Air Quality: A Comprehensive Analysis of Past & Predictions in the Future

Student-1 Name: Adam Che Nazahatuhisamudin

Student-1 PSU ID : 942760910

Project Start Date : December 1, 2024

Abstract

Air quality, a critical factor affecting public health and environmental stability, has emerged as a global concern. Pollutants such as PM2.5, SO2, and O3 significantly contribute to respiratory and cardiovascular diseases, as well as environmental degradation. This study focuses on understanding these pollutants by analyzing their historical trends, spatial variations, and future projections. Leveraging comprehensive datasets from the United States Environmental Protection Agency (EPA) spanning 1990 to 2022, this research employs advanced methodologies including regression analysis and machine learning. The study aims to identify the effectiveness of environmental policies, highlight pollution hotspots, and provide actionable recommendations. By integrating temporal and spatial analyses, it sheds light on the improvements achieved and challenges remaining in air quality management. Ultimately, the findings seek to guide policymakers and environmental agencies in formulating targeted strategies to enhance public health and sustainability.

1.0 Introduction

Air quality plays a dual role in shaping public health and environmental sustainability, influencing both individual well-being and broader ecological systems. Pollutants such as PM2.5, SO2, and O3, often byproducts of industrial activities and urbanization, have far-reaching impacts, contributing to respiratory diseases, cardiovascular conditions, and environmental degradation. Despite ongoing regulatory measures and advancements in emission control technologies, air pollution continues to pose significant challenges due to its complex interplay with geographic, climatic, and industrial factors. These pollutants stem from various sources, including industrial activities, vehicular emissions, and household energy use, highlighting the multifaceted nature of air quality management.

The burden of air pollution is not evenly distributed. Urban and industrial regions often experience higher concentrations of pollutants, disproportionately affecting vulnerable populations and ecosystems. This spatial disparity raises concerns about the effectiveness of existing environmental policies and their ability to address localized pollution sources. Additionally, the variability in pollutant trends over time highlights the need for a comprehensive understanding of both temporal and spatial dynamics.

This paper examines historical trends, regional variations, and future projections of air pollutants through advanced analytical methods, including regression analysis and predictive modeling. By focusing on PM2.5, SO2, and O3, the study aims to evaluate the impact of current environmental policies and provide insights into targeted strategies for improving air quality. The findings are intended to contribute to ongoing discussions about air quality management, emphasizing the importance of equitable and effective interventions to mitigate the adverse effects of pollution. Ultimately, this research contributes to the broader effort of enhancing sustainability and protecting public health through improved air quality management.

2.0 Research Methodology

2.1 *Data Collection*

A. Data Extraction

- The data for this study will be collected from the United States Environmental Protection Agency (EPA) in CSV format, containing detailed air quality measurements from 1990 to 2022. These datasets will include several key variables: pollutant concentrations (PM2.5, SO2, O3), year-wise trends, mean pollutant levels, and geographic distributions across Core-Based Statistical Areas (CBSAs). Additionally, the dataset will capture metadata such as sampling frequency and regional identifiers to provide contextual insights into pollutant variations. The raw data will be imported into a pandas DataFrame for initial cleaning and structuring. Essential libraries such as numpy and matplotlib will be employed for numerical calculations and data visualizations, respectively. The extraction process will also involve identifying and removing missing values and irrelevant columns to ensure the integrity of the analysis. To maintain consistency across variables, measurement units will be standardized. Debugging parameters will be strategically applied to streamline error tracing and ensure efficient data handling during extraction.

B. Data Cleaning and Processing

- The data cleaning process will begin with forward-filling missing values in the ‘Core Based Statistical Area’ column to ensure completeness. Numeric columns representing pollutant levels over the years (1990–2022) will be converted to numeric data types, and any rows with missing values will be dropped to ensure accuracy. This will ensure that only complete data is utilized for the analysis, minimizing inaccuracies. Descriptive statistics will be computed to analyze the central tendencies and variability of the pollutants, providing an overview of their distribution. Metrics such as count, mean, variance, and interquartile range (IQR) will be calculated to summarize the data comprehensively. Variables with empty data set due to (Not a Number) NaN or null data after preprocessing will not be used for predictions.

