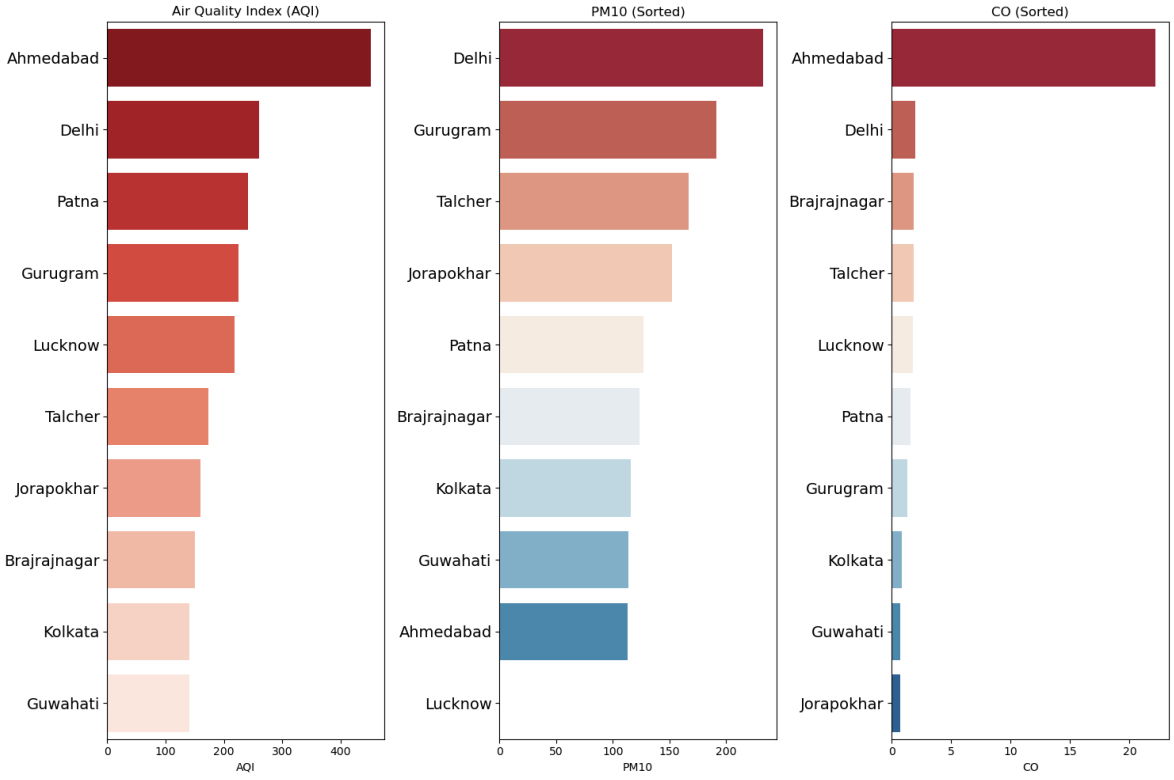


Figure 2: City-wise comparison shows Ahmedabad and Delhi as the most polluted, with high AQI, PM10, and CO levels.

Figure 3 heatmap displays the correlation between numeric features and the target variable, AQI. PM2.5, PM10, and CO exhibit the strongest positive correlation with AQI, indicating their significant influence on air quality. Additionally, there is a strong correlation between PM10 and PM2.5 (0.85), and NOx and NO. Most other features display weaker correlations, either with each other or with AQI, suggesting they have less impact on the overall air quality index. This analysis highlights PM2.5, PM10, and CO as key predictors for AQI.

