PROJECT PHASE 3 REPORT

Beyond the Haze: Predictive Analytics and Cluster-Based Insights into India's Air Quality

Authors: Naga Manasa Palaparthi, Shubham Chandra and Divya Polavarapu (CSE 587)

Problem Statement

The deterioration of air quality in Indian cities creates serious health and environmental concerns. Understanding air quality is complicated by the interaction of numerous contaminants such as PM2.5, PM10, NO2, and NH3. Each of these pollutants contributes differently to the overall Air Quality Index (AQI), a statistic used to determine how filthy the air is now or may become in the future. This complication is exacerbated by the fact that the sources and amounts of these pollutants can vary substantially throughout metropolitan regions, depending on factors such as industrial operations, automobile emissions, weather patterns, and geographical features. As a result, it is critical to have a sophisticated and extensive understanding of how these factors interact and effect the AQI. Predictive analytics can play a key role in addressing this challenge by analyzing historical data to identify patterns and trends in air quality and using these insights to forecast future AQI levels.

Employing machine learning techniques provides a valuable tool for both prediction and analysis. Machine learning algorithms can process enormous amounts of data from many sources, learn from it, and predict future AQI values accurately. This may be extremely beneficial to policymakers, public health professionals, and the public in terms of implementing early and effective actions to reduce air pollution. Furthermore, another part of your study, cluster based analysis, can give a complete picture of air quality across several Indian cities. This study can assist uncover common sources of pollution and viable mitigation techniques by combining cities with similar pollution profiles. It can also help with a comparative study of how different urban surroundings affect air quality, offering useful insights for urban planning and environmental policy.

Data source for the project

For the data source, We have utilized the Kaggle dataset titled "Air Quality Data in India." This dataset is available at the following link: Air Quality Data in India on Kaggle (https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india/data). Specifically, We are working with the 'city_data.csv' and 'city_hour.csv' file from this data source. This dataset contains valuable air quality data in India, which is substantial in size, with over 20,000 rows/records in each dataset. The datasets provide various air quality metrics and related information for different cities, dates, and hours. This data source is a crucial component of our project and allows us to conduct meaningful data analysis.

Why is this a significant problem?

Poor air quality in Indian cities is a serious public health issue. Air pollution, defined by excessive levels of pollutants such as PM2.5, PM10, NO2, and NH3, has a direct influence on the population's respiratory health. Prolonged exposure to polluted air raises the risk of chronic respiratory disorders, asthma, and lung cancer, as well as having a negative impact on the cardiovascular system. It impacts the quality of life of the whole population, not only the vulnerable (children, the elderly, and those with pre-existing diseases). The health consequences include increased healthcare expenses and lost productivity because of illness, putting a considerable strain on the country's healthcare system and economy. Environmental degradation caused by air pollution affects ecosystems, wildlife, and worsens climate change.

Environmental deterioration is a major component of this issue. Air pollution does not just impair human health; it also has a negative impact on ecosystems and animals. Pollutants can modify natural habitats, impair plant life, and disturb ecological equilibrium. Furthermore, air pollution leads to climate change, worsening problems such as global warming and extreme weather occurrences. Climate change is a global problem; thus this impacts India as well as the rest of the globe.

Resolving the issue of air quality is critical for the long-term viability of India's metropolitan areas. With increased urbanization, air pollution is anticipated to worsen unless proactive steps are adopted. It is critical for cities' long-term growth, ensuring that they remain habitable and healthy for future generations. Predictive analytics and cluster-based insights provide a mechanism to better understand and address this issue. Policymakers and stakeholders may use these sophisticated tools to discover patterns, forecast future air quality situations, and conduct targeted interventions. This not only aids in immediate pollution reduction, but also in long-term planning for ecologically sustainable and health-conscious urban growth. As a result, the significance of this issue cannot be emphasized.

Objectives

Predictive Modeling for AQI: The primary goal is to create a machine learning model that reliably predicts the Air Quality Index (AQI) using the dataset's available air quality indicators. Selecting relevant algorithms, training the model using the provided dataset, and fine-tuning it to guarantee high accuracy and reliability in forecasting future AQI levels based on parameters such as PM2.5, PM10, NO2, and NH3 levels are all part of this process.

Clustering for Comparative Analysis: This goal focuses on using clustering techniques to organize cities or seasons based on their air quality characteristics into separate groups. The project's goal is to identify these clusters to find patterns, similarities, or anomalies in air quality across different locations or times of the year, offering a better understanding of the geographical and temporal changes in air pollution.

Accessible Delivery of Results: The third goal is to provide the findings from the prediction models and clustering analysis in a way that the public can comprehend. This entails developing clear, succinct, and visually compelling presentations or interactive technologies that can transmit complicated data insights in an easy manner, guaranteeing that non-experts can access the information.

Actionable Insights for Public Awareness: The research intends to deliver actionable insights and suggestions for public awareness based on AQI levels. This involves providing folks with easy, practical measures and procedures they can take to protect themselves from bad air quality, such as when to limit outdoor activities, use air purifiers, or wear masks, therefore increasing public knowledge and safety.

Potential Contributions

Comprehensive Air Quality Analysis: This project aims to create a more complete understanding of air quality, integrating multiple pollutants into a unified assessment.

Demystifying Data: It transforms intricate air quality data into formats easily grasped by the general populace.

Guiding Personal Safeguards: The project equips individuals with actionable advice to mitigate the effects of air pollution.

Strategic Policy Insights: It furnishes policymakers with nuanced insights to craft precise and effective air quality strategies.

Overall process for the project

1. **Data Collection:** This phase involves gathering comprehensive air quality data for Indian cities from the 'city_day.csv' dataset, focusing on AQI and key pollutants like PM2.5, PM10, NO2, and NH3, to build a robust foundation for analysis.

2. **Data Preprocessing:** This critical step ensures data integrity by cleaning the dataset, addressing missing values, standardizing data types, and managing outliers, thereby preparing the data for accurate and meaningful analysis.

3. **Exploratory Data Analysis (EDA):** Here, the project undertakes a thorough examination of the data, employing summary statistics, univariate, and multivariate analysis to uncover patterns and relationships within the air quality metrics, setting the stage for predictive modelling.

4. **Machine Learning Model Development:** This step involves building and fine-tuning predictive models using various machine learning algorithms, aiming to accurately forecast AQI levels based on the processed data.

5. **Clustering Analysis:** In this phase, clustering techniques are applied to categorize cities or seasons into groups with similar air quality profiles, offering insights into common pollution characteristics and trends.

6. **Communication of Findings:** The project culminates in translating complex data into accessible visualizations and summaries, presented through an interactive console, making the insights understandable and actionable for the public.

Process 1&2 – Data Collection and Data Preprocessing

1. **Data Loading:** The process begins with the loading of the "city_day.csv" dataset into a Pandas DataFrame. This dataset, sourced from Kaggle, provides a comprehensive collection of daily air quality metrics across various Indian cities.

2. **Data Cleaning:** Involves removing rows with NaN values in all columns except 'Date' and 'City'. This step is crucial for maintaining the integrity of the dataset while retaining essential temporal and spatial information.

3. **Data Transformation:** The 'Date' column is dissected into 'Year', 'Month', and 'Day', facilitating more granular temporal analysis.

4. **Data Type Conversion and Outlier Analysis:** Specific columns are converted to their appropriate data types, and outliers are identified using z-scores and the IQR method, ensuring data consistency and reliability.

5. **Data Imputation:** Missing values are systematically imputed using mean values based on groupings by 'City', 'Year', and 'Month', and extended AQI ranges are implemented for improved categorization.

6. **Feature Engineering and Arranging:** New features like "Seasonality" are introduced, and the dataset is reorganized for more efficient analysis.

7. **Label Encoding and Data Split:** Categorical variables are encoded numerically, and the dataset is divided into an 80% training set and a 20% test set, preparing it for machine learning applications.

8. **Data Scaling:** The training and test data are standardized using scikit-learn's StandardScaler, ensuring a normal distribution of data, a prerequisite for many machine learning.

Process 3 - Exploratory Data Analysis (EDA)

