Final Project
Applied Competitive Lab in Data Science

Daniel Ohana
Avraham Meisel
Ben Moyal
Adi Shnaidman

April 25, 2024

1 Introduction

The United States, with its diverse landscapes, climates and population, is particularly susceptible to wildfires. These natural disasters pose significant threats to both human life, wildlife and the environment. Understanding the causes of these fires is crucial for effective prevention and management. The aim of this project is to model and predict the cause of wildfires in the US using a comprehensive dataset.

The project at hand involves the analysis of a spatial database of wildfires that occurred in the United States from 1992 to 2015. This data publication, whose source is the national Fire Program Analysis (FPA) system, includes wildfire records obtained from the reporting systems of federal, state, and local fire organizations.

The dataset, referred to as the Fire Program Analysis fire-occurrence database (FPA FOD), includes 1.88 million geo-referenced wildfire records, representing a total of 140 million acres burned over a 24-year period.

The dataset includes a variety of features, such as unique identifiers (FOD ID, FPA ID), source system type and name, reporting agency and unit identifiers and names, local fire report ID, local incident ID, fire code, fire name, incident identifiers and names from ICS-209 reports and MTBS perimeter dataset, complex name, fire year, discovery date, day of year and time, statistical cause code and description, containment date and much more.

The goal of the project is to develop a model that can accurately predict the cause of these fires, optimizing the weighted ROC AUC score for all classes in a One vs Rest approach. The project will involve data exploration, feature engineering, data pre-processing, among other steps, to achieve this goal. The insights gained from this project could significantly enhance fire management strategies and prevention measures.

2 Data Exploration and Feature Engineering

During the data exploration phase, various analyses were conducted to gain insights into the underlying patterns and relationships within the dataset. Here, some of the key visualizations that aided in this exploration and led to initial conclusions are presented:

2.1 Discovery Hour - Bar Plots

After binary features were engineered to indicate whether a fire occurred within specific time periods of the day, bar plots were created to visualize the distribution of fire start reasons across different discovery hours. These plots offered insights into whether certain times of the day are more susceptible to specific causes of fire incidents.