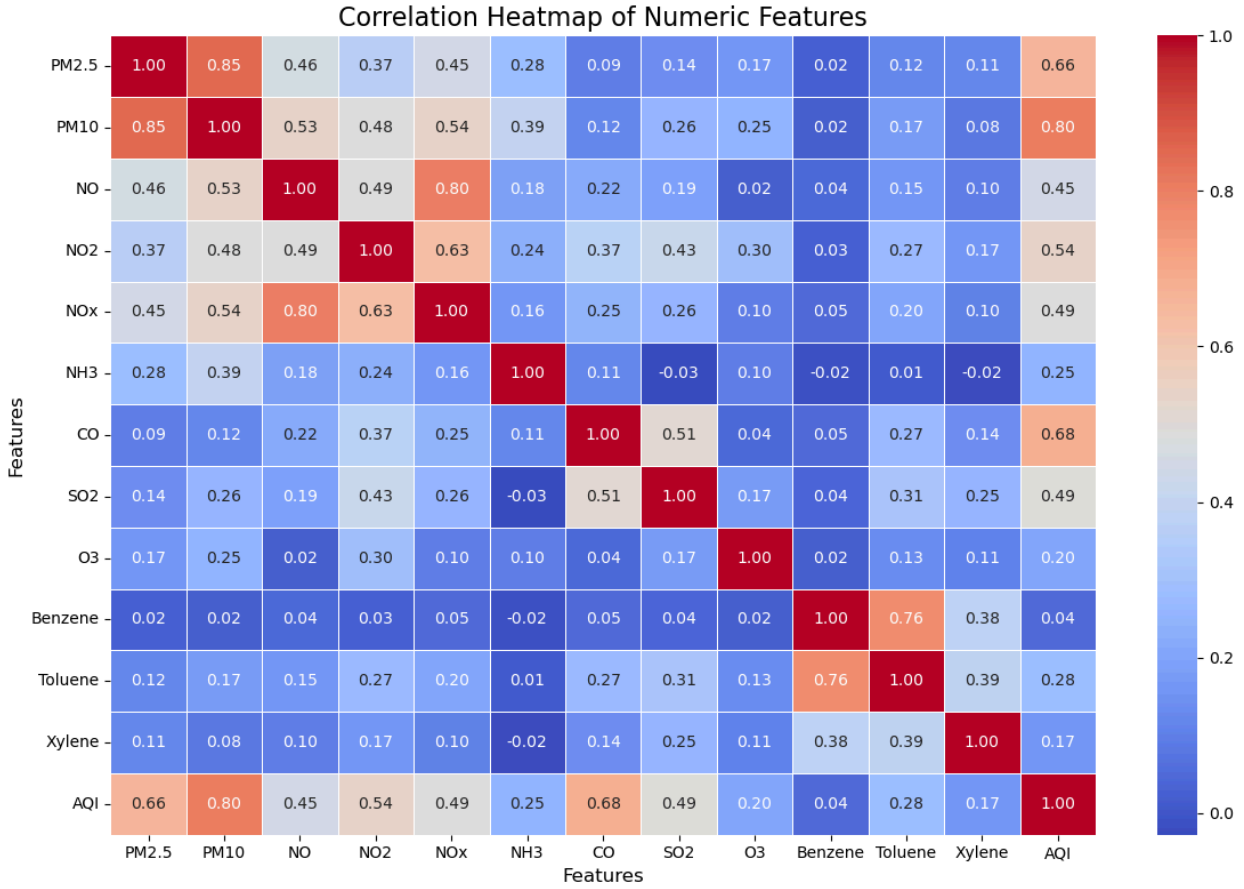


Figure 3: Heatmap showing correlations between pollutants and AQI, with PM10 and PM2.5 as key contributors.

The autocorrelation (ACF) and partial autocorrelation (PACF) graphs after differencing reveal important time series characteristics. The ACF plot shows a strong correlation at lag 1, gradually declining across subsequent lags, which indicates a significant relationship between the target value and its immediate past values. Meanwhile, the PACF plot shows a sharp drop-off after lag 1, suggesting that the correlation with other lags can be explained by the immediate previous time step. These observations justify the inclusion of lagged features (e.g., lag-1, lag-2) and rolling average features in the model, capturing autoregressive patterns and smoothing temporal variations for better predictive performance.

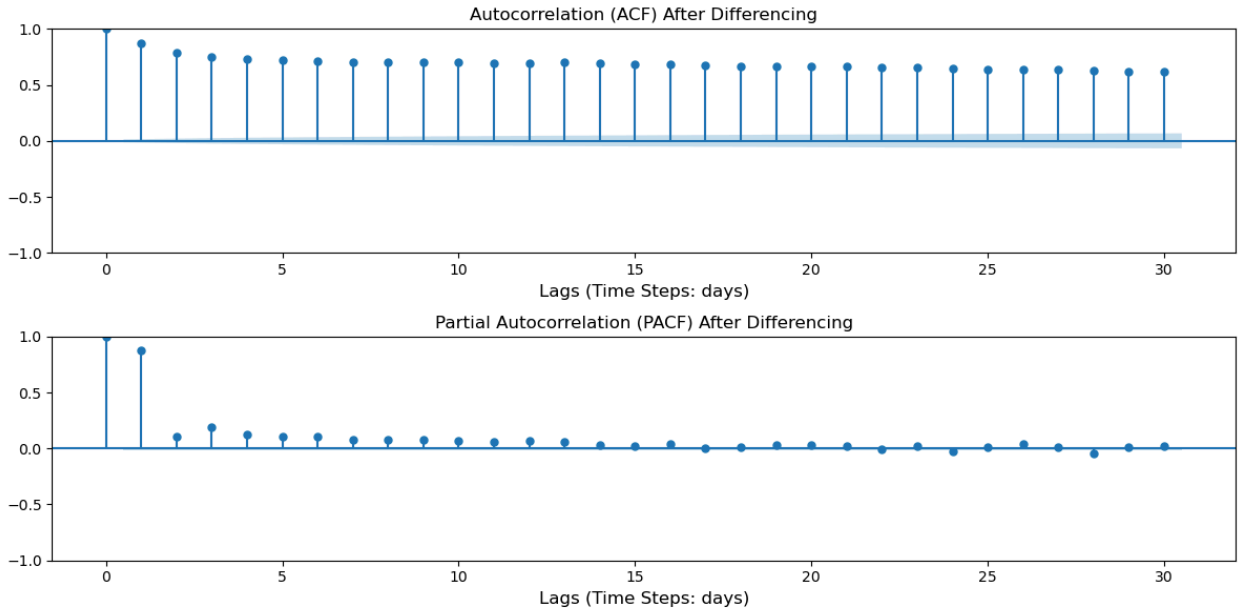


Figure 4: The autocorrelation plot shows significant correlations at multiple lags, and the partial autocorrelation plot highlights a strong lag-1 relationship

# Methods

### Splitting

The dataset was split into training, validation, and test sets using a temporal split to preserve the chronological order of the data, which is crucial for time series analysis. The split proportions were 60% for training, 20% for validation, and 20% for testing. This temporal splitting approach avoids data leakage by ensuring that future information does not influence the training process.