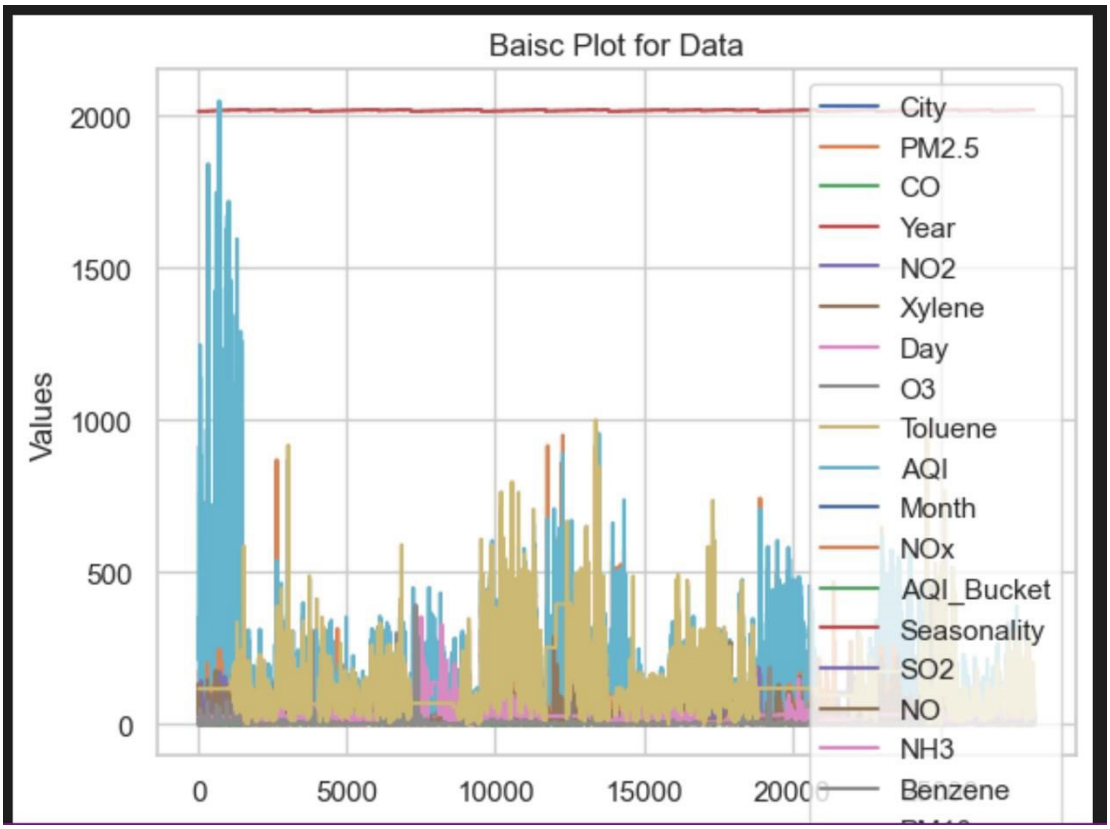
Dataset Overview: The EDA begins with a descriptive analysis, revealing a dataset of 28,157 entries and 19 columns, all non-null, indicating a clean dataset ready for further exploration.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28157 entries, 0 to 28156
Data columns (total 19 columns):
 #    Column         Non-Null Count   Dtype
---   ------         --------------   -----
 0    City           28157 non-null   int64
 1    Year           28157 non-null   int64
 2    Month          28157 non-null   int64
 3    Day            28157 non-null   int64
 4    Seasonality    28157 non-null   int64
 5    PM2.5          28157 non-null   float64
 6    PM10           28157 non-null   float64
 7    NO             28157 non-null   float64
 8    NO2            28157 non-null   float64
 9    NOx            28157 non-null   float64
 10   NH3            28157 non-null   float64
 11   CO             28157 non-null   float64
 12   SO2            28157 non-null   float64
 13   O3             28157 non-null   float64
 14   Benzene        28157 non-null   float64
 15   Toluene        28157 non-null   float64
 16   Xylene         28157 non-null   float64
 17   AQI            28157 non-null   float64
 18   AQI_Bucket     28157 non-null   int64
dtypes: float64(13), int64(6)
memory usage: 4.1 MB
```

Statistical Summary: Provides detailed statistical insights into the dataset, encompassing mean, standard deviation, and quartiles, highlighting the diversity of air quality indicators.

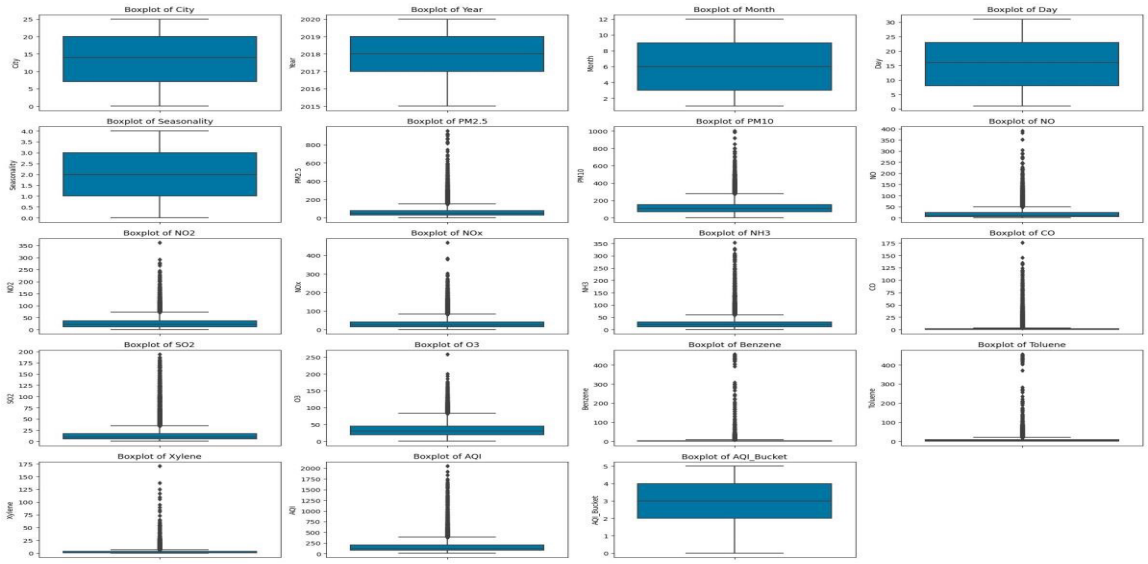
	City	Year	Month	Day	Seasonality	PM2.5	PM10	NO	NO2
count	28157.000000	28157.000000	28157.000000	28157.000000	28157.000000	28157.000000	28157.000000	28157.000000	28157.000000
mean	13.020954	2017.917108	6.256100	15.757716	2.098057	66.450276	121.725411	18.975098	28.294443
std	7.475270	1.541680	3.442144	8.813237	1.330389	62.673440	84.790508	24.270416	23.857254
min	0.000000	2015.000000	1.000000	1.000000	0.000000	0.040000	0.010000	0.020000	0.010000
25%	7.000000	2017.000000	3.000000	8.000000	1.000000	30.162000	68.591271	5.900000	12.060000
50%	14.000000	2018.000000	6.000000	16.000000	2.000000	48.180000	103.200000	10.590000	22.680000
75%	20.000000	2019.000000	9.000000	23.000000	3.000000	78.550000	151.570000	23.190000	36.410000
max	25.000000	2020.000000	12.000000	31.000000	4.000000	949.990000	1000.000000	390.680000	362.210000

Visual Analysis: Initial plots offer a high-level overview of the numerical data, but more detailed graphs like categorical distributions, feature box plots, histograms, and various AQI-related visualizations offer deeper insights into the data's characteristics.

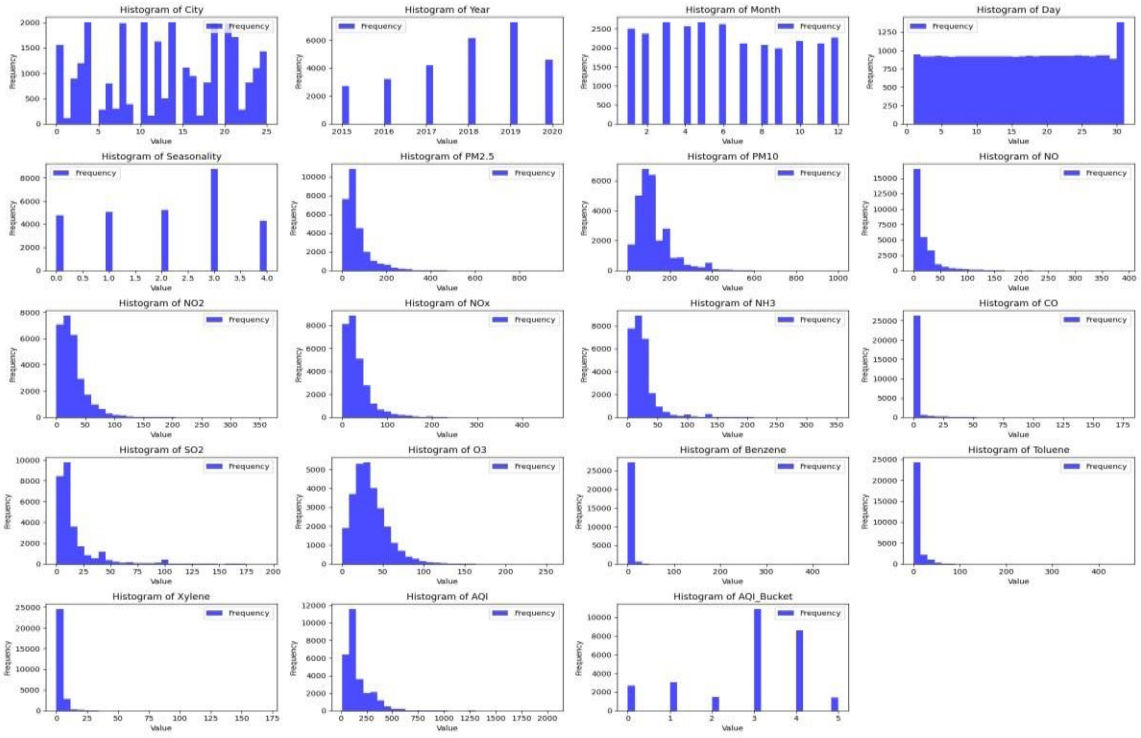


Feature Relationships: Scatter plots, kernel density plots, and Pearson correlation heatmaps are employed to analyze relationships between various features, identifying potential correlations and patterns. It uses box plots, histograms, AQI by month stacked bar chart etc.

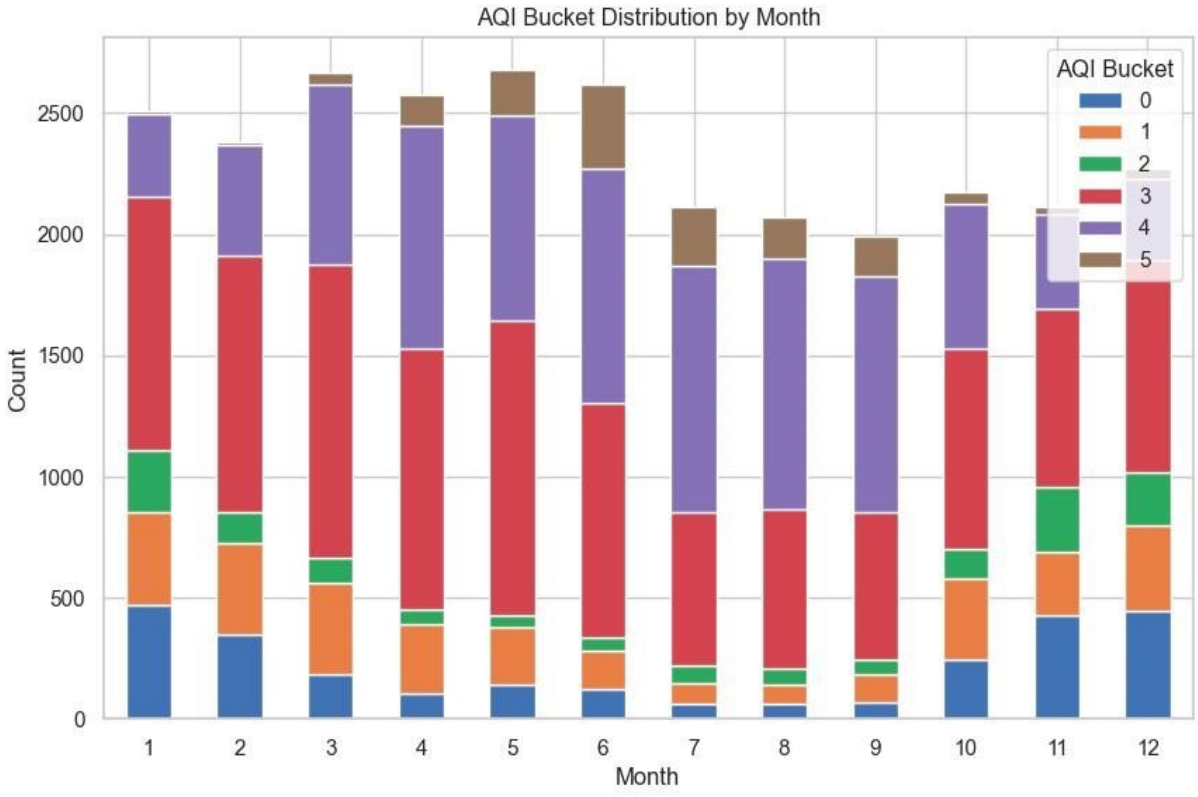
1. Box plot for AQI features



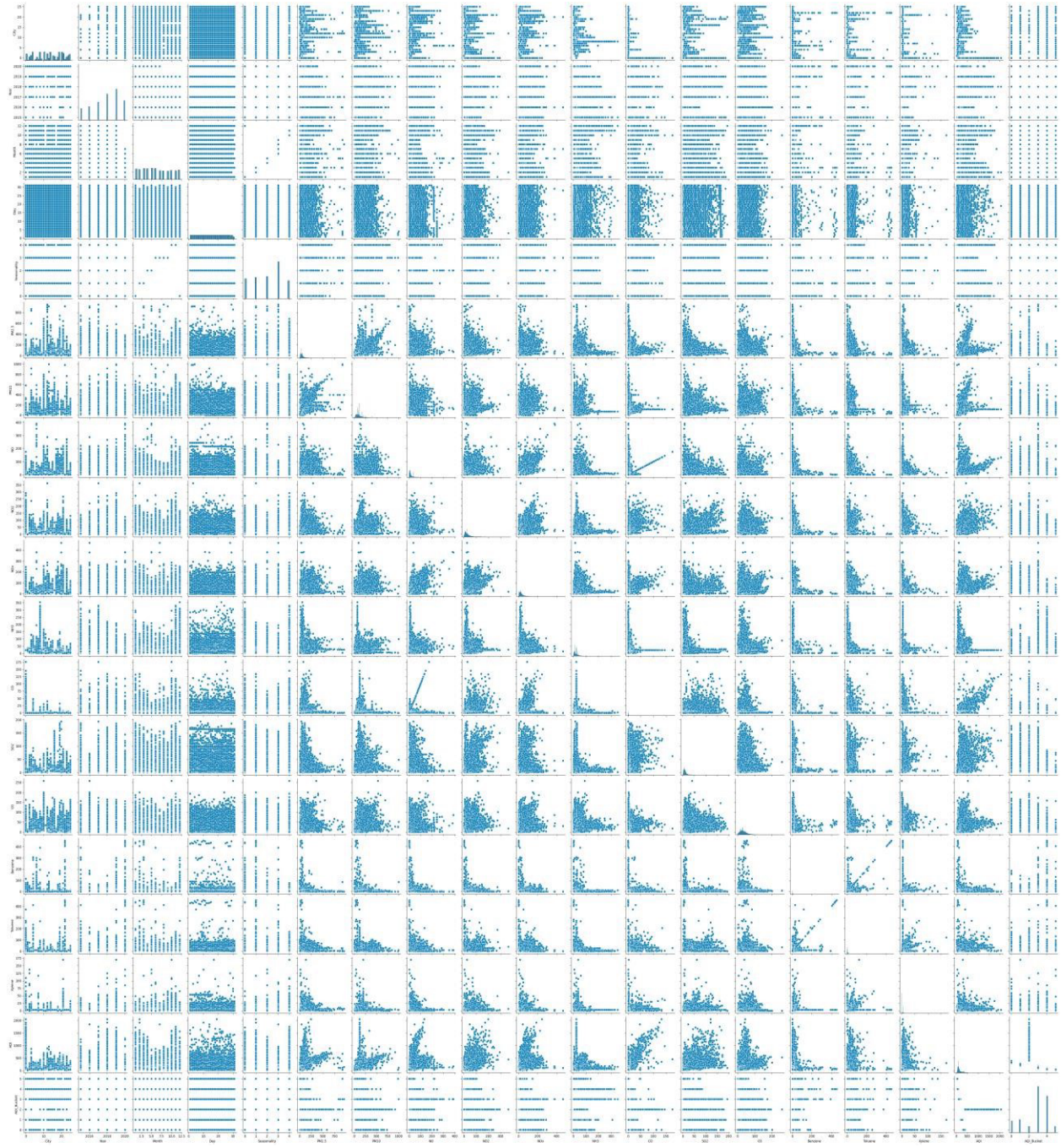
2. Histograms for AQI features



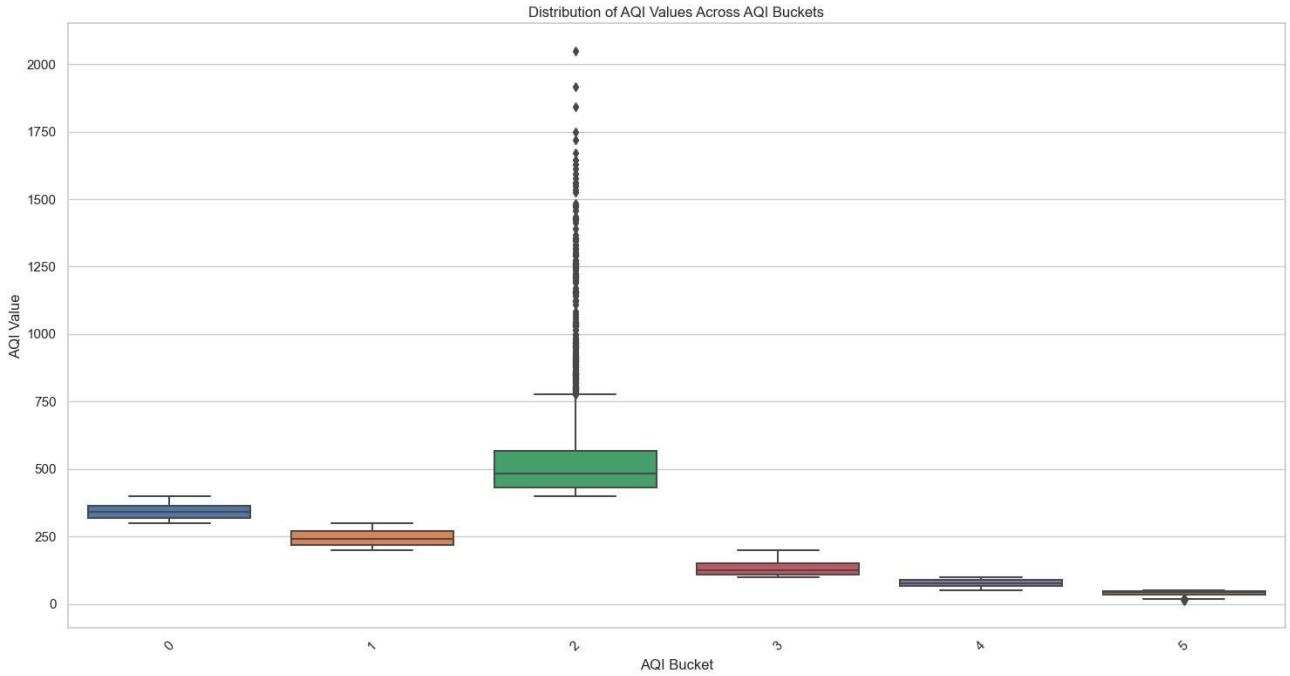
3. AQI by Month Stacked Bar Chart



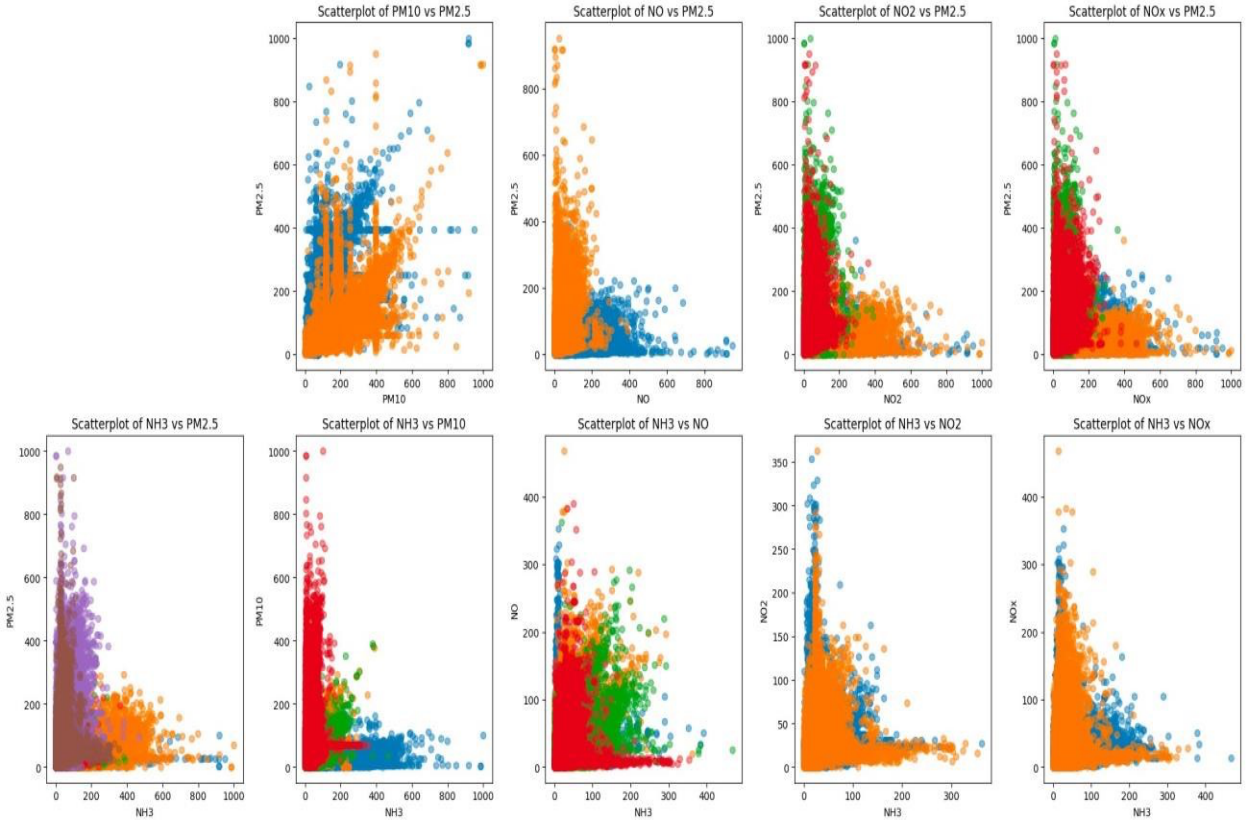
4. Pair plot



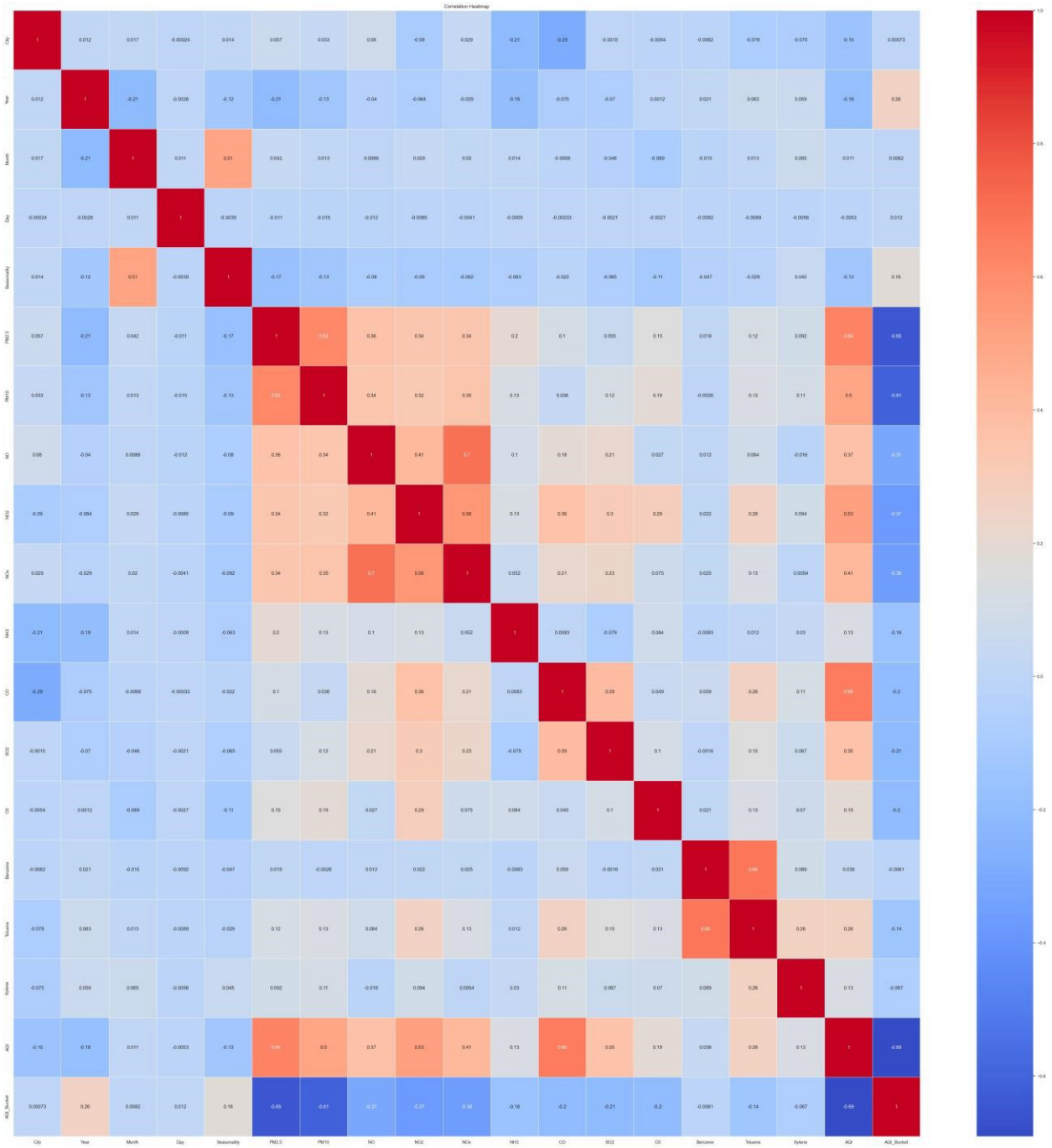
5. AQI Ranges by AQI Bucket



6. Scatter Plots of Key Features

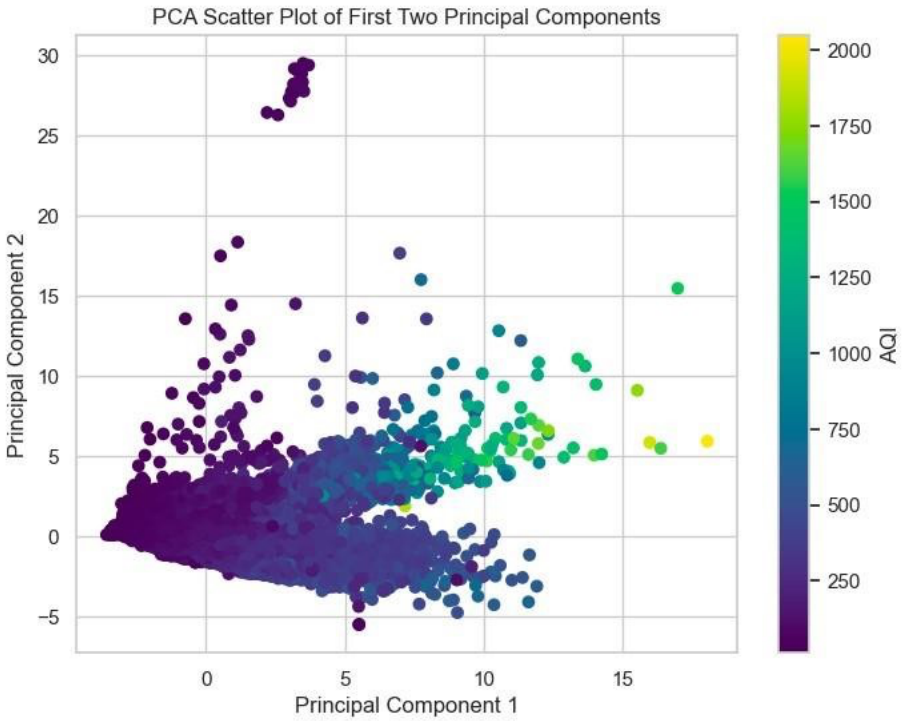


7. Pearson Correlation Heatmap

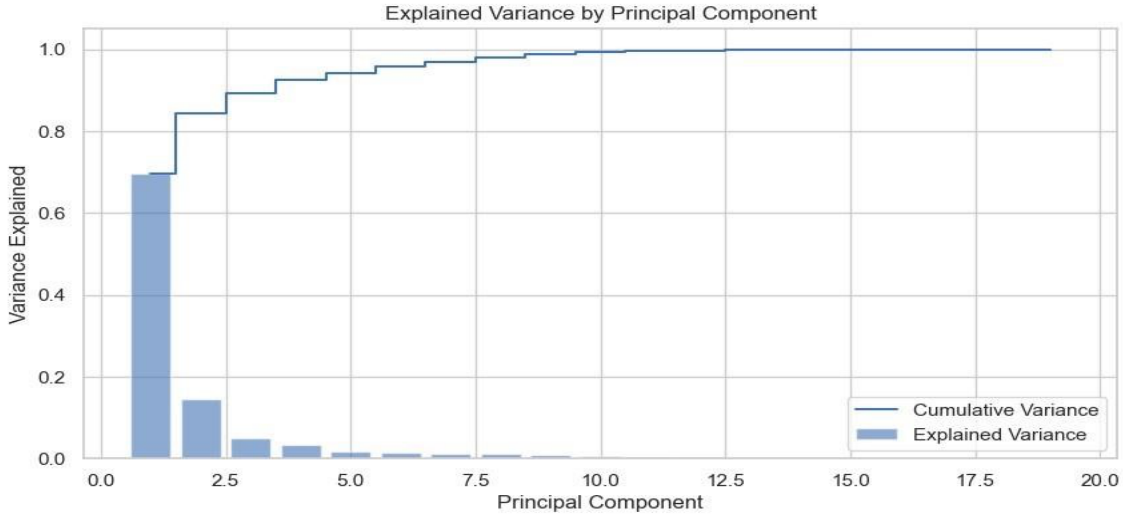


PCA Analysis: Principal Component Analysis (PCA) is utilized for dimensionality reduction, with scatter plots, elbow plots, and variance charts providing insights into the principal components that capture most of the dataset's variance.

8. PCA Scatter Plot of First Two Principal Components



9. Explained and Cumulative Variance by Principal Components



Subsequently, following the preprocessing phase, we will proceed to utilize the processed datasets, namely "clean_air.csv" and "clean_air_hour.csv" for the upcoming stages of the project.

Process 4 - Implementation of Machine Learning algorithms

Feature Selection: The project began with an ANOVA F-test to evaluate the significance of each feature, using F-scores to assess variance and P-values to test the null hypothesis. A RandomForestClassifier further identified crucial features, focusing on those contributing most to model accuracy. This dual approach ensured a data-driven foundation for feature importance.

	F_Scores	P_Values	Importance
PM2.5	6752.133070	0.000000e+00	0.256332
PM10	3652.051058	0.000000e+00	0.176499
CO	2834.236856	0.000000e+00	0.104179
NO2	1897.656065	0.000000e+00	0.046748
NOx	1210.943865	0.000000e+00	0.041208
NO	1018.362870	0.000000e+00	0.046487
SO2	977.027603	0.000000e+00	0.045688
Year	511.830879	0.000000e+00	0.023576
O3	431.913901	0.000000e+00	0.053132
Toluene	412.606686	0.000000e+00	0.032857
Seasonality	233.274804	6.780665e-245	0.012019
City	217.624651	1.335339e-228	0.025623
NH3	177.319836	1.882239e-186	0.039499
Xylene	62.133753	1.172252e-64	0.022864
Month	42.978322	2.743983e-44	0.018990
Benzene	9.990069	1.446681e-09	0.029721
Day	2.667942	2.041514e-02	0.024578

Dimensionality Reduction: To concentrate on the most informative features, StandardScaler was used for normalization before applying Principal Component Analysis (PCA). PCA streamlined the dataset, retaining principal components that captured 95% of the variance, effectively reducing complexity while preserving essential information.

	F_Scores	P_Values	Importance
PM2.5	6752.133070	0.000000e+00	0.256332
PM10	3652.051058	0.000000e+00	0.176499
CO	2834.236856	0.000000e+00	0.104179
NO2	1897.656065	0.000000e+00	0.046748
NOx	1210.943865	0.000000e+00	0.041208
NO	1018.362870	0.000000e+00	0.046487
SO2	977.027603	0.000000e+00	0.045688

Correlation Analysis: A thorough correlation analysis was conducted, leading to the removal of features with a correlation above 0.95 to eliminate redundancy. This step was crucial in ensuring the model's predictors were independent, enhancing the robustness of the predictive analysis.

Final Feature Set: The culmination of these steps resulted in a carefully curated set of features, optimized for predictive modeling. This final feature set was divided into training and test datasets, setting the stage for effective model training with a 20% test size.

Algorithm 1 – Linear Regression:

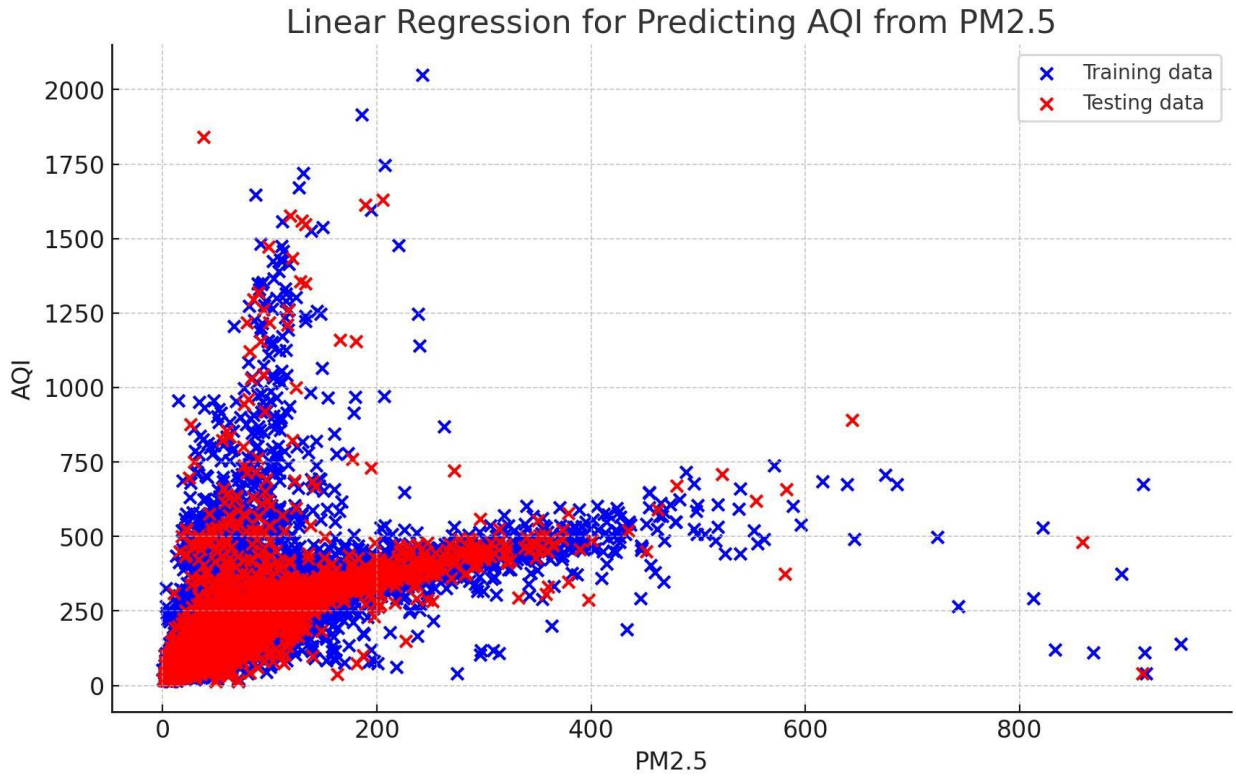
