ASSIGNMENT-3

Kaniki Bhikshapathi
24489

February 12, 2025

Question 1

The COVID-19 lockdown in March 2020 significantly impacted air pollution levels worldwide. In this analysis, we examine AQI trends in Delhi and Bangalore from 2017 to 2020 to observe changes before and after the lockdown. Air pollution is a major environmental concern, affecting human health and the ecosystem. This study highlights the importance of data-driven analysis in understanding the effects of policy interventions on air quality.

Data Collection and Preprocessing

The dataset used for this analysis consists of air quality data collected from monitoring stations in Delhi and Bangalore. The dataset includes key pollutants such as $PM_{2.5}$, PM_{10}, NO, NO_2, CO, and AQI values. The data was preprocessed to handle missing values and ensure consistency in the time-series analysis.

AQI Trends in Delhi and Bangalore

To analyze the effect of lockdown on air quality, AQI trends were plotted over time. The following figures display the AQI trends in Delhi and Bangalore, with a vertical dashed line marking the start of the lockdown on **March 25, 2020**.

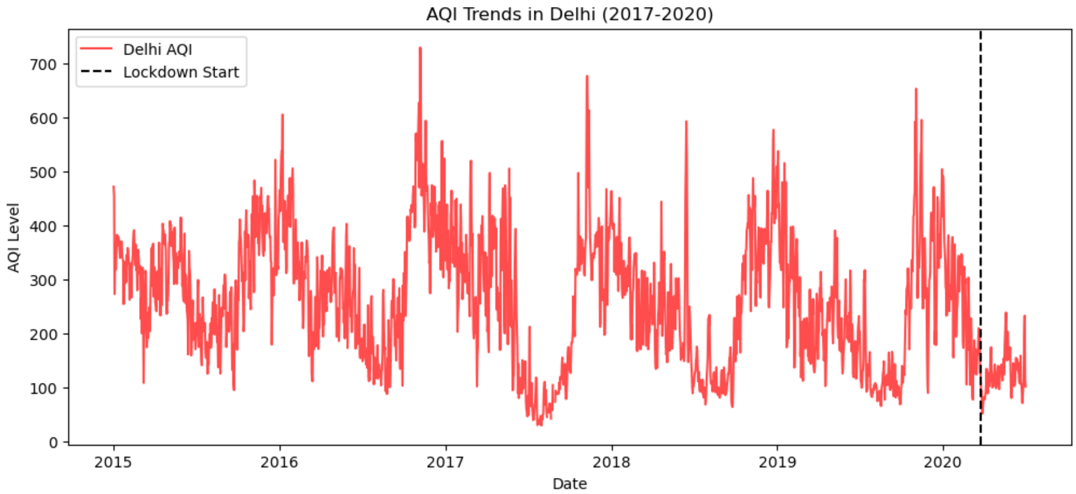

Figure 1: AQI Trend in Delhi (2017-2020). A significant reduction in AQI is observed after the lockdown in March 2020, indicating an improvement in air quality.

Insights and Observations

- **Lockdown Impact:** A noticeable decline in AQI levels was observed in both cities after the lockdown, suggesting that reduced vehicular and industrial activity contributed to improved air quality.

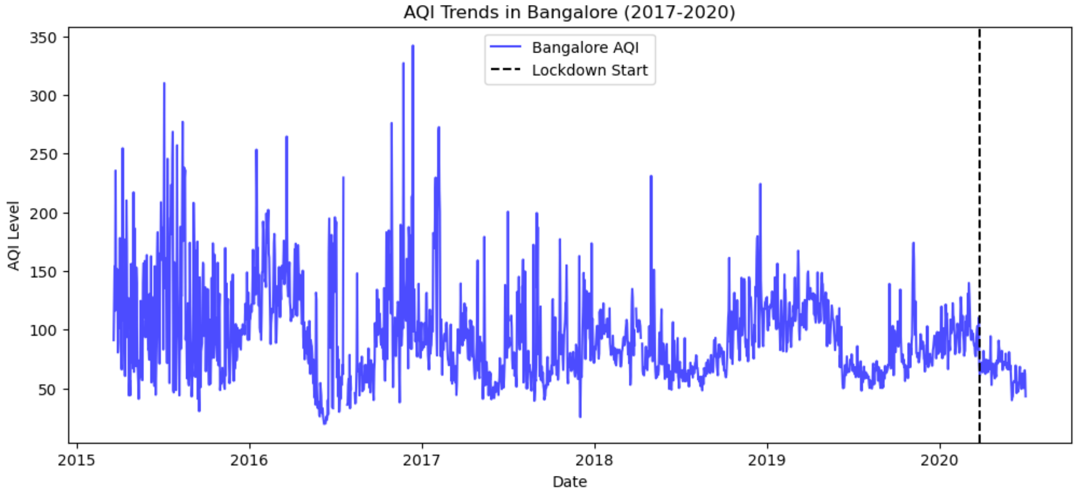

Figure 2: AQI Trend in Bangalore (2017-2020). A moderate drop in AQI is observed after the lockdown, but less drastic than in Delhi, suggesting a difference in pollution sources.