**Preprocessing**

To enhance the predictive power of the model, additional time-based features were engineered, including lagged variables (e.g., AQI_lag_1 and AQI_lag_2), which represent the AQI values from one and two time steps earlier, and rolling averages (e.g., AQI_rolling_avg_2 and AQI_rolling_avg_3), which capture the average AQI over the last two and three time steps. These features help the model capture temporal dependencies and smooth variations in the data.

**CV and ML Pipeline**

After splitting, a preprocessing pipeline was applied, including Min-Max Scaling for numerical features (e.g., PM2.5, NO2, AQI lags) and one-hot encoding for categorical features (e.g., City), ensuring the data was appropriately transformed for machine learning models.

The machine learning pipeline further incorporates advanced model tuning and evaluation using GridSearchCV and TimeSeriesSplit to optimize hyperparameters and assess the model's performance on time-based cross-validation folds. This approach ensures that the model is trained and evaluated in a time-consistent manner. The pipeline integrates the preprocessor with the machine learning model using scikit-learn's Pipeline, streamlining the process of training and evaluation. Models such as Random Forest, XGBoost, ElasticNet, and Support Vector Regression were trained using this pipeline, with hyperparameter optimization tailored to each algorithm. Evaluation metrics, including mean squared error (MSE) and $R^2$, were computed for validation and test sets, providing insights into model accuracy and predictive performance. Although showing both MSE and $R^2$, we used MSE for evaluation as $R^2$ baseline is 0, which does not show the feature importance clearly.

The table below shows the ML algorithms and the parameters tuned. The bolded values are the best parameter tuned for the best models for each algorithm.

| ML Algorithms | Parameters tuned |
|---|---|
| Random Forest | Max_depth: [1, 2, 5, 10, 20, 30, **50**, None]<br>Max_features: ['auto', 'sqrt', 'log2', **None**] |
| XGBoost | Max_depth: [**3**, 5, 7, 10, 20, 50]<br>Reg_alpha: [0, 0.01, **0.1**, 1]<br>Reg_lambda: [1, 5, 10, **20**] |
| Elastic net | Alpha: [0.001, **0.01**, 0.1, 1, 10]<br>L1_ratio: [0.1, **0.3**, 0.5, 0.7, 1.0] |
| SVR | C: [0.01, 0.1, 1, **10**, 100]<br>Gamma: [0.001, **0.1**, 1, 1000, 100000] |

**Uncertainty**

Mean predictions: [153.984 124.977 156.086 146.187 162.256]

Uncertainty (std): [ 97.39201242  74.68968711  98.54257721 122.65424771 106.41273553]

The uncertainty results highlight the model's variability and uncertainty. RMSE values across splits range from ~25 to ~42, indicating sensitivity to data partitioning and variability in performance. The mean predictions suggest reasonable accuracy, but the high standard deviations (e.g., ~97 for some test samples) reveal inconsistency in predictions, likely due to challenging or noisy data. These findings suggest that while the model performs well on average, certain data points exhibit higher prediction uncertainty, pointing to potential areas for improvement in feature engineering or data quality.

# Results

Figure 5 and 6 summarize the performance of different machine learning models—Random Forest, XGBoost, ElasticNet, and SVR—evaluated using Mean Squared Error (MSE) and $R^2$ scores. The first graph, depicting MSE values, highlights that Random Forest and XGBoost achieve the lowest MSE, with Random Forest performing marginally better. ElasticNet, by contrast, has the highest MSE, indicating weaker predictive performance. The second graph, showing $R^2$ scores, confirms the findings, with Random Forest and XGBoost achieving $R^2$ values over 0.9, significantly outperforming ElasticNet and SVR. The minimal standard deviations (almost imperceptible) reflect consistent performance across runs. Compared to the baseline MSE of 8097 and $R^2$ of 0, the models—especially Random Forest—demonstrate substantial improvements. Using both of them can let us determine the best model from two perspectives.