2.2 Data Analysis

A. Simple Linear Regression

- Descriptive statistics will be computed to analyse the central tendencies and variability of the pollutants, providing an overview of their distribution. Metrics such as count, mean, variance, and interquartile range (IQR) will be calculated to summarize the data comprehensively. Regression models will then be implemented to explore the relationship between pollutant levels and influencing factors such as geography and time. Ordinary Least Squares (OLS) methodology will be used to fit the models, generating outputs including coefficients, R-squared values, and p-values. These statistical summaries will form the foundation for understanding pollutant trends and their influencing factors. Scatter plots and temporal trend lines will be generated, providing visual representations of the identified patterns and validating the relationships uncovered during the calculations.

B. Scatter Plot

- Scatter plots will be created to highlight the relationship between pollutant levels and years. Regression lines will visually indicate whether pollutant levels are increasing or decreasing. Temporal trend lines will further be generated to validate the identified patterns and provide visual insights into the data trends.

C. Spatial Analysis

- Spatial analysis will focus on comparing pollutant levels across different Core-Based Statistical Areas (CBSAs) for a specific year, such as 2022. Bar plots will rank CBSAs by pollutant levels, identifying hotspots for air pollution. These spatial visualizations will provide critical insights into regional disparities in air quality and will inform localized mitigation strategies.

D. Predictive Analysis

- Predictive modeling will be conducted using a machine learning approach with a train-test split method, focusing specifically on SO2. Yearly averages of SO2 levels will be used as the target variable, while corresponding years will serve as the feature variable. The data will be split into 80% for training and 20% for testing to ensure robust model validation. A linear regression model will be trained on the training dataset and tested on unseen data. Predictions will be made on the test set, and performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared values will be computed to evaluate the model's accuracy. Scatter plots and regression lines will be used to visualize the model's predictions against actual test data, offering insights into its predictive reliability.