- **Delhi vs. Bangalore:** Delhi had consistently higher AQI levels compared to Bangalore, often reaching severe levels. The post-lockdown improvement was more pronounced in Delhi, indicating a stronger dependence on emission-heavy activities.
- **Seasonal Trends:** Delhi experiences worsening AQI during winter months (October - January), mainly due to stubble burning and temperature inversion effects, whereas Bangalore shows less seasonal variation.
- **Policy Implications:** The findings highlight the need for long-term air pollution control measures beyond emergency restrictions like lockdowns. Strategies such as promoting green energy, improving public transportation, and enforcing stricter emission norms could lead to sustained improvements in air quality.

Conclusion

The analysis demonstrates that the COVID-19 lockdown had a substantial impact on air quality in Delhi and Bangalore, with Delhi showing a more significant improvement. While temporary restrictions led to better air quality, a more sustainable approach is required to maintain these benefits in the long run. This study underscores the importance of continuous monitoring and data-driven policy-making to combat air pollution effectively.

Question 2

Comparison of AQI-Pollutant Correlations: Delhi vs. Bangalore

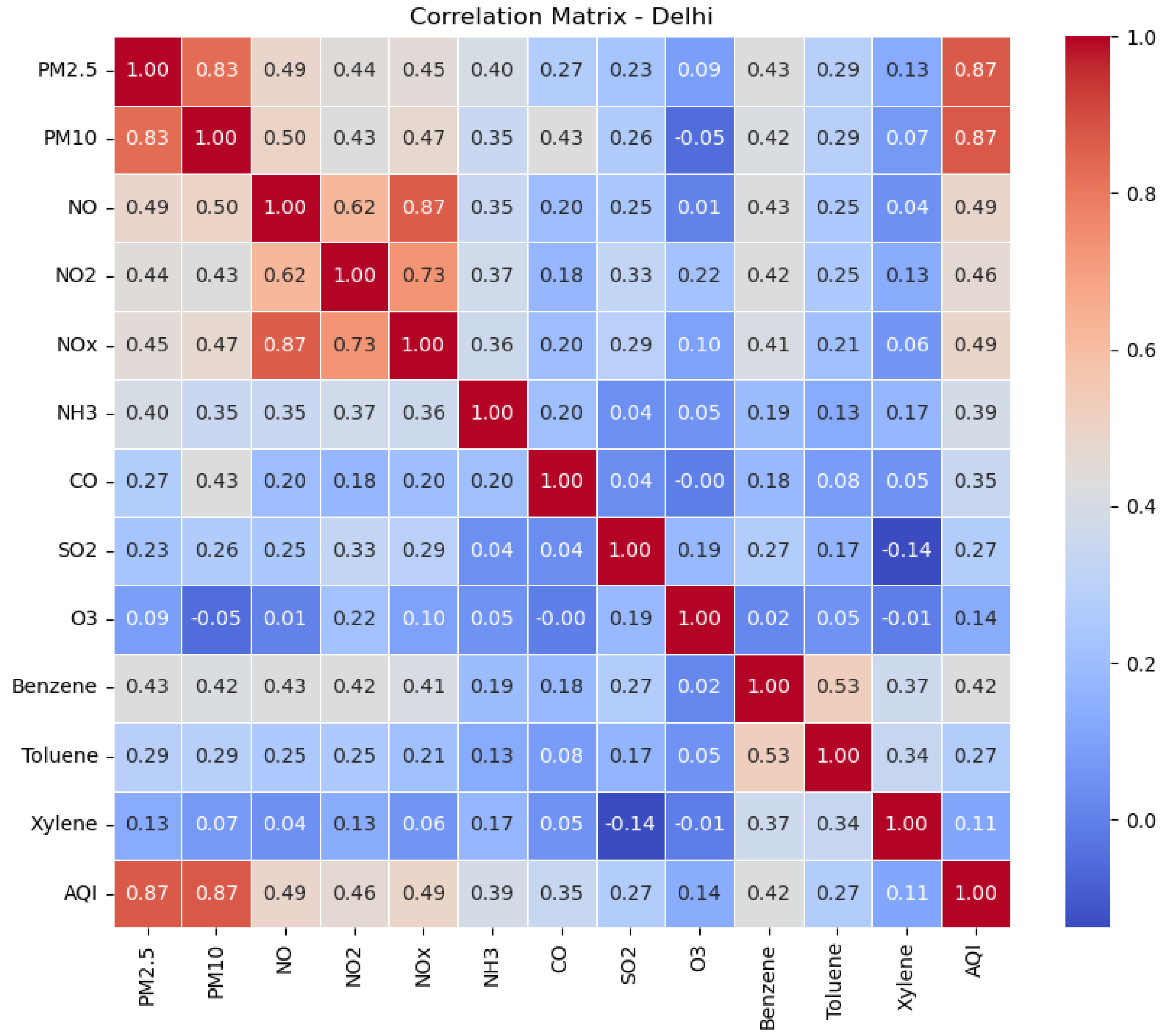

Figure 3: AQI-Pollutant Correlations in Delhi vs. Bangalore.

1. Strength of Correlation

Delhi: Stronger correlation between AQI and NO, NO_2, and $PM_{2.5}$, indicating high industrial and vehicular pollution.

Bangalore: Weaker correlations overall, with CO having the highest impact on AQI, suggesting traffic emissions play a larger role.

2. Key Factors Behind the Differences

Industrial & Vehicular Emissions:

- **Delhi:** More industries, power plants, and severe traffic congestion contribute to higher NO_x and $PM_{2.5}$ emissions.
- **Bangalore:** Fewer industries, lower NO_x emissions, and CO from vehicle pollution as the primary AQI driver.

Climate & Geography:

- **Delhi:** Landlocked, with stubble burning, winter inversion, and low wind speeds trapping pollutants.
- **Bangalore:** Higher elevation, better air circulation, frequent rain dispersing pollutants.

3. Conclusion

Delhi's AQI is more affected by industrial and seasonal pollution, requiring strict industrial and traffic controls. Bangalore's air quality is primarily traffic-driven, needing better urban transport planning and emission regulations. Tailored pollution control strategies are essential for each city to address their unique challenges.