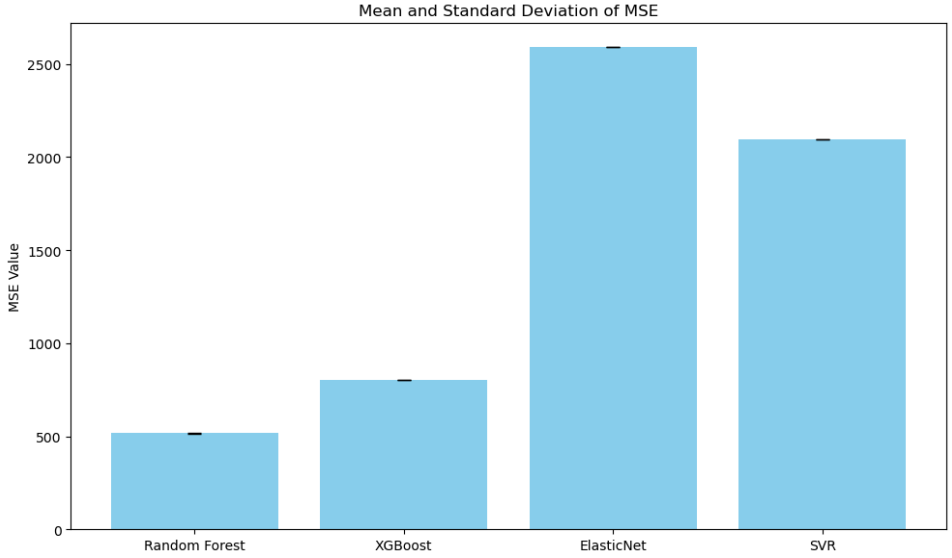


Figure 5: Mean and Standard Deviation of MSE for four models. Random Forest and XGBoost achieved the lowest MSE, while ElasticNet had the highest MSE.

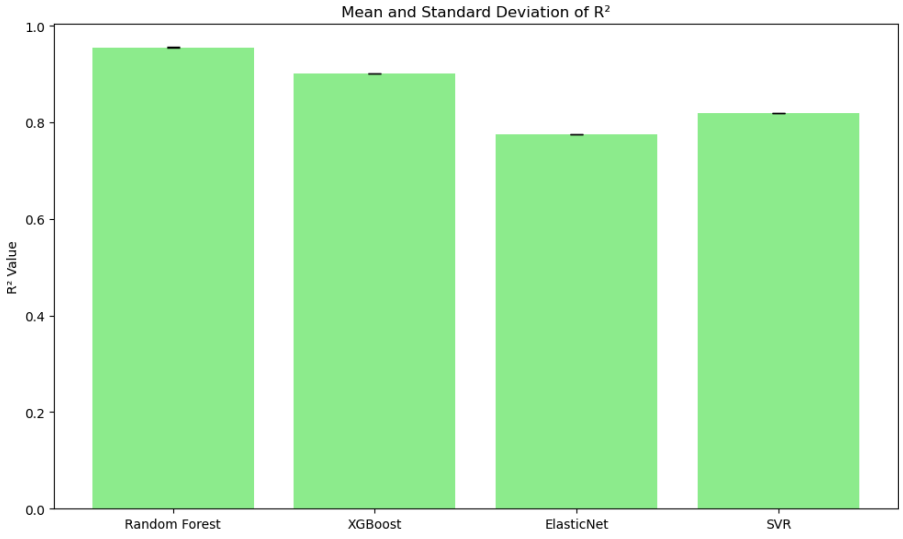


Figure 6: Mean and Standard Deviation of $R^2$ for four models. Random Forest and XGBoost performed best with $R^2$ values near 0.95, while ElasticNet had the lowest $R^2$.

The Random Forest model with random state = 322 demonstrated strong predictive performance, as shown by the True vs. Predicted scatter plot. The majority of points align closely along the red diagonal line, indicating that the model's predictions are highly accurate with minimal deviations. The overall tight clustering of points around the line highlights the model's effectiveness in capturing the patterns in the data and reducing prediction errors.