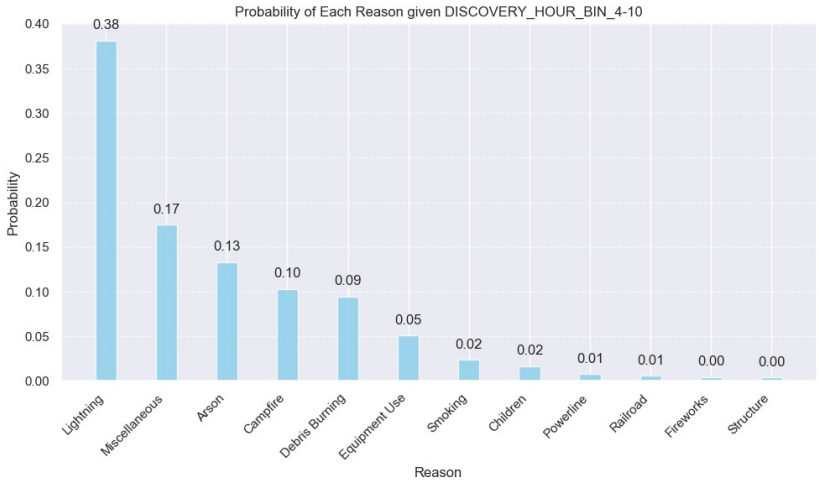

Figure 1: Distribution of fire start reasons between 04:00-10:00

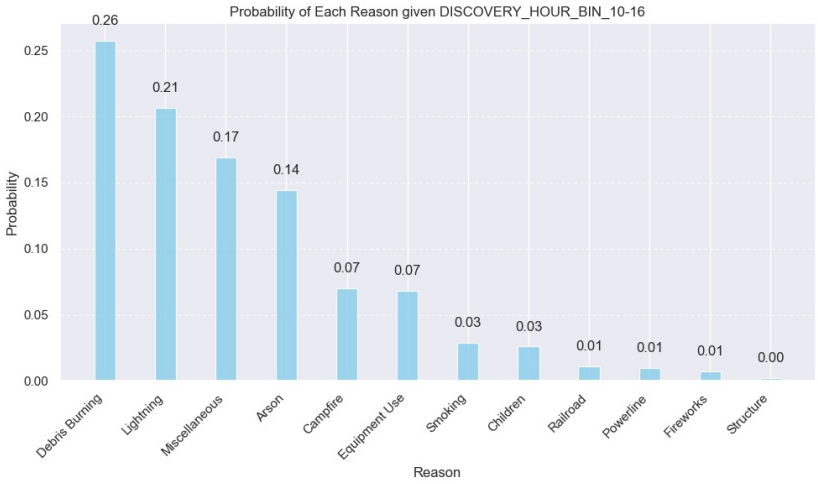

Figure 2: Distribution of fire start reasons between 10:00-16:00

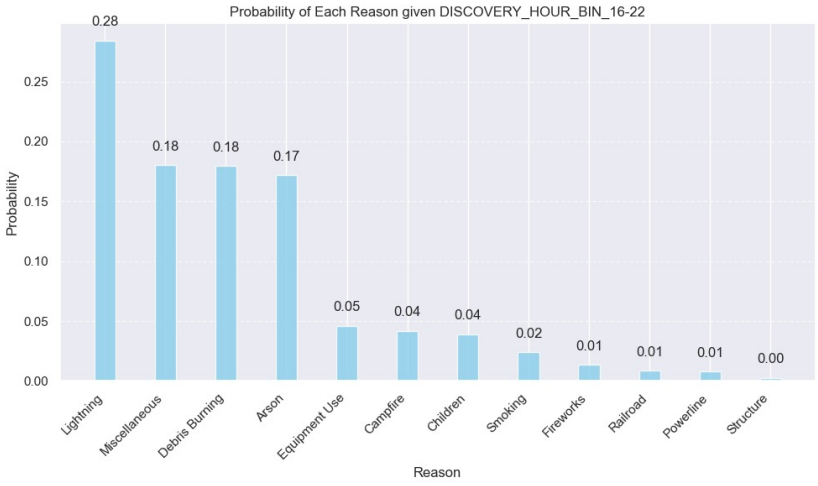

Figure 3: Distribution of fire start reasons between 16:00-22:00

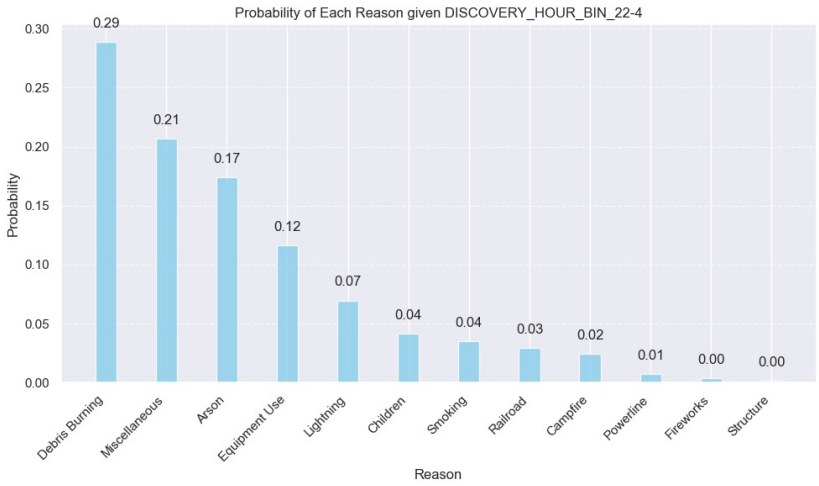

Figure 4: Distribution of fire start reasons between 22:00-04:00

As shown in the graphs above, certain causes are more probable at different times of the day. For instance, while a fire is much more likely to be caused by lightning than by any other cause throughout most hours of the day (sometimes more than twice as probable as any other cause), between 10:00 and 16:00, a fire is more likely to be caused by Debris Burning.

2.2 Fire Causes over Day of Year

Histograms depicting fire causes over the day of the year were created to observe the frequency of specific fire causes on different days of the year (DOY). These histograms offer a visual representation of the temporal distribution of fire incidents, enabling the identification of patterns and trends in fire causes throughout the year. By analyzing the frequency of fire causes across different DOY, insights into the seasonal variations and temporal dynamics of fire occurrence are sought.

It should be noted that the 4th of July falls on the 185th day of the year in non-leap years and the 186th day in leap years. The times around these days are prone to different types of fires, some of which will be discussed below.

Graphs with a distribution similar to uniform distribution, from which fewer inferences can be drawn, are included in the appendix (see Figures 5.3).

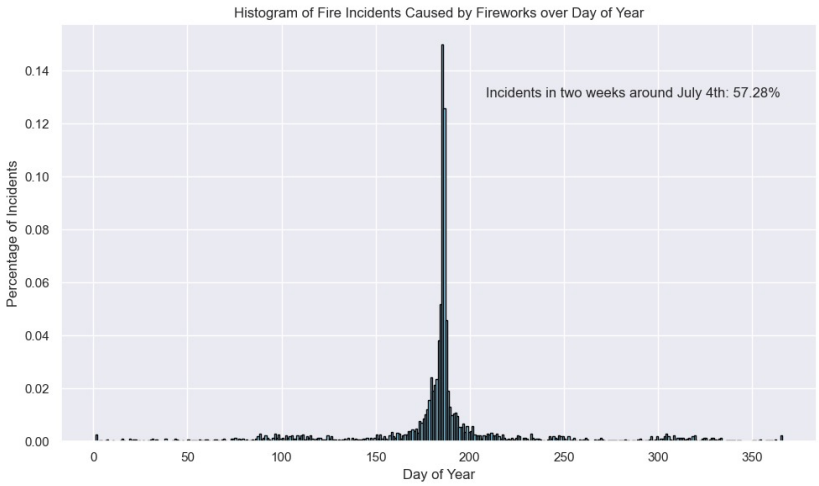

Figure 5: Fire incidents caused by fireworks over DOY

This graph illustrates the distribution of fire incidents caused by fireworks throughout the year. Each bar represents the percentage of incidents occurring on a specific day of the year.

As shown in the graph, and as we suspected, there are significantly more firework incidents occurring around July 4th. The label positioned in the top right corner highlights the proportion of incidents that took place within a two-week period around July 4th - 57%. This visualization offers insights into the temporal patterns of fire incidents related to fireworks, with a particular focus on their occurrence in proximity to Independence Day celebrations.

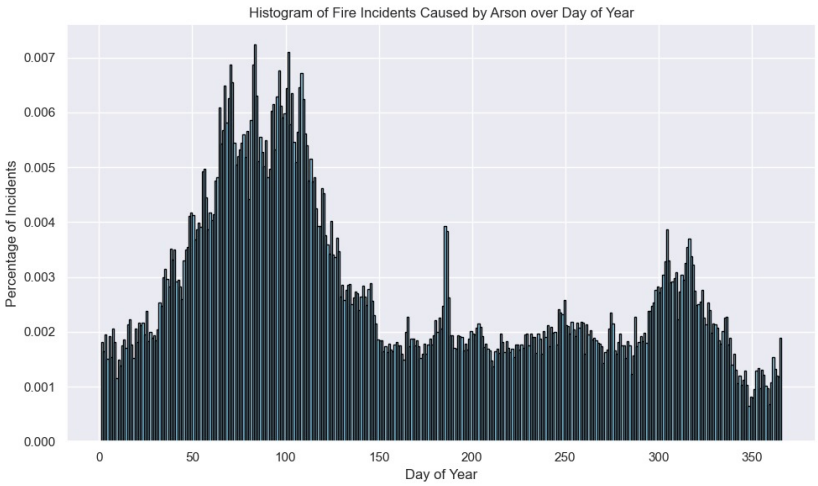

Figure 6: Fire incidents caused by arson over DOY

Arson exhibit higher frequency during the months of March and April compared to the rest of the year. This suggests a potential seasonal pattern in arson incidents.