3.0 Results

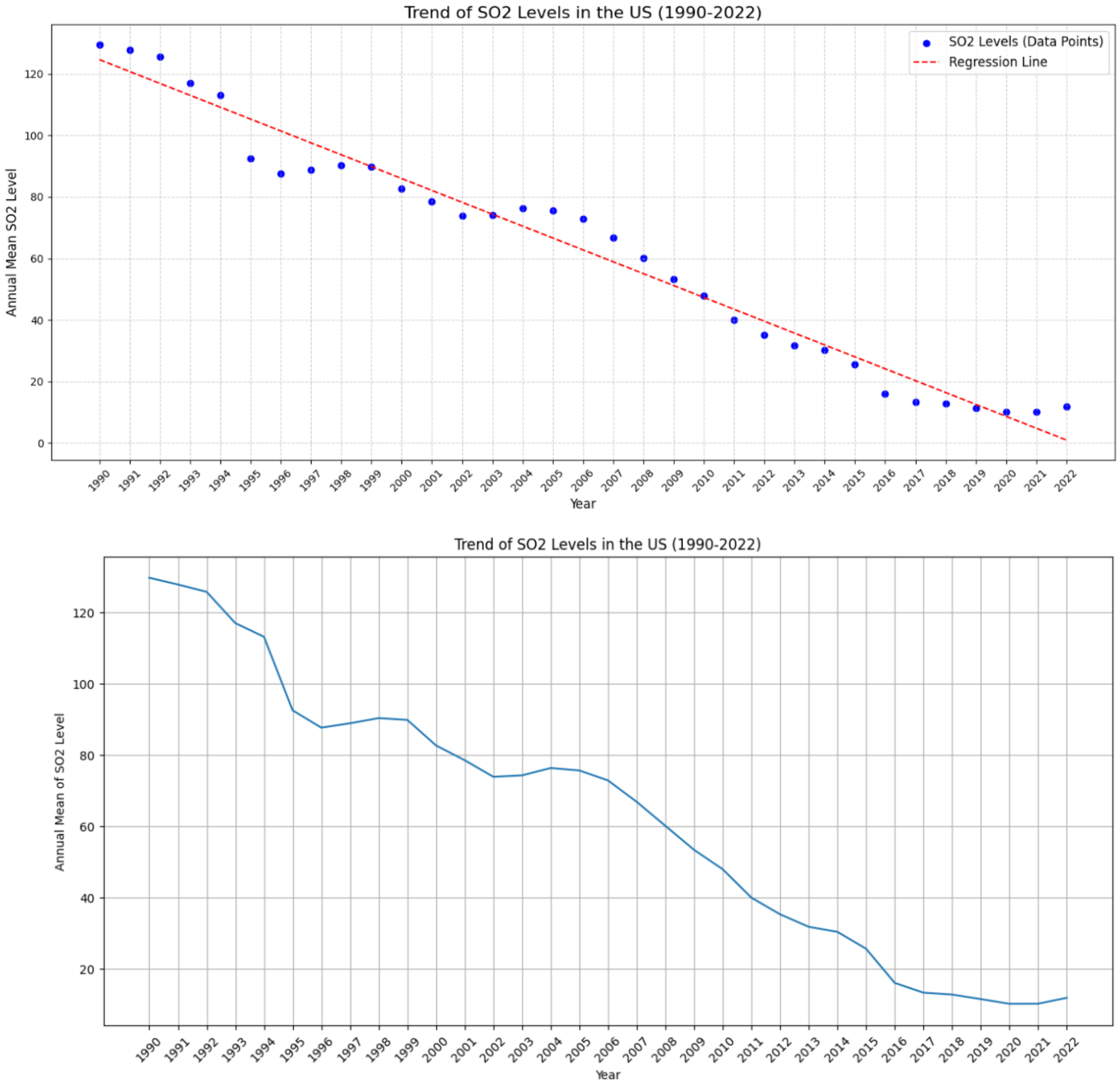


Figure 1: Trend of SO2 Levels in the US

Figure 1 displays the scatter plot displayed SO2 levels across years, with a regression line overlaying the data points to highlight the overall trend. The regression analysis revealed a consistent downward slope, indicating a substantial reduction in SO2 levels over time. This decline suggested that policy measures and advancements in emission control technologies were effective during the analyzed period. The significant decrease in SO2 levels aligned with major regulatory milestones, such as the Clean Air Act amendments and improvements in industrial emission standards. These interventions appeared to have successfully curbed SO2 emissions.

