Wildfire Detection and Analysis

using NASA FIRMS Data

Manal BABKHOUTI

July 16, 2025

Contents

1 Introduction

1.1 Context and Motivation

The increasing frequency and severity of wildfires over recent years have raised global concerns. These events, often intensified by climate change, result in the destruction of ecosystems, significant greenhouse gas emissions, and threats to human lives and infrastructure. To address these challenges, there is a growing need for tools that can help us analyze, monitor, and understand fire activity at scale.

This project focuses on exploring real satellite data provided by NASA's FIRMS (Fire Information for Resource Management System), specifically using the MODIS C6.1 fire archive dataset. This dataset contains millions of fire detection records collected via satellite instruments such as MODIS aboard the Aqua and Terra satellites. Each row in the dataset represents a detected fire event with attributes such as location (latitude, longitude), detection time, satellite source, brightness, confidence score, and fire radiative power (FRP).

By applying a data science approach, this project aims to go beyond static reports and try to extract patterns, trends, and actionable insights using tools such as data visualization, clustering algorithms, and classification models. The motivation is both scientific and environmental: using data to better understand where, when, and how fire events occur and how that knowledge could eventually support smarter environmental management.

1.2 Objectives

The main objective of this project was to analyze wildfire detection data from NASA's FIRMS platform using modern data science techniques. The primary aim was to investigate **when**, **where**, and **how** fire events occurred globally throughout the year 2024, by leveraging the MODIS satellite dataset. This involved identifying spatial and temporal patterns, analyzing the intensity and confidence of fire detections, and evaluating the feasibility of **predicting fire confidence levels** using satellite-captured variables.

Additionally, the project aimed to explore **unsupervised learning techniques**—including clustering and dimensionality reduction—to determine whether fire events could be grouped into coherent categories without using labeled data. This dual approach, combining supervised and unsupervised analysis, provided a more holistic understanding of global fire dynamics.

2 Data Understanding

2.1 Dataset Description

The dataset we worked with is titled, and it contains over 4.8 million fire detection records from the year 2024. Each row corresponds to a thermal anomaly picked up by NASA's MODIS instrument, which is mounted on the Terra and Aqua satellites. The data is extremely detailed, covering not just the spatial location and time of detection, but also characteristics like brightness, confidence level, fire intensity, and whether the fire was observed during the day or night.

Below is a preview of the first few rows of the dataset as loaded in the Jupyter Notebook:

Shape: (4866204, 15)

	latitude	longitude	brightness	scan	track	acq_date	acq_time	satellite	instrument	confidence	version	bright_t31	frp	daynight	type
0	48.4610	38.7808	311.1	1.9	1.3	2024-01-01	35	Aqua	MODIS	82	61.03	271.2	40.5	N	2
1	-15.2828	132.3189	377.9	1.6	1.2	2024-01-01	39	Terra	MODIS	94	61.03	295.9	258.2	D	0
2	-15.0487	132.6429	327.9	1.5	1.2	2024-01-01	39	Terra	MODIS	69	61.03	294.0	26.7	D	0
3	-15.2938	132.3170	324.8	1.6	1.2	2024-01-01	39	Terra	MODIS	51	61.03	291.5	16.3	D	0
4	-15.0463	132.6290	330.5	1.5	1.2	2024-01-01	39	Terra	MODIS	77	61.03	294.8	32.7	D	0

Figure 1: Sample of 2024 MODIS fire data with key variables

We can already see key variables like latitude, longitude, brightness, acq_date, satellite, and confidence. These are the backbone of the entire analysis especially confidence, which we later use to classify and evaluate fire severity.

To understand the overall structure of the dataset, we checked the number of entries, the types of data, and the presence of missing values:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4866204 entries, 0 to 4866203
Data columns (total 15 columns):
 #   Column      Dtype  
---  ------      -----  
 0   latitude    float64
 1   longitude   float64
 2   brightness  float64
 3   scan        float64
 4   track       float64
 5   acq_date    object 
 6   acq_time    int64  
 7   satellite   object 
 8   instrument  object 
 9   confidence  int64  
 10  version     float64
 11  bright_t31  float64
 12  frp         float64
 13  daynight    object 
 14  type        int64  
dtypes: float64(8), int64(3), object(4)
memory usage: 556.9+ MB

Missing values per column:
 latitude      0
longitude     0
brightness    0
scan          0
track         0
acq_date      0
acq_time      0
satellite     0
instrument    0
confidence    0
version       0
bright_t31    0
frp           0
daynight      0
type          0
dtype: int64
```

Figure 2: Overview of dataset structure, data types and missing values

The dataset includes exactly 4,866,204 rows and 15 columns. What's really convenient is that there are no missing values in any of the fields which makes the preprocessing stage a lot smoother. Most columns are either floats (like brightness, frp, or bright_t31)

or objects like dates or satellite names. No column looked suspicious or broken at this stage.

To get a better feel for the range of values in each numerical column, we also used .describe() on the dataset. Here are the summary statistics:

	latitude	longitude	brightness	scan	track	acq_time	confidence	version	bright_t31	frp	type
count	4.866204e+06	4.866204e+06	4.866204e+06	4.866204e+06	4.866204e+06	4.866204e+06	4.866204e+06	4866204.00	4.866204e+06	4.866204e+06	4.866204e+06
mean	4.551616e+00	1.235429e+01	3.256077e+02	1.567897e+00	1.198683e+00	1.161760e+03	6.708279e+01	61.03	3.005342e+02	4.240300e+01	8.928006e-02
std	2.353193e+01	6.567742e+01	1.747237e+01	7.724767e-01	2.380633e-01	5.286347e+02	2.105241e+01	0.00	8.369543e+00	1.090580e+02	4.248024e-01
min	-7.498340e+01	-1.788230e+02	3.000000e+02	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	61.03	2.644000e+02	0.000000e+00	0.000000e+00
25%	-1.277660e+01	-4.923720e+01	3.152000e+02	1.100000e+00	1.000000e+00	7.530000e+02	5.400000e+01	61.03	2.954000e+02	1.080000e+01	0.000000e+00
50%	-3.755600e+00	2.113960e+01	3.224000e+02	1.200000e+00	1.100000e+00	1.232000e+03	6.900000e+01	61.03	3.009000e+02	1.930000e+01	0.000000e+00
75%	1.341220e+01	3.560200e+01	3.316000e+02	1.800000e+00	1.300000e+00	1.430000e+03	8.200000e+01	61.03	3.059000e+02	3.950000e+01	0.000000e+00
max	8.562080e+01	1.796477e+02	5.100000e+02	4.800000e+00	2.000000e+00	2.359000e+03	1.000000e+02	61.03	4.001000e+02	1.349030e+04	3.000000e+00

Figure 3: Summary statistics of numerical variables

From this, we can tell a few interesting things. The brightness values range from around 300 to over 500 Kelvin, with an average close to 325 K. The frp, which represents the Fire Radiative Power (basically the intensity of the fire), has some extremely high values — up to over 13,000 — but the median is below 2, which tells us that those intense cases are rare and probably outliers. The confidence score ranges from 0 to 100, with most values sitting above 80, meaning the detections are generally reliable.

Finally, we looked at the unique values for some categorical features like satellite, instrument, day/night, and fire type:

```
Satellites: ['Aqua' 'Terra']
Instruments: ['MODIS']
Day/Night: ['N' 'D']
Fire Types: [2 0 3 1]
```

Figure 4: Unique values in categorical features

The fires were all detected by MODIS — no surprise there, since it's the instrument used in this MODIS-specific archive. Both Terra and Aqua satellites contribute to the data, and detections happen during both day and night. The type column includes values like 0, 1, 2, and 3, but most detections seem to be of type 0, which corresponds to the standard MODIS fire pixels.

Overall, the dataset is rich, clean, and gives us all the ingredients we need to dive into a deeper spatio-temporal and intensity-based analysis of global fire activity.

2.2 Source and Features

The dataset was sourced from NASA's FIRMS platform, specifically the MODIS Collection 6.1 archive for 2024. It provides fire detection data captured by thermal sensors onboard the Terra and Aqua satellites.

The key features used in this project include spatial coordinates (latitude, longitude), timestamps (acq_date, acq_time), and detection-specific metrics such as brightness, frp (fire radiative power), and confidence. Additional fields like satellite, instrument, daynight, and type offer contextual information that helps interpret variations in detection patterns.

These variables were selected and used directly in the analysis to explore temporal and geographic distributions, compare satellite performance, and extract meaningful patterns through clustering and classification.

3 Exploratory Data Analysis

3.1 Temporal Analysis

To understand how fire activity evolved over time in 2024, I started by plotting the number of fire detections per day. This revealed clear fluctuations in fire volume throughout the year, with noticeable peaks and quiet periods that reflect real-world seasonal dynamics.

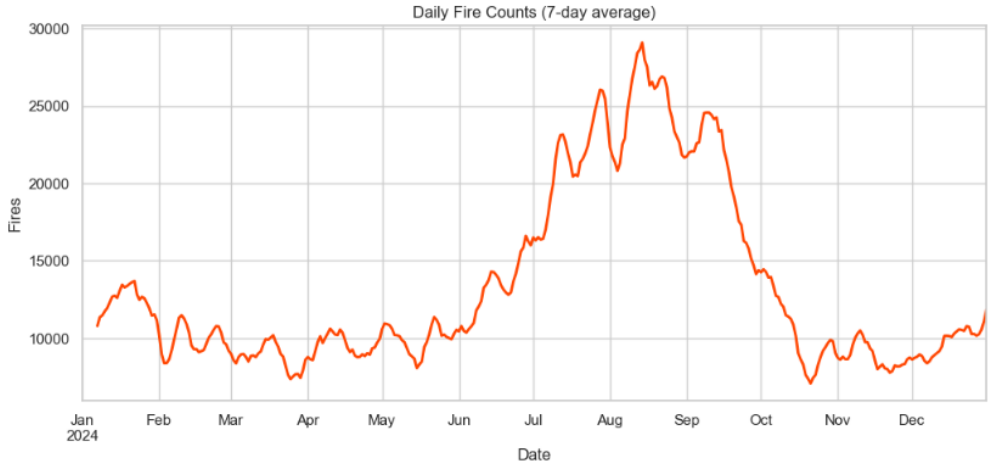

Figure 5: Daily fire counts across 2024

This figure shows daily variations in fire activity across 2024. We observe several intense waves of detections, especially during the middle of the year, which likely correspond to dry seasons in fire-prone regions.

To get a more global overview, I aggregated fire events by month. This helped visualize seasonal trends more clearly and confirmed that fire detections were far from evenly distributed throughout the year. Some months showed a sharp increase in activity, while others remained relatively calm.

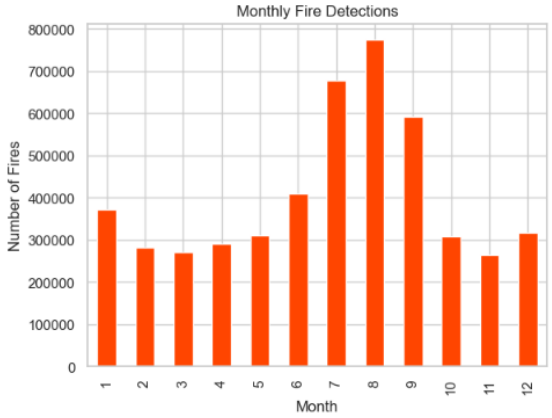

Figure 6: Monthly fire counts in 2024

This bar chart shows that fire activity peaks during certain months — often mid-year — which aligns with known dry periods in several regions monitored by MODIS.

Depending on the focus, I also looked into fire detection by time of day, using the hour extracted from the acquisition time. Although this doesn't reflect the real start of a fire, it helps understand satellite coverage and detection timing. Differences between day and night detections, as well as satellite pass timing, became more apparent through this lens.

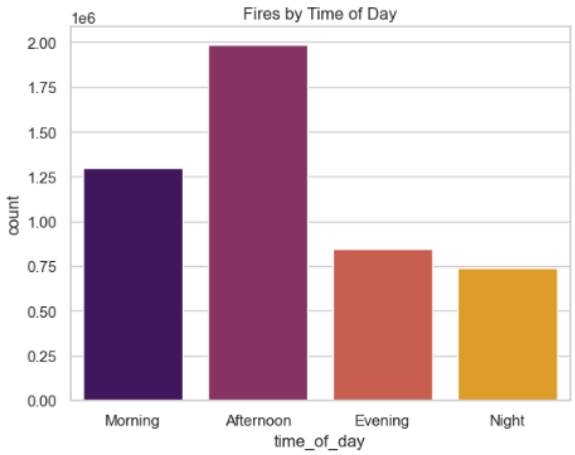

Figure 7: Fire detections by time of day

This histogram reveals the time-of-day distribution of fire detections. Satellite overpass times influence the shape of this curve, with certain hours seeing systematically more detections.