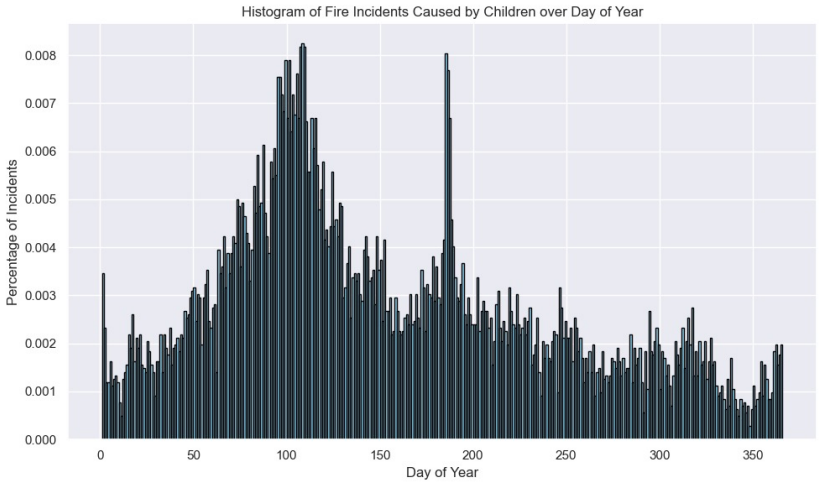

Figure 7: Fire incidents caused by children over DOY

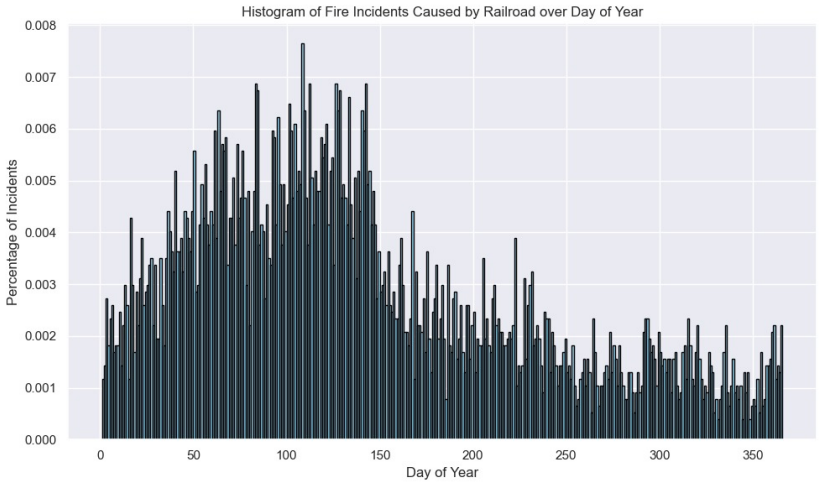

Figure 8: Fire incidents caused by railroad over DOY

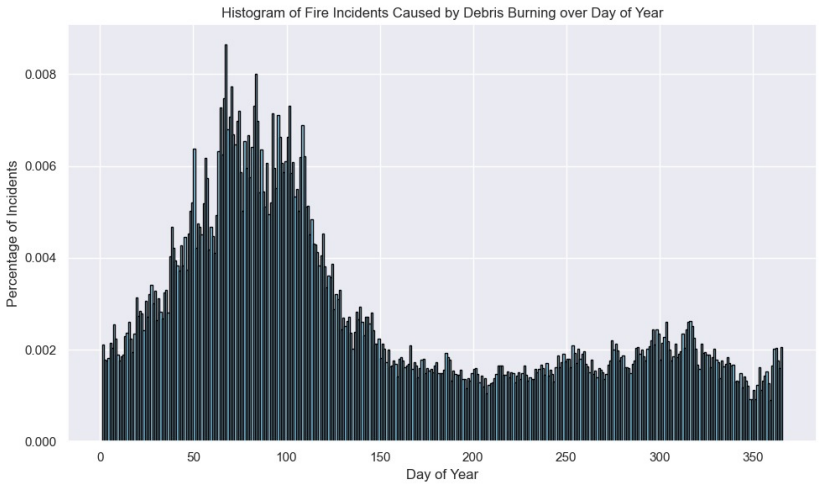

Figure 9: Fire incidents caused by debris burning over DOY

Debris Burning, Children, and Railroad: Similar to arson, incidents related to debris burning, children, and railroad occur more frequently in the months of March and April. This trend aligns with the increased activity in outdoor settings during the spring months.

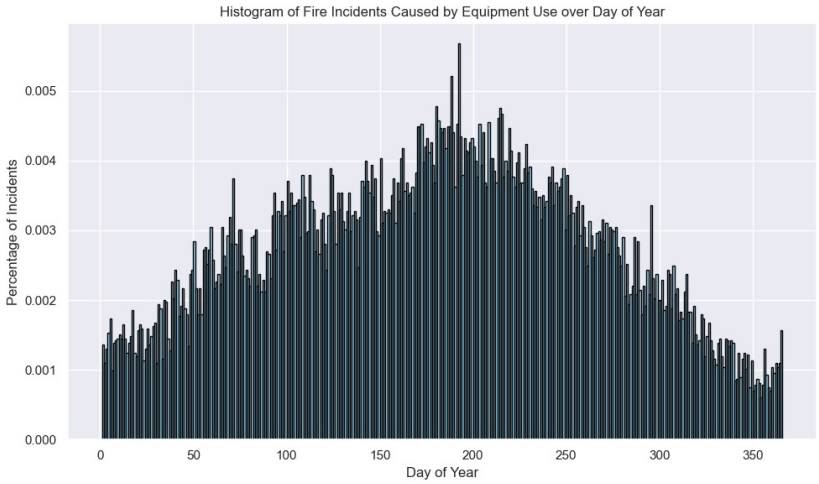

Figure 10: Fire incidents caused by equipment use over DOY

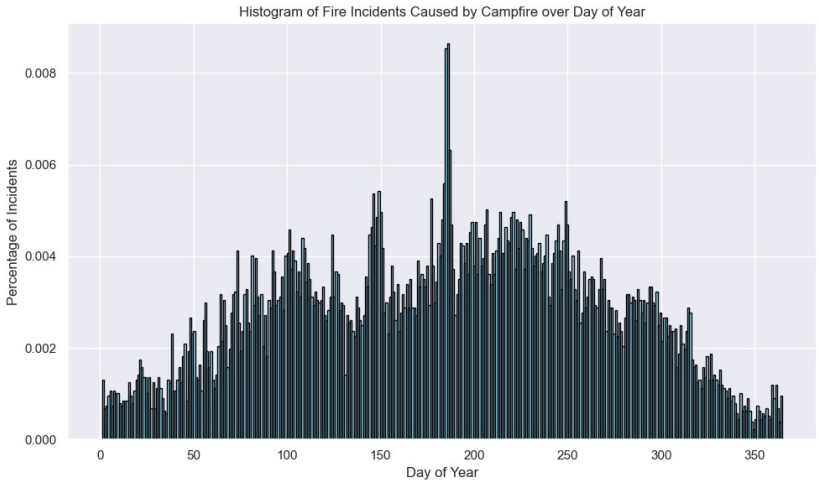

Figure 11: Fire incidents caused by campfire over DOY

Equipment Use and campfire: Fire incidents caused by equipment use and camping show a slight increase during the mid-year months compared to the beginning and end of the year. This suggests a potential seasonal variation in campfire and the use of equipment that may contribute to fire incidents.