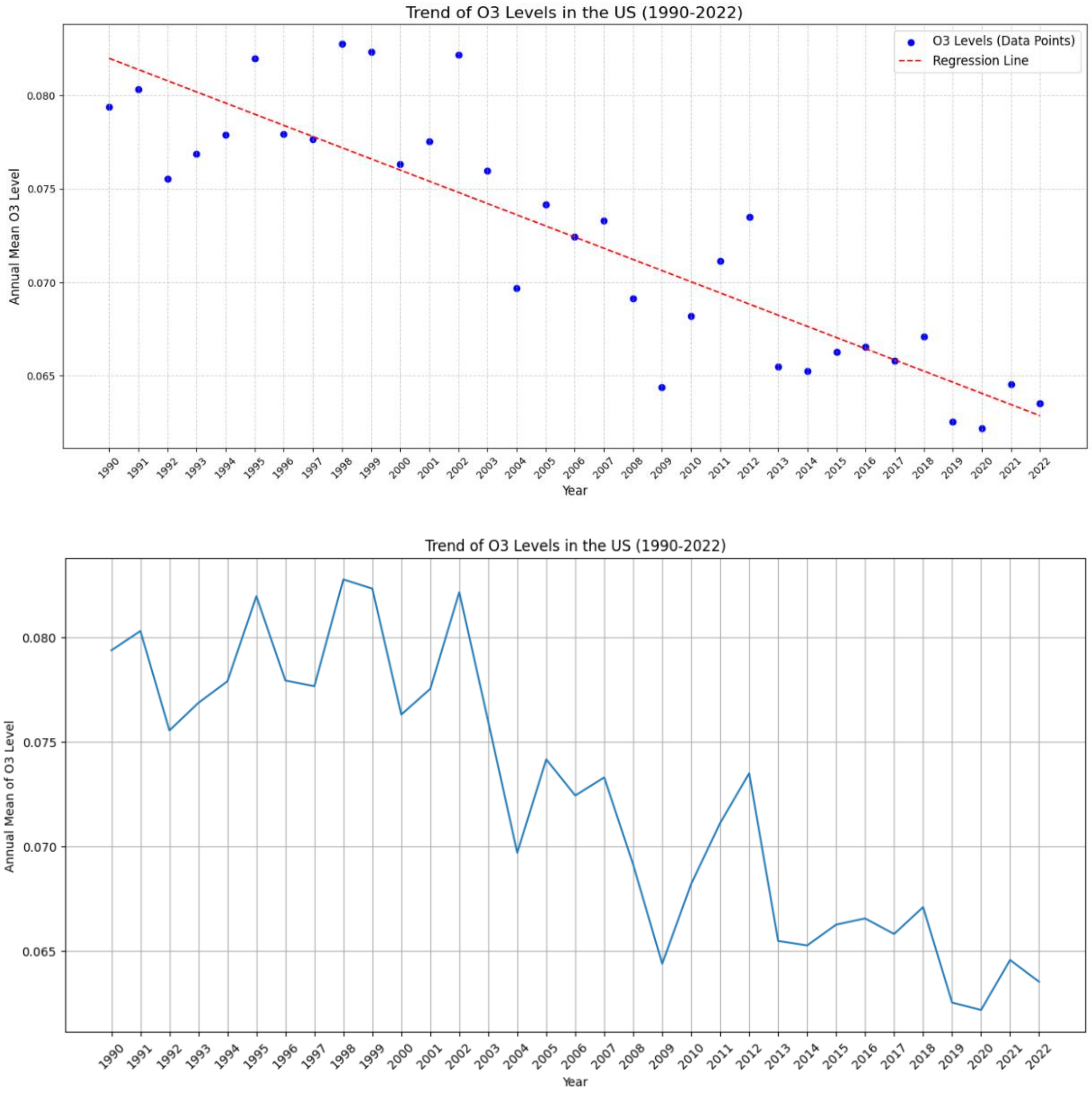


Figure 2: Trend of O3 Levels in the US

Figure 2 displays the scatter plot and regression line for O3 levels which also showed a downward trend, although the slope was less pronounced compared to SO2. This indicated a slower, more gradual improvement in controlling ozone concentrations over time. However, the variability in O3 levels suggested that geographical and seasonal factors played a significant role in influencing concentrations.