This temporal perspective gave a solid foundation for the rest of the analysis — especially when comparing regions, satellite instruments, or confidence levels over time. In addition to temporal patterns across the year, I also compared the volume of detections made during the day versus at night. The dataset includes a daynight column that distinguishes whether the fire was observed under daylight (D) or nighttime (N) conditions. When grouped by month, a clear difference emerges: daytime detections consistently outnumber nighttime ones by a large margin. This is expected, as satellite fire detection is typically more effective in daylight, and some thermal anomalies may go unnoticed or be less frequent at night due to atmospheric and visibility constraints. Interestingly, while both trends follow a similar seasonal shape, the amplitude is significantly greater during the day, especially around the peak months of July and August.

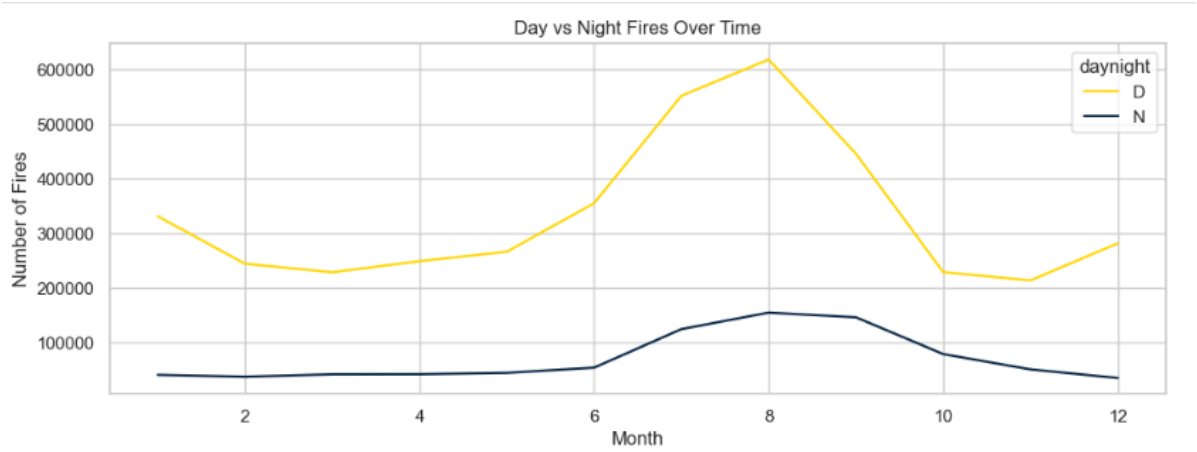

Figure 8: Monthly fire trends by day and night

This figure shows that the seasonal increase in fire activity is much more pronounced during the day than at night, particularly mid-year. To complement the monthly comparison of day and night activity, I also looked at the overall distribution of detections across the two categories.

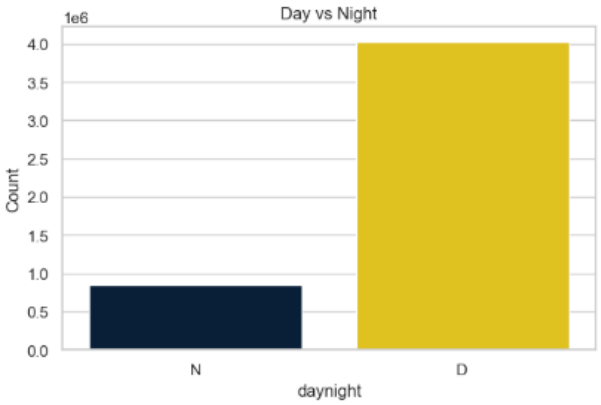

Figure 9: Total fire detections by day and night

The majority of detections were made during daylight hours, with daytime fires accounting for over four million records compared to less than one million at night. This reinforces the idea that satellites are more effective at capturing surface thermal anomalies during the day, likely due to better signal clarity and fewer atmospheric obstructions.

To visualize this disparity globally, I mapped the fire locations by time of detection.

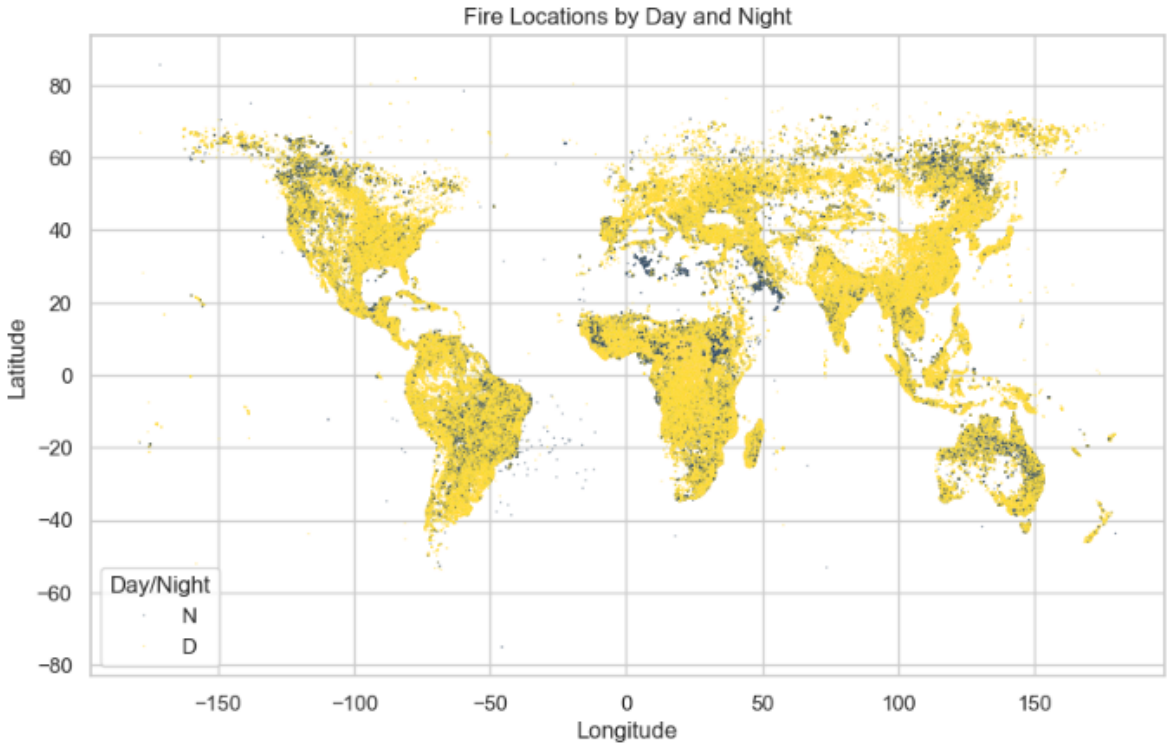

Figure 10: Global fire detections by day and night

The spatial spread is largely consistent with the earlier global map, but this visualization adds an extra layer: it shows that night detections are more sporadic and scattered, whereas daytime coverage is denser and more evenly distributed. Some regions, such as parts of central Africa and northern Australia, still show clusters of nighttime activity, which could reflect satellite overpass schedules or specific land and atmospheric conditions.

Lastly, I explored how fire intensity as measured by Fire Radiative Power (FRP) varies by time of day. I grouped the data by hour and computed the average FRP in each time slot.

The result was striking. While most of the day shows relatively stable FRP values, a sharp spike appears in the late evening, particularly around hour 22 (10 PM UTC). This could be linked to satellite detection angles or post-sunset thermal contrasts that make certain fires stand out more clearly. In any case, it highlights that not all hours are equal when it comes to capturing fire intensity, and this kind of variation should be kept in mind when comparing temporal trends.

3.2 Geospatial Analysis

To understand the spatial distribution of fire events, I started by plotting all fire detections on a latitude–longitude scatter plot. This gave an immediate visual sense of where fires tend to occur across the globe.