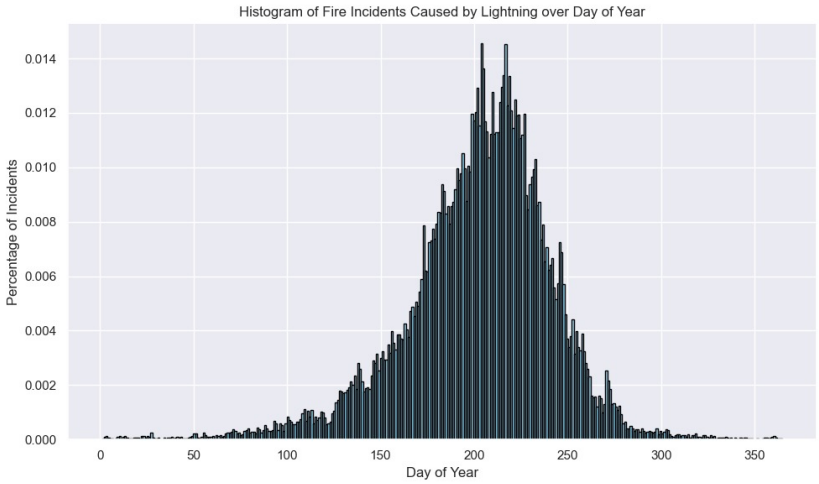

Figure 12: Fire incidents caused by lightning over DOY

Lightning: Most fires caused by lightning occur between May and September. This finding indicates a peak in lightning-related fire incidents during the warmer months, which is consistent with typical weather patterns associated with thunderstorms.

2.3 Cause Distribution of Fires Over Time

Density graphs have proven to be invaluable tools in understanding the distribution of fire incident causes across various time periods.

The insights gained from density graphs depicting fire incident causes categorized by discovery hour, discovery month, hours to contain the fire, and the year of fire occurrence were explored. These visualizations offer a comprehensive view of how fire incident causes vary over different time frames, providing valuable insights into temporal patterns and potential contributing factors.

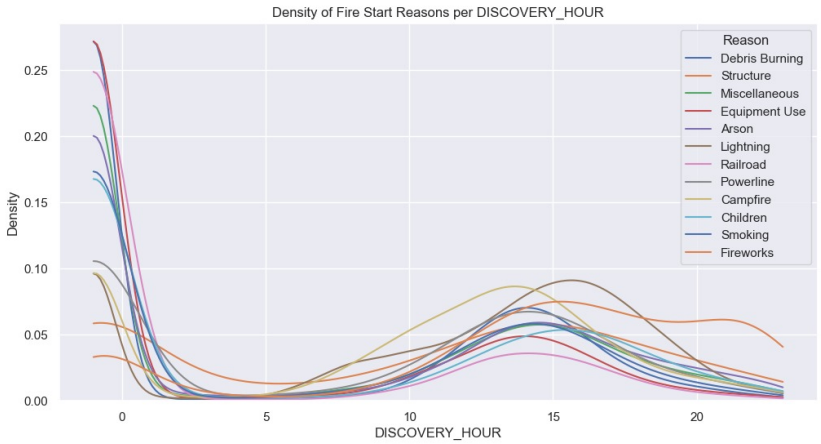

Figure 13: Density of causes per discovery hour

In general, the density of causes based on the hour at which the fire was discovered exhibits some similarities. However, subtle yet potentially meaningful differences were observed. These nuances could hold importance for predictive models. Therefore, this feature was incorporated into the analysis.

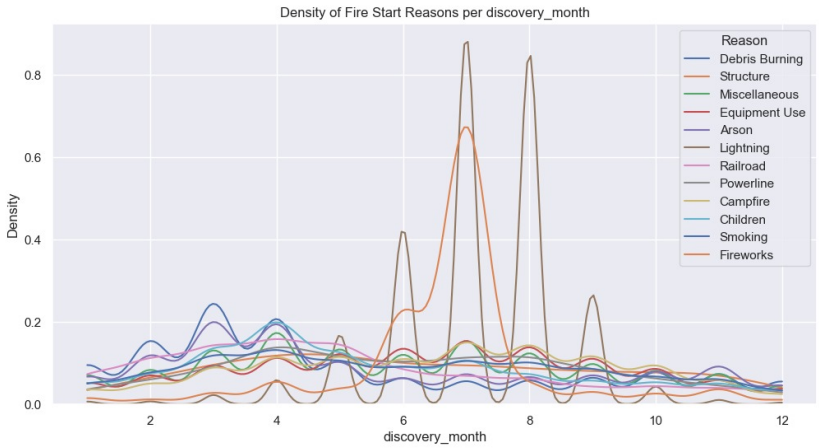

Figure 14: Density of causes per month

It's evident that the month when a fire is detected impacts the distribution of its causes. For instance, fires triggered by lightning or construction show a notable increase in density during the months of June to August. Conversely, incidents like malicious arson or waste burning exhibit a sharp density spike from February to April. Based on these observations, we've chosen to create a new feature called "Season of the Year," which indicates the season in which the fire occurred.

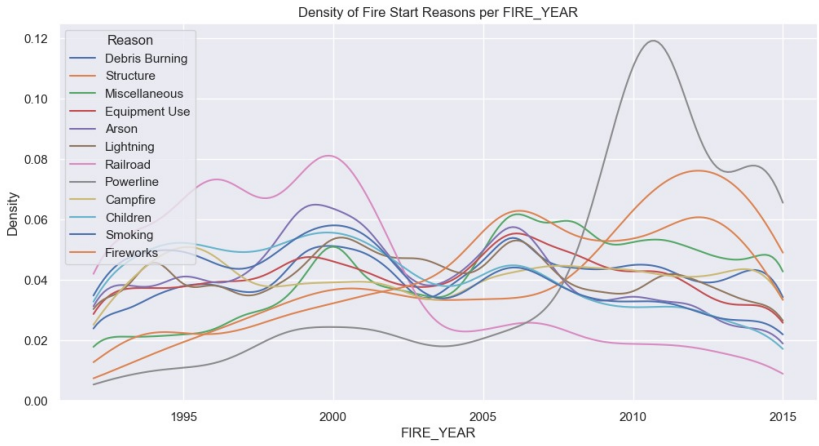

Figure 15: Density of causes per fire year

Significant shifts in the distribution of fire causes across various years are evident. Yet, incorporating a feature derived from this data poses challenges. For instance, if we were to predict a sample from the year 2000 using such a feature, it would introduce information leakage. This is because the feature would rely on fires occurring in subsequent years. Consequently, we've chosen to omit this feature from our model to maintain the integrity of our predictions.