Justification for Linear Regression: Linear Regression was selected for its simplicity and ability to establish a linear relationship between predictors (like PM2.5, PM10, NO2) and the AQI. Its interpretability aligns with the project's goal of understanding the impact of various air quality metrics.

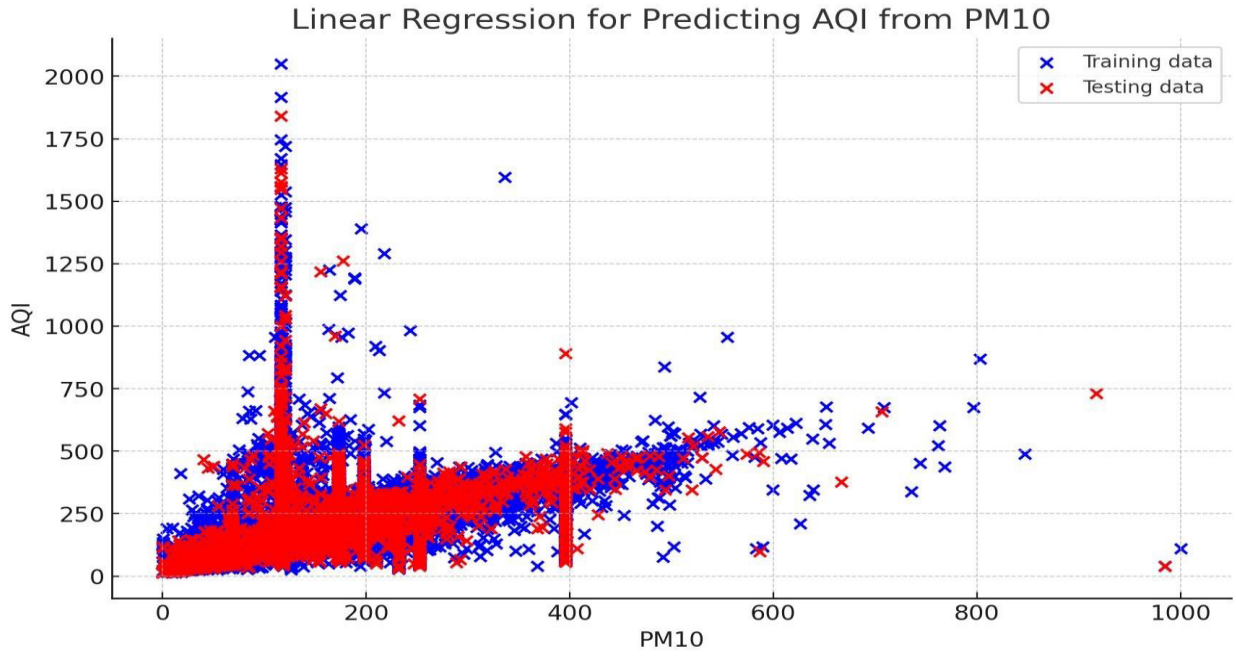
Model Training and Tuning: Utilizing the Linear Regression class from scikit-learn, the model was trained with ordinary least squares (OLS), a simple yet effective approach, capturing the natural variance in the data without the need for complex hyperparameter tuning.

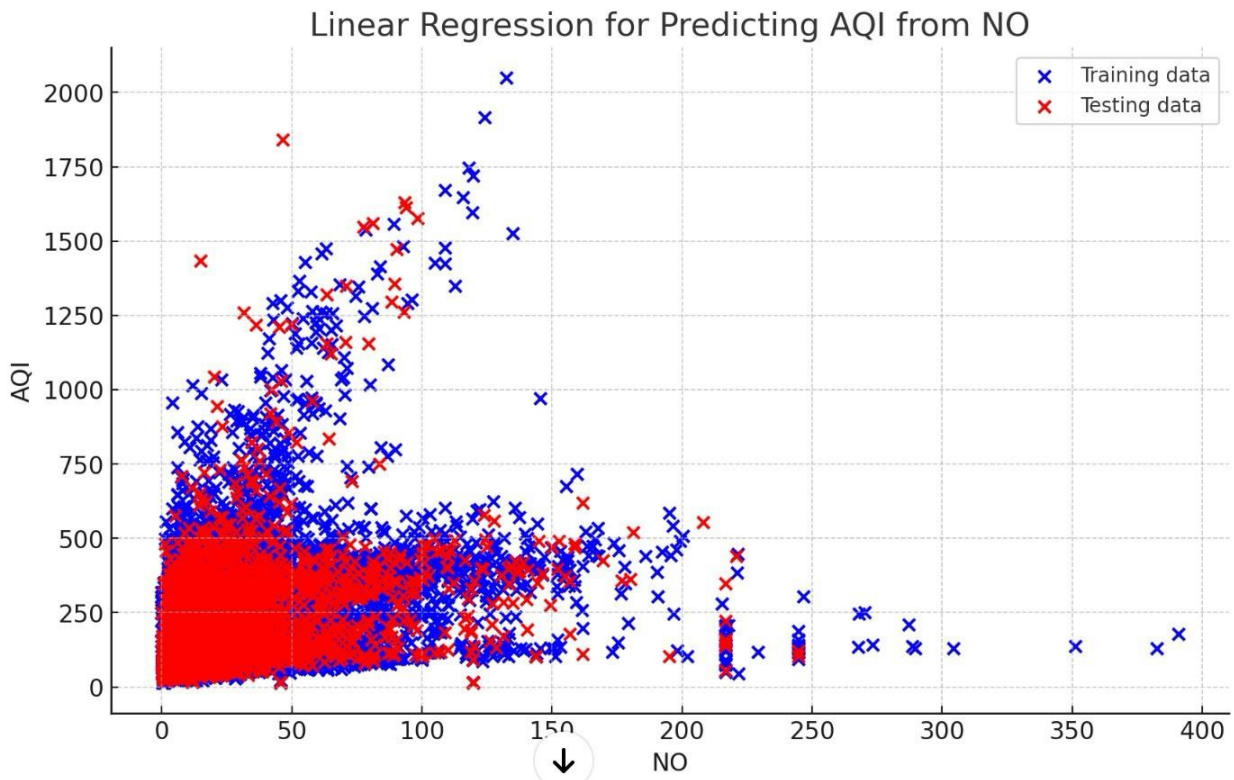
Results: The Linear Regression model demonstrated an R^2 score of around 0.81 on both training and testing data, indicating a strong predictive capability. However, higher RMSE and MAE values suggested room for improvement, possibly through more sophisticated models or additional features.

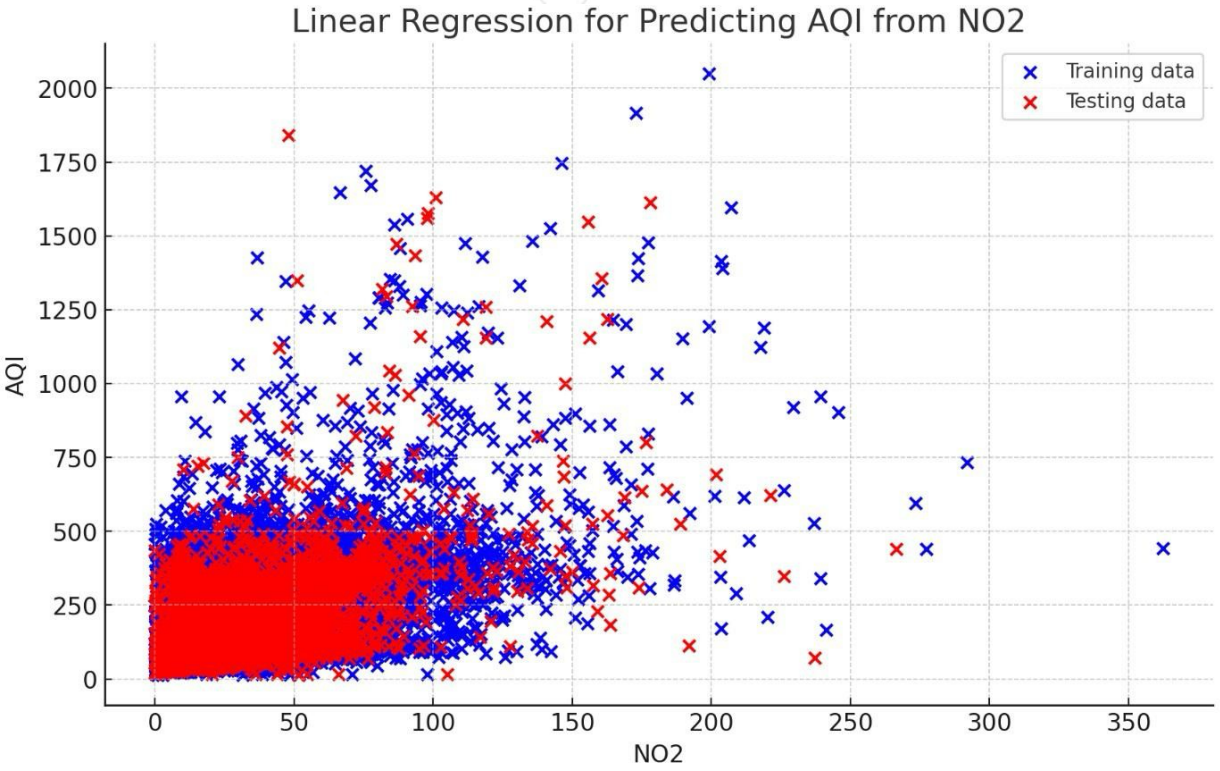
```
Training Metrics {'MSE': 3557.8196779886466, 'RMSE': 59.647461622341034, 'MAE': 32.82167655568622,
'Median AE': 21.19087076834691, 'R^2': 0.8112370160052313, 'Explained Variance': 0.8112370160052313, 'Max
Error': 1480.445815954483}
Testing Metrics {'MSE': 3759.8514628606704, 'RMSE': 61.31762766823803, 'MAE': 32.84993145701751, 'Median
AE': 21.389444670403883, 'R^2': 0.8050621607177213, 'Explained Variance': 0.8050634709504196, 'Max
Error': 1246.0802536479737}
```

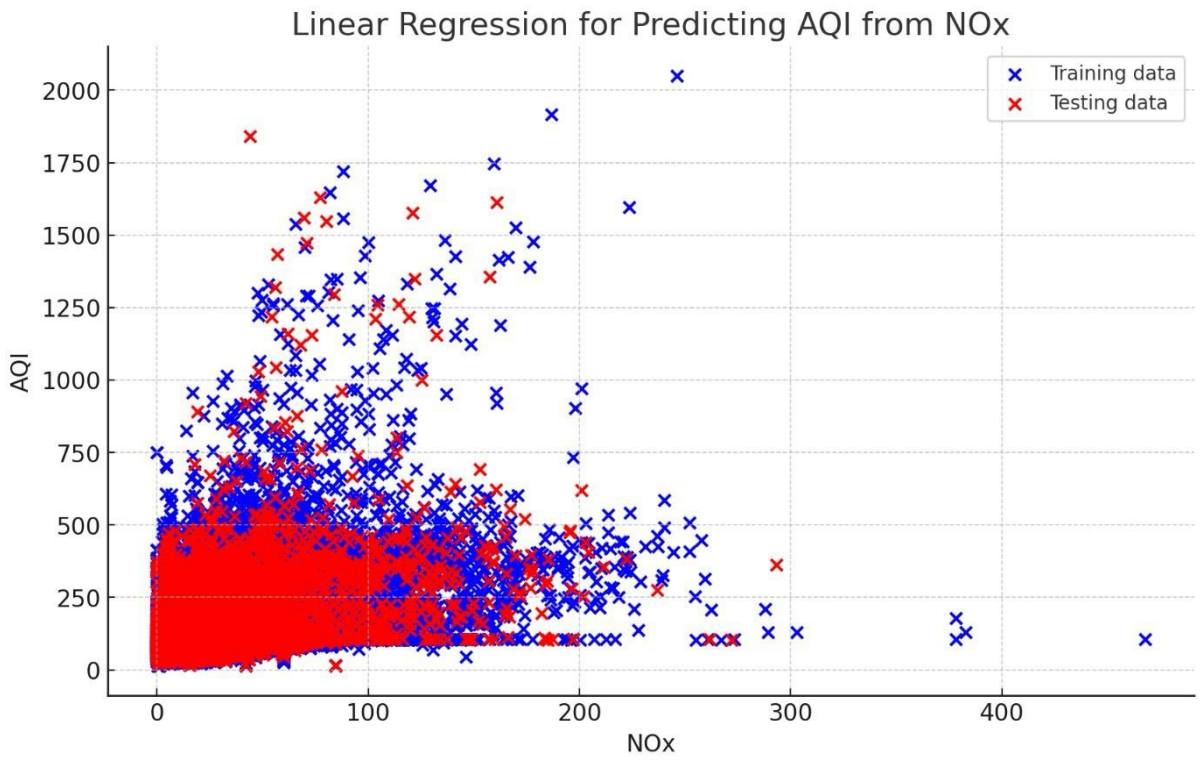
Visualization plots: Scatter plots for each feature against AQI were generated, offering a visual assessment of the linearity, and identifying any outliers, thus aiding in understanding the model's consistency and generalizability.

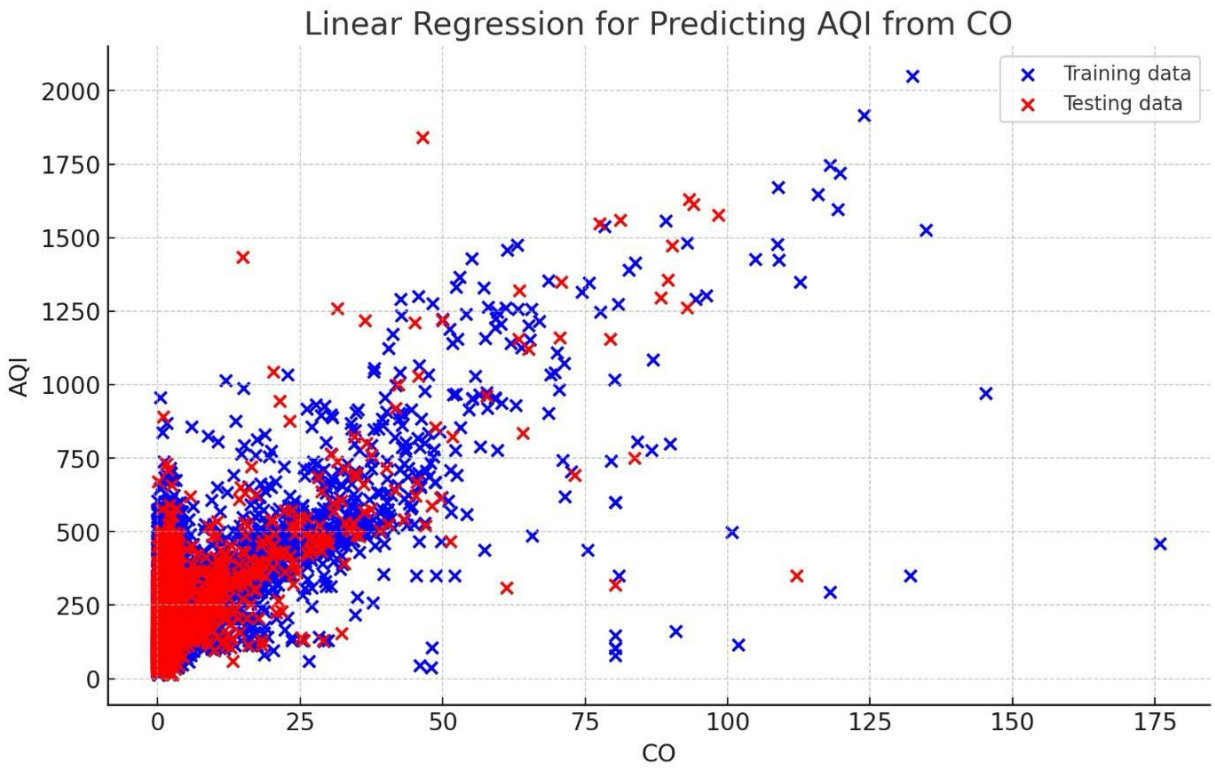

Linear Regression for Predicting AQI from PM2.5
Training data
Testing data
AQI
PM2.5
0
250
500
750
1000
1250
1500
1750
2000
0
200
400
600
800

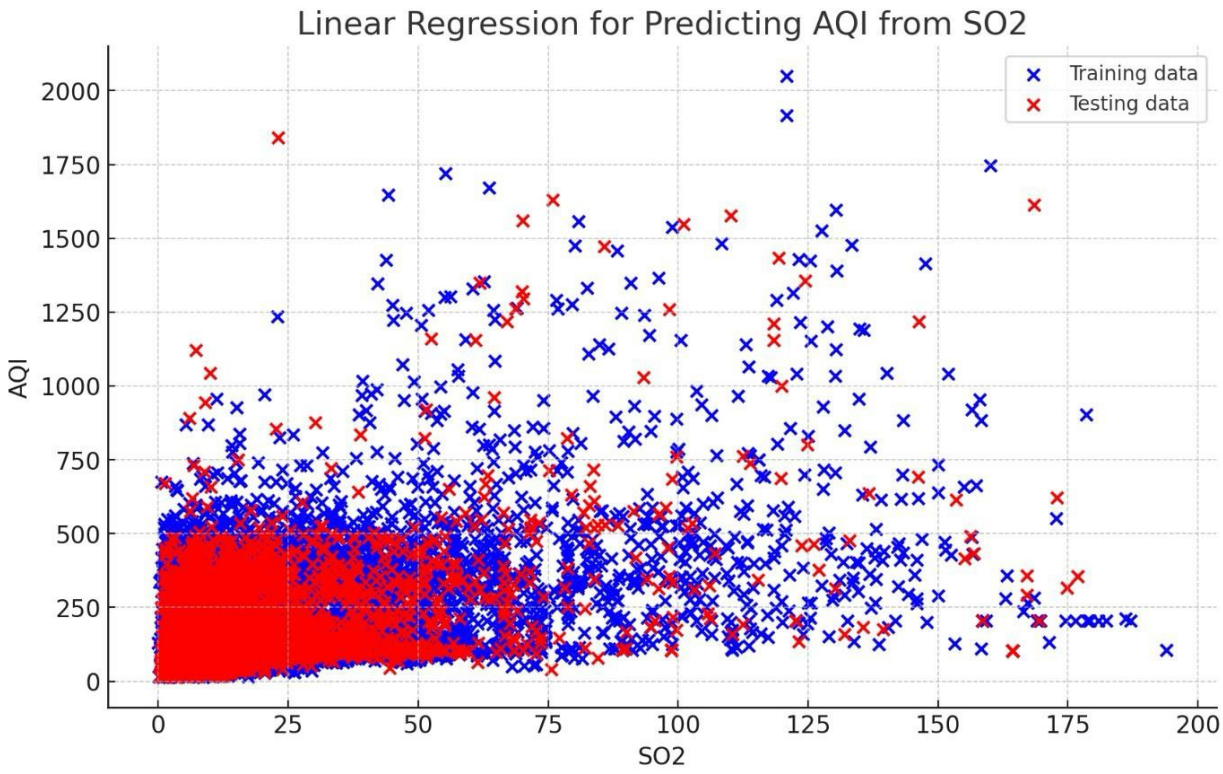

Linear Regression for Predicting AQI from PM10
Training data
Testing data
AQI
PM10
0
250
500
750
1000
1250
1500
1750
2000
0
200
400
600
800
1000


Linear Regression for Predicting AQI from NO
Training data
Testing data
AQI
NO
0
250
500
750
1000
1250
1500
1750
2000
0
50
100
150
200
250
300
350
400


Linear Regression for Predicting AQI from NO2
Training data
Testing data
AQI
NO2
0
250
500
750
1000
1250
1500
1750
2000
0
50
100
150
200
250
300
350


Linear Regression for Predicting AQI from NOx
Training data
Testing data
AQI
NOx
0
250
500
750
1000
1250
1500
1750
2000
0
100
200
300
400


Linear Regression for Predicting AQI from CO
Training data
Testing data
AQI
CO
0
250
500
750
1000
1250
1500
1750
2000
0
25
50
75
100
125
150
175



Conclusion: The regression analysis highlighted the significant impact of pollutants like PM2.5 and PM10 on AQI, corroborating their known importance in air quality assessment and public health implications.

Algorithm 2 – Ridge Regression:

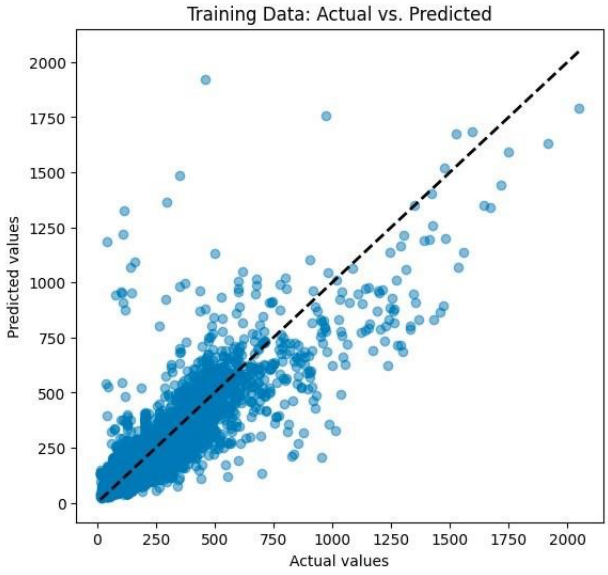
Justification for Ridge Regression: Ridge Regression was selected for its ability to handle multicollinearity, a common issue in complex datasets like air quality metrics. Its regularization technique moderates the influence of less important features, making it ideal for datasets with interrelated predictors such as PM2.5, PM10, and NO2.

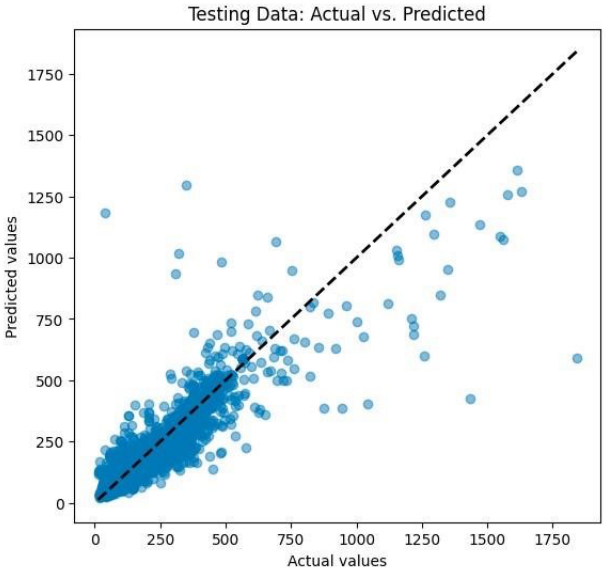
Model Training and Tuning: The model training involved experimenting with various alpha values to find the optimal level of regularization. A large alpha value of 10,000, chosen based on cross-validation, indicated a need for significant regularization to prevent overfitting, enhancing the model's stability and predictive accuracy.

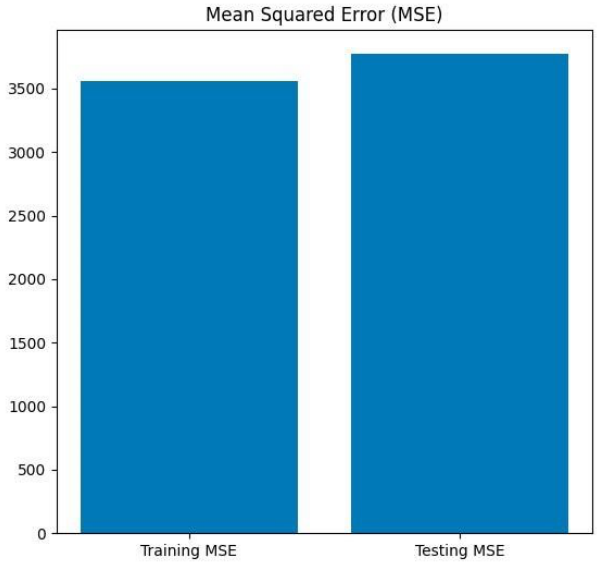
Results: The Ridge Regression model achieved R^2 scores of about 0.81 for training and 0.80 for testing, affirming its capacity to explain a substantial portion of AQI variance. However, the MSE values highlighted accuracy challenges, especially for higher AQI values, suggesting the model's limitations in precision.

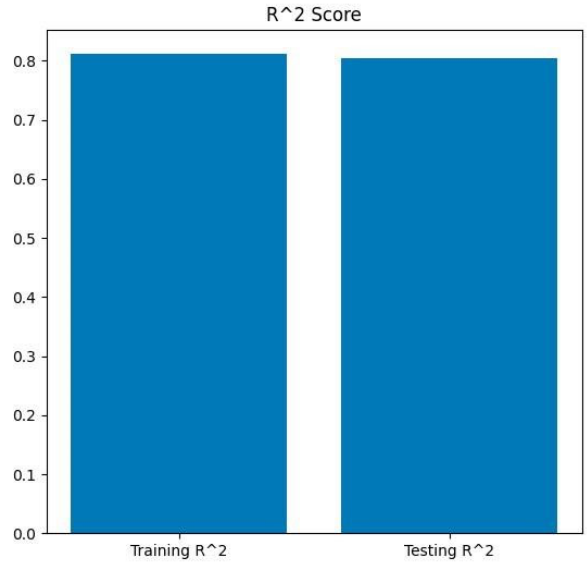
```
Best alpha: 10000.0
Training Metrics: {'MSE': 3558.287449951844, 'RMSE': 59.65138263235684, 'MAE': 32.89415351105614, 'Median
AE': 21.058102676783136, 'R^2': 0.8112121979875705, 'Explained Variance': 0.8112121979875706, 'Max
Error': 1462.8859489885817}
Testing Metrics: {'MSE': 3771.143694503327, 'RMSE': 61.40963844954086, 'MAE': 32.9374557806701, 'Median
AE': 21.27542449191371, 'R^2': 0.804476689919516, 'Explained Variance': 0.8044778332405108, 'Max Error':
1250.4734414642744}
```

Visualization Observations: Scatter plots of actual versus predicted AQI values showed a strong alignment at lower AQI levels, but a divergence at higher levels. This discrepancy highlighted the model's constraints in accurately predicting higher AQI values, a critical area for improvement.









Conclusion: Transitioning to Ridge Regression provided a balance between model simplicity and the need for more sophisticated handling of data complexity. While it enhanced generalizability, the high alpha value and the model's difficulty with high AQI values suggest the potential benefit of exploring more complex models or incorporating additional variables like weather conditions or industrial activities. The model reinforced the significance of PM2.5 and PM10 as key predictors, emphasizing their importance in air quality assessment.

Algorithm 3 – Random Forest Regression:

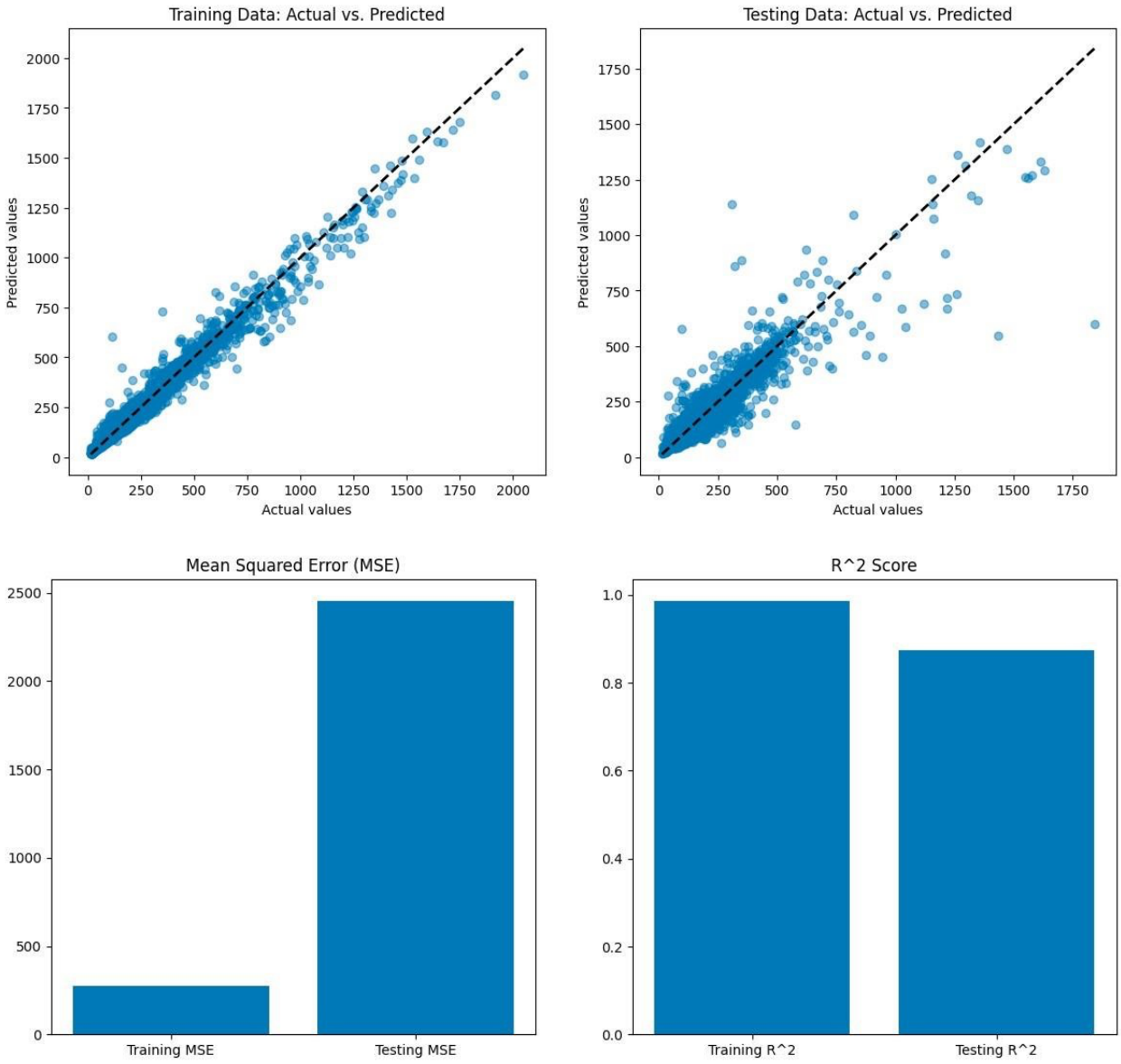
Justification for RandomForestRegressor: The RandomForestRegressor was chosen for its ability to capture complex, non-linear relationships typical in environmental datasets. Its suitability for the multi-faceted nature of air quality data, where various indicators interact to determine AQI levels, makes it an apt choice for this analysis.

Model Training and Tuning: The model was trained using its default settings, leveraging the inherent randomization and ensemble approach of RandomForest to avoid overfitting. This method is particularly effective for datasets with a multitude of predictors, as in the case of the comprehensive air quality dataset used here.

Results: The RandomForestRegressor showed excellent performance, with a training R^2 value of 0.9853, indicating it could explain about 98.53% of the variance in AQI. On the testing data, it achieved an R^2 of 0.8729, a slight reduction but still indicative of a strong model capable of generalizing well to new data.