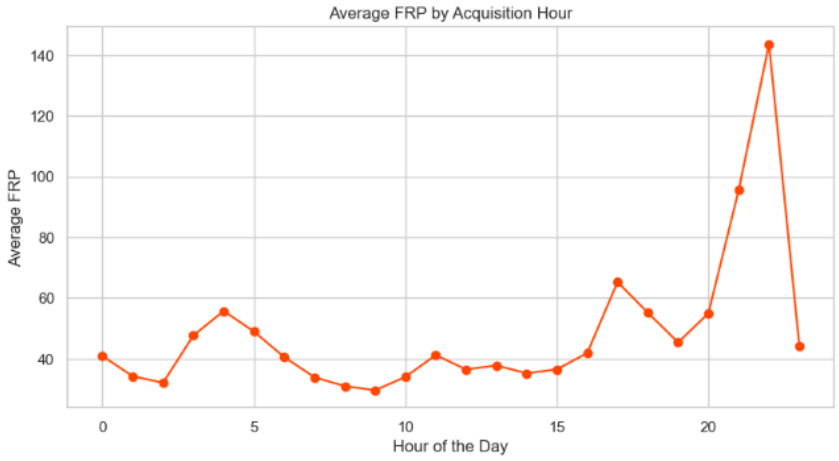

Figure 11: Average fire radiative power (FRP) by hour of day

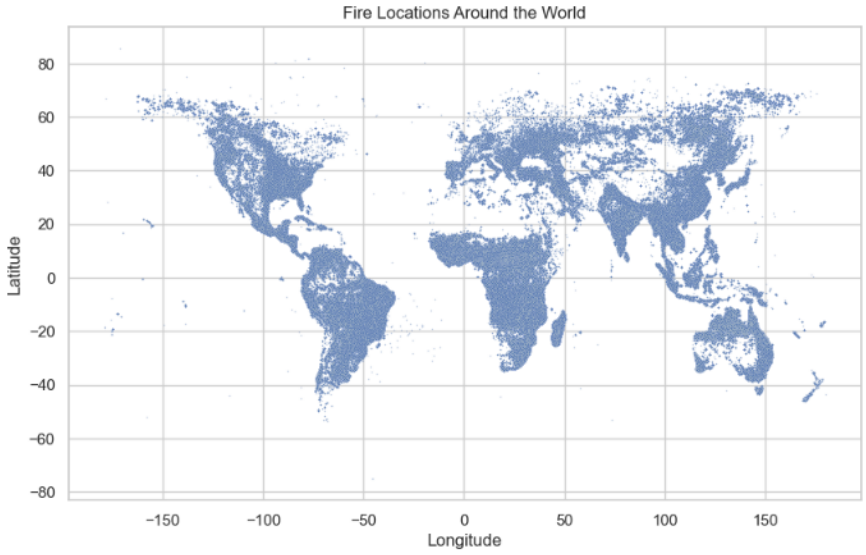

Figure 12: Global fire locations based on MODIS detections

This scatterplot reveals that fire activity is concentrated in specific regions — particularly across central Africa, parts of South America, Southeast Asia, and northern Australia. These areas align with known tropical and subtropical zones where vegetation and climate conditions support frequent burning.

To go further, I plotted a density map using a hexbin representation. This made it easier to see which regions consistently experience high volumes of fire activity, without being overwhelmed by individual points.

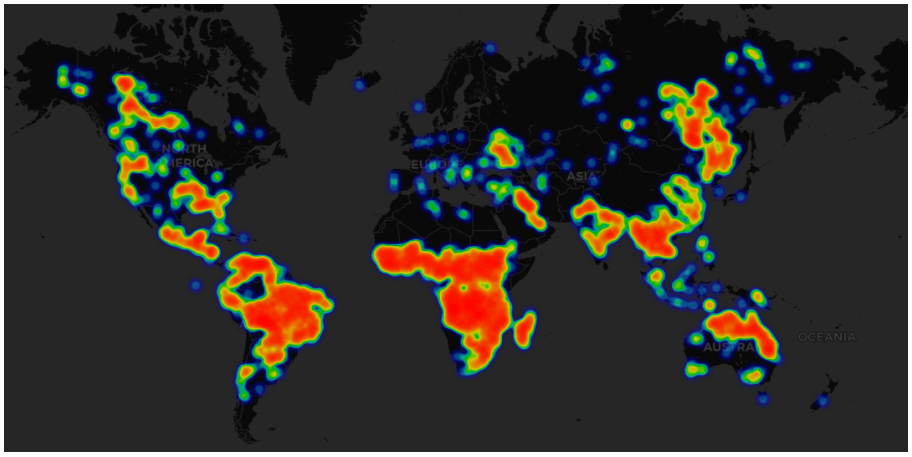

Figure 13: Fire density map highlighting global hotspots

The hexbin map highlights large-scale fire zones more clearly. Central Africa stands out with the highest density, especially in the equatorial belt. Fire clusters are also visible in Brazil, the Indian subcontinent, and portions of Indonesia.

Since the dataset spans the entire planet, but some zones are far more active than others, I also zoomed in on specific regions at different moments in the analysis — especially when trying to interpret temporal peaks or unusual fire patterns.

If needed, I filtered by confidence score to remove low-certainty detections and get a sharper view of confirmed fire zones. This helped distinguish between real activity and noise in sparsely populated areas.

Together, these visualizations formed a foundation for later clustering and classification work. They also helped confirm that fire activity is not randomly distributed but tied to climate, land use, and ecological zones.

Beyond the global overview, I wanted to understand how fire behavior varied across latitudes and continents — not just in terms of where fires happen, but also how frequent and how intense they are.

To do that, I grouped the data into latitude bands and calculated the number of fires in each.

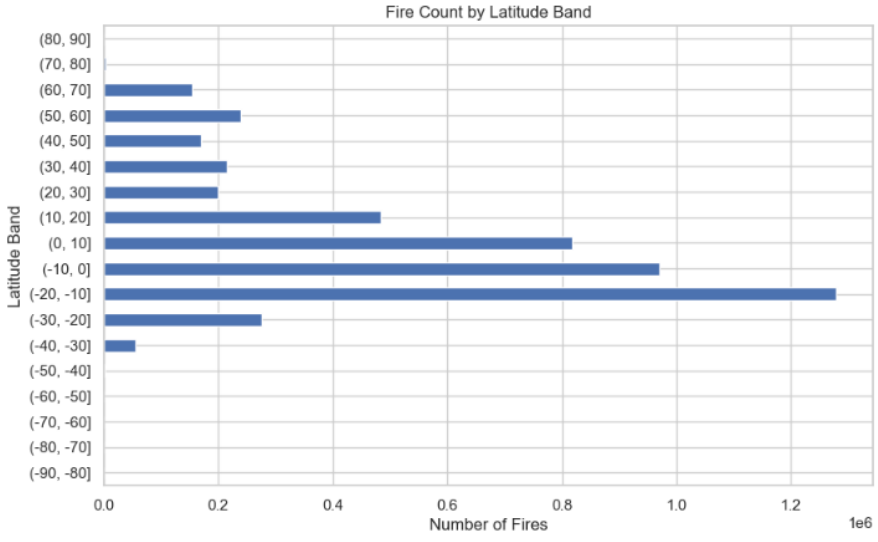

Figure 14: Fire count by latitude band

This clearly shows that most fire detections occur between -20° and +10° latitude — areas that include tropical ecosystems such as the Amazon rainforest, the Congo Basin, and parts of Southeast Asia. These regions combine high vegetation density with dry periods that favor frequent fire activity.

But raw fire counts only tell part of the story. I also looked at the average Fire Radiative Power (FRP) by latitude band, to get a sense of where fires burn most intensely.

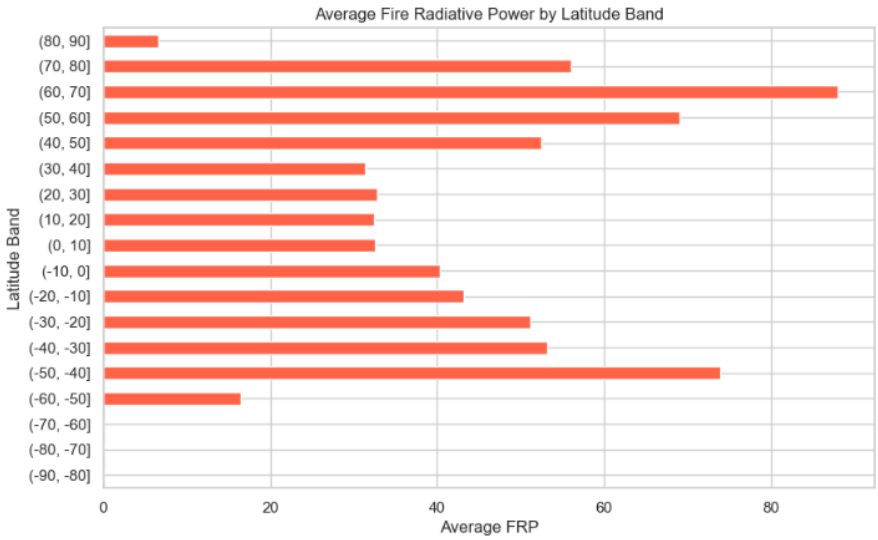

Figure 15: Average fire radiative power (FRP) by latitude band

Interestingly, some of the most powerful fires appear in the 40° to 70° bands in the northern hemisphere — despite those areas having lower overall fire counts. This suggests that while fires are less frequent there, the ones that do occur can release significantly more energy, potentially due to the nature of the vegetation or land use.

To explore this idea further, I analyzed FRP distributions by continent.

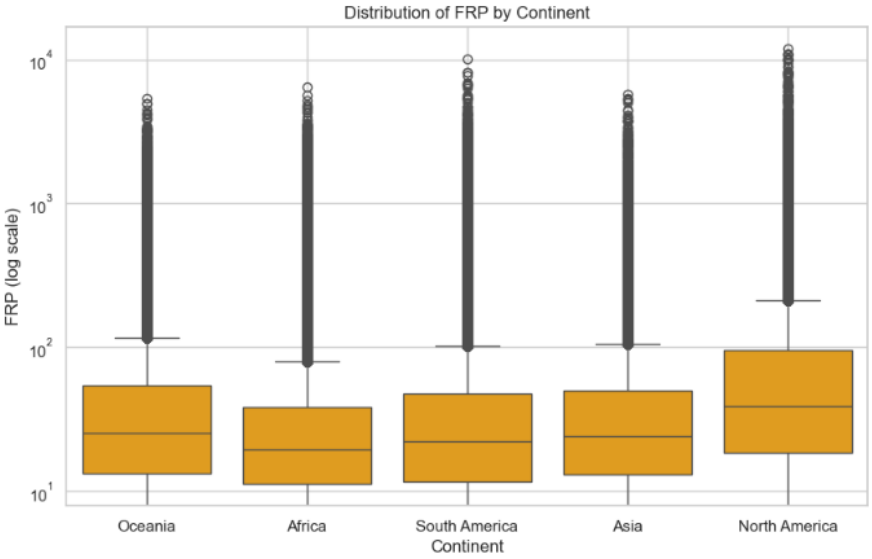

Figure 16: FRP distribution by continent