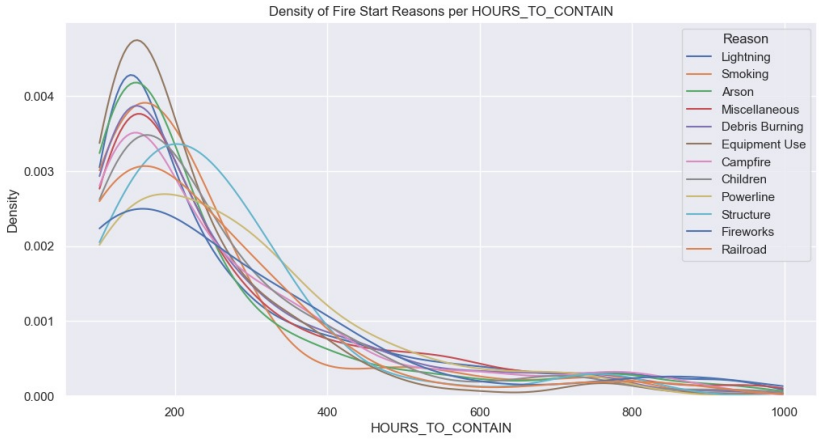

Figure 16: Density of causes per hours to contain

The cause densities, when correlated with the containment time of the fire, display resemblances. However, although no additional feature was identified to significantly enhance the analysis from this distribution, subtle yet potentially meaningful differences were observed. These nuances could hold importance for predictive models. Therefore, this feature was chosen to be incorporated into the analysis.

2.4 Day & Night Labeling

Utilizing the Astral library, along with latitude and longitude for the location, state for the time zone, and date for the specific date and hour, the hours of dawn and dusk for the fire's location and date were obtained. By comparing these hours with the time of fire discovery, fires were categorized as occurring during daylight if they were discovered between dusk and dawn, or during darkness otherwise. This approach allowed the time of day when fires occurred to be discerned, providing valuable insights into fire incident patterns.

2.5 Seasons Labeling

Seasonal variations in environmental conditions, human activities, vegetation dynamics, and cultural practices all influence the likelihood of fire causes. By leveraging these seasonal nuances, classifiers can more accurately identify the reasons behind fire outbreaks, enhancing overall classification accuracy.

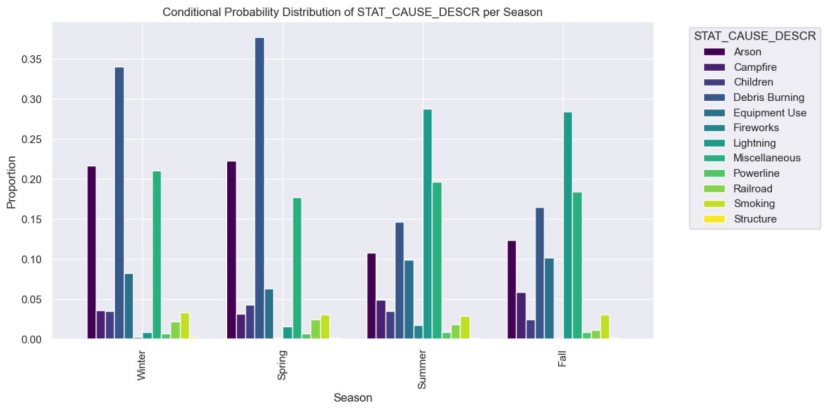

Figure 17: Conditional distribution of each state_cause_descr per season

2.6 Location Analysis - Nearest Neighbors

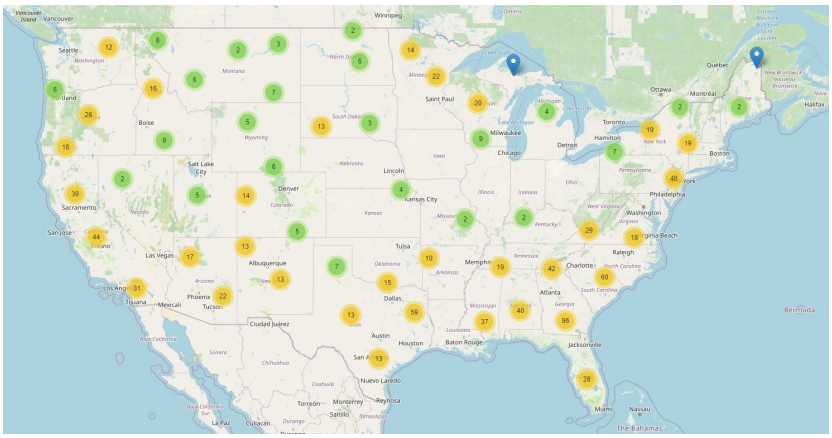

Figure 18: KNN visualization over the US map over 1000 samples

An additional feature was constructed, derived from the output of the K-Nearest Neighbors (KNN) algorithm. The appropriate value of k is first determined through k-fold cross-validation over the validation set. This process

ensures that the model generalizes well to unseen data. Once the optimal value of k is identified, the KNN algorithm is trained on the training data. The output of the algorithm for each data point in the training set is then used as a new feature. This feature captures the relationship between the data point and its nearest neighbors in the feature space. By incorporating the KNN feature, the model gains additional information about the local structure of the data, potentially improving its predictive performance.

2.7 Time-Based Feature Engineering - Time to Containment

An examination of the dataset revealed a subset of features related to time:

- `CONT TIME`
- `CONT DOY`
- `CONT DATE`
- `DISCOVERY TIME`
- `DISCOVERY DOY`
- `DISCOVERY DATE`

It was observed that the features `CONT DATE` and `DISCOVERY DATE` are formatted in the Julian style and so they were converted.