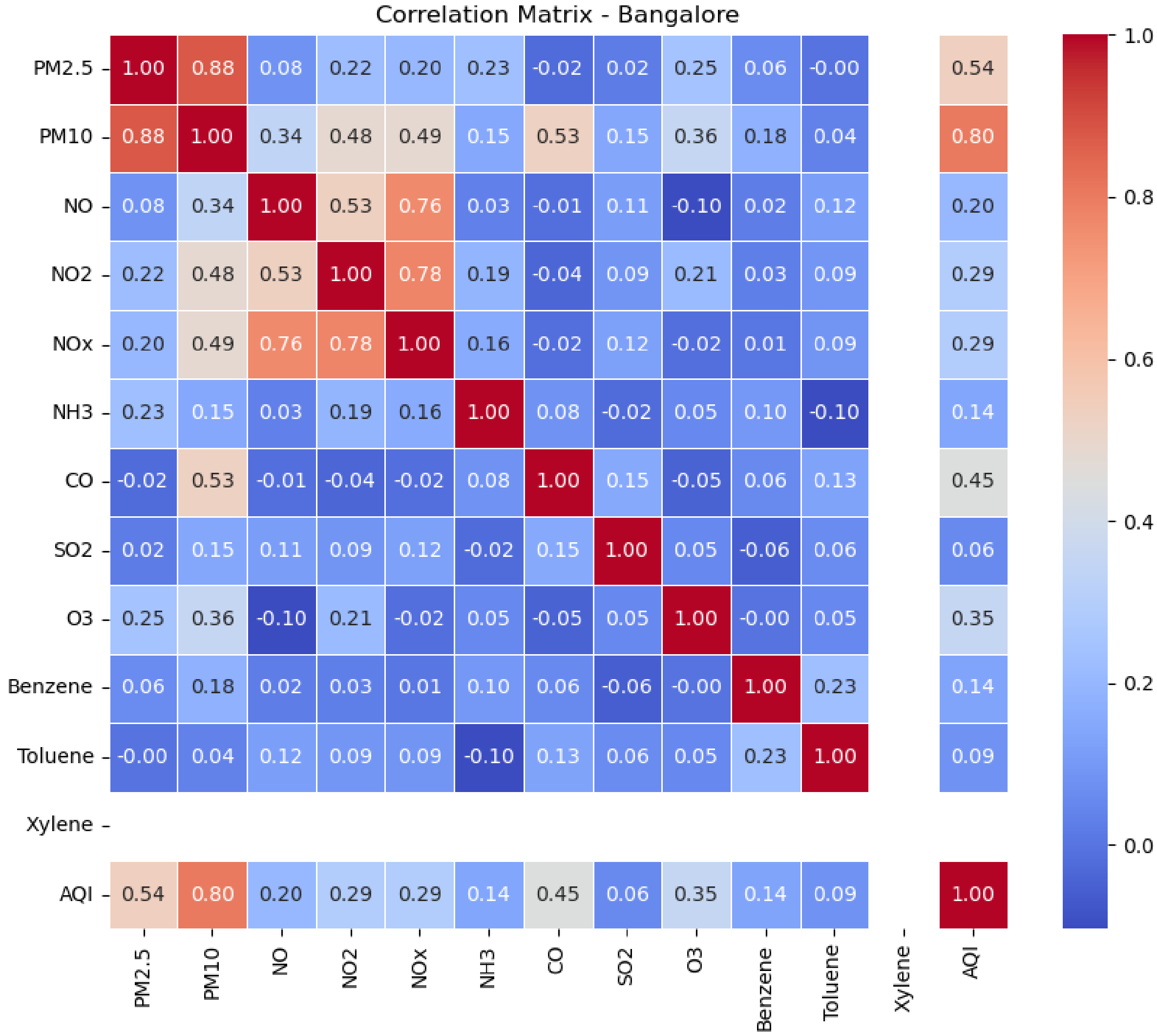

Figure 4: AQI-Pollutant Correlations in Delhi vs. Bangalore.

Question 3

Normalization and Its Effect

Introduction

Normalization is essential in data analysis to ensure all features contribute equally, especially when dealing with pollutants like $\mathbf{PM_{2.5}}$**, CO,** $\mathbf{NO_2}$**, and** $\mathbf{O_3}$, which have different scales. For instance:

- $\mathbf{PM_{2.5}}$ ranges from **10 to 500** $\mu g/m^3$
- **CO** ranges from **0.1 to 5** ppm

Without normalization, models may be biased toward features with larger numerical values.

Process of Normalization

1. Min-Max Scaling

Min-Max Scaling **rescales values** between $[0, 1]$ to maintain proportional differences:

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

Example: If $\mathbf{PM_{2.5}}$ ranges from **10 to 500**, and a specific measurement is **250**:

$$PM_{2.5}' = \frac{250 - 10}{500 - 10} = 0.49$$

Best for: When values need to be **bounded**. **Limitation:** Sensitive to **outliers**.

2. Standardization (Z-score Normalization)

Standardization centers data with **mean = 0** and **standard deviation = 1**:

$$X' = \frac{X - \mu}{\sigma}$$

Example: For $\mathbf{NO_2}$ with **mean = 40, std dev = 15**, and a value of **55**:

$$NO_2' = \frac{55 - 40}{15} = 1.0$$

Best for: Normally distributed data, less sensitive to outliers. **Limitation:** Does not **bound values**.

Importance of Normalization

- **Prevents bias** from large-scale features dominating models.
