Data Dictionary: Air Quality, Weather, and Wildfire Dataset

Dataset Overview

- **Total Records:** 19,802 observations
- **Total Features:** 44 columns
- **Time Period:** 2024 (full year)
- **Geographic Scope:** United States (multiple states)

Feature Definitions

Column Name	Data Type	Description	Unit/Format	Range/Values	Category
Unnamed: 0	Integer	Row index	-	0 to 19,801	Identifier
date	String	Observation date	YYYY-MM-DD	2024-01-01 to 2024-12-31	Temporal
site_id	String	Unique monitoring site identifier	XX-XXX-XXXX	Various	Identifier
latitude	Float	Station latitude coordinate	Decimal degrees	21.32 to 47.57	Geographic
longitude	Float	Station longitude coordinate	Decimal degrees	-158.11 to -67.87	Geographic
state_name	String	US state name	Text	Various states	Geographic
county_name	String	County name	Text	Various counties	Geographic
city_name	String	City name	Text	Various cities	Geographic
site_name	String	Monitoring site name	Text	Various site names	Geographic

Air Quality Measurements (Target Variables)

Column Name	Data Type	Description	Unit/Format	Range/Values	Category
PM25	Float	Fine particulate matter (≤2.5 micrometers)	µg/m³	-2.90 to 89.10	Air Quality
CO	Float	Carbon monoxide concentration	ppm	-0.40 to 4.90	Air Quality
O3	Float	Ozone concentration	ppm	0.00 to 0.19	Air Quality
NO2	Float	Nitrogen dioxide concentration	ppb	-0.60 to 80.00	Air Quality
SO2	Float	Sulfur dioxide concentration	ppb	-1.00 to 234.00	Air Quality

Air Quality Index (AQI) Values

Column Name	Data Type	Description	Unit/Format	Range/Values	Category
AQI_PM25	Float	Air Quality Index for PM2.5	Index (0-500)	0 to 295	Air Quality Index
AQI_CO	Float	Air Quality Index for CO	Index (0-500)	0 to 52	Air Quality Index
AQI_O3	Float	Air Quality Index for O3	Index (0-500)	0 to 179	Air Quality Index
AQI_NO2	Integer	Air Quality Index for NO2	Index (0-500)	0 to 75	Air Quality Index
AQI_SO2	Float	Air Quality Index for SO2	Index (0-500)	0 to 305	Air Quality Index
AQI	Float	Overall Air Quality Index (max of all pollutant AQIs)	Index (0-500)	12 to 291	Air Quality Index

AQI Categories:

- 0-50: Good (Green)
- 51-100: Moderate (Yellow)

- 101-150: Unhealthy for Sensitive Groups (Orange)
- 151-200: Unhealthy (Red)
- 201-300: Very Unhealthy (Purple)
- 301-500: Hazardous (Maroon)

Weather Features

Column Name	Data Type	Description	Unit/Format	Range/Values	Category
temperature_2m_mean	Float	Mean daily temperature at 2m above ground	°Celsius	-28.40 to 41.20	Weather
temperature_2m_max	Float	Maximum daily temperature at 2m above ground	°Celsius	-24.30 to 48.20	Weather
temperature_2m_min	Float	Minimum daily temperature at 2m above ground	°Celsius	-35.00 to 34.50	Weather
relative_humidity_2m_mean	Float	Mean daily relative humidity at 2m	Percentage (%)	10.00 to 100.00	Weather
wind_speed_10m_mean	Float	Mean daily wind speed at 10m above ground	km/h	1.60 to 50.60	Weather
wind_direction_10m_dominant	Float	Dominant wind direction at 10m	Degrees (0-360)	0 to 359	Weather
precipitation_sum	Float	Total daily precipitation	mm	0.00 to 142.60	Weather
precipitation_hours	Float	Hours of precipitation	Hours	0.00 to 24.00	Weather
et0_fao_evapotranspiration	Float	Reference evapotranspiration (FAO Penman-Monteith)	mm	0.00 to 10.34	Weather

| weather_code | String | WMO weather code (https://www.nodc.noaa.gov/archive/arc0021/0002199/1.1/data/0-data/HTML/WMO-CODE/WMO4677.HTM) or description | Code/Text | Various codes | Weather |