Leveraging these features, a new attribute named `HOURS TO CONT` was formulated. This involved calculating the duration, in hours, from the moment a fire was discovered to when it was successfully contained. The rationale behind this feature was the assumption that the duration of containment could reflect the complexity of the fire. It was hypothesized that fires caused by similar factors would typically present comparable challenges in terms of containment duration.

3 Data Pre-Processing

3.1 Dummy Features

Dummy features are used to convert categorical variables into a binary numerical variable, a format that's easier for a machine learning model to work with. The following features were transformed into dummy features:

- `season` feature, extracted from the `DISCOVERY_DATE` feature, as elaborated on in the previous section.
- `month_discovery` feature, extracted from the `DISCOVERY_DATE` feature.
- `classified_as` feature, created with a KNN model, as elaborated on previously.
- `FIRESIZECLASS` feature.

3.2 Feature Dropping

3.2.1 Values Issues

An examination was conducted on the features that exhibited the highest number of missing values, which include:

- `MTBSFIRENAME`
- `ICS209INCIDENTNUMBER`
- `ICS209NAME`
- `MTBSID`
- `COMPLEXNAME`

A thorough investigation was carried out to ascertain if there existed any bias in the distribution between samples that were devoid of a value and those that possessed one. A significant correlation was observed between the receipt of a non-null value and the occurrence of a lightning strike. As a result, it became apparent that these features were susceptible to leakage, and hence, the decision was made not to utilize them.

Moreover, three additional features exist, each exhibiting a unique value for every sample:

- `FOD_ID`
- `FPA_ID`
- `OBJECT_ID`

Consequently, these features were discarded as they failed to provide any valuable information for analysis. This decision was guided by the principle of maintaining the integrity and usefulness of the data for the modeling process.

3.2.2 Redundant Features

In the process of refining the model, a conscious decision was made to exclude certain features, namely:

- `STATE`
- `FIPS_NAME`
- `FIPS_CODE`
- `COUNTY`

This decision was predicated on the utilization of longitude and latitude data to encapsulate location information. By integrating longitude and latitude coordinates, a more precise depiction of the geographical location of each fire incident is achieved. Consequently, the incorporation of these location-centric features would introduce redundancy without contributing substantial new information to the model. This streamlined approach not only mitigates model complexity but also circumvents potential multicollinearity issues that may arise from incorporating closely related features. Therefore, by concentrating solely on longitude and latitude for location data, the efficiency and effectiveness of the model are optimized while preserving its predictive accuracy.

3.2.3 Potential Leakage

NWCG_REPORTING_AGENCY feature

In the course of evaluating this feature, apprehensions arose regarding a potential leakage issue, due to the participation of 11 distinct agencies in scrutinizing fire incidents. The involvement of a multitude of agencies in these investigations engenders the prospect that each agency may concentrate exclusively on fires instigated by causes that align with their specific interests. As a result, armed with the knowledge of the investigating agency, one could inadvertently deduce the cause of the fire. This presents a potential pitfall as it could lead to biased predictions. Hence, caution was exercised to mitigate this risk.

For instance, for the IA agency, the graph illustrates that the overwhelming majority of fire cases they examined were attributed to lightning.

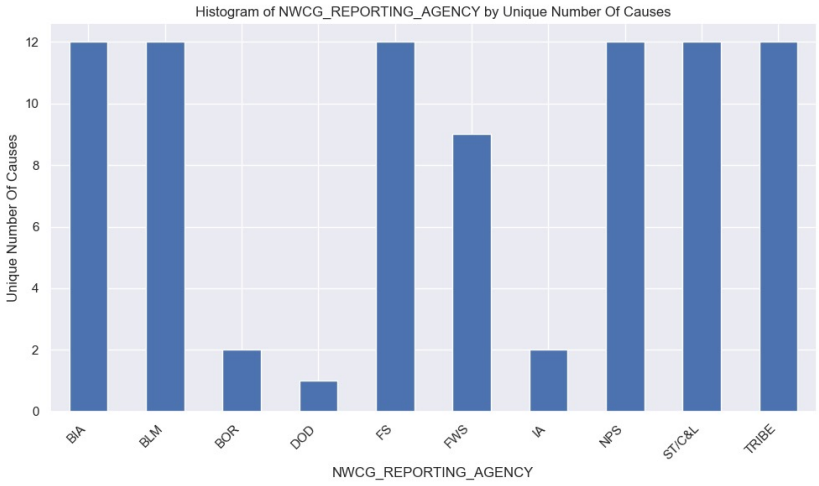

Figure 19: Reporting agency by unique number of causes

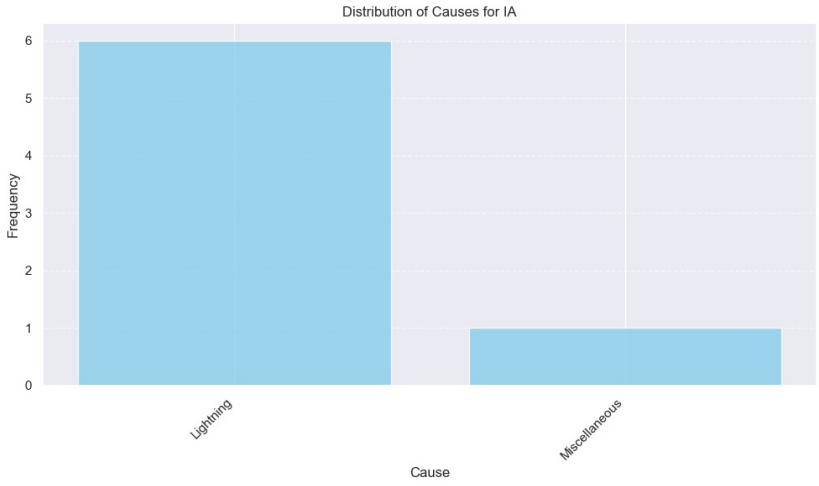

Figure 20: IA fire causes

SOURCE_SYSTEM feature

Much like reporting agencies, this feature may also result in leakage for similar reasons.

