AQI TRENDS OVER THE YEARS (2015-2020)

This project aimed to analyze air quality data from 26 cities in India, spanning from January 2015 to July 2020. The primary objectives were to explore the data for insights, visualize pollution trends, identify key patterns, and develop a predictive model for AQI (Air Quality Index).

Techniques Used

Data Cleaning and Preprocessing

- Handled missing values by replacing them with the mean of respective columns (PM2.5, PM10, AQI).
- Converted date column to datetime format for temporal analysis.
- Sorted data chronologically to facilitate time series analysis.

Exploratory Data Analysis (EDA)

- Computed average AQI values for each city.
- Visualized the top 10 cities with the highest pollution levels using bar plots.
- Calculated and visualized the average increase in AQI over time for each city.
- Identified and visualized the top 5 months with the highest AQI for major cities such as Delhi, Mumbai, Kolkata, Chennai, Bengaluru, and Bhopal.
- Produced descriptive statistics and correlation matrix for key pollutants.
- Generated histograms to show the distribution of PM2.5, PM10, and AQI levels.
- Plotted time series to illustrate AQI trends over the specified period.
- Created scatter plots to explore relationships between pollutants and AQI.
- Analyzed trends using moving averages.

Visualization

- Employed Seaborn and Matplotlib to create various plots, enhancing the visual appeal with palettes like viridis and styles like ggplot.

Machine Learning

- Developed a Linear Regression model to predict AQI based on PM2.5 and PM10 levels.

- Split the dataset into training and testing sets to evaluate model performance.
- Assessed model accuracy using Mean Squared Error (MSE) and R-squared (R^2) metrics.
- Visualized actual vs. predicted AQI values to assess the model's predictive capability.

Process Followed

Data Loading and Inspection

- Loaded the dataset and performed initial inspection to understand its structure and contents.

Data Cleaning

- Addressed missing values by imputing with column means.
- Converted and formatted date information for temporal analysis.

Exploratory Data Analysis (EDA)

- Calculated descriptive statistics to summarize the data.
- Generated correlation matrix to understand relationships between variables.
- Visualized distributions and relationships using histograms and scatter plots.

Trend Analysis

- Analyzed and visualized the average AQI by city.
- Identified cities with the highest pollution levels and the greatest average increase in AQI over time.
- Explored seasonal pollution trends in major cities.

Predictive Modeling

- Developed a Linear Regression model to predict AQI.
- Evaluated model performance using appropriate metrics and visualizations.

Visualization and Reporting

- Created comprehensive visualizations to communicate findings.
- Documented the entire process, techniques used, and insights gained.

Findings

Top Polluted Cities

- Identified the top 10 cities with the highest average AQI, with Delhi leading the list, indicating severe pollution levels.

Average Increase in Pollution

- Highlighted cities with the highest average increase in AQI, suggesting worsening air quality over time in these areas.

Descriptive Insights

- Provided a statistical summary and correlation analysis of key pollutants, revealing strong correlations between PM2.5, PM10, and AQI.

Time Series Analysis

- Visualized AQI trends over time, identifying periods of high pollution and observing overall patterns.

Predictive Modeling

- Developed a Linear Regression model to predict AQI based on PM2.5 and PM10 levels with reasonable accuracy (evaluated using MSE and R-squared).

Conclusion

This project provided valuable insights into air quality trends across major Indian cities. By leveraging data cleaning, EDA, visualization, and machine learning techniques, we identified key pollution patterns and developed a predictive model for AQI. The findings underscore the critical air quality issues faced by cities like Delhi and highlight the need for effective pollution control measures. The comprehensive analysis and visualizations can inform policymakers and stakeholders in devising strategies to improve air quality.