Numeric Variables (14+):

Numeric Variables (14+):

- PM2.5, *PM10*, O3, NO2, CO, SO2
- Weather – Humidity, temperature, wind speed, wind direction, precipitation
- Distance to fire
- Fire brightness, confidence, FRP
- Fires within radii (50/100/200km)
- Fire intensity score
- Day of week, month

Categorical Variables (6+):

- Fire proximity category (Very Close/Close/Moderate/Far)
- PM2.5 category (Good/Moderate/Unhealthy...)
- Has nearby fire (Yes/No)
- Weekend (Yes/No)
- Station name (if you keep it)
- Month (can treat as categorical)

Goal: predict air quality degradation from wildfire proximity

Key Questions:

- How does fire proximity affect PM2.5 levels?
- Which pollutants are most impacted by wildfires?
- Do weather conditions moderate fire impacts on air quality?
- Can we cluster regions by air quality response patterns?

Linear Regression – predict PM2.5 from fire distance and conditions, weather
KNN – predict air quality from similar conditions
K-Means Clustering – cluster monitoring stations by pollution patterns, interpret clusters (e.g., heavily impacted, moderately impacted)
PCA – reduce pollutants to 2-3 principal components

For modeling – we need to bootstrap observations, do Durbin-Watson test to check for autocorrelation when we're checking our regression assumptions

```
# Each row represents: ONE STATION, ONE DATE/TIME
final_df columns:
- date
- station_lat, station_lon, station_name  # Where air quality measured
- PM25, PM10, O3, NO2, CO, SO2           # Air quality AT STATION
- temperature, humidity, pressure        # Weather AT STATION ← KEY!
- wind_speed, wind_direction             # Wind AT STATION ← KEY!
```

- distance_to_fire_km # Distance from station to nearest fire
- fire_brightness, fire_frp # Characteristics of that fire
- fires_within_50km, fires_within_100km # Fire density around station

Fire data comes from NASA's Visible Infrared Imaging Radiometer Suite (VIIRS), which provides near real-time data on active fires. NRT active fire data is distributed by NASA's Fire Information for Resource Management System (FIRMS).
Source: https://www.earthdata.nasa.gov/data/tools/firms/active-fire-data-attributes-modis-viirs

Attribute	Short Description	Long Description
Latitude	Latitude	Center of nominal 375 m fire pixel
Longitude	Longitude	Center of nominal 375 m fire pixel
Bright_ti4 / Brightness (in web services)	Brightness temperature I-4	VIIRS I-4 channel brightness temperature of the fire pixel measured in Kelvin
Scan	Along Scan pixel size	The algorithm produces approximately 375 m pixels at nadir. Scan and track reflect actual pixel size
Track	Along Track pixel size	The algorithm produces approximately 375 m pixels at nadir. Scan and track reflect actual pixel size
Acq_Date	Acquisition Date	Date of VIIRS acquisition

Acq_Time	Acquisition Time	Time of acquisition/overpass of the satellite (in UTC)
Satellite	Satellite	N= Suomi National Polar-orbiting Partnership (Suomi NPP), N20=NOAA-20 (designated JPSS-1 prior to launch), N21=NOAA-21 (designated JPSS-2 prior to launch)
Confidence	Confidence	This value is based on a collection of intermediate algorithm quantities used in the detection process. It is intended to help users gauge the quality of individual hotspot/fire pixels. Confidence values are set to low, nominal and high. Low confidence daytime fire pixels are typically associated with areas of sun glint and lower relative temperature anomaly (<15K) in the mid-infrared channel I4. Nominal confidence pixels are those free of potential sun glint contamination during the day and marked by strong (>15K) temperature anomaly in either day or nighttime data. High confidence fire pixels are associated with day or nighttime saturated pixels. Please note: Low confidence nighttime pixels occur only over the geographic area extending from 11deg E to 110 deg W and 7 deg N to 55 deg S. This area describes the region of influence of the South Atlantic Magnetic Anomaly which can cause spurious brightness temperatures in the mid-infrared channel I4 leading to potential false positive alarms. These have been removed from the NRT data distributed by FIRMS.

Version	Version (Collection and source)	Version identifies the collection (e.g. VIIRS Collection 1) and source of data processing: Near Real-Time (NRT suffix added to collection) or Standard Processing (collection only) "1.0NRT" - Collection 1 NRT processing "1.0" - Collection 1 Standard processing
Bright_ti5 / Brightness_2 (in web services)	Brightness temperature I-5	I-5 Channel brightness temperature of the fire pixel measured in Kelvin
FRP	Fire Radiative Power	FRP depicts the pixel-integrated fire radiative power in MW (megawatts). FRP depicts the pixel-integrated fire radiative power in MW (megawatts). Given the unique spatial and spectral resolution of the data, the VIIRS 375 m fire detection algorithm was customized and tuned in order to optimize its response over small fires while balancing the occurrence of false alarms. Frequent saturation of the mid-infrared I4 channel (3.55-3.93 µm) driving the detection of active fires requires additional tests and procedures to avoid pixel classification errors. As a result, sub-pixel fire characterization (e.g., fire radiative power [FRP] retrieval) is only viable across small and/or low-intensity fires. Systematic FRP retrievals are based on a hybrid approach combining 375 and 750 m data. In fact, starting in 2015 the algorithm incorporated additional VIIRS channel M13 (3.973-4.128 µm) 750 m data in both aggregated and unaggregated format.
DayNight	Day or Night	D= Daytime fire, N= Nighttime fire

Data Dictionary

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 (~1000m^2 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