The boxplot shows that most fires are relatively low in intensity, but every continent

contains extreme outliers — some of them very high. South America and Oceania show particularly wide ranges, with many events crossing into the upper tail of the FRP scale.

Next, I focused on night fire activity. I calculated which countries had the highest share of fires detected at night rather than during the day.

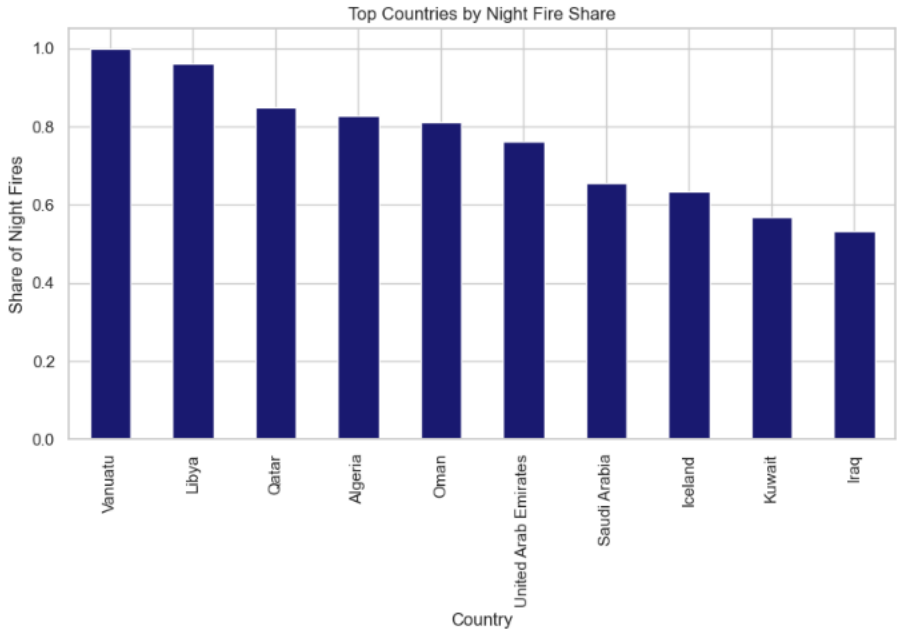

Figure 17: Top countries by night fire share

Countries like Vanuatu, Libya, and Qatar stood out for having mostly night detections. This could be linked to their geography, atmospheric conditions, or the timing of satellite overpasses — especially for smaller or drier regions.

Of course, when looking at total fire volume rather than night share, the picture changes.

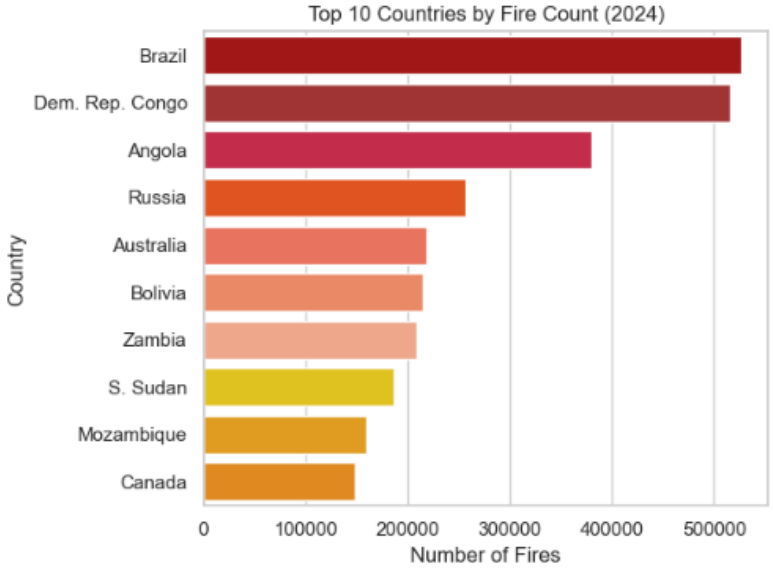

Figure 18: Top 10 countries by total fire count in 2024

Brazil, the Democratic Republic of Congo, and Angola came out on top, consistent with what we saw earlier in the global scatterplots and density maps. These countries represent major fire zones year-round.

To close this part of the analysis, I zoomed in on those three countries and mapped fire locations individually — again separating day and night detections.

Figure 19: Fire detections in Brazil by day and night

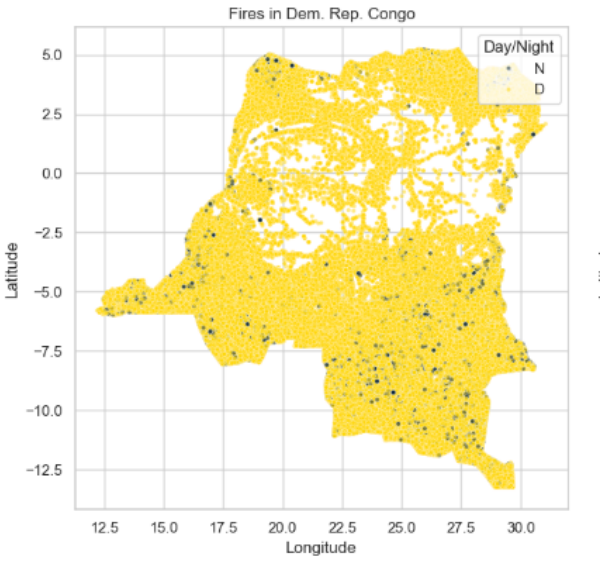

Figure 20: Fire detections in Congo by day and night

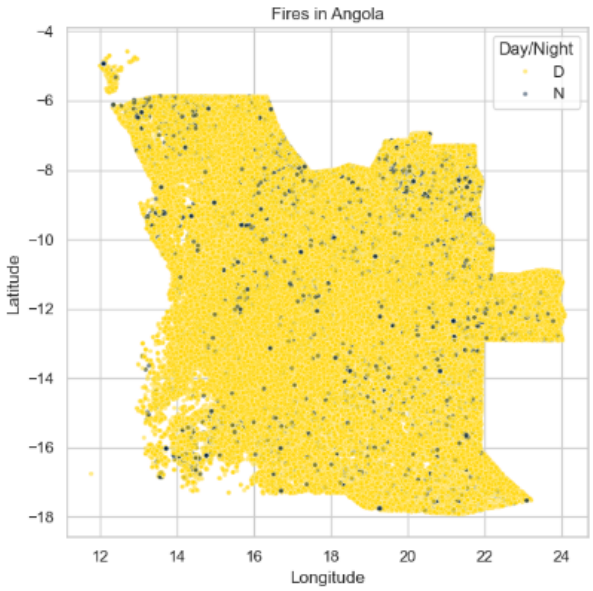

Figure 21: Fire detections in Angola by day and night

The detailed maps show how fire patterns shift within each country. Some areas burn more frequently at night, others have wide swaths of land covered by persistent daytime fires. These regional variations offer valuable context when thinking about fire risk and satellite-based monitoring strategies.

Given that Brazil ranks first in total fire detections, I explored its activity more closely. The monthly trend clearly shows a sharp rise starting in June, peaking dramatically in August, and then declining toward the end of the year. This seasonal pattern reflects the country's dry period, especially in regions like the Amazon and the Cerrado.

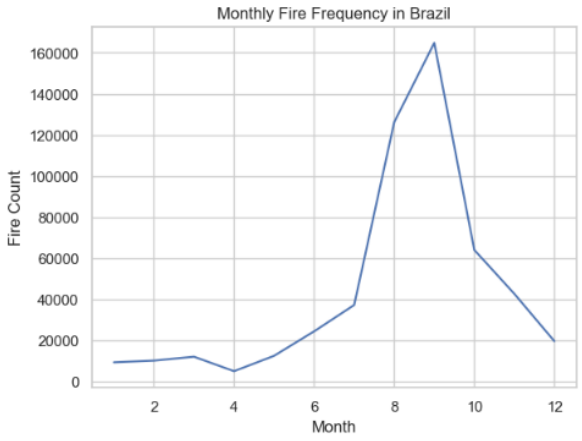

Figure 22: Monthly fire frequency in Brazil

To understand where exactly these fires occur, I generated a density heatmap of the detections within Brazilian territory. The map reveals strong clustering across the north and central-west, with concentrated activity in areas known for agricultural expansion and deforestation.

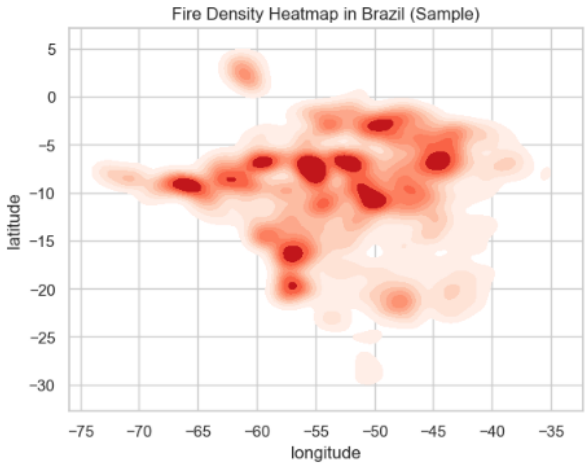

Figure 23: Fire density heatmap in Brazil

These patterns confirm Brazil's central role in global fire dynamics, both in terms of frequency and spatial impact.

3.3 Confidence and Satellite Comparison

To evaluate the reliability and characteristics of the recorded fire events, I analyzed two key columns: confidence and satellite.

The confidence score represents the certainty of the detection algorithm, ranging from 0 to 100.

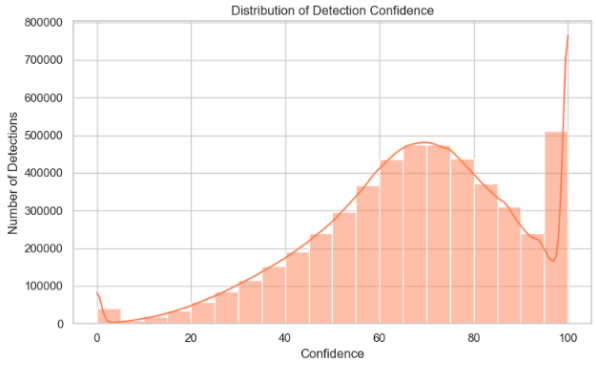

Figure 24: Distribution of detection Confidence

As seen in the histogram above, the distribution is skewed towards higher values, with a significant concentration near 100. This suggests that most detections are highly reliable. A few low-confidence readings are present, but they represent a small fraction and can be filtered later for more conservative analyses.