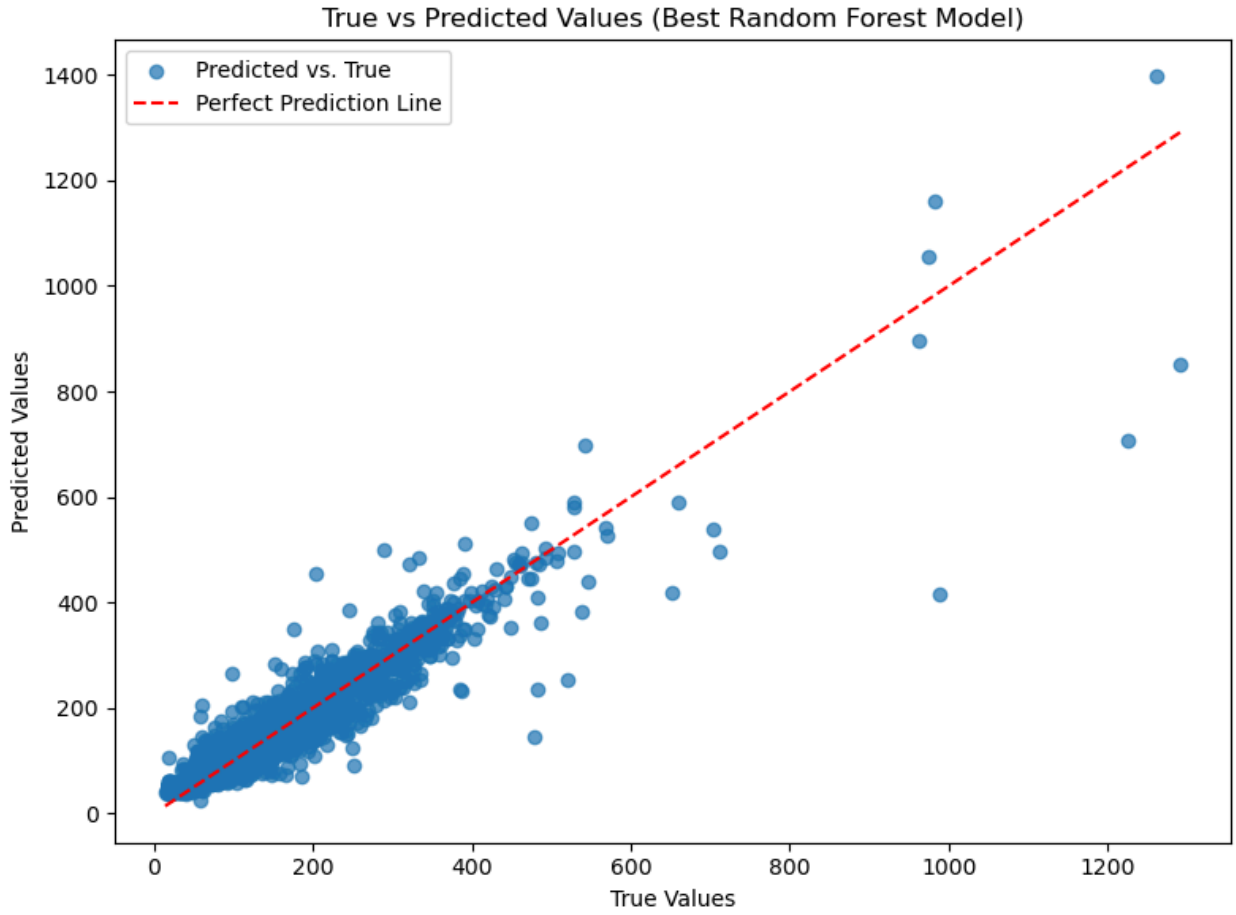


Figure 7: The scatter plot shows the true vs. predicted values from the Random Forest model. The closeness of the points to the perfect prediction line demonstrates the model's performance.

# Interpretation

## SHAP

The SHAP plots provide insights into the feature contributions to the model's predictions. The summary plot (top) shows that AQI_lag_1, PM2.5, and PM10 have the most significant impact on the model's output, with higher values of these features generally increasing predictions. The color gradient indicates feature values, where red represents high values and blue represents low values. The bar plot confirms this ranking, displaying the average absolute SHAP values, with AQI_lag_1 as the most influential feature, followed by PM2.5 and PM10. These findings highlight the importance of lagged and pollutant features in predicting AQI.