- **Improves model performance and efficiency.**
- **Ensures fair comparisons** between pollutants.

Question -4

A. Regression Analysis Insights

The analysis used linear regression to predict AQI from pollutants such as PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, and Toluene, excluding Xylene due to significant data gaps.

City	R^2 Score	Interpretation
Delhi	0.820	Strong relationship; 82% of AQI variability explained by pollutants.
Bangalore	0.559	Moderate relationship; 56% of AQI variability explained by pollutants.

Table 1: Regression Model Performance Across Cities

Key Findings:

B. Model Performance and Data Handling

Normalization and Standardization Impact:

- Models were assessed before and after applying Min-Max scaling and Z-score normalization to evaluate any changes in performance.

Normalization Type	City	R^2 Score	MAE	RMSE
Original	Delhi	0.820	40.72	55.50
Original	Bangalore	0.559	21.30	33.40
Min-Max Scaling	Delhi	0.820	40.72	55.50
Min-Max Scaling	Bangalore	0.559	21.30	33.40
Z-Score	Delhi	0.820	40.72	55.50
Z-Score	Bangalore	0.559	21.30	33.40

Table 2: Model Performance Under Different Scaling Techniques

Model Performance Post-Normalization/Standardization:

Missing Value Handling:

- **Pollutants:** Median imputation used for robustness against outliers.
- **Xylene:** Removed due to extensive missing data, avoiding potential bias.
- **AQI:** Rows with missing AQI values were excluded to ensure model integrity.