Next, I explored the satellite variable, which indicates which MODIS satellite — Aqua or Terra — made the detection.

The bar plot above shows that Aqua contributed a noticeably larger share of detections compared to Terra. This difference could be due to varying orbital schedules or sensor configurations, and may impact certain temporal analyses depending on which satellite passed over a region at a given time.

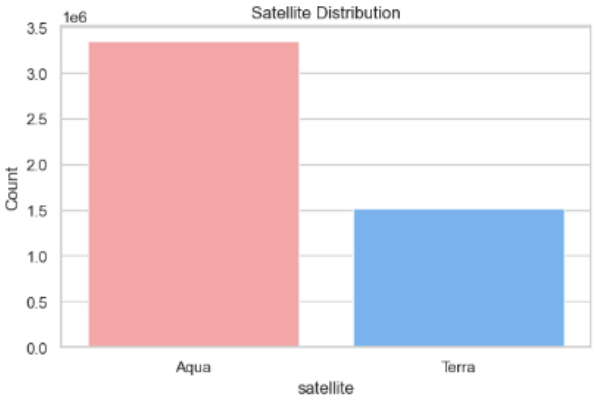

Figure 25: Satellite Distribution

These insights about confidence and satellite source are crucial for deciding whether to filter the data before modeling or visualization. For example, some later plots only include observations with confidence ≥ 50 or 75 to reduce noise from uncertain detections.

3.4 Feature Distributions and Correlations

To gain a deeper understanding of the numeric features present in the dataset, I plotted the distribution of each of them side by side. This provides a quick glance at the nature of the variables — whether they are skewed, multimodal, or uniformly spread.

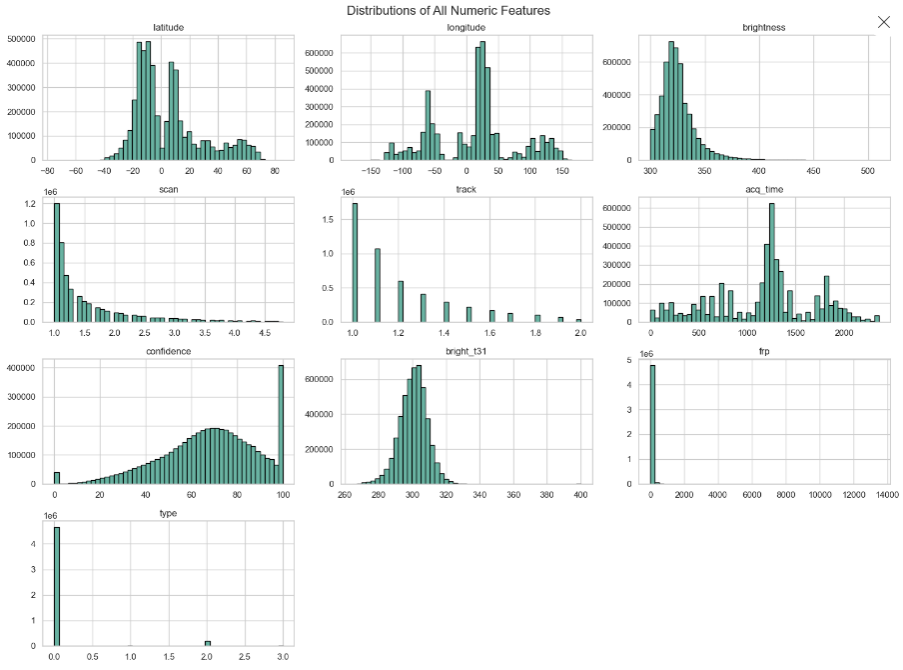

Figure 26: Distribution of Numeric features

From this overview, we can see that variables like brightness, frp, and confidence are strongly right-skewed, indicating the presence of extreme values or rare intense fire events. The bright_t31 feature, however, is much more bell-shaped and centered, while scan and track are highly concentrated at lower values, reflecting the satellite's detection resolution.

To assess the relationships between these variables, I also computed a correlation matrix. This helps identify potential redundancy, collinearity, or interesting associations among features.

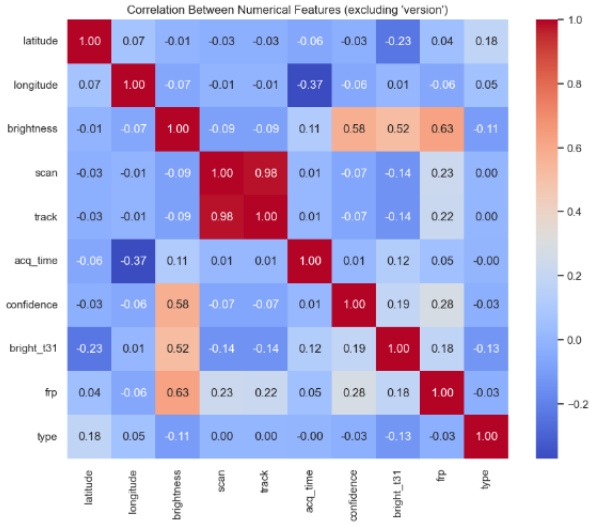

Figure 27: Correlation between Numeric features

Some strong positive correlations emerged — most notably between scan and track, which is expected since both relate to the satellite's spatial coverage. Also, brightness and frp are strongly correlated, as both are proxies for fire intensity. Meanwhile, confidence shows moderate correlations with both brightness and frp, suggesting it might partially reflect fire intensity. Latitude and longitude are largely uncorrelated with other features, reinforcing their role as geographic identifiers.

This statistical perspective helped decide which features were meaningful to retain, transform, or eventually feed into the machine learning models later in the pipeline.

4 Data Preprocessing

4.1 Cleaning and Preparation

The dataset was clean overall, with no missing values in any of the 15 columns. That allowed me to skip the usual imputation or removal steps and focus directly on preparing the structure of the data for analysis.

The first thing I did was convert the acq_date column into datetime format so I could later extract time-based features like the month or group by date. This also made plotting over time easier. The acq_time column, which was originally an integer in HHMM format, also needed to be padded and transformed. I used it to extract the hour of detection when I needed to study fire activity across different times of day.