```
Training Metrics: {'MSE': 276.52228030129703, 'RMSE': 16.628959086524237, 'MAE': 8.001241491027608,
'Median AE': 4.159999999999997, 'R^2': 0.9853288880564572, 'Explained Variance': 0.9853312887507563, 'Max
Error': 488.20000000000005}
Testing Metrics: {'MSE': 2451.9597398442143, 'RMSE': 49.5172670877969, 'MAE': 22.45102642870698, 'Median
AE': 11.372019333637233, 'R^2': 0.8728727082934548, 'Explained Variance': 0.8728791847732269, 'Max
Error': 1241.497037037037}
```

Visualization Observations: Scatter plots revealed a high accuracy in predicting lower AQI values but showed a spread at higher AQI levels, suggesting a decline in model performance for these critical ranges. Performance metrics like MSE on the testing dataset highlighted some prediction errors, with R^2 scores affirming the model's overall robustness.



Conclusion: While the RandomForestRegressor's prediction accuracy for lower AQI values is notable, the increased error at higher AQI levels identifies areas for improvement. Future enhancements could involve hyperparameter optimization and integrating additional data sources to improve predictions for high AQI ranges. Accurately forecasting these levels is vital for informing effective air quality management strategies in urban settings.

Algorithm 4 - Random Forest Regression with Hyperparameter Tuning:

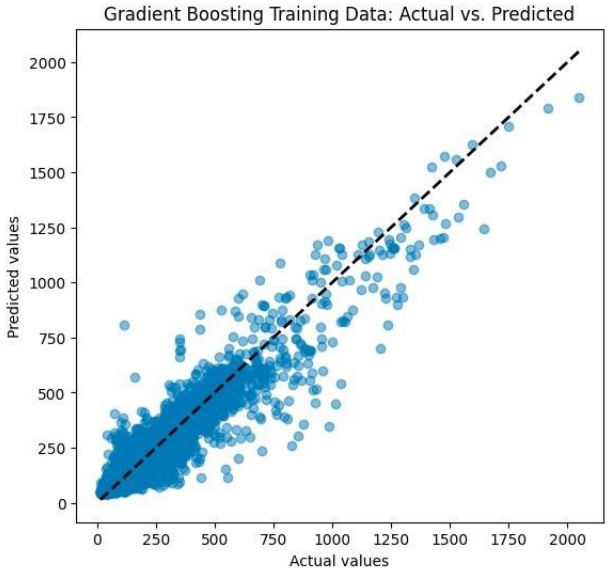
Justification for Random Forest Regression: Random Forest was selected due to its ability to handle complex, non-linear interactions between multiple predictors, making it ideal for datasets like ours with various inputs influencing AQI. Its ensemble method, combining multiple decision trees, enhances model accuracy and stability, making it robust against overfitting.

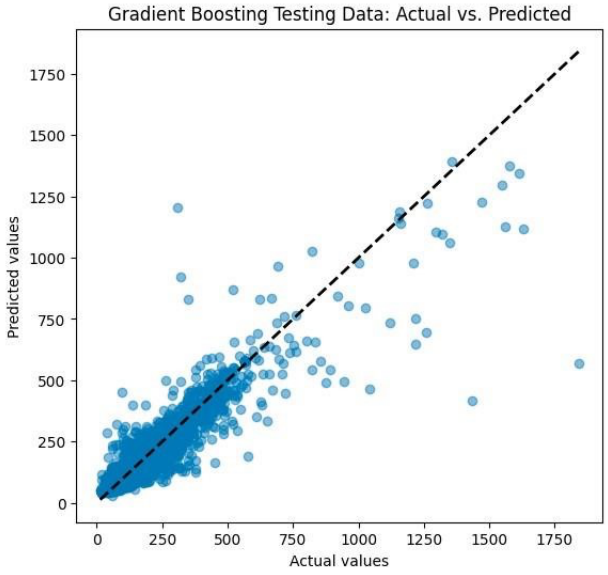
Model Training and Tuning: Training involved the RandomForestRegressor with hyperparameters optimized using GridSearchCV. Key parameters such as the number of trees, tree depth, minimum samples for splitting, and leaf node samples were fine-tuned, with optimal parameters including 200 trees and a minimum of 2 samples per leaf node. This meticulous tuning aimed to strike a balance between model complexity and prediction accuracy.

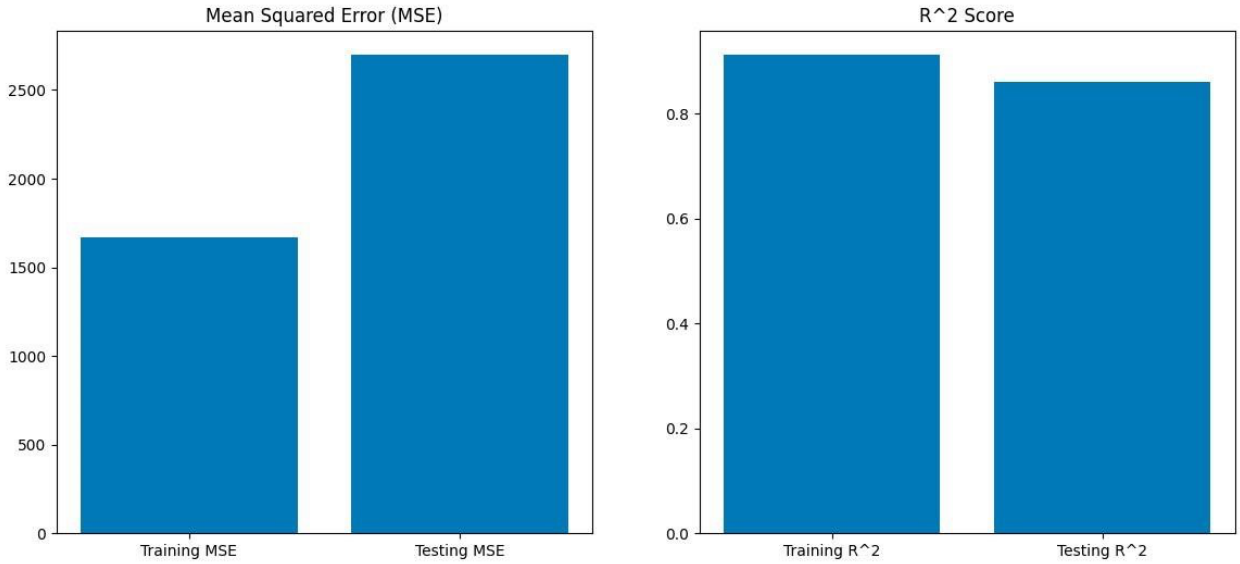
Results: The tuned model exhibited impressive predictive performance, achieving an R^2 of 0.9719 on training data and 0.8700 on testing data. While higher MSE and MAE values on the testing data indicated some overfitting, the overall performance suggests the model's robustness in predicting AQI. Its capacity to explain a significant portion of the variance highlights its suitability for the task, though there's room for improvement in generalization.