Fire Features

Column Name	Data Type	Description	Unit/Format	Range/Values	Category
distance_to_fire_km	Float	Distance from station to nearest fire	Kilometers	0.00 to 999.00 (NaN = no fire)	Fire
fire_brightness	Float	Mid-infrared brightness temperature (Channel I-4, ~4µm)	Kelvin	295.01 to 367.00	Fire
fire_frp	Float	Fire Radiative Power - energy released by fire	Megawatts (MW)	0.00 to 363.68	Fire
fires_within_50km	Integer	Count of fires within 50km radius	Count	0 to 150+	Fire
fires_within_100km	Integer	Count of fires within 100km radius	Count	0 to 300+	Fire
has_nearby_fire	Integer	Binary flag for fire presence	0 = No, 1 = Yes	0 or 1	Fire

Fire Brightness Notes:

- Typical range for active fires: 300-500K
- Higher values indicate more intense fires
- Standard fire detection metric from satellite data (MODIS/VIIRS)

Fire FRP Notes:

- Directly measures the energy released by fire
- Better correlates with smoke emissions than brightness
- Higher FRP = more intense fire and greater air quality impact

Temporal Features

Column Name	Data Type	Description	Unit/Format	Range/Values	Category
datetime	String	Full datetime (same as date in this dataset)	YYYY-MM-DD	2024-01-01 to 2024-12-31	Temporal
month	Integer	Month of year	1-12	1 (January) to 12 (December)	Temporal
day_of_week	Integer	Day of week	0-6	0 (Monday) to 6 (Sunday)	Temporal
is_weekend	Integer	Weekend indicator	0 = Weekday, 1 = Weekend	0 or 1	Temporal
season	String	Meteorological season	Text	winter, spring, summer, fall	Temporal
wildfire_season	Integer	Wildfire season indicator (June-October)	0 = No, 1 = Yes	0 or 1	Temporal

Season Definitions:

- Winter: December, January, February
- Spring: March, April, May
- Summer: June, July, August
- Fall: September, October, November

Categorical Features (Engineered)

Column Name	Data Type	Description	Unit/Format	Range/Values	Category

fire_distance_category	String	Categorical fire proximity	Text	no_fire, very_close (<25km), close (25-50km), moderate (50-100km), far (>100km)	Fire Category
fire_intensity	String	Categorical fire intensity based on FRP	Text	no_fire, low (<10 MW), moderate (10-50 MW), high (50-100 MW), extreme (>100 MW)	Fire Category

Data Quality Notes

Missing Values

- **Fire features** (distance_to_fire_km, fire_brightness, fire_frp): NaN indicates no fire detected on that date
- **Weather features**: Minimal missing values (data from Open-Meteo API)
- **Air quality features**: Some negative values present (sensor calibration artifacts)

Data Sources

1. **Air Quality Data:** EPA Air Quality System (AQS) - Daily summary data
2. **Weather Data:** Open-Meteo Historical Weather API
3. **Fire Data:** NASA FIRMS (Fire Information for Resource Management System) - MODIS/VIIRS satellite data
4. **Geographic Data:** US Census Bureau TIGER/Line shapefiles (2023)

Important Considerations

Negative Values in Air Quality:

- Some pollutant measurements show negative values (e.g., PM25: -2.90, CO: -0.40)
- These are sensor artifacts/calibration issues from the original EPA data
- Consider filtering or setting to 0 during modeling

Fire Detection Limitations:

- Satellite fire detection has a minimum fire size threshold (~1000m² for MODIS)
- Cloud cover can obscure fire detection
- Temporal resolution: Satellite passes 2-4 times per day

Temporal Coverage:

- Dataset covers full year 2024
- Daily temporal resolution
- Some sites may have gaps in monitoring