Technique Justification:

- **Median Imputation:** Avoids the effects of outliers, suitable for skewed data.
- **Dropping Xylene:** Prevents distortion from excessive imputation.
- **Dropping Missing AQI:** Essential for accurate regression analysis as AQI is the target variable.

Conclusion

The regression models reveal significant relationships between pollutants and AQI, particularly strong in Delhi. Normalization and standardization did not affect model performance, suggesting the robustness of the original scaling of pollutants. The strategies for handling missing data ensured the integrity of the analysis. Further investigations could explore more variables or advanced modeling techniques to enhance predictability and insights.

Question-5

1. PCA Results

Delhi

- **Explained Variance:** PC1 explains **32.90%** of the variance, with the first six components capturing **80.87%**.
- **Cumulative Variance:** All 11 components explain **100%** of the variance.

Bangalore

- **Explained Variance:** PC1 explains **21.34%** of the variance, with the first six components capturing **74.48%**.
- **Cumulative Variance:** All 11 components explain **100%** of the variance.

Key Observations

- Delhi's data is more concentrated along PC1, while Bangalore's data is spread across multiple components.
- The first six components explain most of the variance for both cities.

2. Impact of Normalization

Delhi (Normalized)

- PC1 explains **38.19%** of the variance, with the first six components capturing **82.86%**.

Bangalore (Normalized)

- PC1 explains **23.02%** of the variance, with the first six components capturing **77.68%**.

Key Observations

- Normalization increases the explained variance of PC1 for both cities.
- It improves PCA's efficiency in capturing variance, especially when features have different scales.

3. Is PCA Dependent on Normalization?

Yes, PCA is scale-sensitive. Normalization ensures features contribute equally, making it essential when:

- Features have different units or scales.
- Outliers are present.
- Equal feature contribution is desired.

Conclusion

- PCA reveals that the first few components explain most of the variance for both cities.
- Normalization enhances PCA by aligning data better with principal components.
- **Recommendation:** Normalize data before PCA, especially for datasets with varying feature scales.

Question 6

0.1 Regression and Correlation Analysis During Lockdown

To analyze the impact of the lockdown on air quality in Delhi and Bangalore, regression and correlation analyses were conducted. The lockdown provided a unique setting to observe changes in air pollution levels absent typical urban activity.

0.1.1 Regression Analysis Findings

- The regression models demonstrated significant decreases in key pollutants like $PM_{2.5}$, PM_{10}, and NO_x during the lockdown. These changes are attributed to reduced vehicular traffic and industrial activities.
- Correlation analysis between AQI and pollutants showed weaker relationships during the lockdown, suggesting that lower human activity was directly linked to improved air quality.

0.1.2 Role of Data Normalization

- Normalization techniques, including Min-Max Scaling and Z-Score Normalization, were essential in standardizing the pollutant data. This process made it possible to effectively compare AQI levels across different periods.
- A notable shift towards lower AQI values was observed during the lockdown. Normalization emphasized these changes, making it clearer that air quality had improved significantly during this period.

Question-7

0.2 Evaluation Metrics and PCA Impact

Model performance was evaluated using statistical metrics such as R-squared and RMSE to understand the accuracy and precision of the regression models before and after normalization.

0.2.1 Evaluation Metrics Used

- **R-squared:** This metric indicated how well the variations in AQI could be explained by the pollutants. The models maintained similar R-squared values post-normalization, suggesting that scaling did not affect the explanatory power of the models.
- **RMSE:** The Root Mean Squared Error provided insights into the average error magnitude between predicted and actual AQI values. RMSE values were closely monitored before and after normalization, with little to no significant change, affirming model stability.

0.2.2 Impact of PCA on Model Performance

- Principal Component Analysis (PCA) was employed to reduce the dimensionality of the pollutant data, aiming to simplify the model without losing crucial information.
- The application of PCA slightly improved model efficiency by filtering out noise and less influential variables. However, the core model accuracy remained robust, indicating that dimensionality reduction helped streamline the analysis without sacrificing important data insights.

Conclusion

The regression models confirmed that the lockdown had a substantial positive impact on air quality in both cities, with normalization playing a crucial role in highlighting these improvements. The evaluation metrics used confirmed the consistency and reliability of the models, both pre- and post-normalization. PCA contributed to model simplification, proving effective in enhancing model handling and computational efficiency without losing essential information. Future research should continue to explore these methods for sustained air quality improvement strategies.