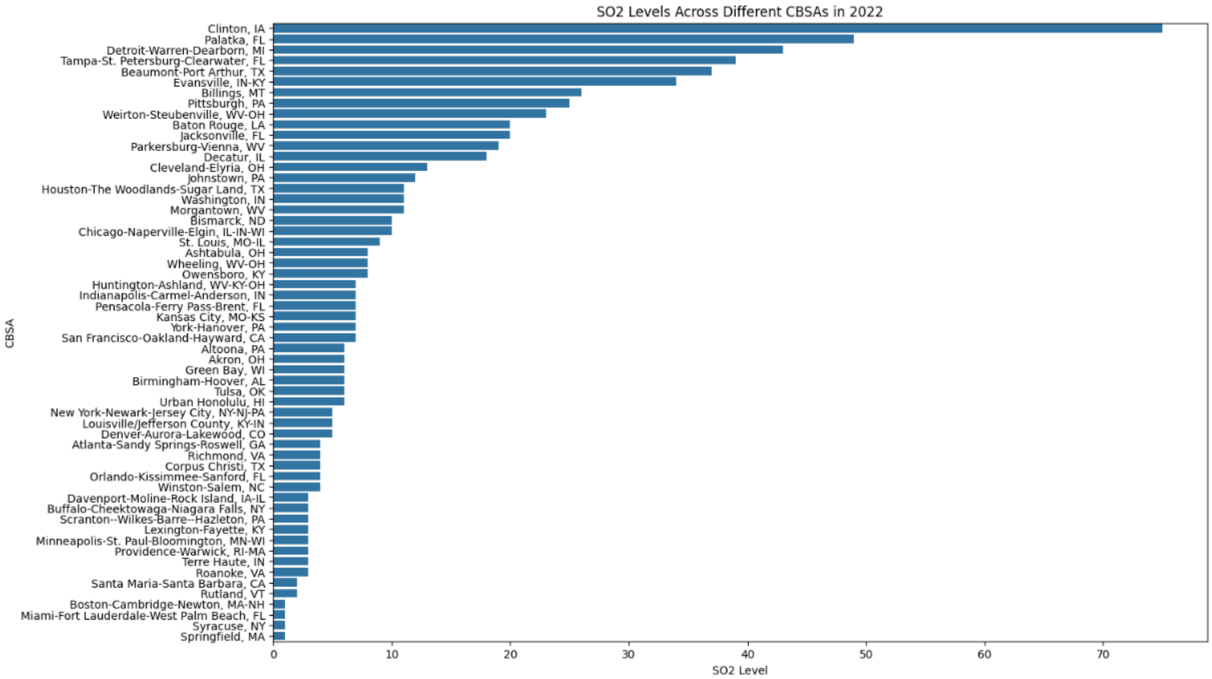


Figure 3: SO2 Levels Across Different States

Figure 3 showed the bar plot illustrating the spatial distribution of SO2 levels across different Core-Based Statistical Areas (CBSAs) in 2022. Significant disparities were evident, with cities like Clinton, IA, and Detroit-Warren-Dearborn, MI, exhibiting the highest SO2 levels.

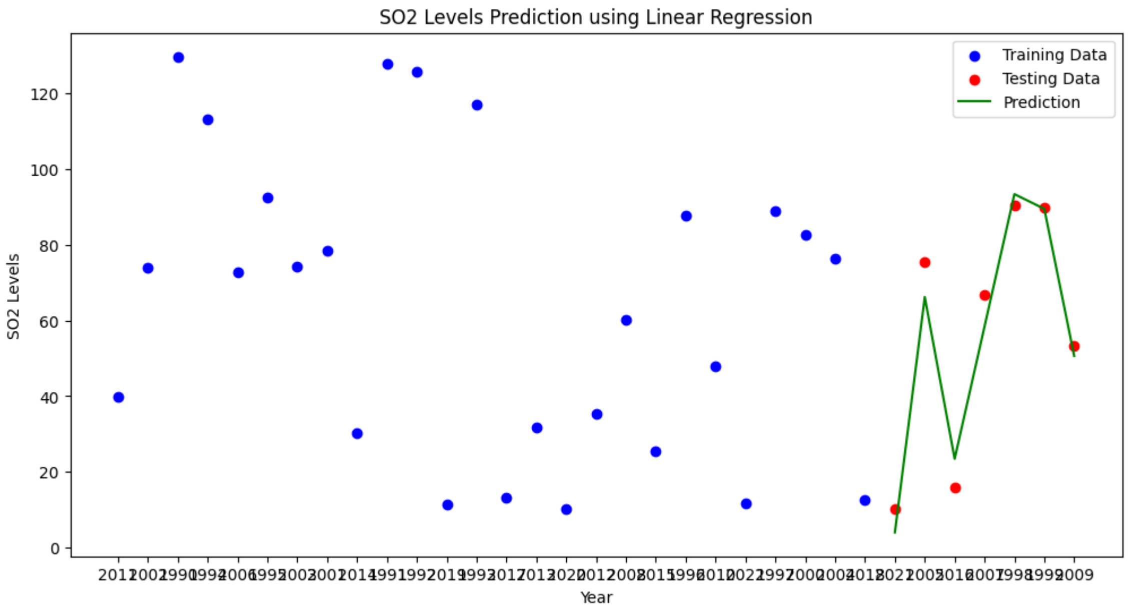
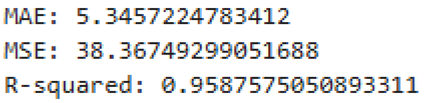