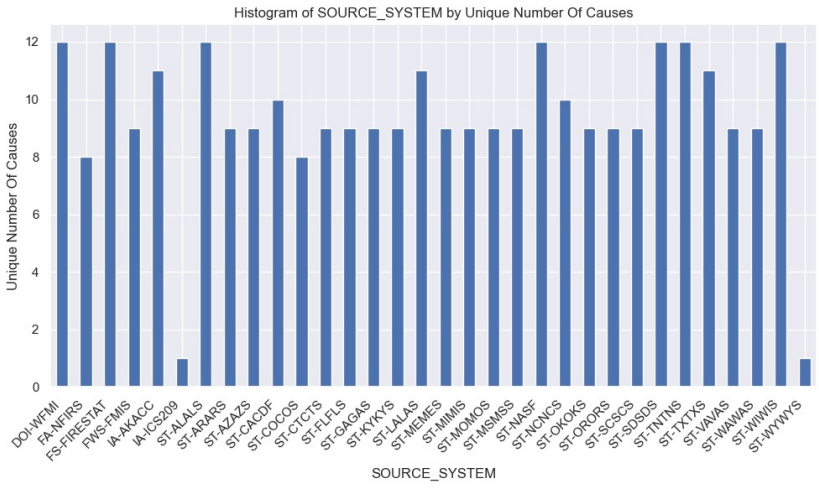

Figure 21: Source system by unique number of causes

For instance, for the ST-WYWYS agency, the graph illustrates that the overwhelming majority of fire cases they examined were attributed to lightning.

Figure 22: ST-WYWYS fire causes

4 Data Enrichment

4.1 Land Cover Data

In the study, data from the 2016 National Land Cover Database (NLCD), a resource provided by the U.S. Geological Survey, was incorporated. The NLCD offers comprehensive information on land cover and changes in land cover across the United States. It provides a series of map products that depict land cover and changes in land cover across nine epochs, spanning from 2001 to 2021. These map products include, but are not limited to, land cover, land cover change index, and urban imperviousness.

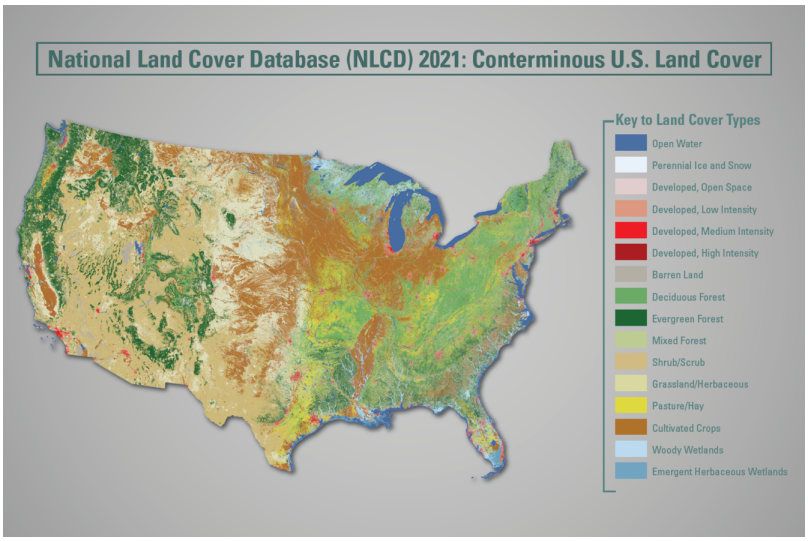

Figure 23: 2021 NLCD Land Cover Data - Contiguous US

The objective was to augment the dataset with the land cover attribute, guided by the hypothesis that the nature of the land and its vegetation can have a substantial influence on the likelihood of fires. The geospatial data from the 2016 epoch was selected as it aligns most closely with the end date of the dataset.

The NLCD is a complex database that houses a vast amount of data. In the initial attempt to utilize this data, the actual spatial data was downloaded and an attempt was made to manipulate it using various Geographic Information System (GIS) software, but the efforts were unsuccessful. Subsequently, the `pygeohydro` package was turned to, which facilitates direct querying of several geospatial data web services, including the NLCD. Using the `nlcd_bycoords` method, the NLCD land cover code for each coordinate was queried.

Before this, the `pyproj` package was employed to convert the geographical data from the North American Datum 83 (NAD83) coordinate system (ESPG:4269) to the coordinate system utilized by the NLCD - WGS84 (ESPG:4326).

The initial attempts to query this field for over 500,000 samples were unsuccessful, leading to the implementation of a batch method with a retry mechanism. Regrettably, despite numerous attempts, a sufficient amount of data could not be queried to incorporate this feature into the models. However, it is believed that this feature could significantly enhance future models. The full code used for this trial at enrichment can be found in the included `preprocess.py` file.

5 Model Selection

In the Model Selection section, various machine learning models were explored and evaluated to determine the most suitable approach for predicting fire incidents. This model selection was done to assess their performance in capturing the complex patterns within the dataset and making accurate predictions.

5.1 Models

5.1.1 Random Forest with Grid Search

Random Forest was utilized as a predictive model for fire incident prediction. To enhance the performance of the Random Forest model, a Grid Search technique was employed.
The Grid Search hyperparameter optimization technique systematically explores a predefined grid of hyperparameters to identify the optimal combination that yields the best performance. In this case, the combination of hyperparameters, including the number of trees, maximum depth of trees, and minimum samples per leaf, was fine-tuned using Grid Search.
The parameter space considered was as follows:
- Minimum Samples per Leaf: [16, 25, 50]
- Minimum Samples per Split: [80, 120, 250]
- Number of Estimators: [30, 40]

The best model identified through Grid Search had the following hyperparameters:
- Minimum Samples per Leaf: 16
- Minimum Samples per Split: 80
- Number of Estimators: 40

The ROC-AUC score of this best model was 0.81.

5.1.2 XGBoost with Grid Search

Similar to the Random Forest approach, XGBoost was utilized as a predictive model for fire incident prediction. To enhance the performance of the XGBoost model, Grid Search was once again employed.

This time, Grid Search was utilized to fine-tune the hyperparameters of XGBoost, including the learning rate, maximum depth of trees, and the number of boosting rounds.
The hyperparameter space considered was as follows:
- Number of Estimators: [30, 40]
- Maximum Depth: [5, 10]
- Learning Rate: [0.1, 0.3]
- Subsample: [0.1, 0.3]
- Colsample by Tree: [0.1, 0.3]

The best model identified through Grid Search had the following hyperparameters:
- Number of Estimators: 40

- Maximum Depth: 5
- Learning Rate: 0.1
- Subsample: 0.1
- Colsample by Tree: 0.1
The ROC-AUC score of this best model was 0.77.