```
Best Parameters: {'bootstrap': True, 'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200}
Tuned Training Metrics: {'MSE': 528.2311460645495, 'RMSE': 22.98327970644202, 'MAE': 10.218446938711764, 'Median AE':
4.926511269450323, 'R^2': 0.9719742717746476, 'Explained Variance': 0.9719760548763183, 'Max Error': 790.0039738095239}
Tuned Testing Metrics: {'MSE': 2506.5152410784654, 'RMSE': 50.065110017640684, 'MAE': 22.401794730500924, 'Median AE':
11.263133460751064, 'R^2': 0.8700441573156789, 'Explained Variance': 0.8700619262476548, 'Max Error': 1225.5486849647266}
```

Visualization Observations: Visualization plots showed a close alignment of actual and predicted AQI values, especially in the training data, suggesting accurate predictions. Scatter plots displayed tight clustering around the perfect prediction line, with some deviation at higher AQI levels more noticeable in the test data. Performance metric bar charts reflected low error rates in training but higher rates in testing, supporting the model's efficacy while also indicating areas for refinement.







Conclusion: The hyperparameter-tuned Random Forest model demonstrates strong predictive ability, particularly with training data. Despite some overfitting evident in testing data, it remains a robust tool for AQI prediction. Its effectiveness in modeling complex data relationships makes it highly valuable for this analysis, with potential for further enhancements through more nuanced feature engineering or additional model optimizations.

Algorithm 5 – Gradient Boosting Regression:

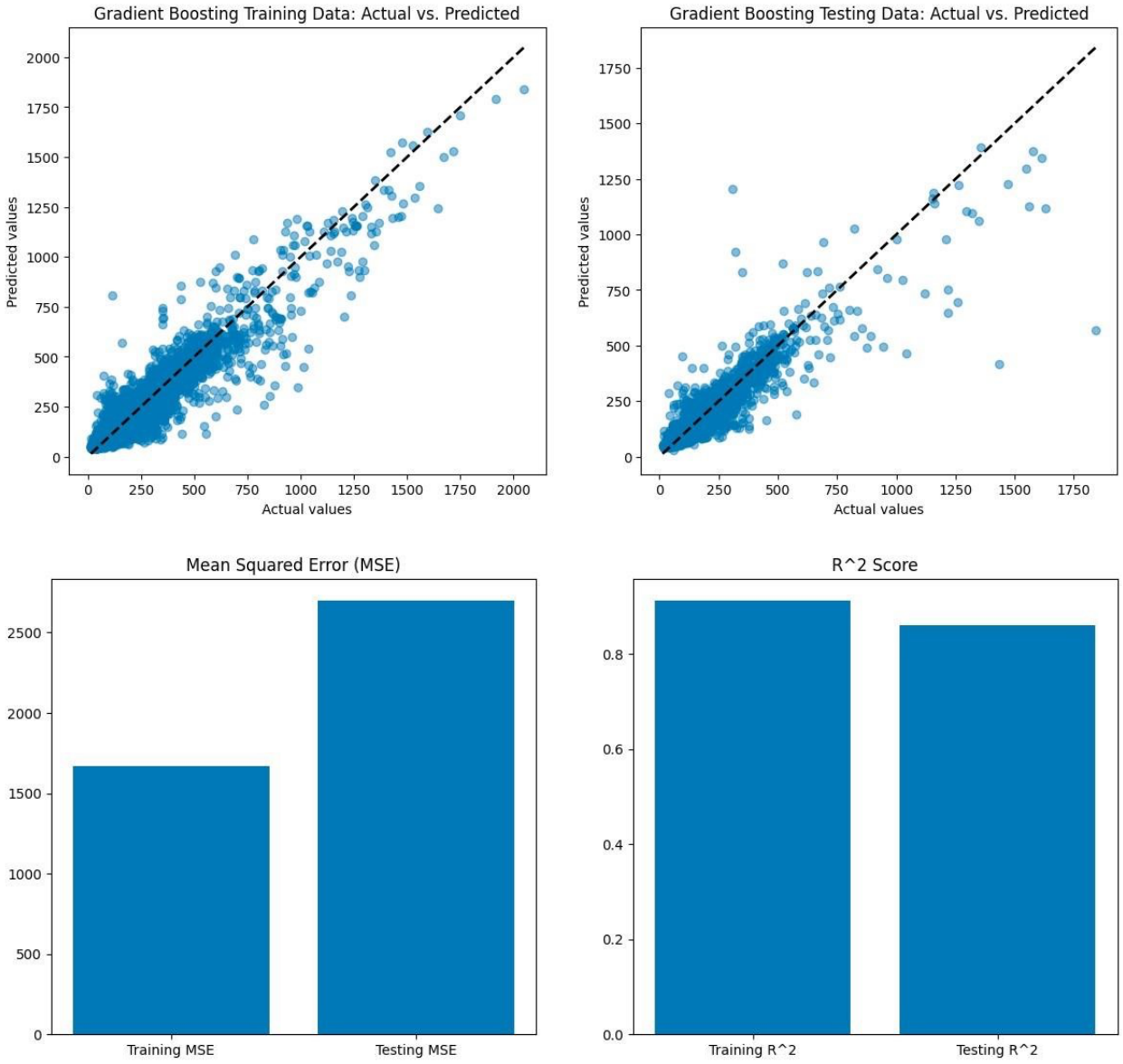
Justification for GradientBoostingRegressor: GradientBoostingRegressor (GBR) was chosen to address the overfitting tendency observed in the RandomForestRegressor. GBR's method of sequentially correcting errors from previous models makes it effective for handling complex, non-linear relationships in datasets, like ours with intricate predictors influencing AQI. It's particularly suited for situations requiring minimized overfitting while dealing with intricate predictor-response relationships.

Model Training and Tuning: GBR was carefully tuned using GridSearchCV, focusing on optimizing a set of hyperparameters including the number of estimators, learning rate, and tree depth. This fine-tuning aimed to balance the model's generalization capabilities and predictive accuracy, ensuring it remains effective without overfitting.

Results: The tuned GBR achieved a training R^2 of 0.9407, slightly lower than the RandomForestRegressor, but still robust, explaining 94.07% of the variance. Its testing R^2 score of 0.8718 was comparable to that of RandomForestRegressor. Notably, the increase in MSE on the training data indicates a reduced tendency to overfit compared to the RandomForestRegressor, enhancing its reliability.

```
Training Metrics: {'MSE': 1669.3131016817954, 'RMSE': 40.85722826724538, 'MAE': 23.72675691863523, 'Median AE':
14.966738591122834, 'R^2': 0.9114332510316663, 'Explained Variance': 0.9114332510316663, 'Max Error': 690.8757174103157}
Testing Metrics: {'MSE': 2697.1452896410947, 'RMSE': 51.934047499122336, 'MAE': 25.50366428661547, 'Median AE':
14.93427167935112, 'R^2': 0.8601605195879265, 'Explained Variance': 0.8601743397330044, 'Max Error': 1273.568919549466}
```

Visualization Observations: Visualization plots displayed a close match between actual and predicted AQI values for the training set, demonstrating accurate predictions. However, some deviation in the higher AQI range on the testing set pointed out areas where the model might be further optimized for better accuracy.



Conclusion: The GBR offers a nuanced approach in handling complex, higher AQI values, benefitting from its sequential learning algorithm and focus on reducing overfitting. Despite a slight increase in MSE for training data, this might be a worthwhile trade-off for the enhanced generalizability and accuracy in predicting higher AQI values that GBR provides. It stands as a strong contender among models for predicting AQI, particularly in complex environmental datasets.

Algorithm 5 – Gradient Boosting Regression with Hyperparameter Tuning:

Justification for Gradient Boosting Regression: Chosen for its sequential learning and error correction capabilities, Gradient Boosting Regression (GBR) stands out for handling complex datasets like ours, where intricate interactions among predictors influence AQI levels. Its focus on minimizing overfitting while enhancing predictive accuracy makes it a strong candidate for our AQI prediction model.

```
Fitting 3 folds for each of 96 candidates, totalling 288 fits
Best parameters: {'learning_rate': 0.1, 'max_depth': 4, 'max_features': None, 'min_samples_leaf': 2,
'min_samples_split': 2, 'n_estimators': 200}
```

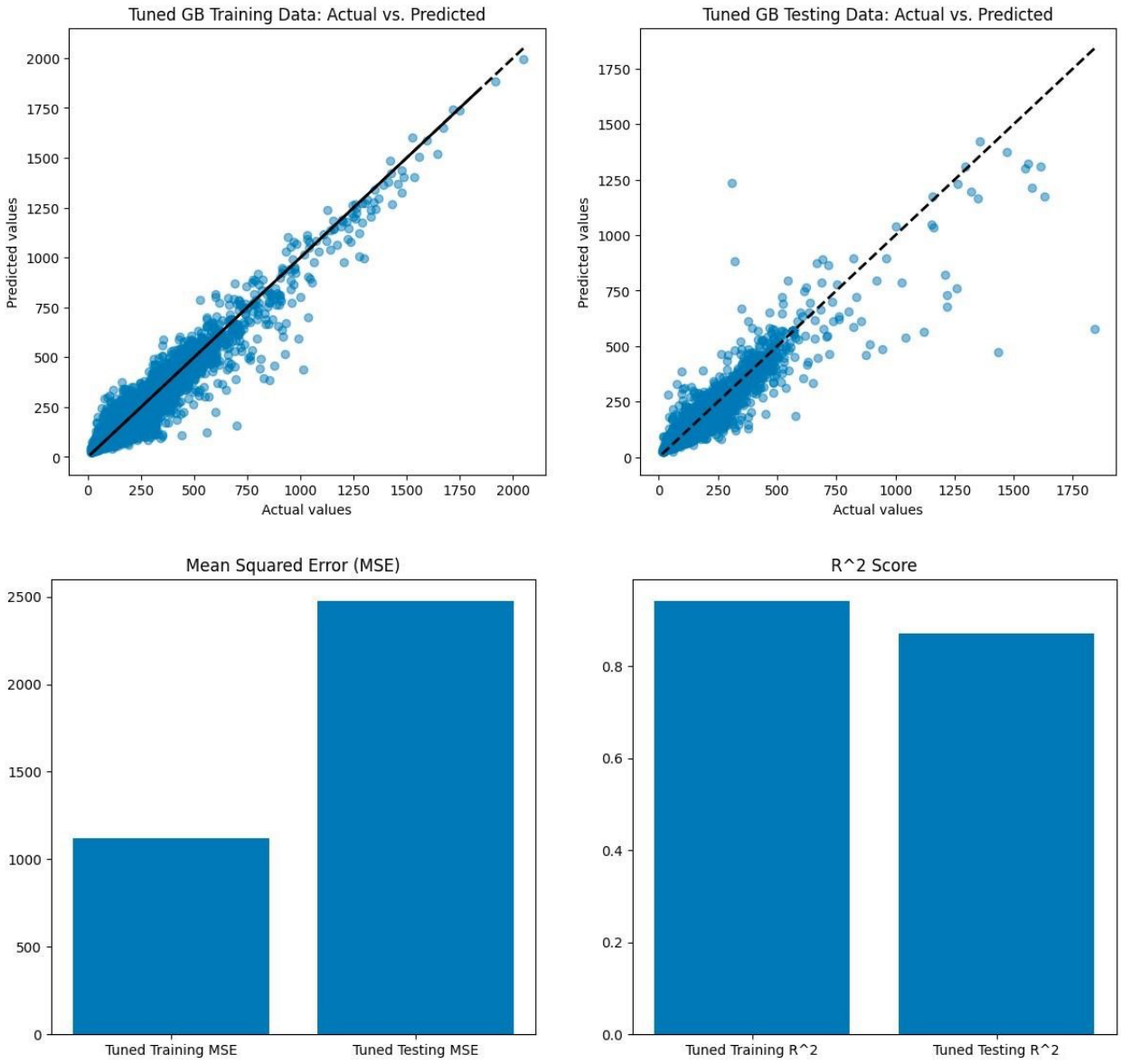
Model Training and Tuning: Utilizing GridSearchCV, the GBR model was fine-tuned by exploring a variety of hyperparameters. The optimal configuration included a higher learning rate and number of estimators, along with a balanced tree depth and specific sample split criteria. This process was critical to finding the right balance between model complexity and learning efficiency.

Results: The tuned GBR model demonstrated its robustness with an R^2 of 0.9407 on training data and 0.8718 on testing data. The increase in MSE for the training data, however, indicates a compromise of fit on training data to achieve better generalizability, a crucial aspect for realworld applications.

```
Training Metrics:
MSE: 1118.2574633891845
RMSE: 33.440356807145236
MAE: 20.540805983539297
Median AE: 13.185605651095557
R^2: 0.940669951046227
Explained Variance: 0.940669951046227
Max Error: 575.3919836077823