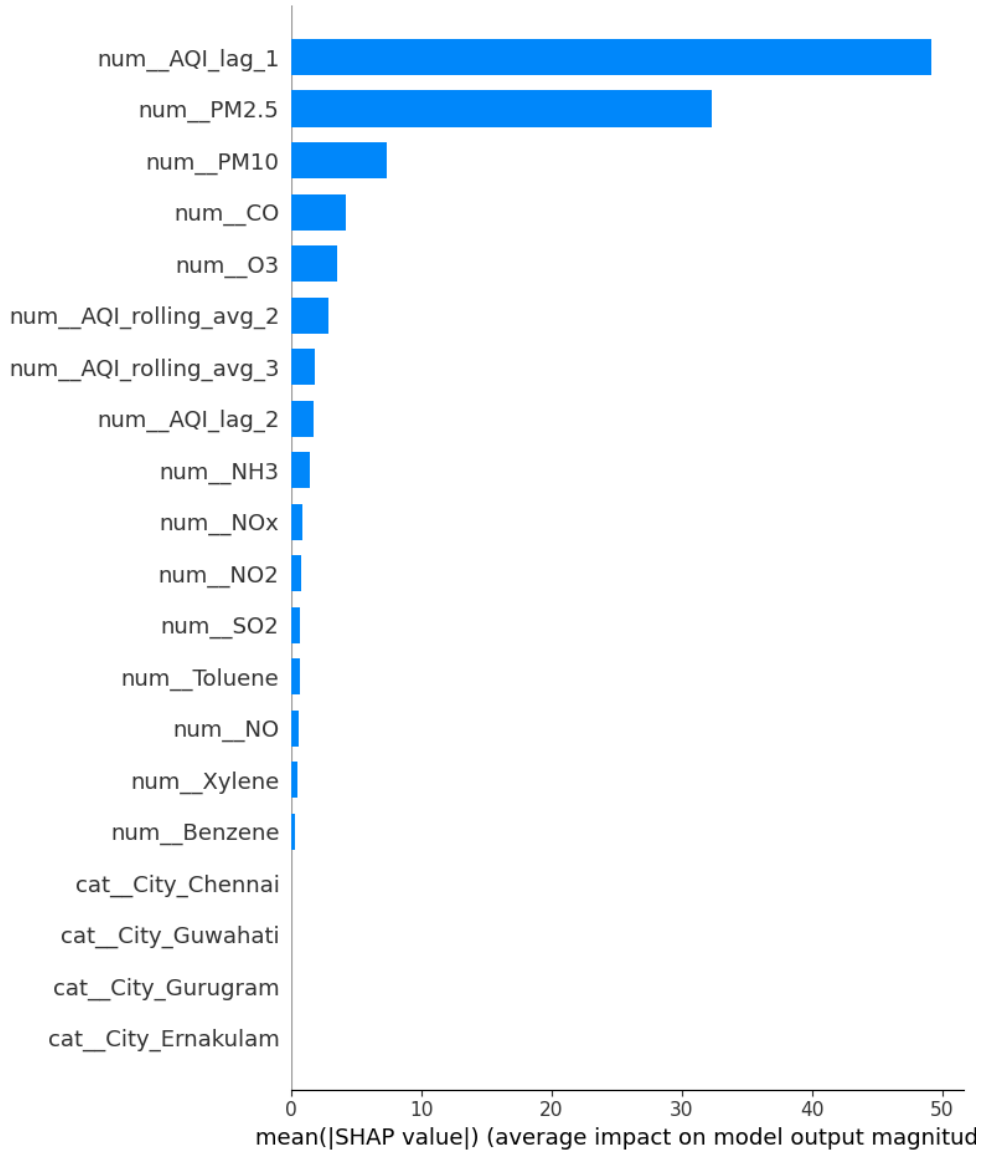


Figure 8: SHAP Feature Importance Plot

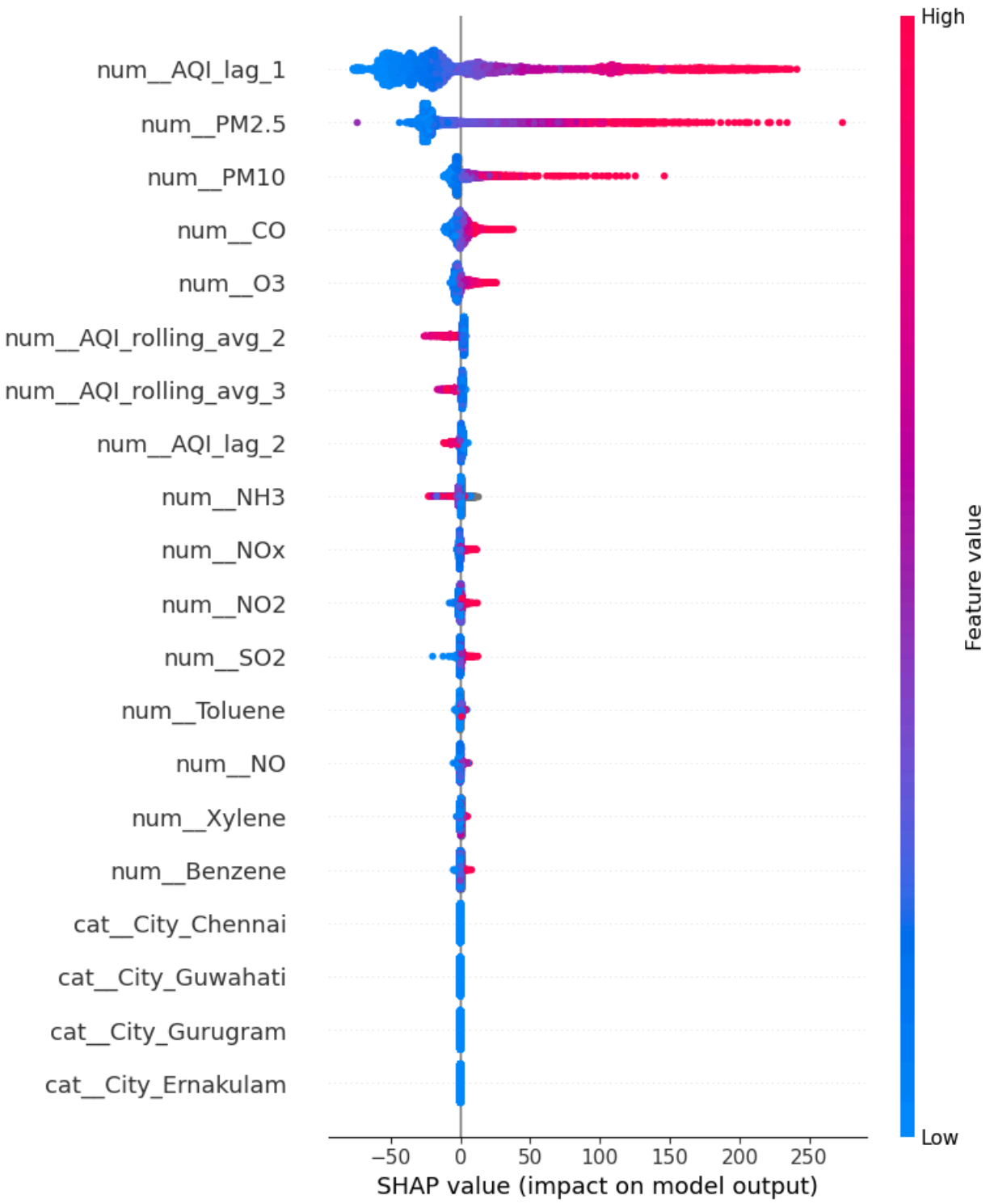


Figure 9: SHAP Mean Impact Plot

**Permutation Importance**

The Permutation Importance graph evaluates the contribution of each feature to the model's performance by measuring the increase in Mean Squared Error (MSE) after shuffling each feature. Features with larger MSE increases are considered more important as they disrupt predictions when perturbed. The top-ranked features: AQI_lag_1, PM2.5, and PM10 show the most significant impact, indicating their critical role in predicting the target variable. This aligns with their high importance in previous SHAP analysis. Conversely, many categorical city features and less influential pollutants (e.g., SO2, Toluene) have minimal impact, with MSE values close to the baseline score, suggesting their lower contribution. The consistent results across multiple runs reinforce the robustness of this analysis.

