|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Summary Statistics** | | | | | | | |
| variable | missing | zeros | mean | std | min | max | unique |
| id | 0 | 1 | 120946.8 | 70521.97 | 0 | 243928 | 239663 |
| year | 0 | 0 | 2004 | 0 | 2004 | 2004 | 1 |
| doy | 0 | 13 | 163.3998 | 90.17132 | 0 | 366 | 367 |
| longitude | 0 | 0 | -98.175 | 13.77921 | -124.2862 | -67.2274 | 136434 |
| latitude | 0 | 0 | 35.21794 | 5.565285 | 25.2192 | 48.9998 | 106695 |
| grid10k | 0 | 0 | 54111.09 | 29614.02 | 2065 | 131898 | 7530 |
| covertype | 0 | 5594 | 25.47199 | 36.72419 | 0 | 95 | 12 |
| fuelcode | 0 | 75028 | 601.3591 | 691.3018 | 0 | 2950 | 32 |
| area\_burned | 0 | 111449 | 33436.01 | 31173.51 | 0 | 62500 | 2 |
| prefire\_fuel | 0 | 75029 | 2553.278 | 2881.367 | 0 | 13337.5108 | 9182 |
| consumed\_fuel | 0 | 75029 | 1482.814 | 1603.583 | 0 | 9600.08329 | 11699 |
| ECO2 | 0 | 75029 | 2413.393 | 2599.511 | 0 | 14918.5294 | 11699 |
| ECO | 0 | 75029 | 145.5681 | 175.4834 | 0 | 1276.81108 | 11699 |
| ECH4 | 0 | 75029 | 6.40991 | 9.324988 | 0 | 71.808623 | 11699 |
| EPM2\_5 | 0 | 75029 | 23.71143 | 29.57503 | 0 | 218.881899 | 11699 |
| cwd\_frac | 0 | 132065 | 0.036084 | 0.071237 | 0 | 0.464559 | 342 |
| duff\_frac | 0 | 132065 | 0.055813 | 0.085311 | 0 | 0.787142 | 342 |
| fuel\_moisture\_class | 0 | 0 | 2.737711 | 0.8606 | 1 | 4 | 4 |
| burn\_source | 0 | 0 | 1.916616 | 0.945714 | 1 | 4 | 4 |
| burnday\_source | 0 | 0 | 25.90159 | 25.7994 | 12 | 81 | 5 |
| BSEV | 0 | 67547 | 1.430071 | 1.120584 | 0 | 6 | 7 |
| BSEV\_flag | 0 | 155236 | 0.352274 | 0.47768 | 0 | 1 | 2 |
| fire\_date | 0 | <NA> | 6/11/2004 | NaN | 12/31/2003 | 12/31/2004 | 367 |
| bi\_value | 0 | 40950 | 37.02481 | 26.6636 | 0 | 120 | 119 |
| fm100\_value | 0 | 0 | 12.32157 | 4.800205 | 2.3 | 28.700001 | 250 |
| pet\_value | 0 | 2 | 4.894996 | 2.087405 | 0 | 12.7 | 123 |
| fm1000\_value | 0 | 0 | 14.75171 | 5.120754 | 2.9 | 34.299999 | 294 |
| pr\_value | 0 | 192508 | 1.143498 | 5.265034 | 0 | 124.599998 | 535 |
| rmax\_value | 0 | 0 | 74.52294 | 24.29095 | 10 | 100 | 840 |
| rmin\_value | 0 | 0 | 30.4454 | 14.7056 | 2.5 | 94.5 | 776 |
| sph\_value | 0 | 0 | 0.006975 | 0.004144 | 0.00043 | 0.02096 | 1830 |
| srad\_value | 0 | 0 | 240.6275 | 73.33466 | 19.6 | 428.800018 | 3428 |
| tmmn\_value | 0 | 0 | 283.2673 | 7.482869 | 247 | 304.799988 | 467 |
| th\_value | 0 | 238 | 189.8934 | 97.88463 | 0 | 359 | 360 |
| tmmx\_value | 0 | 0 | 297.9314 | 7.622406 | 256.100006 | 321.200012 | 535 |
| vpd\_value | 0 | 88 | 1.300629 | 0.808449 | 0 | 6.91 | 481 |
| vs\_value | 0 | 0 | 4.044534 | 1.562581 | 0.5 | 15 | 122 |

**Basic Identifiers**

**id**

* No missing or zero values
* Values range from 0 to 243,928
* 239,663 unique values
* Clean and consistent identifier

**year**

* All values are 2004
* No variability, as expected for a single-year dataset

**doy (Day of Year)**

* 13 missing values (zero)
* Values range from 0 to 366 (366 in 2004 calendar year)
* Valid range for a leap year
* 367 unique values

**Location Data**

**longitude / latitude**

* No missing values
* Values fall within realistic geographic bounds for the United States

**grid10k**

* No missing values
* 7,530 unique grid codes

**covertype**

* 5,594 missing values
* 12 unique classes

**Fire Metrics**

**Fuelcode**

* 75,028 zero values
* Values range from 0 to 2950
* 32 unique codes observed in total
* Some of the most frequent non-zero values include 1, 2, 1140, 1180, 1500, and 1600
* Codes like 1240, 1320, 1340, and 1910 are associated with the maximum recorded area burned (62,500), indicating these fuel types may correlate with larger fire events

**area\_burned**

* 111,449 zero values
* Max value is 62,500, which may be a reporting cap
* Most values are concentrated at 0 and 62,500
* 69,833 rows have both prefire\_fuel and area\_burned equal to 0

**prefire\_fuel / consumed\_fuel**

* Consumed fuel is much lower than prefire fuel, which aligns with expectations—only a portion of available fuel is typically burned in a fire event.
* 75,029 zero values each
* Max consumed fuel is 13,337
* A histogram comparison reveals a highly right-skewed distribution, with a majority of values clustered at low amounts.
* The scatter plot between prefire\_fuel and consumed\_fuel shows a clear linear relationship when values are non-zero, indicating that consumed\_fuel tends to increase proportionally with prefire\_fuel.
* 41,616 rows with area\_burned = 0 have significant prefire fuel, fuel consumption, and emissions
* This suggests area\_burned = 0 may not truly mean “no fire”, but rather a data resolution or reporting issue
* The consistent overlap of zero values across prefire\_fuel, consumed\_fuel, and all four emission variables strongly suggests that zero does not represent actual zero values, but rather missing or unrecorded data.