```
df['month'] = pd.to_datetime(df['acq_date']).dt.month
```

Figure 28: Extracting month from acquisition date

```
df['acq_date'] = pd.to_datetime(df['acq_date'])
```

Figure 29: Converting acquisition date to datetime format

To prepare for visualizations and modeling, I checked for anomalies or extreme values. Some variables like frp had very high outliers (over 13,000), but I didn't remove them globally. Instead, I filtered or capped them locally during specific analyses where they skewed the visual scale.

No global filtering was done at this stage, but I did apply local conditions later, such as focusing only on detections with confidence ≥ 50 or working within a specific date range for clearer plots.

This early phase was about making sure the data was readable, structured, and consistent, not forcing unnecessary cleaning, but adjusting just enough to make the rest of the analysis efficient and reliable.

4.2 Feature Engineering

Beyond the raw variables provided in the dataset, I added a few new features to extract more meaningful insights during analysis.

The first was a month feature, derived from the converted date, which allowed me to group and visualize fire activity by season. This helped capture recurring patterns and made comparisons between months more intuitive during the temporal analysis.

In a similar way, I generated an hour column to isolate the time of day when fires were detected. This new feature became especially useful when comparing daytime and nighttime fire distributions, or when studying detection patterns across satellites.

Another feature I introduced was a simplified classification of fire confidence levels. Instead of relying on raw numerical values, I mapped the confidence score into three distinct categories — low, medium, and high — to support clearer labeling and visualization. This made it easier to compare behaviors between different fire intensities and was also used later in clustering and classification steps.

These engineered variables were not part of the original dataset but proved essential in framing the analysis in a way that was both interpretable and operational.

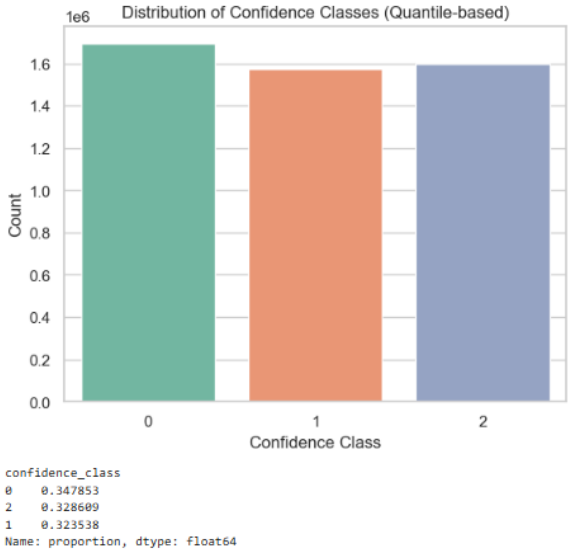

Figure 30: Distribution of fire confidence classes (quantile-based)

5 Machine Learning

5.1 Problem Formulation

The goal was to predict the confidence level associated with each fire detection. Originally, the confidence was a continuous value ranging from 0 to 100, representing the certainty of the satellite's detection algorithm. To make the problem more interpretable and suitable for classification, the score was converted into three classes: low, medium, and high confidence.

Initially, I considered a regression task to predict the Fire Radiative Power (FRP), which reflects the intensity of the fire. However, this variable had extreme variance and a highly skewed distribution, with many outliers. The prediction attempts yielded poor results, and given time constraints, I decided to drop the FRP prediction and focus solely on the classification of confidence which was more robust, interpretable, and aligned better with the rest of the analysis.

5.2 Model Building

The modeling process began with a random sample of 100,000 rows drawn from the cleaned dataset. Categorical variables such as satellite, instrument, and daynight were one-hot encoded to ensure compatibility with scikit-learn models. Additional columns like lat_band and time_of_day, which were derived during earlier preprocessing steps, were carefully removed or encoded to avoid leakage or redundancy.

The input features (X) consisted of all relevant numeric and encoded variables, excluding the target variable confidence_class, which was used as the label (y). The dataset was then split into training and testing subsets using an 80/20 ratio, with stratification to preserve the class distribution across the three confidence levels.

The first model tested was a Random Forest classifier. Hyperparameters such as the number of trees (n_estimators=200), maximum tree depth (max_depth=30), and minimum samples per leaf (min_samples_leaf=4) were selected based on initial experimenta-

tion. The model was trained on the training data and evaluated on the test set using standard classification metrics.

Later, an XGBoost classifier was trained on the same dataset, followed by grid-based hyperparameter tuning. This included tuning parameters such as max_depth, learning_rate, n_estimators, and regularization terms (reg_alpha, reg_lambda). The tuned XGBoost model yielded slightly improved balance across the classes, particularly in terms of precision and recall for the minority class, and was ultimately chosen as the best-performing model.

This modeling pipeline was implemented entirely in Python using the scikit-learn and XGBoost libraries within a JupyterLab environment.

5.3 Evaluation and Results

Each model was evaluated using accuracy, precision, recall, and F1-score. I paid particular attention to class-wise performance, as the dataset was imbalanced, with the "high confidence" class being much more represented than the "low" or "medium" classes.

Across all models tested, the overall accuracy was consistently around 77%, but deeper insights came from the confusion matrices and per-class metrics.

The Random Forest classifier, even before tuning, achieved a respectable performance, especially for class 2 (high confidence), with an F1-score of 0.89. However, it struggled with class 1 (medium confidence), which had both lower recall and higher misclassification rates.

After hyperparameter tuning, the Random Forest showed no significant gain in performance accuracy remained at 77%, and misclassification of medium-confidence detections persisted. The confusion matrix showed persistent confusion between class 1 and both its neighbors.

The XGBoost classifier, on the other hand, showed slightly more balanced results across all classes. Before tuning, it already achieved a macro F1-score of 0.77, on par with the Random Forest. After tuning, the class-wise precision and recall improved particularly for class 1 making the model more reliable for all categories.

Below are the confusion matrix visualizations and classification reports for both models (before and after tuning):

✅ Classification Report:

	precision	recall	f1-score	support
0	0.76	0.71	0.73	6971
1	0.68	0.66	0.67	6473
2	0.85	0.93	0.89	6556
accuracy			0.77	20000
macro avg	0.76	0.77	0.76	20000
weighted avg	0.76	0.77	0.76	20000

✅ Confusion Matrix:

```
[[4952 1624  395]
[1526 4302  645]
[  46  445 6065]]
```

Figure 31: Random Forest – Before Tuning

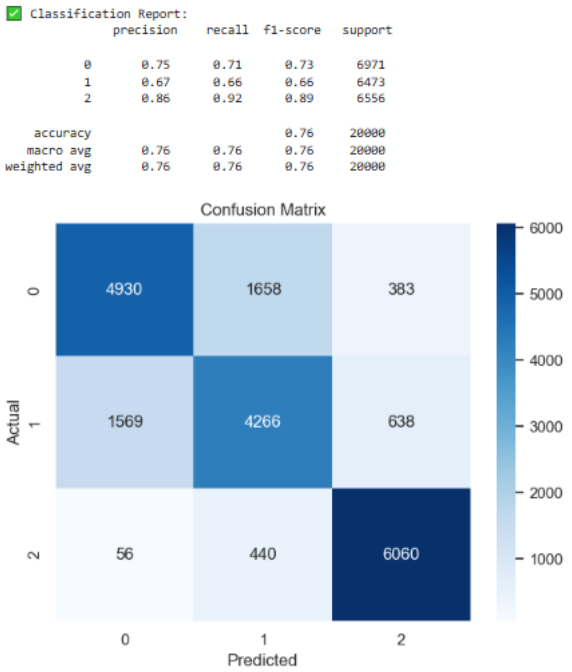

Figure 32: Random Forest – After Tuning

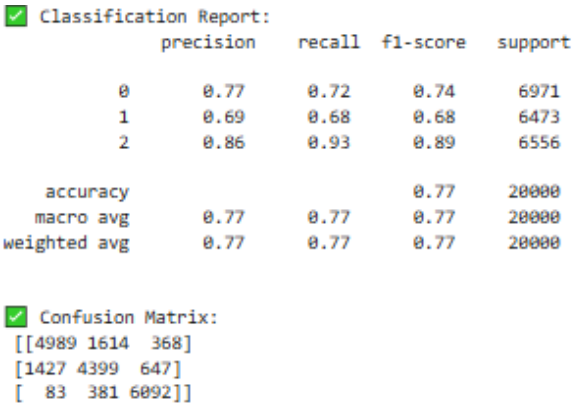

Figure 33: XGBoost – Before Tuning

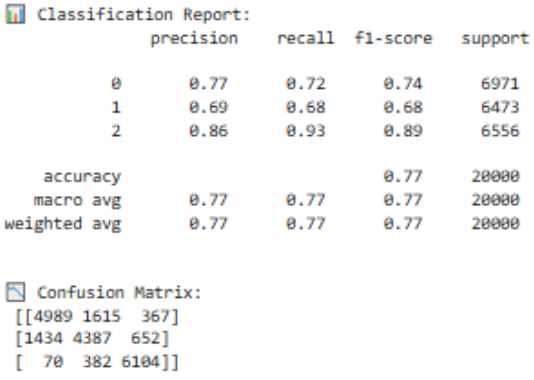

Figure 34: XGBoost – After TuningEnter Caption

6 Clustering and Pattern Discovery

6.1 Dimensionality Reduction

To prepare the data for clustering and visualization, I applied two popular dimensionality reduction techniques: Principal Component Analysis (PCA) and t-SNE. Both were used to project the high-dimensional feature space into two dimensions, allowing easier inspection of the structure and potential separability of the fire confidence classes. The

Figure 35: PCA projection of fire confidence classes (quantile-based)

PCA projection reveals a structured and directional spread of data points, with noticeable banding patterns likely influenced by geographic or temporal variables. Despite this, there is significant overlap between the three confidence classes, especially between classes 0 and 1.

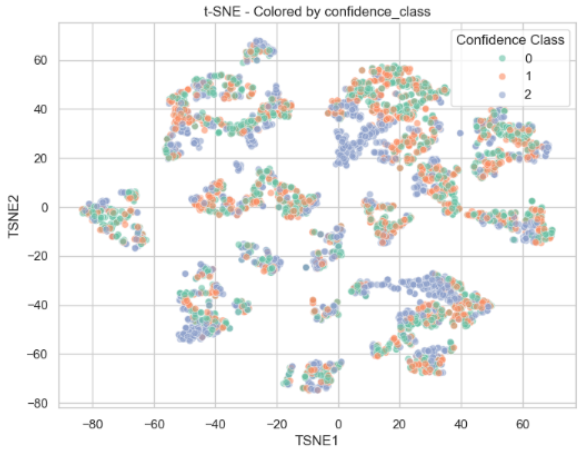

Figure 36: t-SNE projection of fire confidence classes (quantile-based)

The t-SNE visualization offers a more nonlinear view of the data. While the clusters appear more dispersed and discrete, the confidence classes are still intermixed, suggesting limited natural separability based on the available features. This confirms that although the classification model achieved good accuracy, the decision boundaries between classes are not trivially distinguishable in reduced space. That said, while some structure is noticeable, the PCA plot reveals significant overlap between classes, indicating that linear reduction may not clearly separate them.

6.2 Clustering Techniques

To investigate whether meaningful unsupervised groupings existed, I applied KMeans clustering on both the PCA-reduced and t-SNE-reduced datasets. The goal was to see whether clusters would align with the actual fire confidence classes — even without supervision.

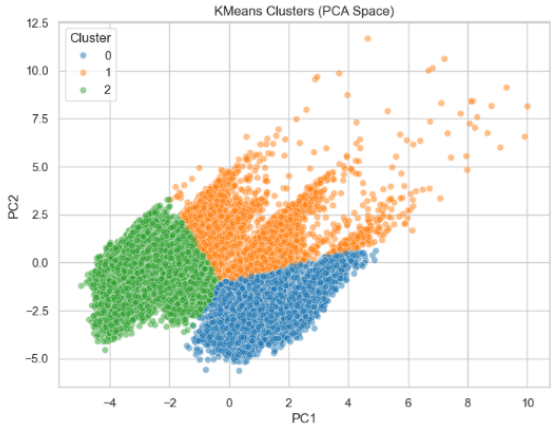

Figure 37: KMeans clustering on PCA-reduced space

The PCA space clusters showed some visually distinct groups. However, the Adjusted Rand Index (ARI) — a measure of agreement between clusters and true labels — was only 0.003, indicating negligible alignment.

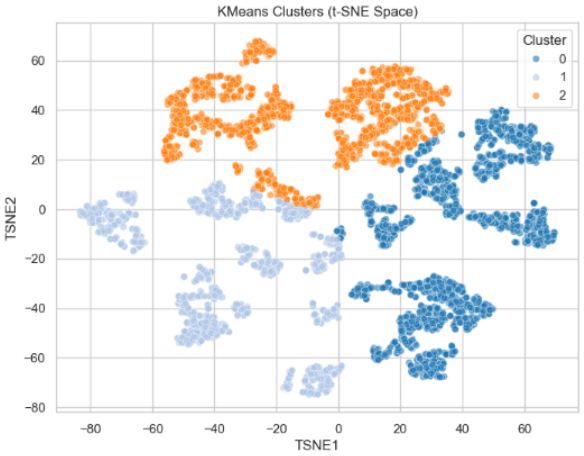

Figure 38: KMeans clustering on t-SNE-reduced space

Likewise, in the t-SNE space, the clusters were visually even more separated, yet the ARI dropped to 0.001, confirming that the separation was not predictive of the confidence labels.

6.3 Insights from Clusters

The result is conclusive: no cluster strongly matches any specific class. Each cluster contains a fairly balanced mix of low, medium, and high-confidence fires — further proving that the clustering structure does not correspond meaningfully to the classification task.

These findings support the conclusion that while clustering reveals interesting spatial or geometric groupings, it does not capture the underlying confidence level logic, which instead requires supervised learning and more targeted feature interpretation.

7 Final Insights and Conclusion

7.1 Key Findings

This project offered a comprehensive exploration of NASA's 2024 MODIS fire detection dataset, blending descriptive analysis, machine learning, and clustering techniques. Several key insights emerged:

- **Temporal trends** revealed strong seasonality in global fire activity, with distinct peaks mid-year. Daytime detections were significantly more frequent than nighttime ones, aligning with known satellite pass schedules and visibility constraints.
- **Geospatial analysis** showed that fire events were highly concentrated in specific regions, especially tropical zones like Central Africa, the Amazon Basin, and Southeast Asia. These findings confirm the role of climate, vegetation, and land use in shaping global fire patterns.
- **Confidence scores** were generally high, with most detections above 80. The classification model successfully leveraged satellite and environmental features to predict confidence categories with good accuracy (77%), especially for high-confidence fires.
- **Clustering analysis** demonstrated that although geometric patterns exist in the reduced feature spaces (via PCA and t-SNE), unsupervised groupings did not align with actual confidence labels. This reinforces the need for supervised approaches when predicting fire certainty.

7.2 Limitations

The analysis faced a number of limitations:

- **Imbalanced classes** in the confidence label made it harder to model the medium-confidence category accurately. The classifier sometimes confused it with the more dominant high-confidence group.
- **No ground truth for fires:** The confidence scores are estimates, not verified truth labels. This limits the interpretation of classification performance, as we do not know the true certainty of each fire event.
- **Limited interpretability of clusters:** The KMeans clusters did not correlate meaningfully with known classes or physical variables, suggesting that unsupervised learning might require more domain-specific features to be effective.
- **Dropped regression task:** An attempt was made to predict Fire Radiative Power (FRP), but results were poor due to its highly skewed distribution.

7.3 Future Work

To build upon this project and deepen the insights, several improvements and extensions can be explored:

- **Improve class balance** by resampling techniques (SMOTE, class weighting) or incorporating additional labeled data from other years or sensors.
- **Experiment with alternative models** like LightGBM, CatBoost, or even neural networks to see if they can better capture the nuanced patterns of medium-confidence fires.
- **Integrate geographic and meteorological context**, such as vegetation cover, precipitation, or land surface temperature, to enrich the predictive features.
- **Explore anomaly detection** to identify outlier fires or unexpected patterns beyond confidence classification.
- **Investigate temporal clustering** over time-series windows to capture evolving fire trends or hotspots.

Overall, this project demonstrated the value of data science in understanding and monitoring fire events at a global scale. It also opened the door to deeper predictive modeling that could eventually support real-world fire risk assessment and management tools. s d'amélioration, perspectives.