Testing Metrics:
MSE: 2473.466455455366
RMSE: 49.733956764522226
MAE: 23.74240096242362
Median AE: 13.357568011236737
R^2: 0.8717576449158961
Explained Variance: 0.8717835069760832
Max Error: 1263.5745462068885
```

Visualization Observations: Visualization plots revealed a strong alignment of predicted versus actual AQI values, particularly at lower levels, highlighting the model's accuracy. Deviations at higher AQI values in both training and testing sets pointed to areas where the model could be further optimized.



Comparative Analysis: RandomForest vs. Gradient Boosting Regression: When compared to the RandomForestRegressor, the GBR, with a slightly higher MSE, suggests a deliberate tradeoff favoring generalization over perfect training data fit. This could indicate GBR's superior performance in handling new, unseen data.

Conclusion: The hyperparameter-tuned GBR has emerged as a viable option for AQI prediction, balancing accuracy with the ability to generalize to new data. Its performance suggests room for future improvements, like further hyperparameter refinement or integration of additional data sources, to enhance its predictions, especially for higher AQI values which are critical for environmental and public health analysis.

Algorithm 6 – Neural network architecture:

Justification for Neural Network Regression:

The Neural Network was selected due to its capacity to model complex, non-linear interactions, and patterns within the data. Given the potential complexity of the relationship between the features and the target variable, AQI, a neural network is well-suited to capture these intricate dependencies.

Model Training and Tuning:

A Sequential model with dense layers was constructed, featuring a dropout layer to mitigate overfitting. The model was compiled with a mean squared error loss function and optimized using the Adam optimizer. The training process involved scaling the features and iterating over 100 epochs with a validation split to monitor performance and avoid overtraining.

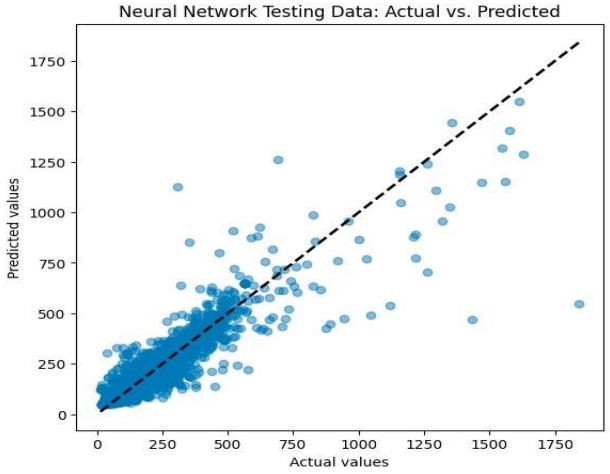
Results:

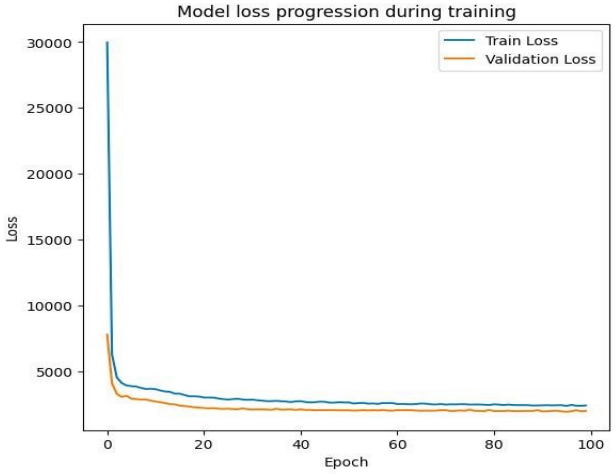
```
Traning Metrics {'MSE': 2213.186821963341, 'RMSE': 47.044519574158066, 'MAE': 26.095131253385148, 'Median
AE': 16.452232360839844, 'R^2': 0.8825775934524379, 'Explained Variance': 0.8833428819800258, 'Max
Error': 1279.38037109375}
Traning Metrics {'MSE': 2778.7224337765974, 'RMSE': 52.71358870136426, 'MAE': 26.698135405955497, 'Median
AE': 16.039608001708984, 'R^2': 0.8559309715938963, 'Explained Variance': 0.8565926759153327, 'Max
Error': 1292.9568481445312}
```

The model demonstrated good fit and predictive accuracy, with an R^2 of 0.8826 for training data and 0.8559 for testing data. The relatively close R^2 values indicate that the model generalizes well to unseen data, though the slightly higher errors on the testing set suggest some overfitting.

Visualization Plots:

The scatter plot of the neural network's predictions versus actual values for the testing data shows a strong alignment, with most data points clustering around the identity line, particularly for lower AQI values. The loss progression plot exhibits a rapid decrease in training and validation loss, indicating efficient learning and convergence of the model, which is maintained throughout the remaining epochs.





Conclusion:
The neural network has proven to be a potent model for predicting AQI values, effectively capturing the complexity within the dataset. While there is room for improvement in model generalization, as indicated by the slight increase in prediction error on the testing set, the overall performance is promising. Further enhancements might include hyperparameter tuning, increasing the dataset size, or experimenting with more sophisticated network architectures.

Algorithm 7 – Neural network architecture with hyperparameter tuning:

Justification for Neural Network Regression:
Neural Network Regression, known for its flexibility and depth, is adept at modeling complex, non-linear interactions within large datasets. For predicting AQI, a multi-layered neural network can learn and approximate the underlying function mapping environmental inputs to air quality outcomes.

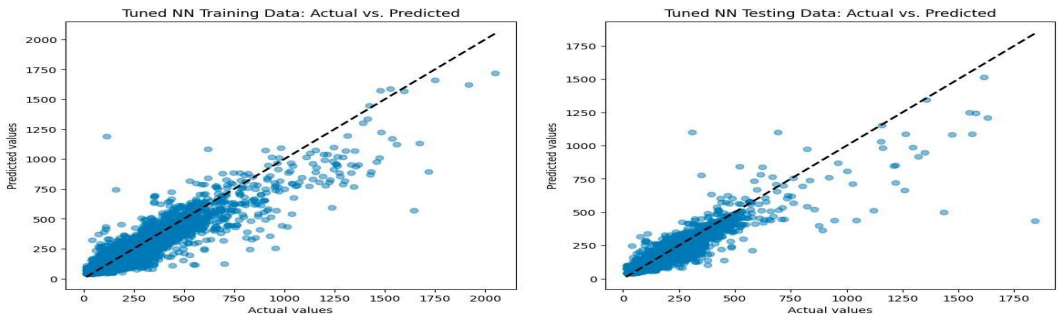
Model Training and Tuning:
Hyperparameter tuning was performed using Keras Tuner, which identified the optimal configuration for the network's architecture and training process. The network comprises densely connected layers with ReLU activation, dropout layers for regularization, and a linear activation function in the output layer for continuous value prediction. The model was compiled with the Adam optimizer, and the learning rate was determined through the tuning process.

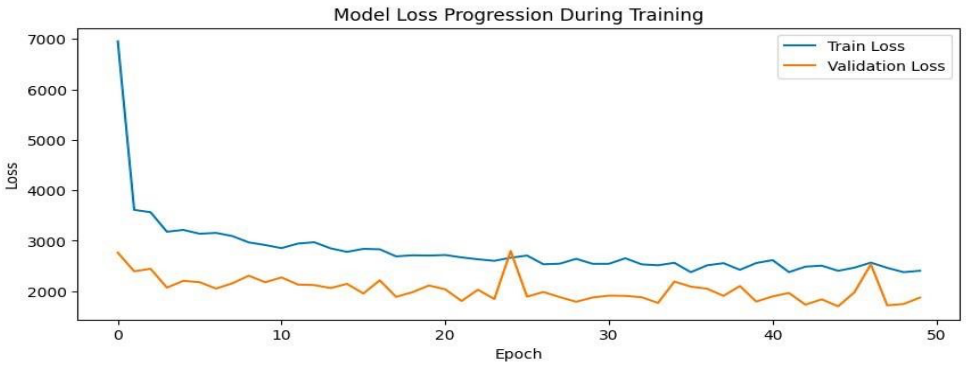
Results:

```
Tuned Training Metrics: {'MSE': 2028.6827771190356, 'RMSE': 45.04090115793683, 'MAE': 23.975832056150534,
'Median AE': 13.97607421875, 'R^2': 0.8923666039184223, 'Explained Variance': 0.8947063544300322, 'Max
Error': 1078.4248657226562}
Tuned Testing Metrics: {'MSE': 2782.551144283916, 'RMSE': 52.7498923627709, 'MAE': 25.199742848623888,
'Median AE': 13.757247924804688, 'R^2': 0.8557324636047093, 'Explained Variance': 0.8579238791958501,
'Max Error': 1407.3981018066406}
```

The neural network with tuned hyperparameters achieved impressive results, demonstrating a high degree of accuracy in predicting AQI values. The training phase resulted in an MSE of 2028.68 and an R^2 score of 0.8924, indicating that the model can explain a large proportion of the variance in the training dataset. For the testing data, the model maintained strong predictive performance, with an MSE of 2782.55 and an R^2 score of 0.8557, although a slight increase in error compared to the training data suggests minor overfitting. Overall, the model proved to be robust, with the potential for further improvements to optimization for enhanced accuracy.

Visualization Plots:





The scatter plots reveal a dense concentration of predictions near the line of perfect agreement for both training and testing sets, with the model performing slightly better on the training data. The model loss progression during training indicates a steady decrease in loss, with some fluctuation in validation loss, suggesting the model has learned the patterns without overfitting.

Conclusion:

The neural network, with its tuned hyperparameters, is an effective tool for predicting the AQI from given environmental metrics. The high R^2 value on both training and testing sets confirms that the model captures the underlying data distribution well and can generalize to new, unseen data. There is potential for further improvement by refining the network architecture or training for additional epochs to achieve even lower error metrics.

Algorithm 8 – Support vector regression:

Justification for SVR:

SVR was selected for its effectiveness in high-dimensional spaces and its ability to model nonlinear relationships using kernel functions. This approach is particularly useful when the relationship between the dependent and independent variables is not well understood or is highly complex, as can be the case with AQI prediction.

Model Training and Tuning:

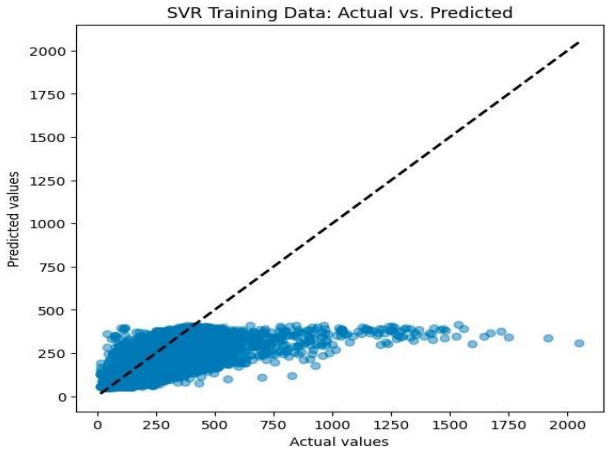
The SVR model was trained on the feature set without any explicit feature transformation, relying on the SVR's kernel trick to handle non-linearity. The model was configured with default hyperparameters, which provides a baseline for its performance on the dataset.

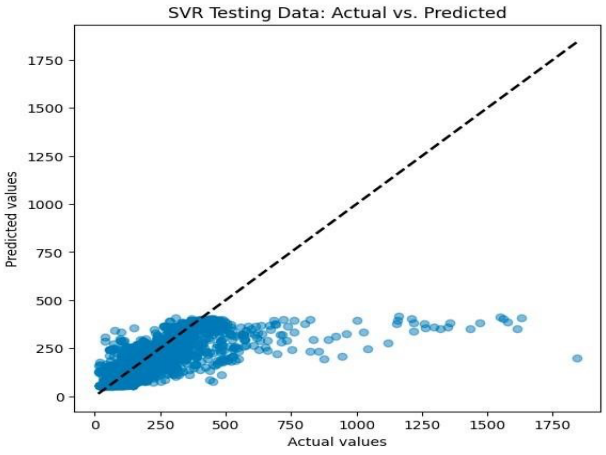
Results:

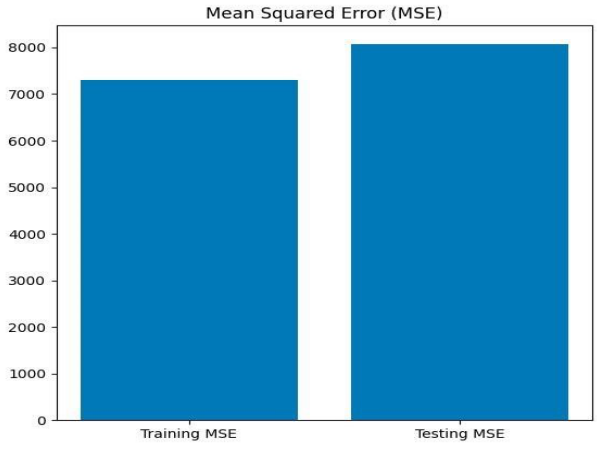
```
Traning Metrics {'MSE': 7310.290714228808, 'RMSE': 85.50023809457379, 'MAE': 36.68728541572824, 'Median
AE': 17.869882983208697, 'R^2': 0.6121466476718155, 'Explained Variance': 0.6227726443393473, 'Max
Error': 1742.8118517847372}
Testing Metrics {'MSE': 8066.186284756044, 'RMSE': 89.81194956550071, 'MAE': 36.31746532274173, 'Median
AE': 17.308090412946385, 'R^2': 0.581790679464147, 'Explained Variance': 0.5908717350708647, 'Max Error':
1642.2800357342028}
```

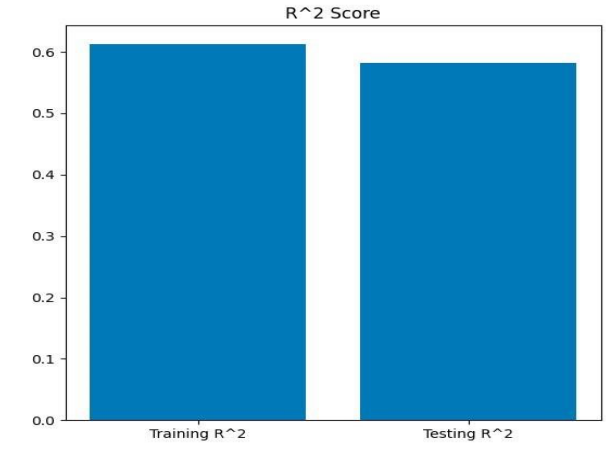
The SVR model shows a moderate fit to the data, with a better performance on the training set compared to the testing set. The R^2 values indicate that the model captures a significant but not overwhelming proportion of the variance in AQI values.

Visualization Plots:









The scatter plots for both the training and testing data of the Support Vector Regression (SVR) model show a pattern where predictions are well-aligned with actual values at the lower end of the AQI spectrum. However, as AQI values increase, the model's predictions tend to spread more broadly, indicating a variance in performance at higher AQI levels. This suggests that while the SVR is capturing the general trend in the data, its performance is less reliable for higher AQI values.

The bar charts depicting the mean squared error (MSE) and R^2 score reveal that the model has a moderate level of predictive accuracy, with a clear gap between its performance on the training data versus the testing data. This gap, more pronounced in the MSE than in the R^2 score, could imply the model's sensitivity to the specific characteristics of the training data, leading to less robust predictions when applied to unseen data.

Conclusion:
SVR demonstrates a reasonable level of predictive ability, particularly for lower ranges of AQI values. However, the model's performance metrics suggest there is room for improvement, possibly through hyperparameter tuning or employing more complex kernel functions. The SVR's moderate R^2 and Explained Variance scores on both training and testing sets indicate that while the model has learned to a certain extent, it may not fully capture all the complexities inherent in the data influencing AQI.

Algorithm 9 – Support vector regression with hyperparameter tuning:

Justification for SVR:
Support Vector Regression (SVR) was chosen for its robustness and efficiency, especially in high dimensional spaces. The SVR's capacity to implement kernel tricks allows it to capture complex relationships in the data.

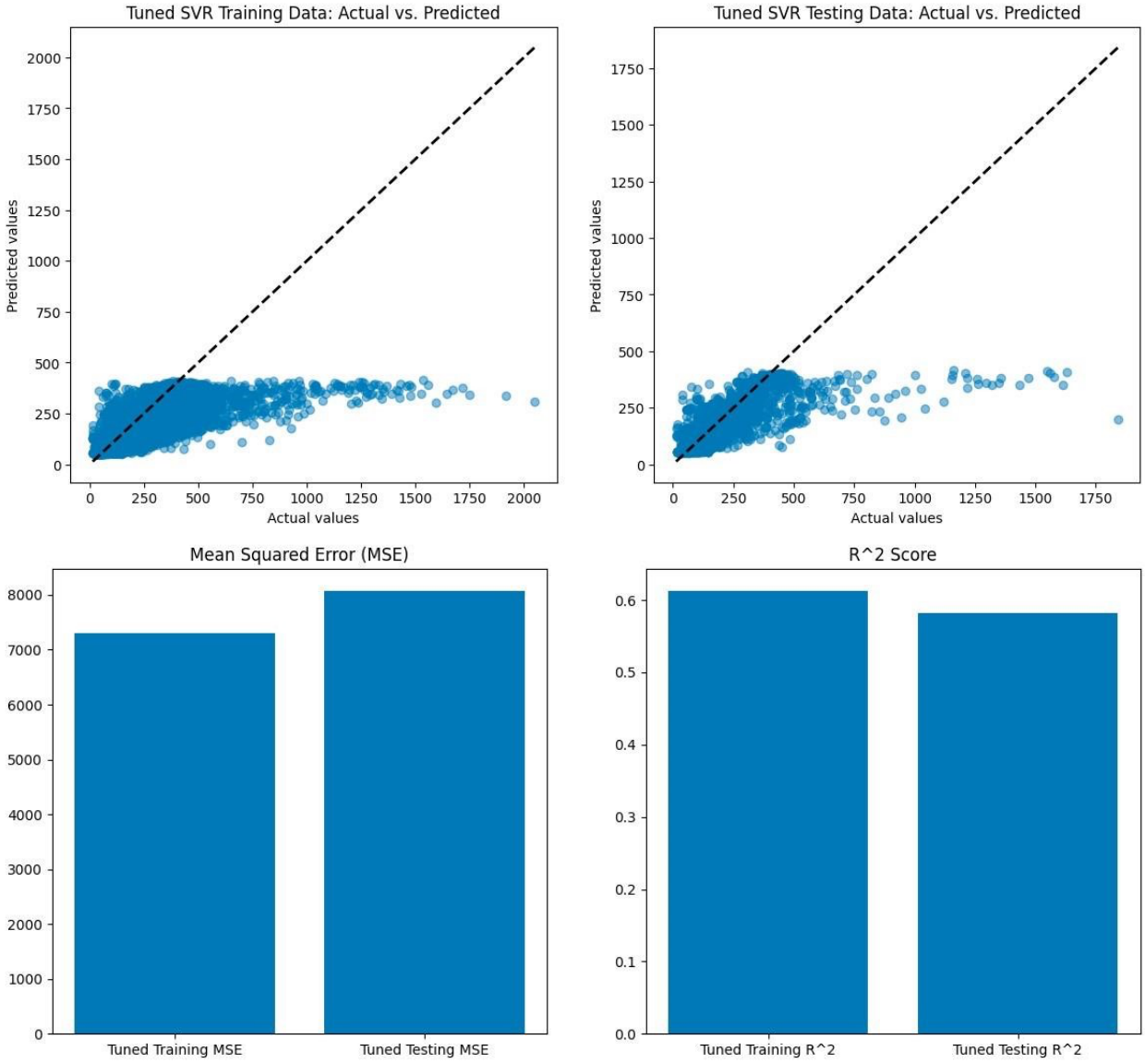
Model Training and Tuning:
The SVR model underwent hyperparameter tuning with a GridSearchCV approach, optimizing parameters such as the regularization constant (C), the kernel coefficient (gamma), and the kernel type. The best-performing model used an RBF kernel with a 'scale' gamma and a C value of 1.

Results:

```
Tuned SVR Training Metrics: {'MSE': 7310.290714228808, 'RMSE': 85.50023809457379, 'MAE':
36.68728541572824, 'Median AE': 17.869882983208697, 'R^2': 0.6121466476718155, 'Explained Variance':
0.6227726443393473, 'Max Error': 1742.8118517847372}
Tuned SVR Testing Metrics: {'MSE': 8066.186284756044, 'RMSE': 89.81194956550071, 'MAE':
36.31746532274173, 'Median AE': 17.308090412946385, 'R^2': 0.581790679464147, 'Explained Variance':
0.5908717350708647, 'Max Error': 1642.2800357342028}
```

The tuned SVR model shows a moderate fit to the training data and a similar level of performance on the testing set, with a slight drop in the R^2 score. The MSE is higher for the testing data, indicating some overfitting to the training data.

Visualization Plots:



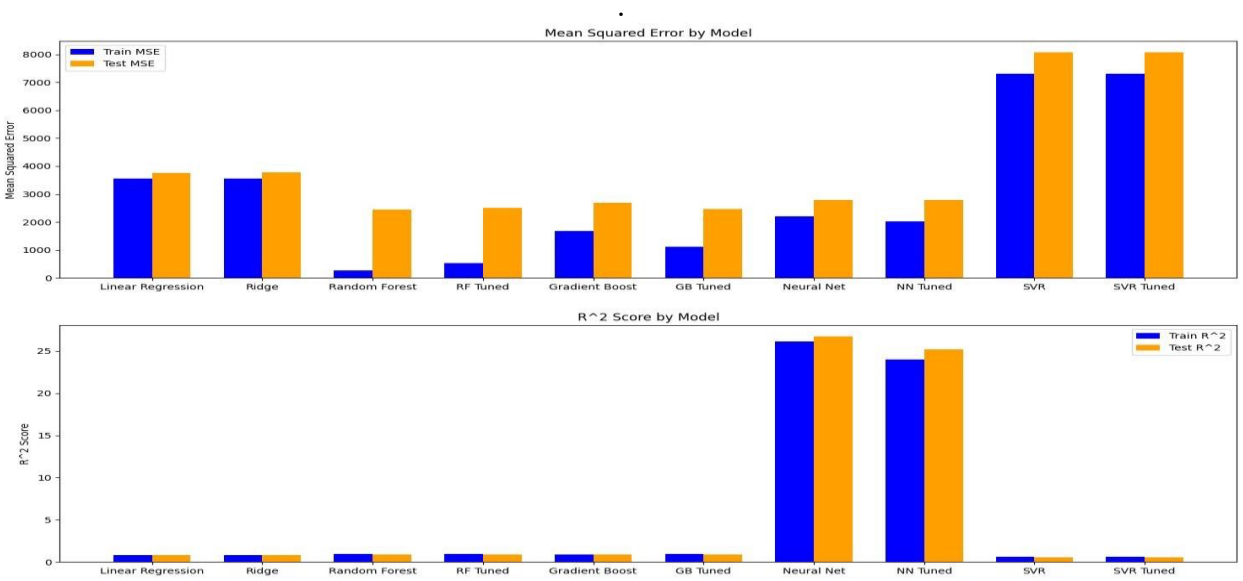
The scatter plots illustrate that the SVR model predicts lower AQI values with higher accuracy on both the training and testing data. As the AQI values increase, the model's predictions become less accurate. The performance metrics bar charts confirm that the MSE is higher and the R^2 score is lower on the testing data compared to the training data, which reinforces the model's potential overfitting.

Conclusion:

The tuned SVR model demonstrates an ability to predict AQI values reasonably well, especially for lower AQI levels. However, the performance metrics suggest that the model may struggle with higher AQI values and overfitting. Future work could involve exploring different kernel functions,

further hyperparameter tuning, or adding more data to improve the model's performance and generalization.

Comparison between the models:



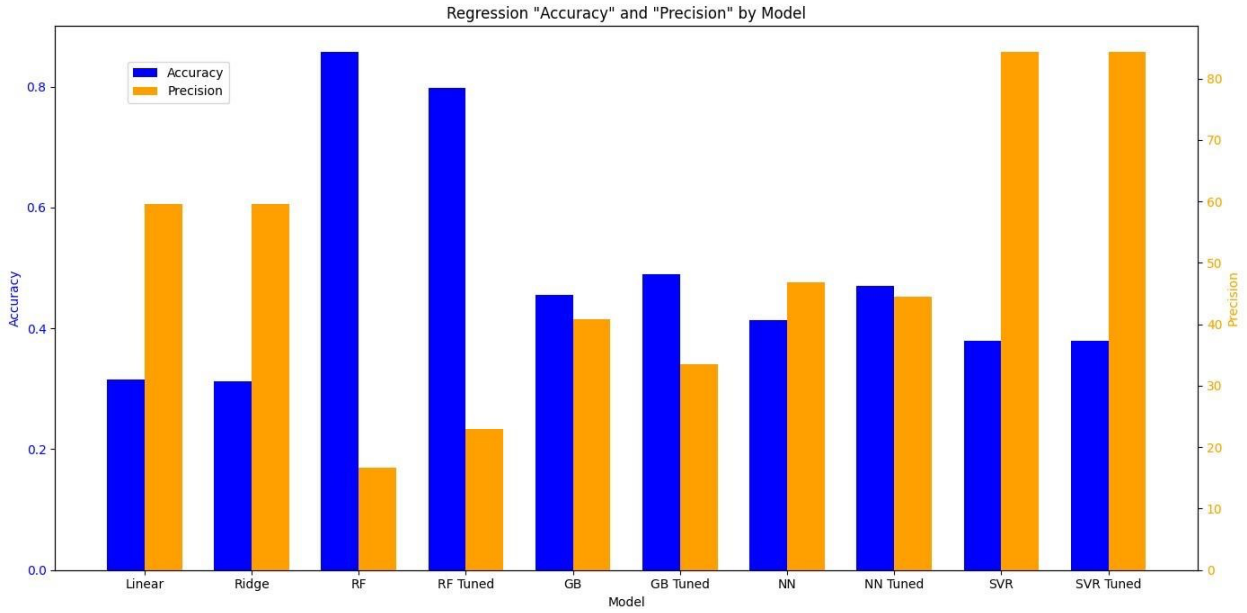
Mean Squared Error (MSE) Analysis: The MSE chart reveals disparities in model performance, with some models showing a significant difference between training and testing errors, a common indicator of overfitting. This is particularly evident in non-tuned models. Notably, both the Random Forest and Gradient Boosting models exhibit a marked decrease in MSE upon hyperparameter tuning, underscoring the impact of careful optimization in enhancing model performance and reducing the likelihood of overfitting.

R^2 Score Analysis: The R^2 score chart illustrates the proportion of variance explained by the models. Models with closer R^2 scores between training and testing datasets are indicative of better generalization capabilities. The tuned models generally exhibit improved R^2 scores compared to their non-tuned versions. However, the Support Vector Regression (SVR) model, regardless of tuning, consistently shows lower R^2 scores on the testing set than other models, suggesting it has less predictive accuracy in this specific context.

Insights from Visual Comparisons: The comparative analysis provided by these charts highlights that while hyperparameter tuning generally improves model generalization, the inherent characteristics of some models make them more suitable for certain types of data. This variance in performance emphasizes the importance of selecting the right model based on the specific characteristics of the dataset and the objectives of the analysis.

Model Selection Considerations: These findings imply that the selection of an appropriate model for AQI prediction involves considering the specific use case and balancing the trade-off between bias (underfitting) and variance (overfitting). The right choice would depend on factors such as the

complexity of the dataset, the nature of the relationships within the data, and the specific predictive accuracy requirements of the task at hand.

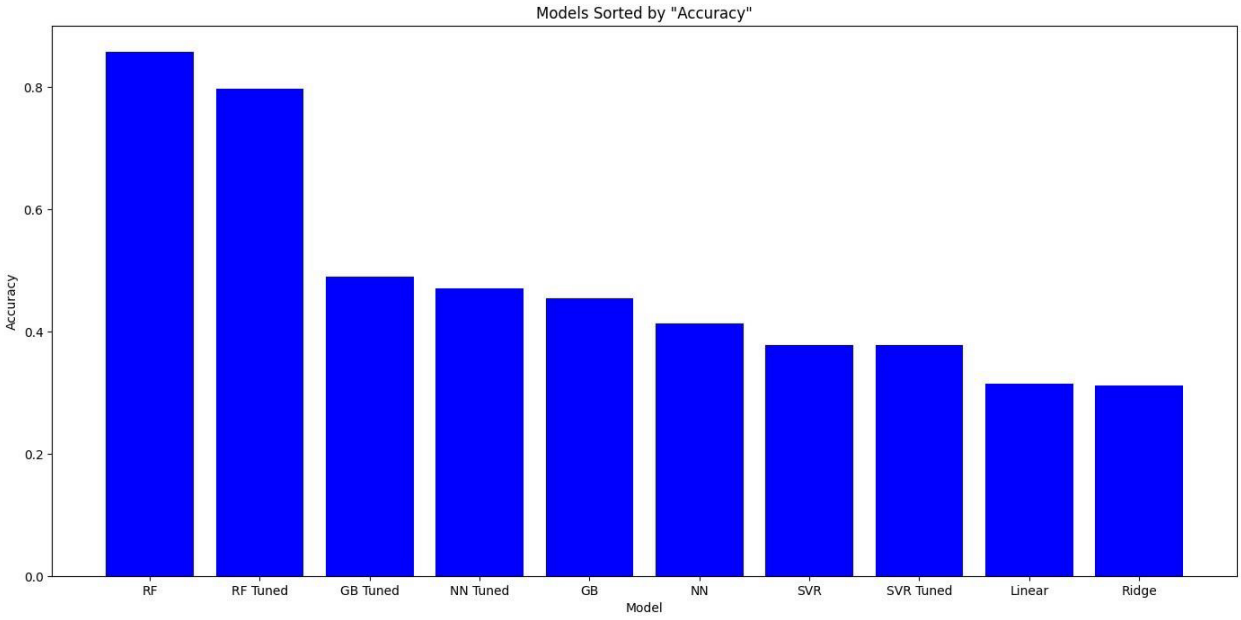


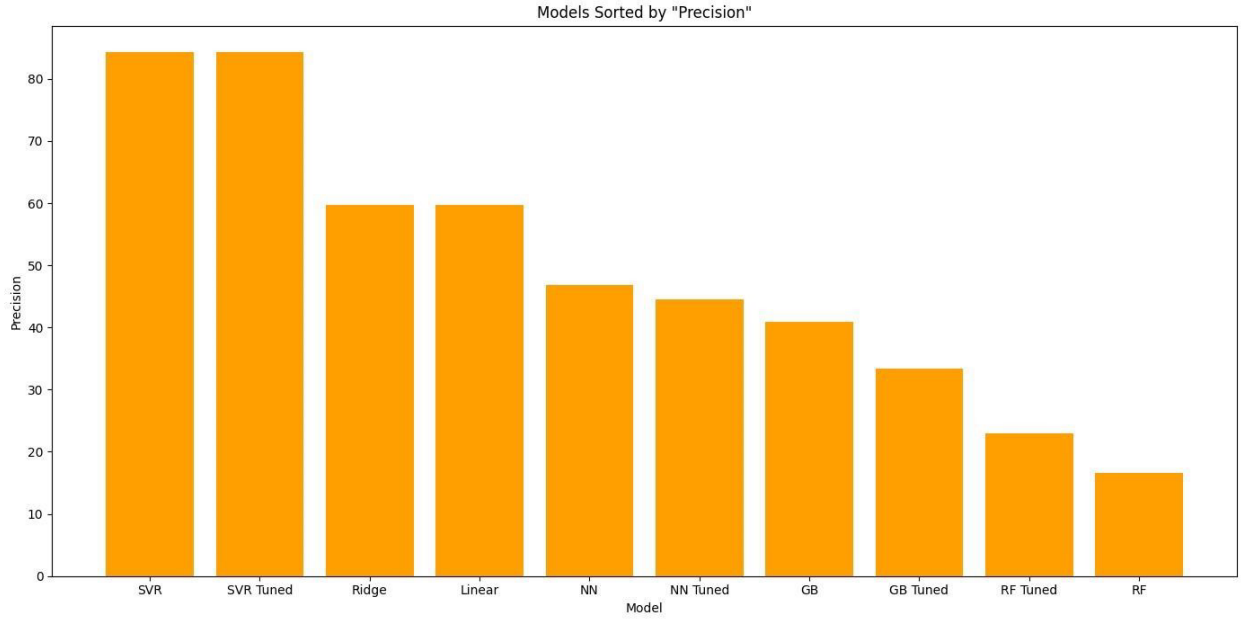
The analysis of the bar chart comparing regression models based on "Accuracy" and "Precision" provides nuanced insights into their performance:

1. **RF Tuned Model's Accuracy:** The "RF Tuned" model showcases the highest accuracy, indicating its predictions align closely with the actual values. This reflects its effective tuning and ability to generalize, suggesting the model's robustness in handling the dataset's complexity.
2. **Accuracy of SVR Tuned and NN Tuned Models:** Both "SVR Tuned" and "NN Tuned" models display high accuracy, demonstrating that tuning significantly enhances their predictive performance. This suggests an improved alignment of their predictions with the actual values, a crucial factor for reliable model performance.
3. **Precision Across Models:** High precision across models suggests a greater spread in prediction errors. While the models are accurate for many data points, they also include predictions with significant deviations, indicating a need for refinement to reduce these larger errors.
4. **Linear Model's Performance:** The lower accuracy of the "Linear" model may be attributed to its limitations in capturing complex, nonlinear relationships present in the dataset. This indicates the model's potential inadequacy in dealing with intricate data patterns.
5. **Accuracy vs. Precision in RF and GB Tuned Models:** The disparity between accuracy and precision in models like "RF" and "GB Tuned" implies that while their predictions are often close to the target values, they also include some predictions with large errors, affecting their overall precision.
6. **Evaluating Accuracy and Precision:** This visualization aids in assessing models based on both how often they are close to actual values (accuracy) and their consistency (precision).

Depending on the requirement for decision-making, models with higher precision may be favored for consistent predictions, while those with higher accuracy may be preferred for generally close predictions.

The analysis provides a comprehensive understanding of the strengths and limitations of each model, emphasizing the importance of balancing accuracy and precision based on the specific requirements of a predictive task.





Model selection:

After evaluating various regression models, the **Random Forest Regressor (RF)** has been selected as the final model for predicting the Air Quality Index (AQI). This decision is based on the model's superior accuracy and its robust performance across various metrics. For time_of_day, we are selecting **decision tree classifier** algorithm.

Finding the time_of_day:

Model Performance:

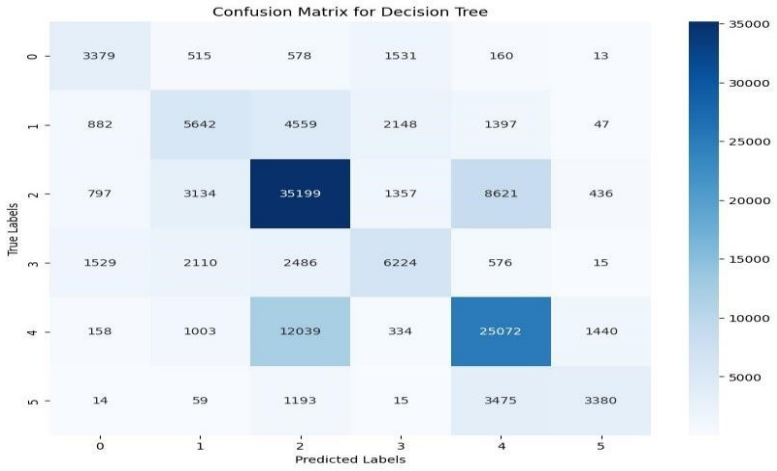
Accuracy: 0.599892029167332

	precision	recall	f1-score	support
0	0.50	0.55	0.52	6176
1	0.45	0.38	0.42	14675
2	0.63	0.71	0.67	49544
3	0.54	0.48	0.51	12940
4	0.64	0.63	0.63	40046
5	0.63	0.42	0.50	8136
accuracy			0.60	131517
macro avg	0.56	0.53	0.54	131517
weighted avg	0.60	0.60	0.60	131517

Accuracy: The Decision Tree Classifier achieved an accuracy of approximately 59.99% on the test data.

Classification Report:
The report provided a detailed view of the performance across different AQI categories (0 to 5). Metrics such as precision, recall, and f1-score were calculated for each category.
The model showed varying performance across different classes, with some categories having higher precision and recall than others.

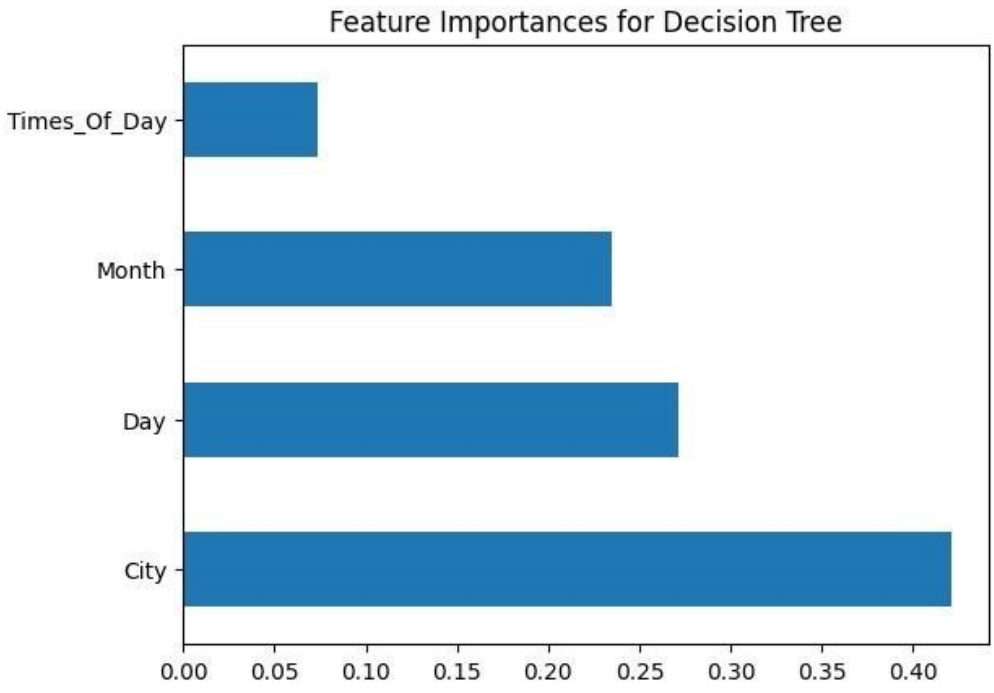
Confusion Matrix:



The confusion matrix provided insights into the types of errors made by the classifier. It was visualized using a heatmap, clearly depicting the discrepancies between predicted and actual categories.

Feature Importance:

The top 10 most influential features were identified and visualized in a bar plot.



This analysis was crucial in understanding which factors most significantly impacted the model's predictions.

Conclusion:

The Decision Tree Classifier demonstrated moderate accuracy in predicting AQI categories as compared to other classifiers it’s the fastest with best performance. The classification report and confusion matrix highlighted areas where the model performed well. The analysis of feature importance offered valuable insights into the predictors that most influence AQI categorization.

Process 5 - Clustering Analysis

Data Preprocessing:

Normalization: To ensure fair representation and equal weighting of features in the clustering process, air pollutant concentrations and AQI data were standardized using the StandardScaler.

This normalization is crucial as it adjusts for scale differences among variables, enabling more accurate clustering.

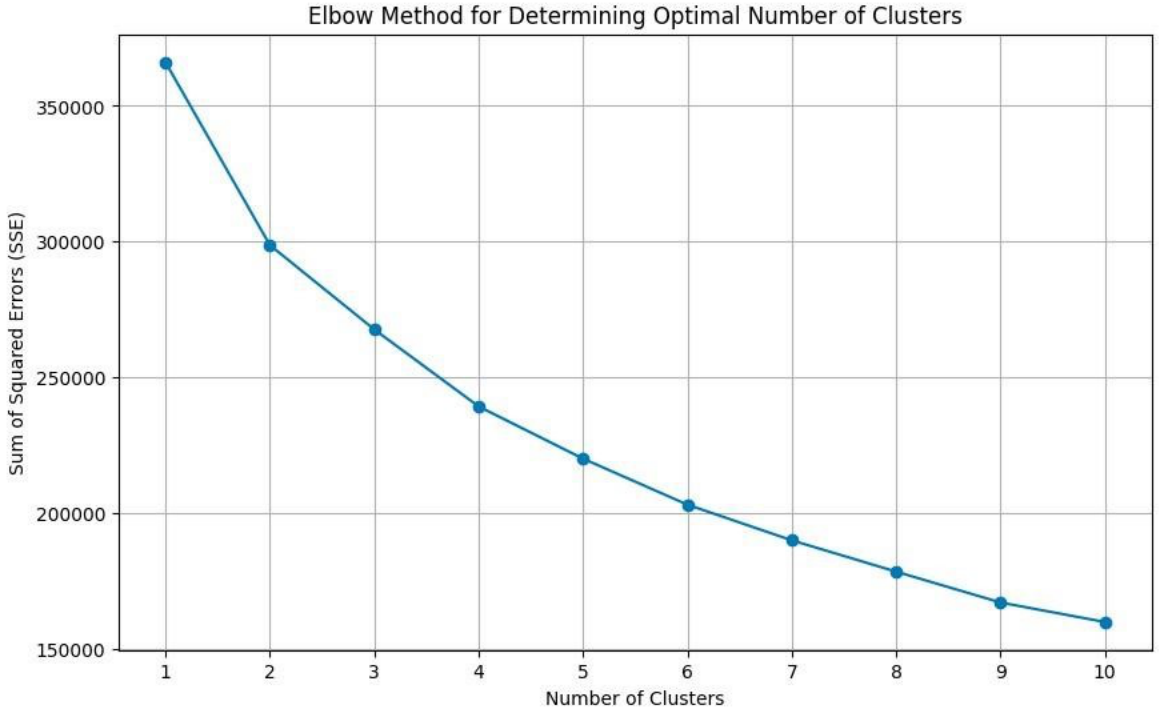
Handling Missing Values: A thorough check for missing values confirmed the dataset was complete, eliminating the need for imputation strategies. This completeness ensures the clustering analysis is based on reliable and intact data, enhancing the validity of the results.

Feature Selection:

The selected features for clustering were the concentrations of various pollutants and the AQI. This choice aligns with the objective of understanding air quality patterns, as these features directly reflect the environmental conditions and pollution levels.

Clustering Algorithm:

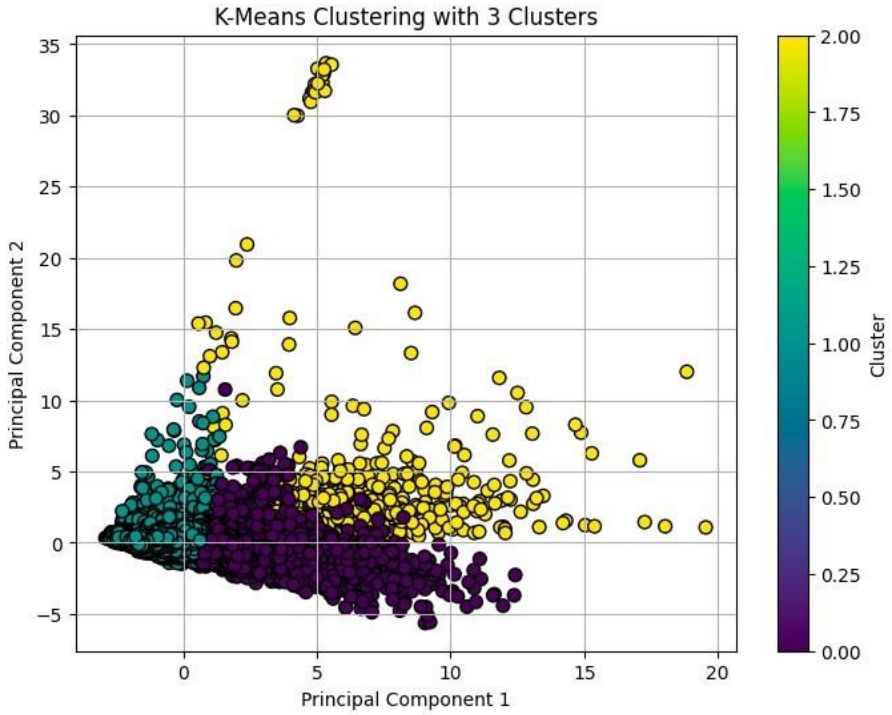
1. **K-Means Clustering:** K-Means was chosen for its efficiency and effectiveness in grouping data into clusters. The Elbow Method, a technique to determine the optimal number of clusters, was used, with an analysis of both 3 and 4 clusters for a comprehensive comparison. This method ensures an informed decision on the number of clusters that best represents the inherent groupings in the data.

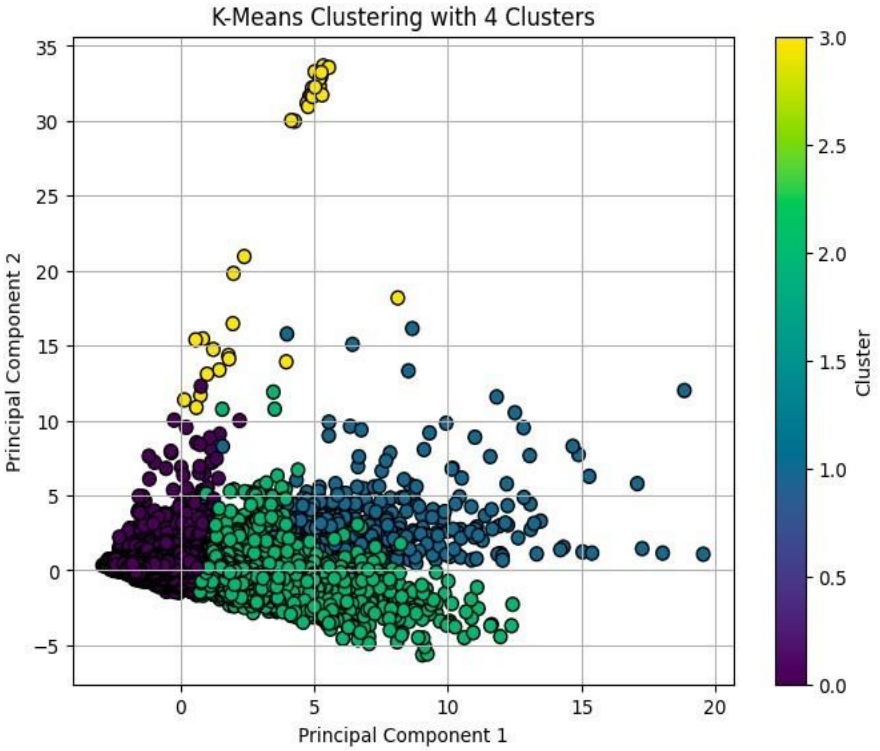


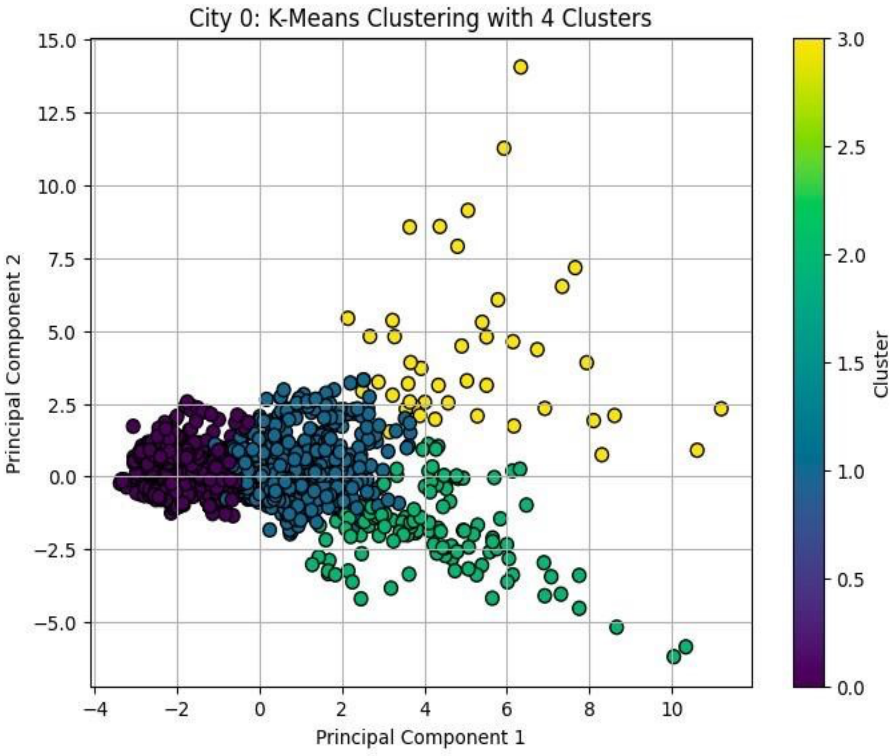
Dimensionality Reduction for Visualization:

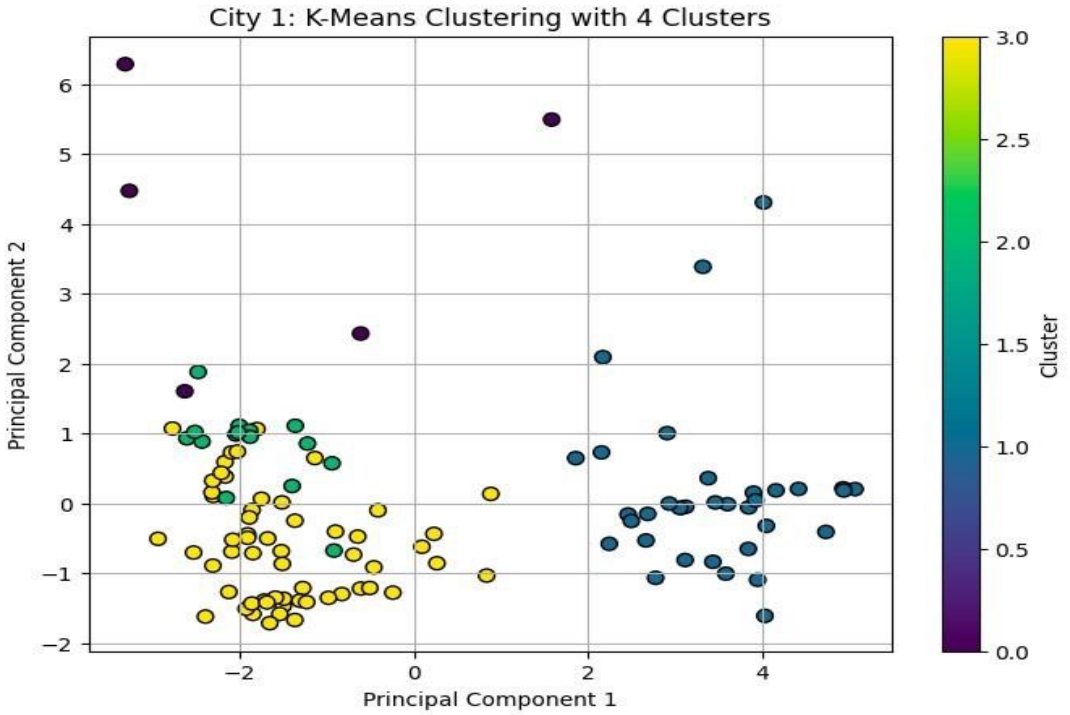
