# Title: Exploring Probability Distributions for Air Quality Data

Subtitle: Statistical Insights into Air Pollution Data Using Probability Distributions

## ISSUE / PROBLEM

Air pollution, particularly carbon monoxide levels, poses a significant environmental and health risk. As part of an analytics team for the U.S. Environmental Protection Agency (EPA), this project investigates air quality data from over 200 sites across different states, counties, and cities. The primary objective is to determine the best-fitting probability distribution for the dataset, calculate z-scores, and identify outliers. This analysis helps pinpoint regions that need support in improving air quality.

## IMPACT

- **Targeted Air Quality Improvements:** Identifying counties and cities with extreme AQI values enables policymakers to allocate resources more effectively.
- **Better Environmental Decision-Making:** Using probability distributions and statistical tests allows data-driven insights into pollution trends.
- **Outlier Detection for Further Investigation:** Sites with abnormally high AQI readings can be further examined for potential sources of pollution.

## RESPONSE

This project follows a structured approach to assess air quality data:

1. **Data Exploration:**
   - A histogram of logarithmically transformed Air Quality Index (AQI) readings (aqi_log) was created to visualize its distribution.
   - The data appears approximately normal with a slight right skew.
2. **Statistical Tests:**
   - The empirical rule was applied to check normality:
   - **76.15%** of data falls within 1 standard deviation (expected: **68%**).
   - **95.77%** falls within 2 standard deviations (expected: **95%**).
   - **99.62%** falls within 3 standard deviations (expected: **99.7%**).
   - While the 1-standard-deviation result deviates slightly, the overall results suggest an approximately normal distribution.
3. **Z-score Analysis & Outlier Detection:**
   - The z-score for each AQI value was computed to identify extreme values.
   - Any values exceeding **±3 standard deviations** from the mean were flagged as potential outliers.

## KEY INSIGHTS

- The AQI data follows an approximately normal distribution but with a slight right skew.
- A significant portion (95.77%) of the data falls within two standard deviations, reinforcing normality assumptions.
- Z-score analysis successfully identifies potential outliers, which could indicate pollution hotspots or data anomalies.