A graph of fuel distribution

AI-generated content may be incorrect.A graph showing a graph of fuel prices

AI-generated content may be incorrect.

**Emissions**

**ECO2, ECO, ECH4, EPM2\_5**

* 75,029 zeros each
* Values fall within plausible environmental ranges, though exact units (e.g., ppm or g/m²) are not confirmed
* ECO emissions appear even when area\_burned is zero in over 41,000 rows, which may suggest that emissions can occur due to smoldering, underreporting of burn area, or other processes not captured by area\_burned
* Strong correlation observed between consumed fuel and emissions, even when area\_burned = 0
* Given the chemical nature of these emission metrics and the observed distribution patterns, it is highly likely that the large number of zero values in ECO2, ECO, ECH4, and EPM2\_5 represent missing data rather than true absence of emissions.

A group of graphs showing different sizes of data

AI-generated content may be incorrect.

**Fuel and Moisture**

**cwd\_frac / duff\_frac (Proportion of large, downed woody material on the forest floor / roportion of the organic soil layer made up of decomposing leaves, needles, and plant material beneath the litter layer)**

* 132,065 zero values each
* Histograms show sharp spikes at zero, followed by right-skewed distributions
* All rows where prefire\_fuel == 0 also have cwd\_frac == 0 and duff\_frac == 0
* This strong alignment suggests that zero values likely represent missing or uncollected data, not true absence of debris or duff

A graph of a graph

AI-generated content may be incorrect.

**fuel\_moisture\_class**

* No missing values
* Values range from 1 to 4
* Indicates categorical classification of moisture levels

**fm100\_value / fm1000\_value**

* No missing values
* Ranges appear appropriate for dead fuel moisture percentages

**Fire Source Information**

**burn\_source**

* No missing values
* Four categorical classifications from 1-4
* Source 1 has the highest fuel and emissions — indicative of uncontrolled natural wildfires.
* Source 2 shows much lower emissions and fuel, suggesting controlled prescribed burns.
* Source 3 has similar emissions to 1 but fewer records — possibly intense but rare events (e.g., accidental or structural fires).
* Source 4 appears more moderate — may include miscellaneous or unknown origin

A screenshot of a computer

AI-generated content may be incorrect.

**burnday\_source**

* No missing values: Every record has a value, indicating this metric is consistently reported.
* Min/Max: Values range from 12 to 81.
* Mean/Std: The average is 25.9 with a standard deviation of 25.8, suggesting considerable variation.
* Unique values: Only 5 distinct values are used across the dataset.

**BSEV / BSEV\_flag**

* BSEV (Burn Severity): Indicates the intensity or ecological impact of the fire on vegetation and soil.
* BSEV\_flag: A binary indicator showing whether the burn severity value is valid or available.
* 67,547 missing values in BSEV
* 155,236 missing values in BSEV\_flag
* These variables are likely only recorded for observations where a fire event actually occurred, making them relevant only to a subset of the data where burning was detected.
* Although summary statistics show higher missing values for BSEV\_flag, further validation reveals that BSEV and BSEV\_flag are always either both present or both null. This discrepancy may stem from how missing values are computed—e.g., BSEV\_flag might contain placeholder or non-standard entries (like empty strings or 'NA') that inflate the apparent count of missing values in summary views but aren't treated as nulls in Spark's actual filtering logic.

**Date**

**fire\_date**

* One missing value
* Dates range from January 1 to December 31, 2004

**Weather Conditions**

**bi\_value**

* No missing values
* Maximum is 120
* burn index

**pr\_value (Precipitation)**

* 192,508 zero values
* Values are right-skewed with most near zero
* Maximum value is 124.6 in mm
* Indicates most fire records occurred during dry conditions

**rmax\_value / rmin\_value**

* No missing or zero values
* Relative humidity values are within normal expected ranges
* Maximum rmax is 100 and rmin is 94.5, showing consistency

**sph\_value (Specific Humidity)**

* No missing values
* Mean: 0.006975 g/kg
* Right-skewed, low absolute values
* Reflects dry atmospheric conditions relevant to fire spread

**srad\_value (Solar Radiation)**

* No missing values
* Mean: 240.6 W/m², Max: 428.8
* High values support rapid vegetation drying and fire ignition

**tmmn\_value / tmmx\_value (Min/Max Temperature)**

* No missing values
* Range: 247–321 K (~−26°C to 48°C)
* Highs and lows critical for fire risk modeling

**th\_value (Wind Direction)**

* 238 missing values
* Covers 0–359° with broad spread

**vpd\_value (Vapor Pressure Deficit)**

* 88 missing values
* Mean: 1.3 kPa
* Higher values → drier conditions → increased fire risk

**vs\_value (Wind Speed)**

* No missing values
* Mean: 4.04 m/s, Max: 15 m/s
* Higher speeds can drastically increase fire spread