Figure 4: SO2 Levels Prediction



```
MAE: 5.3457224783412
MSE: 38.36749299051688
R-squared: 0.9587575050893311
```

Figure 5: MAE, MSE, R^2 Values

The predictive modeling results demonstrated a strong ability to forecast SO2 levels using a linear regression model. The Mean Absolute Error (MAE) of 5.35 indicated that, on average, the predicted SO2 levels deviated by approximately 5.35 units from the actual levels, reflecting a high level of accuracy. Similarly, the Mean Squared Error (MSE) of 38.37, while slightly larger due to the squaring of errors, confirmed consistent predictions with minimal large deviations. The R-squared value of 0.9588 further underscored the robustness of the model, as it explained approximately 95.88% of the variance in SO2 levels. These results suggest that the model effectively captured the underlying trends in SO2 levels over time and provided reliable forecasts.

4.0 Discussion

The analysis revealed a consistent and significant reduction in SO2 levels across the United States from 1990 to 2022, as demonstrated by the regression models and visual plots. This reduction highlighted the effectiveness of regulatory measures, and improvements in industrial emission standards. The predictive modeling results further confirmed this trend, with an R-squared value of 0.9588 indicating that the model accurately captured the relationship between years and SO2 levels. However, the model's residuals suggested the potential influence of unmeasured variables, such as economic growth, industrial activities, or climatic factors, which may have contributed to localized fluctuations in pollution levels. The spatial analysis also underscored significant disparities in SO2 levels across Core-Based Statistical Areas, with regions like Clinton, IA, and Detroit-Warren-Dearborn, MI, exhibiting notably higher levels. This indicated that while national policies have succeeded in reducing overall pollution, targeted interventions are still required to address regional hotspots. The observed trends for O3 levels, though less steep, revealed fluctuations that might be attributed to seasonal or geographical factors, emphasizing the complexity of ozone pollution control. These findings underscore the importance of integrating local-level interventions and accounting for additional variables in future research to improve predictive accuracy and policy relevance.

5.0 Conclusion

From the analysis, it was concluded that air quality in the United States has significantly improved over the past three decades, with substantial reductions in pollutants like SO2 and O3. The results highlighted the success of regulatory measures in mitigating air pollution and improving public health outcomes. However, persistent regional disparities, as identified through spatial analysis, demonstrated the need for localized strategies to address pollution hotspots. The predictive modeling results, with high accuracy and reliability, showcased the potential for using machine learning tools to forecast future trends and inform policy decisions. Despite these advancements, the presence of unexplained variations suggested that additional factors, such as industrial dynamics or climatic changes, should be incorporated into future studies. These findings call for continued investment in environmental monitoring, innovative policies, and research to sustain and further enhance air quality improvements across the nation.

6.0 References

United States Environmental Protection Agency. "Air Quality in Cities and Counties." EPA, https://www.epa.gov/air-trends/air-quality-cities-and-counties. Date Accessed 7th December 2024.

"Types of Pollutants." World Health Organization, World Health Organization, www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/health-impacts/types-of-pollutants. Date Accessed 1st December 2024.

"National Air Quality Initiative." Natural Resources Conservation Service, U.S. Department of Agriculture, www.nrcs.usda.gov/programs-initiatives/eqip-air-quality-initiative#:~:text=The%20National%20Air%20Quality%20Initiative%20assists%20with%20the%20adoption%20of,reducing%20dust%20and%20carbon%20emissions. Date Accessed 1st December 2024.