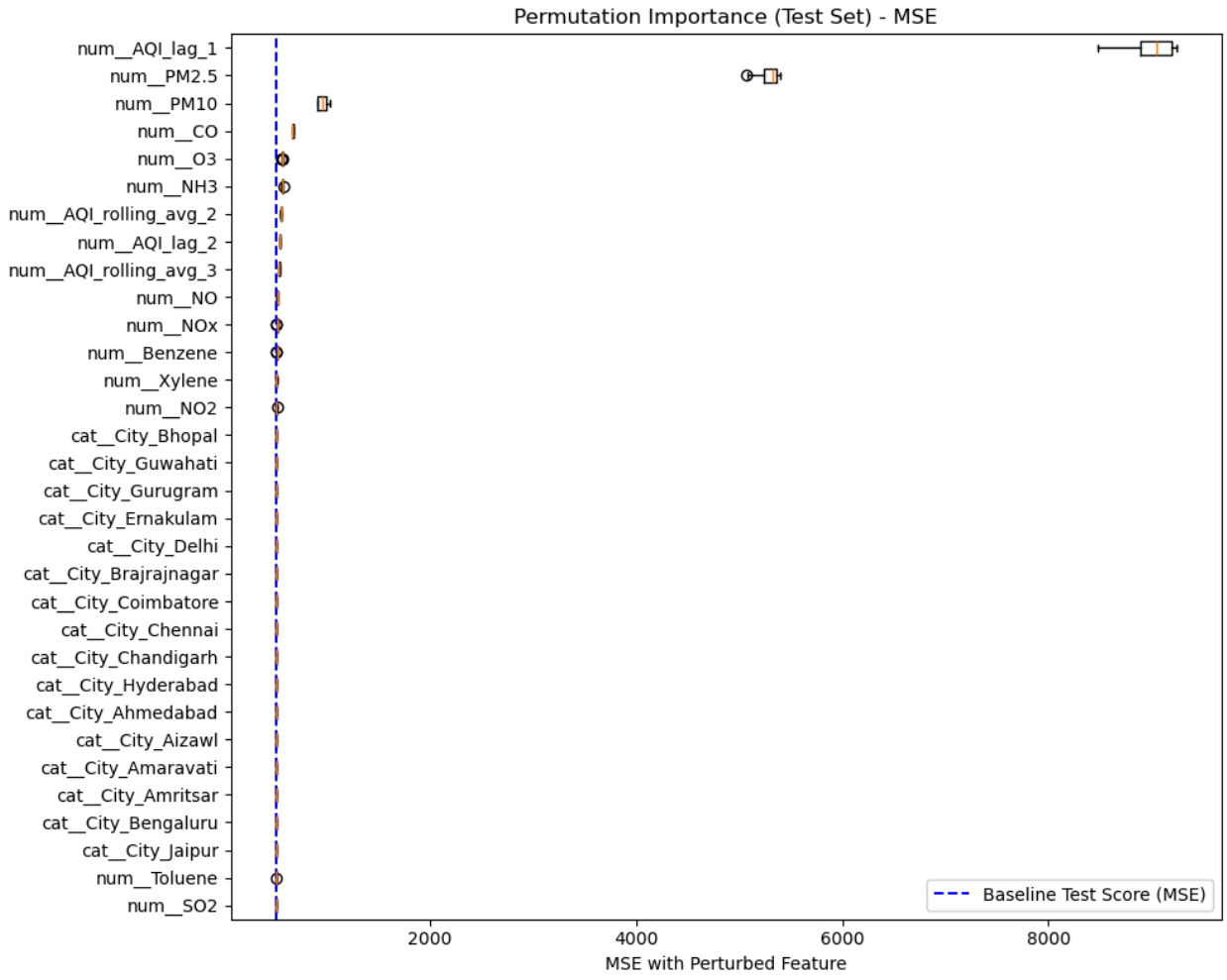


Figure 10: SHAP Force Plot (Index 0)

**Force Plots:**

The SHAP force plots for the three selected indices (0, 100, and 200) visually illustrate how individual features contribute to the predictions of your Random Forest model. Here is the analysis:

1. First Index (63.54):
    - The prediction is 63.54, slightly lower than the base value.
    - The num__AQI_rolling_avg_2 feature (positively contributing) has the largest impact in increasing the prediction.
    - Features such as num__AQI_lag_1 and num__PM2.5 show negative contributions, pulling the prediction down.