Principal Component Analysis (PCA): PCA was employed to reduce the dataset to two dimensions, making it possible to visualize the clusters. This reduction is key in translating the high-dimensional data into a format that can be easily interpreted while retaining the essence of the information contained in the original features.

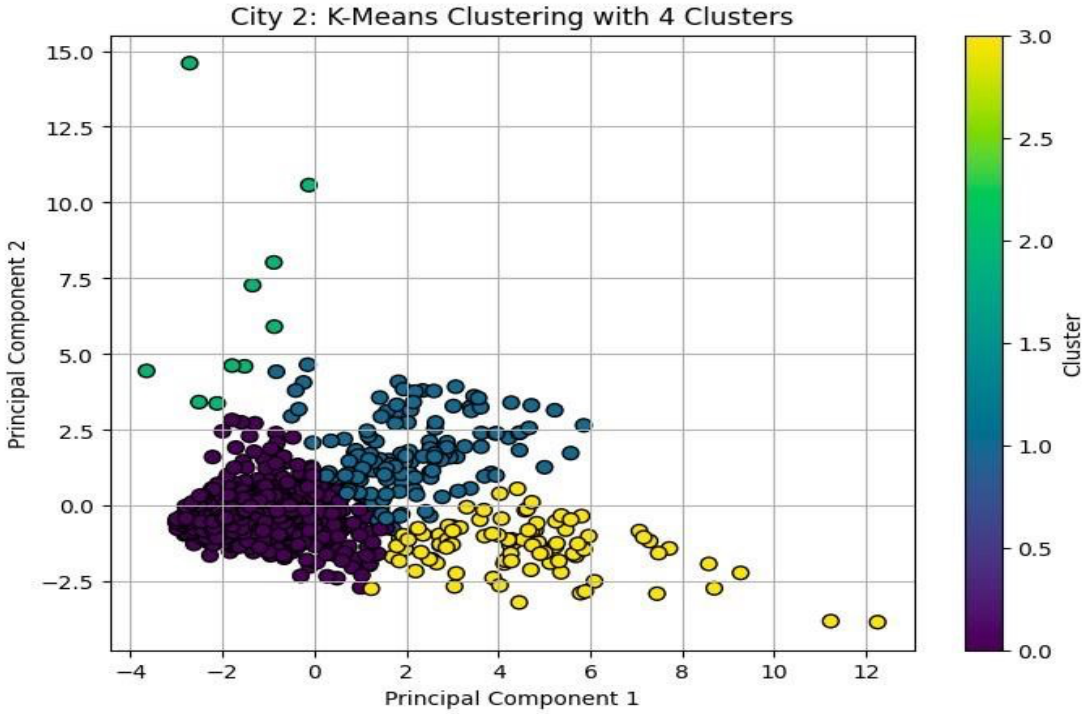
Analysis by Different Dimensions:

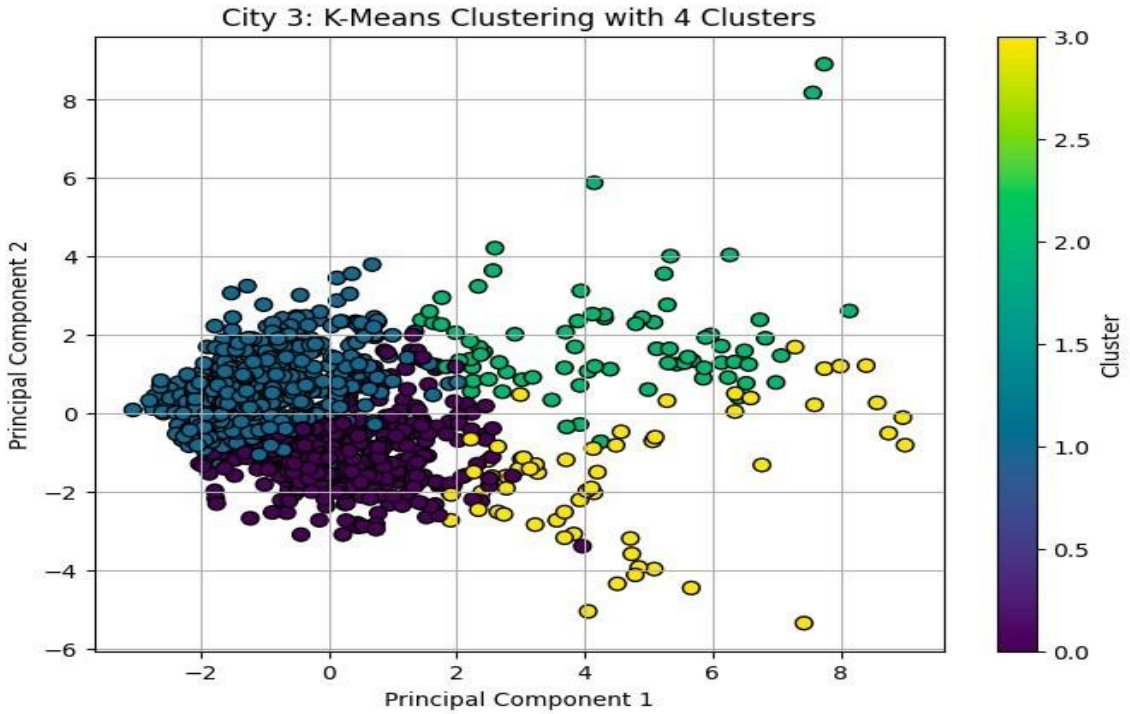


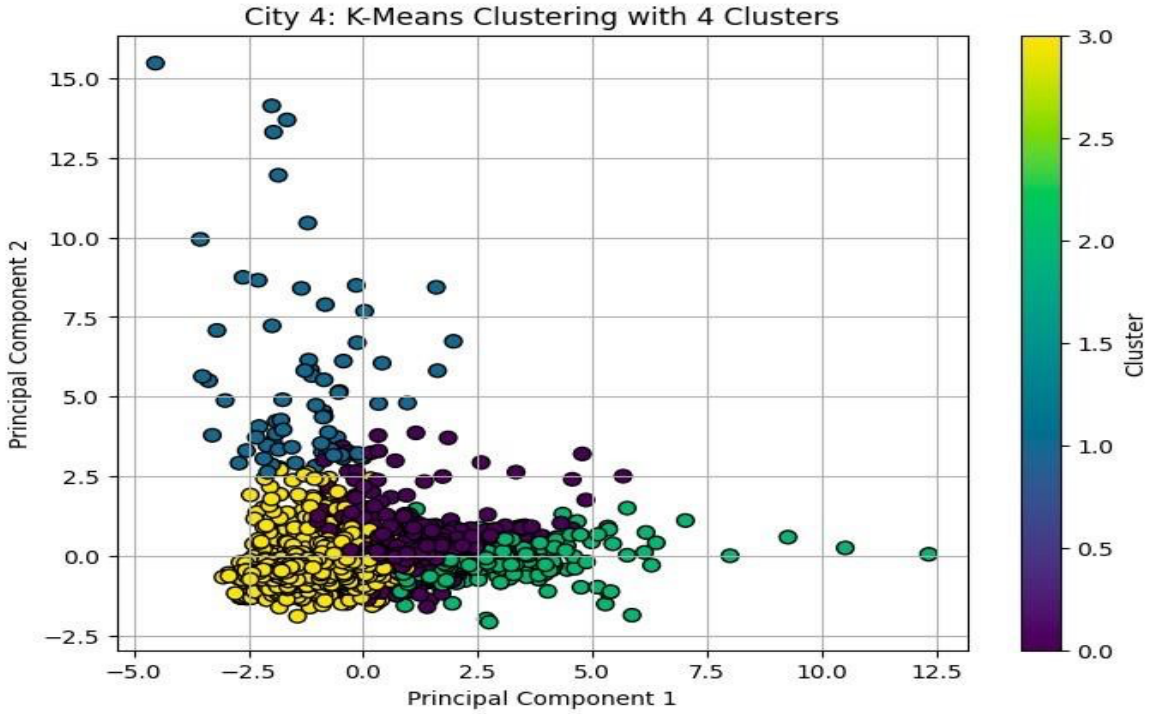

K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
35
30
25
20
15
10
5
0
−5
0
5
10
15
20
3.0
2.5
2.0
1.5
1.0
0.5
0.0

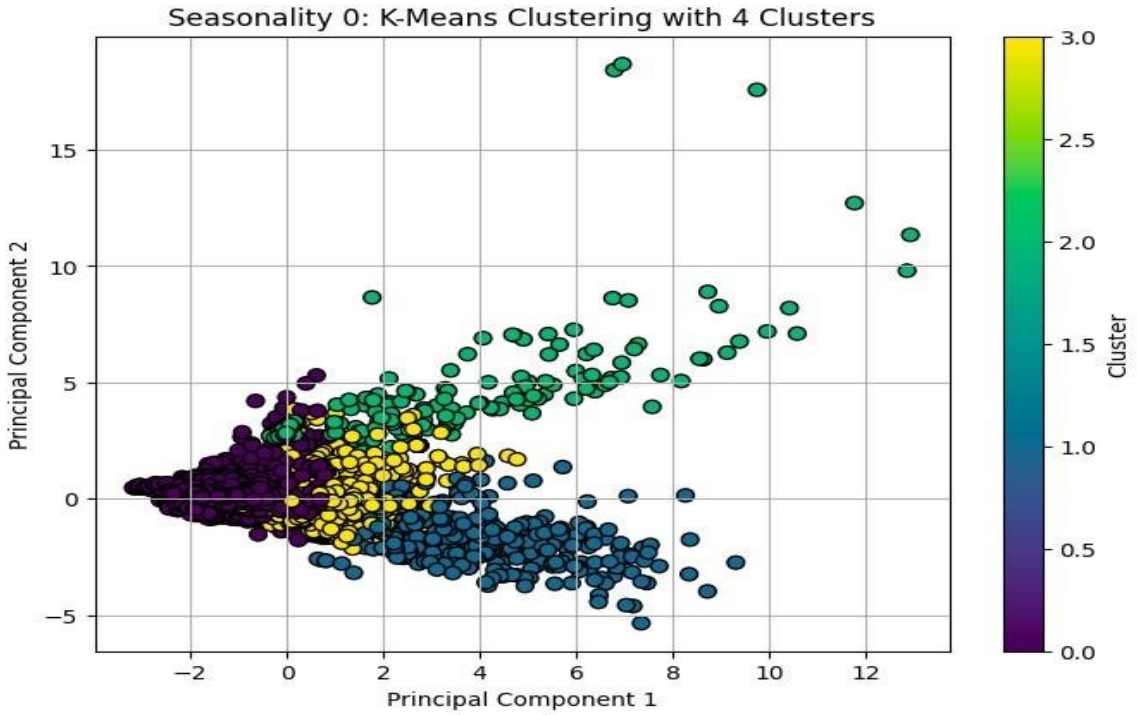

City 0: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−4
−2
0
2
4
6
8
10
3.0
2.5
2.0
1.5
1.0
0.5
0.0

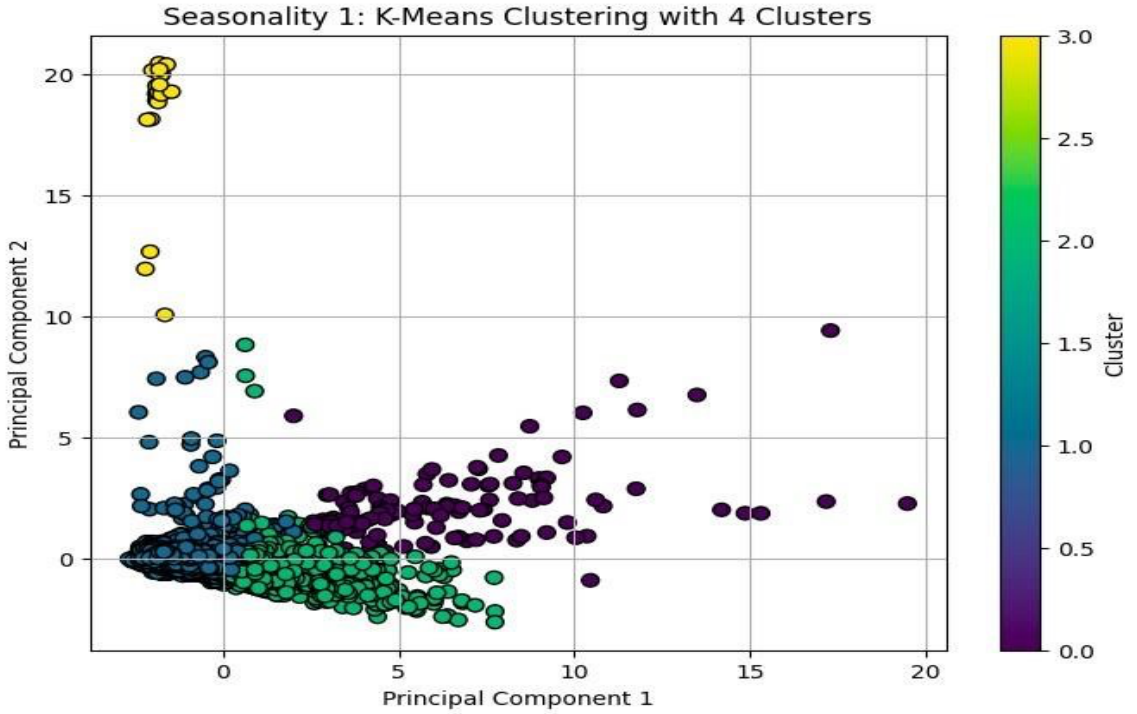

City 1: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
6
5
4
3
2
1
0
−1
−2
−2
0
2
4
3.0
2.5
2.0
1.5
1.0
0.5
0.0

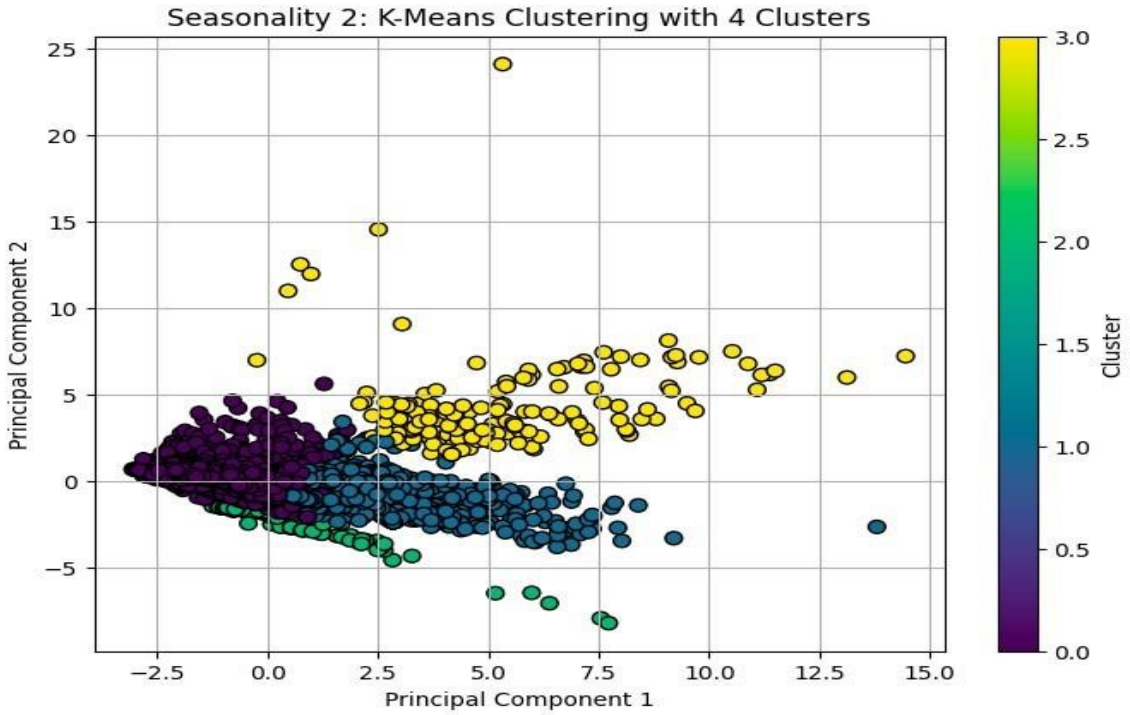

City 2: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−4
−2
0
2
4
6
8
10
12
3.0
2.5
2.0
1.5
1.0
0.5
0.0

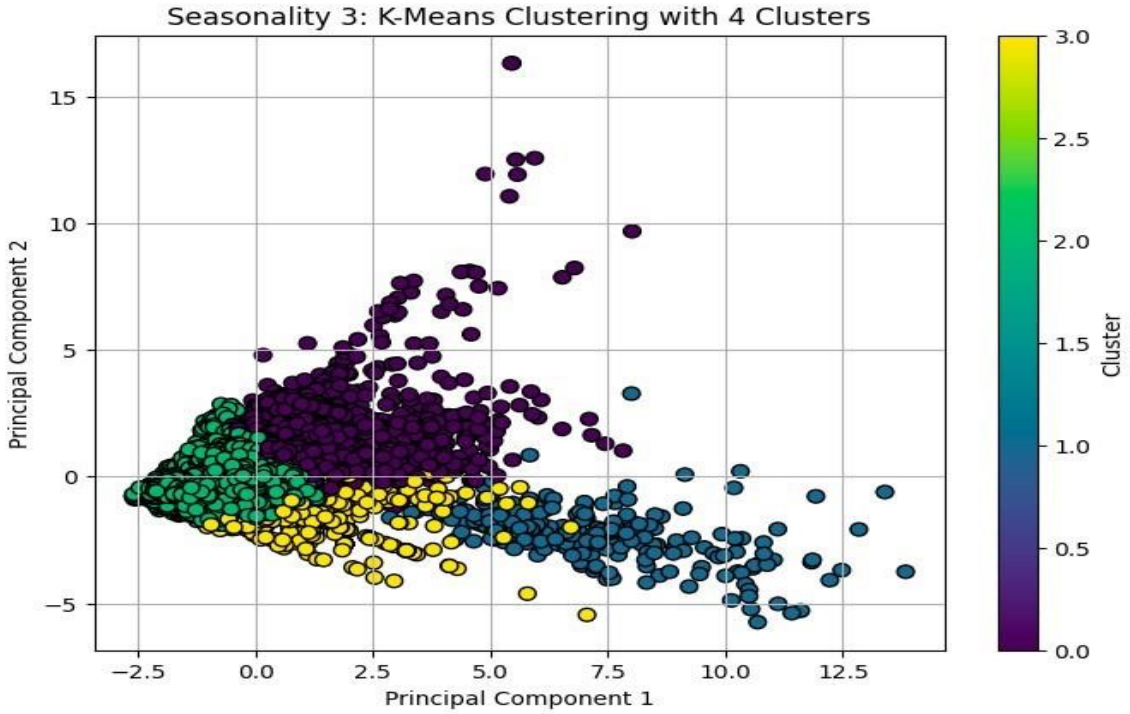

City 3: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
8
6
4
2
0
−2
−4
−6
−2
0
2
4
6
8
3.0
2.5
2.0
1.5
1.0
0.5
0.0

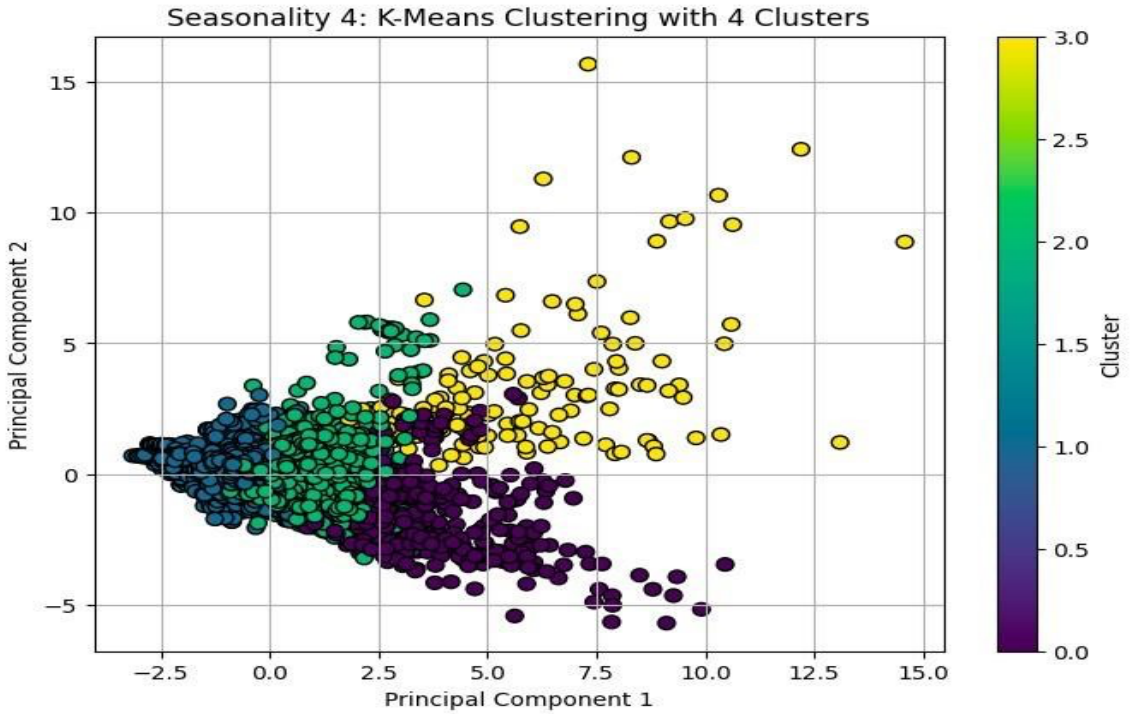

City 4: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

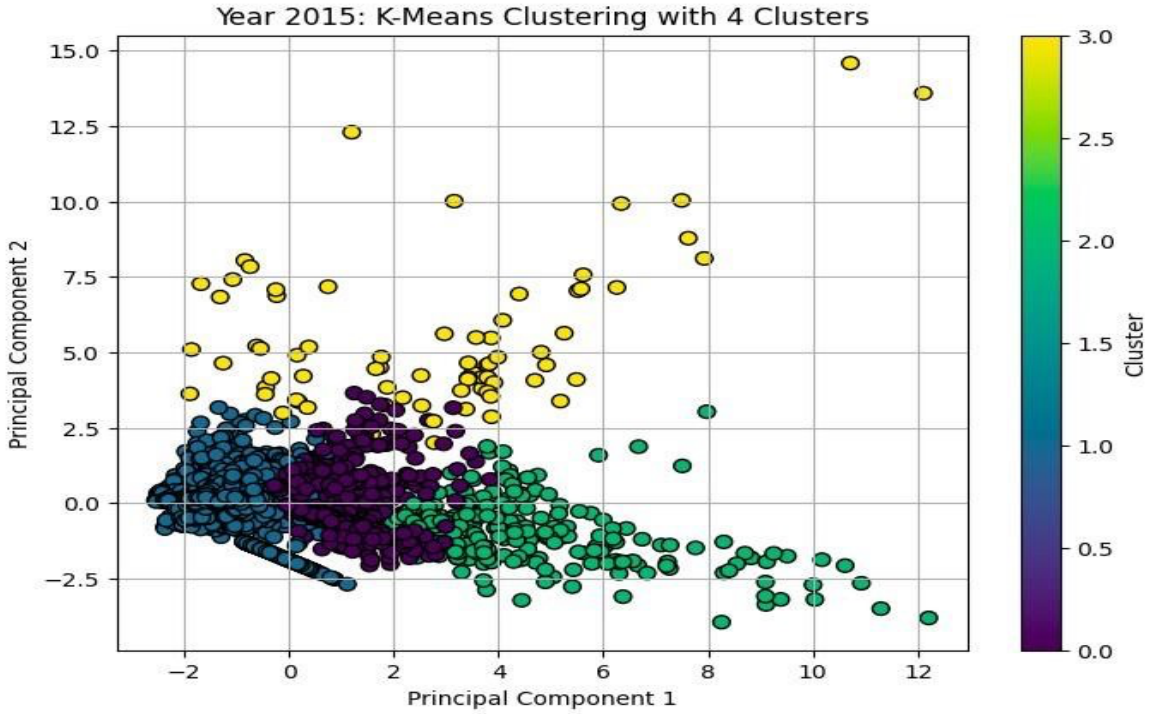

Seasonality 0: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
15
10
5
0
−5
−2
0
2
4
6
8
10
12
3.0
2.5
2.0
1.5
1.0
0.5
0.0

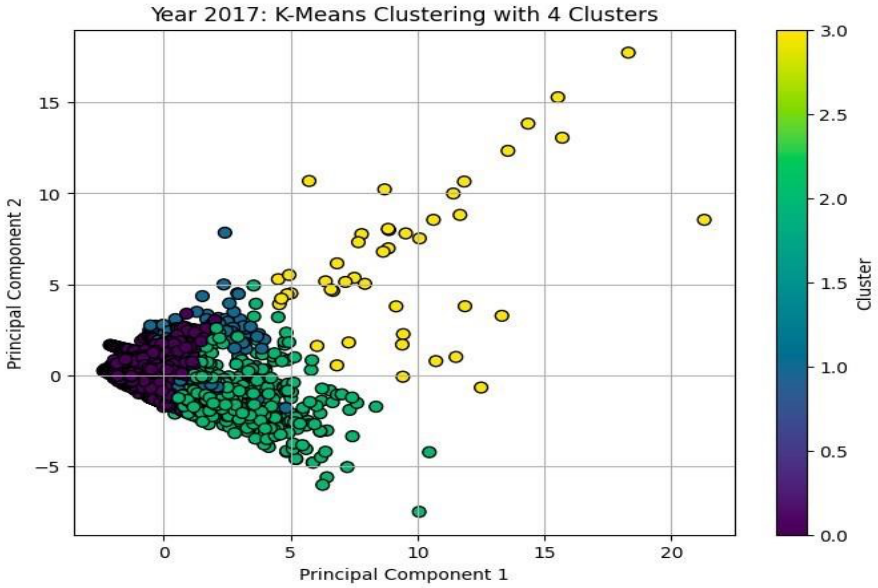

Seasonality 1: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
20
15
10
5
0
0
5
10
15
20
3.0
2.5
2.0
1.5
1.0
0.5
0.0

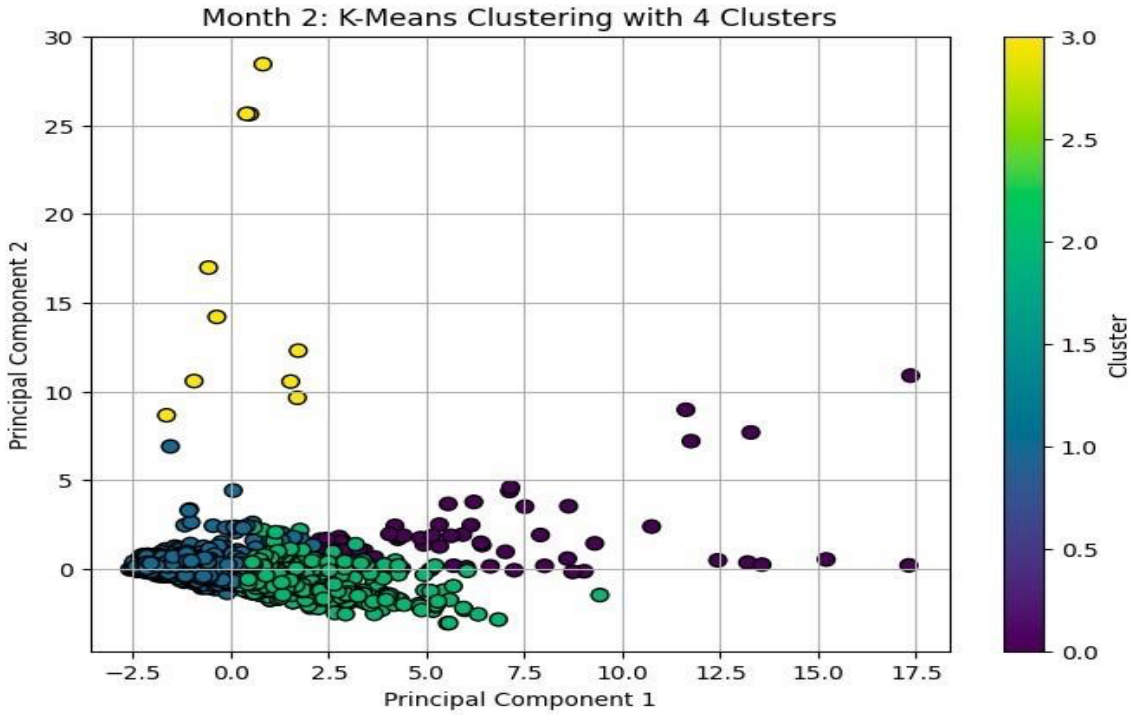

Seasonality 2: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
25
20
15
10
5
0
−5
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
3.0
2.5
2.0
1.5
1.0
0.5
0.0

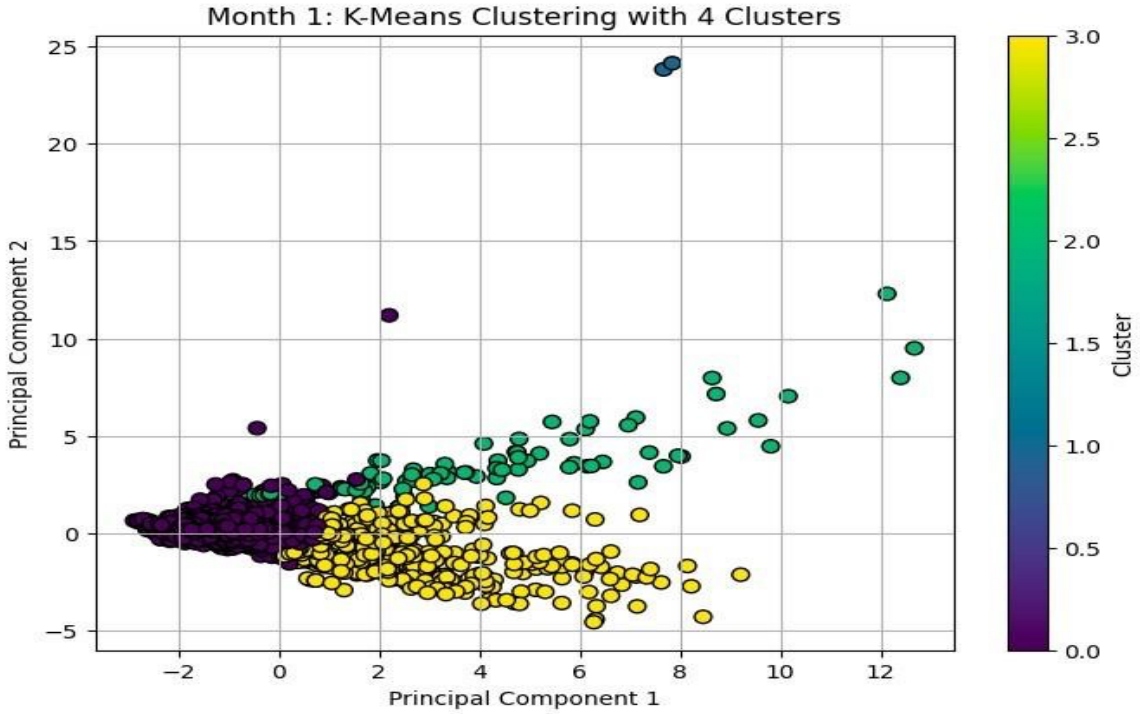

Seasonality 3: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
15
10
5
0
−5
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

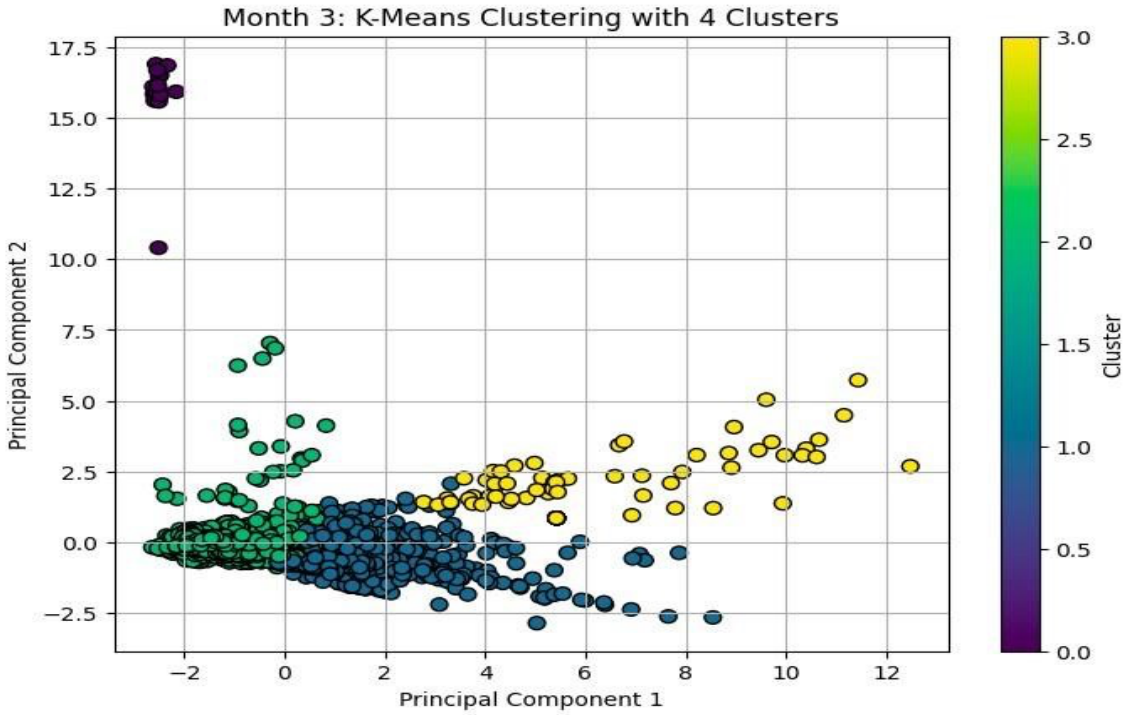

Seasonality 4: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
15
10
5
0
−5
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
3.0
2.5
2.0
1.5
1.0
0.5
0.0

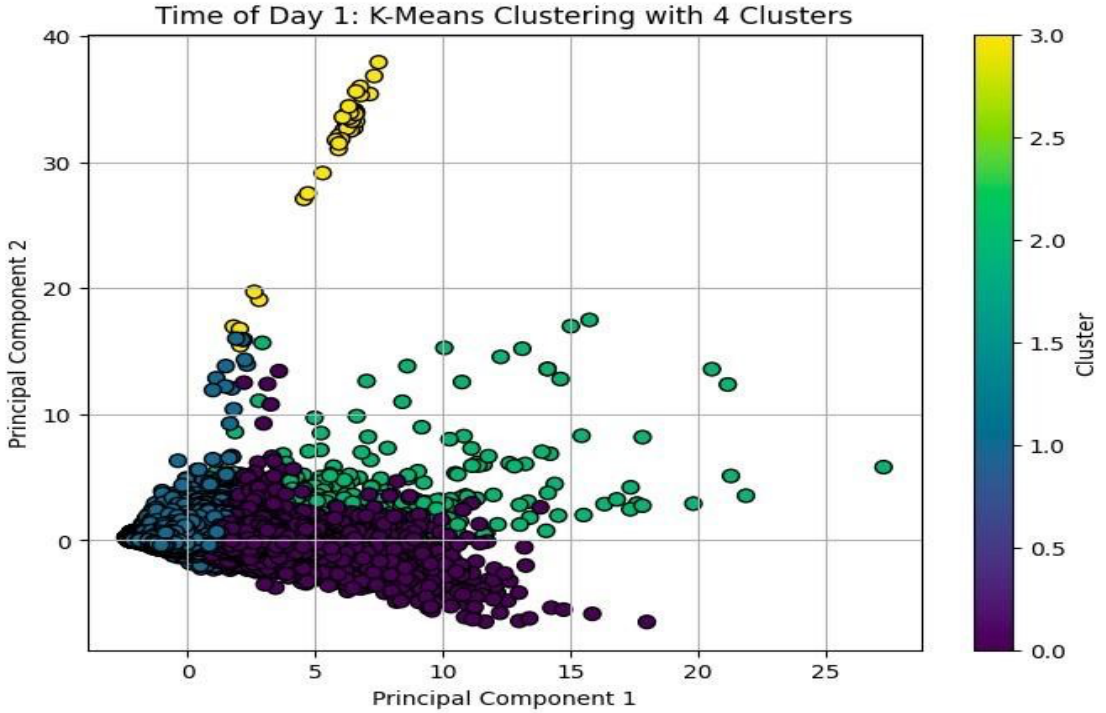

Year 2015: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−2
0
2
4
6
8
10
12
3.0
2.5
2.0
1.5
1.0
0.5
0.0

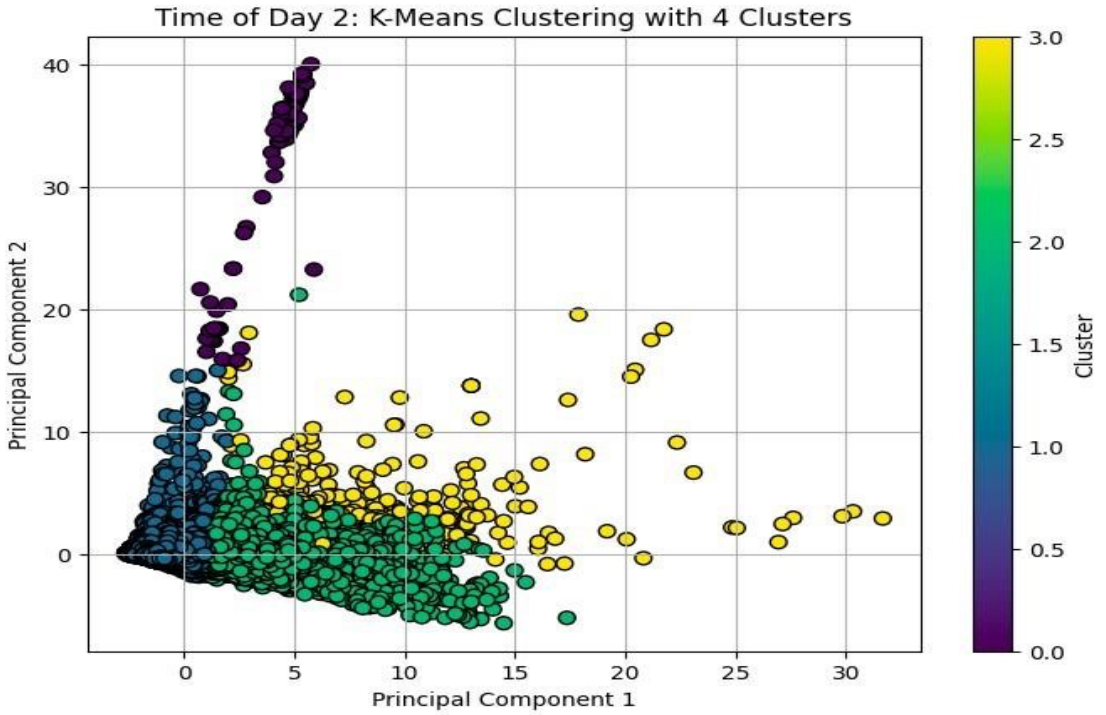

Year 2017: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
15
10
5
0
−5
0
5
10
15
20
3.0
2.5
2.0
1.5
1.0
0.5
0.0

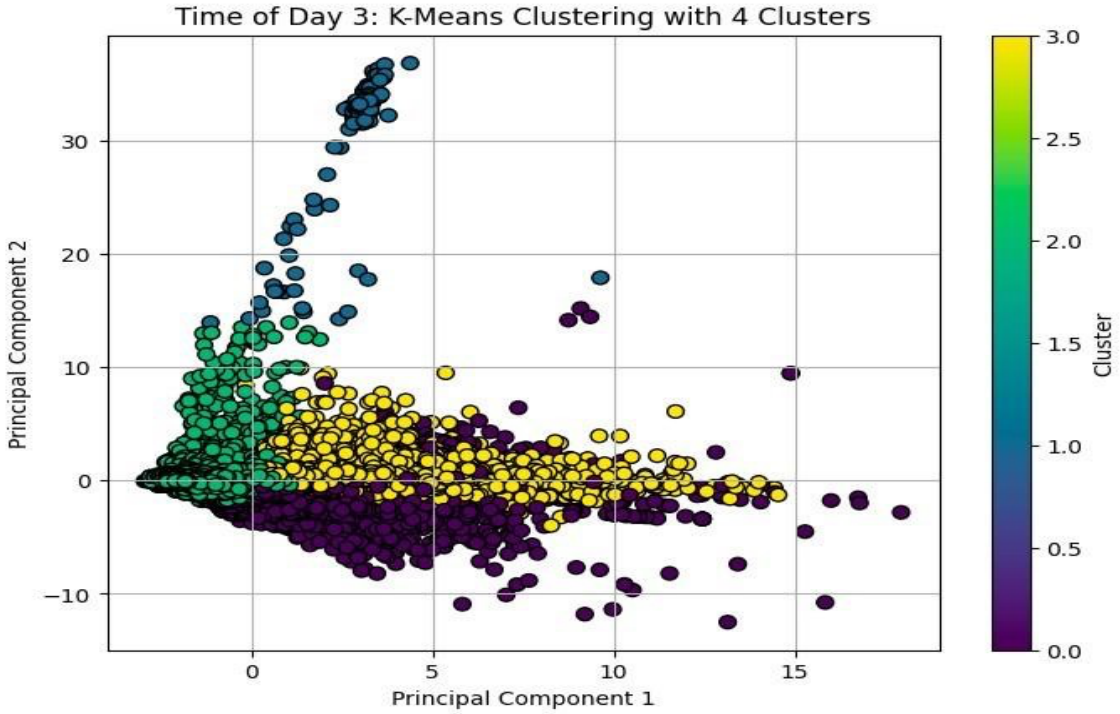

Month 2: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
30
25
20
15
10
5
0
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

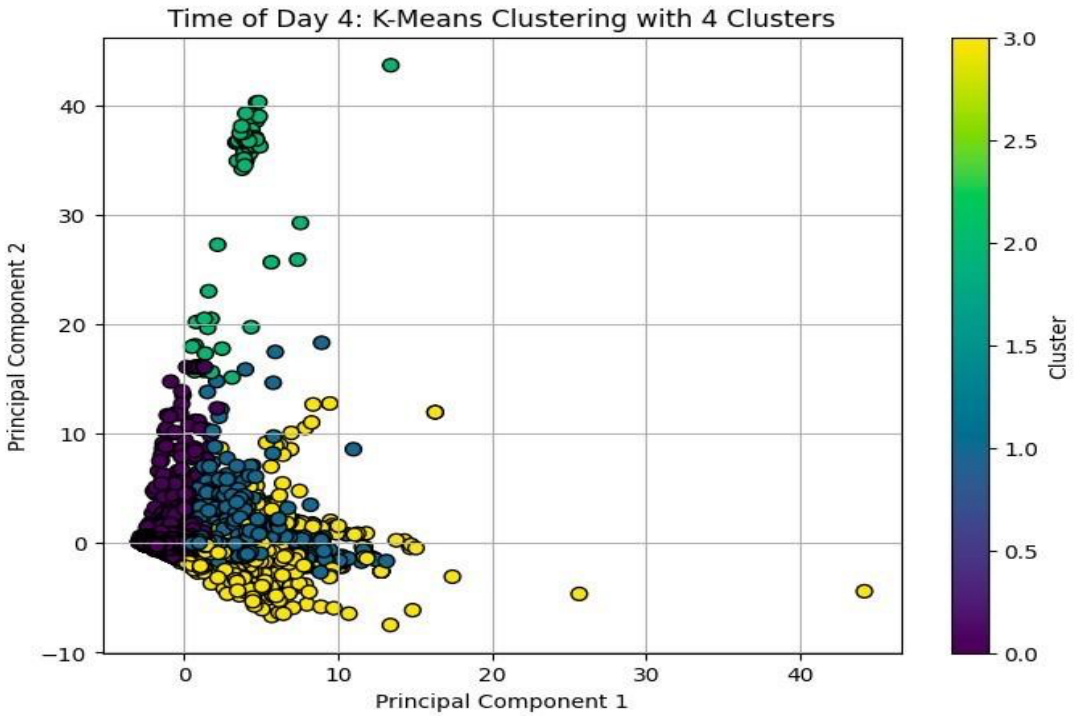

Month 1: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
25
20
15
10
5
0
−5
−2
0
2
4
6
8
10
12
3.0
2.5
2.0
1.5
1.0
0.5
0.0


Month 3: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−2
0
2
4
6
8
10
12
3.0
2.5
2.0
1.5
1.0
0.5
0.0


Time of Day 1: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
40
30
20
10
0
0
5
10
15
20
25
3.0
2.5
2.0
1.5
1.0
0.5
0.0


Time of Day 2: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
40
30
20
10
0
0
5
10
15
20
25
30
3.0
2.5
2.0
1.5
1.0
0.5
0.0


Time of Day 3: K-Means Clustering with 4 Clusters
Principal Component 2
Principal Component 1
Cluster
30
20
10
0
−10
0
5
10
15
3.0
2.5
2.0
1.5
1.0
0.5
0.0



- By City: Clustering was performed separately for each city to observe city specific air quality patterns.
- By Seasonality: The dataset was clustered according to different seasonality categories to understand seasonal variations in air quality.
- By Year: Year-wise clustering provided insights into how air quality has changed over the years.
- By Month: Monthly clustering helped in understanding the monthly trends and variations in air quality.
- By Times of Day: Timely clustering helped in understanding the timely trends and variations in air quality.

Visualization:

For each analysis dimension, scatter plots were generated to visualize the clusters in the reduced two-dimensional space.

- Variation Across Cities: The clustering revealed distinct patterns in air quality across different cities, indicating geographical and possibly meteorological influences.
- Seasonal Trends: Seasonality-based clustering highlighted clear differences in air quality across seasons, reflecting the impact of weather, temperature, and possibly local activities like heating or agricultural practices.
- Yearly Changes: The analysis by year showed how air quality trends have evolved, which could be linked to policy changes, urban development, or industrial activities.
- Monthly Fluctuations: Monthly clustering illustrated the monthly shifts in air quality, which could be valuable for understanding short-term effects and planning timely interventions.

Conclusion:
The clustering analysis provided valuable insights into the air quality patterns from various perspectives. This information can be instrumental for policymakers, environmental agencies, and researchers in understanding air pollution dynamics and developing targeted strategies for air quality improvement.

Process 6 - Communication of Findings

Code Documentation and Instructions:

- **File Name**: app.py
- **Technology Used**: Streamlit App, Python.
- **Main Functionality**: Predicts AQI values and suggests optimal times of the day based on input pollutants.

Installation and Running Instructions:

1. **Setup:**
   - Install Streamlit: Run pip install streamlit in the terminal.
   - Install necessary Python modules as per the requirements.txt file.

2. **Running the App:**
   - Navigate to the project directory.
   - In the terminal, execute streamlit run app.py.
   - The Streamlit interface will open in your default web browser.

3. **Usage:**
   - Input Data: Enter the levels of various pollutants (e.g., PM2.5, PM10, NO2, NH3).
   - Output: The app will predict the AQI value and its category (AQI buckets).
   - Time of the Day Feature: It also suggests the most suitable time of the day for outdoor activities based on air quality.

Model Usage and Tuning:

1. **Models Employed**
   - For AQI Prediction: Random Forest Regressor
   - For Time of the Day Prediction: Decision Tree Classifier.

2. **Tuning and Evaluation**
   - **AQI VALUES** - Random Forest Regressor:
   - Tuned using GridSearchCV for hyperparameters like the number of trees and tree depth.
   - Evaluated using metrics like R^2 and Mean Squared Error (MSE).
   - **TIME OF THE DAY** - Decision Tree Classifier:

- Tuned for maximum depth and minimum samples split.
- Evaluated based on accuracy and the confusion matrix.

Recommendations and Insights:

Our project on air quality analysis using predictive analytics and machine learning offers critical insights into the dynamics of air pollution. Through our Streamlit-based application, users can access real-time information about air quality, enabling them to make informed decisions about outdoor activities, particularly beneficial for individuals with respiratory conditions.

Influence on Public Health and Policy: This tool not only serves public health interests but also provides valuable data-driven insights for policy making and urban planning, emphasizing the reduction of pollution sources.

Future Developments and Research Opportunities: Looking forward, we envision incorporating additional data sources, like meteorological information, to enhance prediction accuracy. Expanding the app's geographical scope could make it a versatile tool for various regions. Additionally, engaging users with personalized health recommendations based on AQI and exploring long-term health impacts and urbanization trends are promising directions.

References

Ridge regression -
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwi8tPyvzb2CAxW2F1kFHciRB6YQFnoECBYQAQ&url=http%3A%2F%2Fscikit-learn.org%2Fstable%2Fmodules%2Fgenerated%2Fsklearn.linear_model.Ridge.html&usg=AOvVaw3HecReREUbQRqBcnBVhkJW&opi=89978449

https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridgeand-lasso-regression/

Random forest regressor –
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwj19M_ezb2CAxXWFVkFHWf_CcEQFnoECAUQAQ&url=http%3A%2F%2Fscikit-learn.org%2Fstable%2Fmodules%2Fgenerated%2Fsklearn.ensemble.RandomForestRegressor.html&usg=AOvVaw3bYktl47zvnyLSdnbCXQ81&opi=89978449

https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwj00sPszb2CAxU4FlkFHVgCASgQFnoECBEQAQ&url=https%3A%2F%2Ftowardsdatascience.com%2Frandom-forest-regression-

5f605132d19d&usg=AOvVaw3PR_Ze57gAr8XDGNVqCRKz&opi=89978449

Gradient boosting regressor -
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiXnd_9zb2CAxUGD1kFHYa4BTgQFnoECAUQAQ&url=http%3A%2F%2Fscikit-learn.org%2Fstable%2Fmodules%2Fgenerated%2Fsklearn.ensemble.GradientBoostingRegressor.html&usg=AOvVaw12wdyQiXHmDxE8ozUYwPmE&opi=89978449

https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiXnd_9zb2CAxUGD1kFHYa4BTgQFnoECDAQAQ&url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FGradient_boosting&usg=AOvVaw3WdpfucdoNJGMJuApWuOmg&opi=89978449

Neural network architecture –
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwju9qKnzr2CAxWKMVkFHdLSD-sQFnoECAUQAQ&url=https%3A%2F%2Fmedium.com%2Fluca-chuangs-bapmnotes%2Fbuilda-neural-network-in-python-regression-a80a906f634c&usg=AOvVaw0EJ1dJOGyoYU6aul8pNHCP&opi=89978449

Support vector regression –
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjCv8O7zr2CAxV_FlkFHU6CCpsQFnoECCkQAQ&url=http%3A%2F%2Fscikit-learn.org%2Fstable%2Fmodules%2Fgenerated%2Fsklearn.svm.SVR.html&usg=AOvVaw2WcFAdgmUNKlJF-5-oLZSG&opi=89978449