Pollution Data Analysis

Exploratory Data Analysis using Python

By: [Muppidi Manaswi]

Objectives

- Load and clean the dataset
- Analyze pollutant trends over time
- Compare pollutant data across stations and states
- Explore correlations between pollutant metrics
- Identify outliers in the data

Data Cleaning Steps

- Loaded dataset with UTF-8 encoding
- Replaced blank spaces with NaN
- Dropped rows with missing pollutant values
- Converted pollutant columns to numeric types
- Removed remaining rows with NaN values

Pollutant Average Over Time

- Line plot of pollutant_avg over last_update
- Reveals trends and patterns in pollution over time

Pollutant Levels by Station

- Boxplot grouped by station
- Highlights variation across different stations

Average Pollutant by State

- Barplot showing mean pollutant_avg per state
- Comparison of pollution levels across states

Correlation Between Metrics

- Heatmap showing correlation among pollutant_min, pollutant_max, and pollutant_avg
- Strong relationships between different pollutant metrics

Pollutant Type Disparities

- Boxplot comparing pollutant_avg across pollutant_id
- Analyzes the distribution of different pollutant types

Pairwise Relationships

- Pairplot to visualize relationships among pollutant metrics
- Helps detect trends and correlations

Comprehensive Heatmap

- Viridis-style heatmap for better contrast
- Reinforces correlation findings from earlier heatmap

Outlier Detection

- Boxplot used to visually identify outliers in pollutant_avg
- Z-score used to programmatically detect outliers
- Outliers flagged with Z-score > 3

Conclusion

- Dataset cleaned and preprocessed successfully
- Insights into pollution trends and disparities identified
- Correlations and outliers discovered through visualization and statistical methods

Thank You

- Questions and Discussion