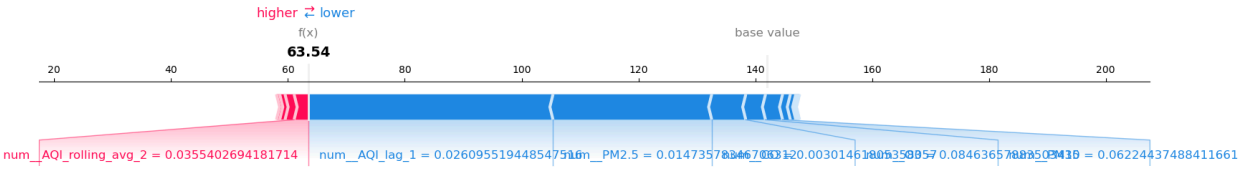


Figure 11: SHAP Force Plot (Index 0)

2. Second Index (58.16):
    - The prediction is 58.16, lower than the base value.
    - Here, num__AQI_lag_1 is the largest positive contributor to the prediction.
    - Features like num__PM2.5 and num__CO slightly increase the prediction, but their contributions are relatively small.
    - Most features pull the prediction down, resulting in a lower prediction value.

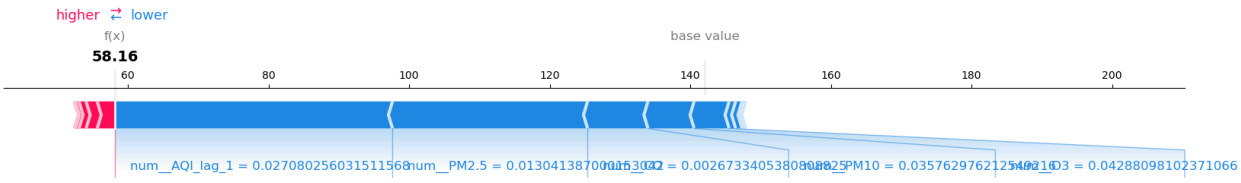


Figure 12: SHAP Force Plot (Index 100)

3. Third Index (177.23):
    - The prediction is 177.23, which is significantly higher than the base value.
    - The largest positive contributors include num__AQI_rolling_avg_2, num__NOx, and num__PM10.
    - These features push the prediction upward, reflecting their strong influence on this instance.
    - Conversely, num__NH3 is the only feature with a slight negative contribution.

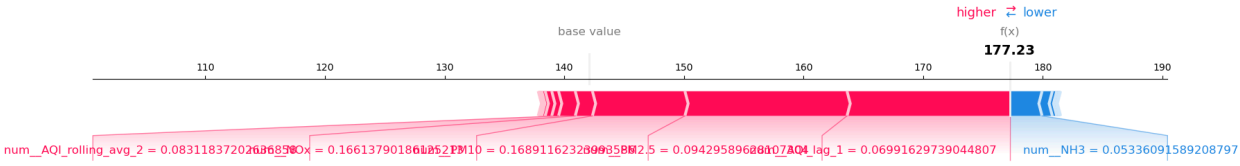


Figure 13: SHAP Force Plot (Index 200)

## Outlook

There is more to explore with data splitting strategies to better evaluate the model's robustness over time. While lag and rolling features helped capture temporal patterns, experimenting with different lag lengths and window sizes could improve results. Overfitting might still be a concern, particularly with complex models like Random Forest, which could be addressed with regularization or feature reduction. Additional data, such as weather, traffic, or industrial activity, could provide richer features and enhance predictive power. Exploring ARIMA, SARIMA, or hybrid forecasting models could improve performance, especially for trends and seasonality.

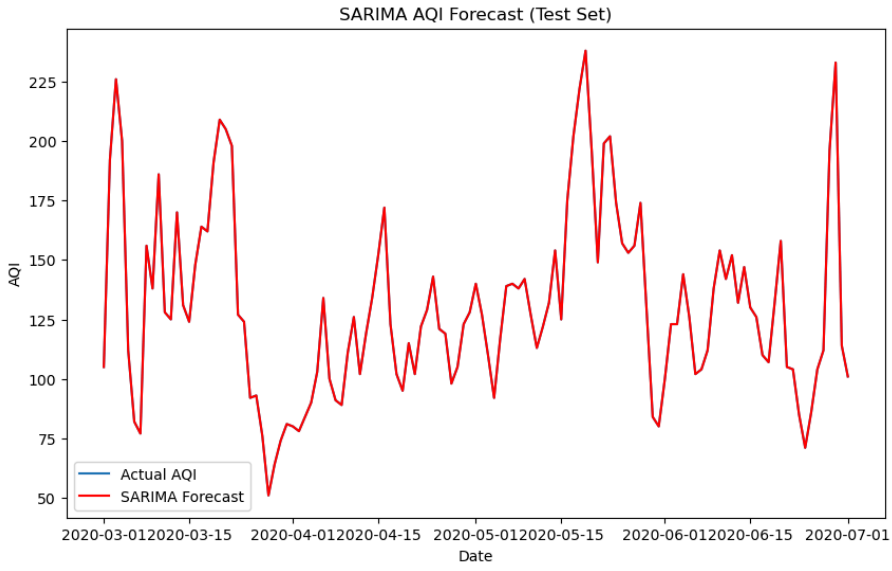


Figure 14: A test on SARIMA Forecast, not compared with other models

**References**

https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india
https://www.kaggle.com/code/parulpandey/breathe-india-covid-19-effect-on-pollution#-1.1-Missing-Values-
https://www.kaggle.com/code/zeynepsyavuz970/air-quality-prediction
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html
https://github.com/YicenYe/Data1030Project