5.1.3 Logistic Regression

Logistic Regression was utilized as a baseline model for fire incident prediction. By employing Logistic Regression, we aimed to establish a baseline level of performance for fire incident prediction and assess the effectiveness of more sophisticated models.
AUC-ROC SCORE: 0.68

5.1.4 Neural Network

Neural Networks were employed as a flexible and powerful model for capturing complex patterns within the dataset. Through training on the dataset, the Neural Network learned to extract relevant features and make predictions regarding fire incidents. The neural network architecture is a feedforward model featuring two densely connected layers, an input layer, a hidden layer with 64 neurons utilizing Rectified Linear Unit (ReLU) activation, and an output layer with 12 units, aligning with the 12 unique categories being predicted. ReLU introduces non-linearity, enabling the model to capture intricate patterns within the data, facilitating effective classification.
AUC-ROC SCORE: 0.58

5.2 Naive Bayes

Naive Bayes is ideal for this task due to its simplicity, efficiency, and effectiveness in handling multi-class classification, especially with 12 categories. It excels when features are conditionally independent given class labels, making it suitable for various applications, including text classification. Moreover, it requires minimal training data, is robust to noise, and offers straightforward result interpretation, making it a practical choice for resource-constrained scenarios.
AUC-ROC SCORE: 0.66

5.3 Minimal Model And Feature Selection

From the results presented above, it's evident that the Random Forest model achieved the highest ROC AUC score. Now, our focus shifts to examining the importance of each feature.

In the graph below, features with importances greater than 0.04 are displayed.
Therefore, our minimal model will consist of the Random Forest algorithm utilizing only those features with importances higher than 0.04 and with the following

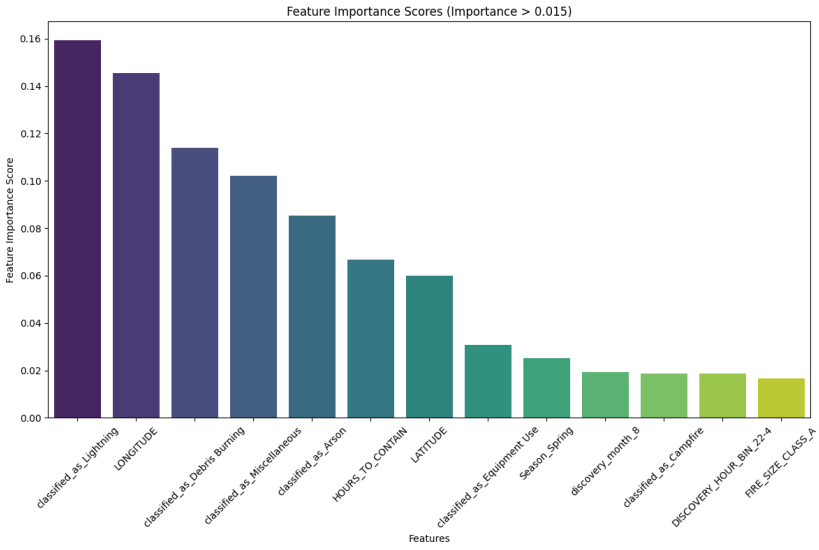

Figure 24: FEATURE IMPORTANCE

hyperparameters:
- Minimum Samples per Leaf: 16
- Minimum Samples per Split: 80
- Number of Estimators: 40
The ROC-AUC score of this model was 0.83.
This indicates that the minimal model performs slightly better than the full model, suggesting that the full model may have overfit to the training data.

Appendix

1. Histograms of Fire Incidents Caused by Various Factors

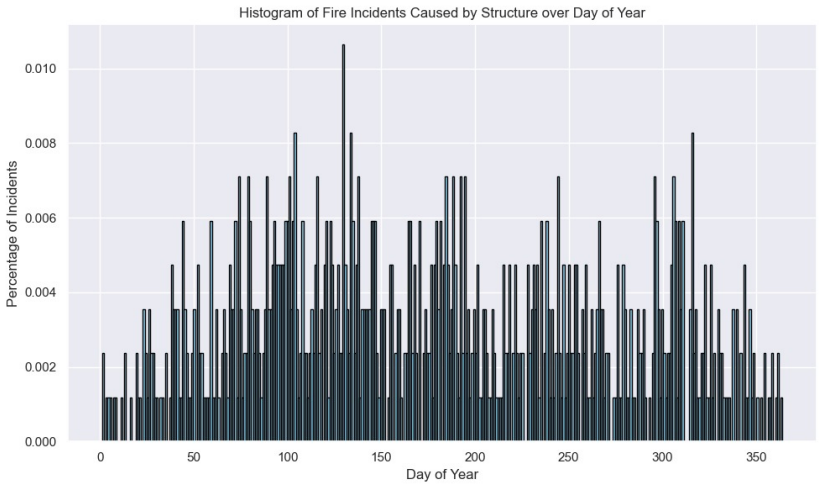

Figure 25: Fire incidents caused by structure over DOY

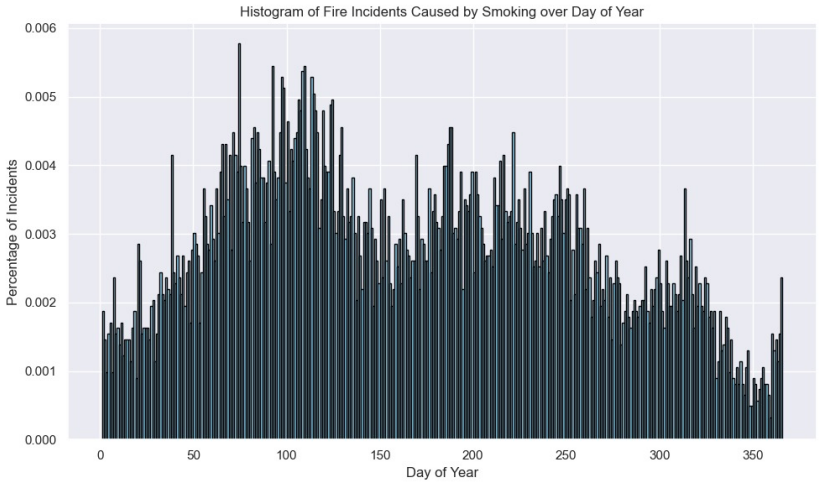

Figure 26: Fire incidents caused by smoking over DOY

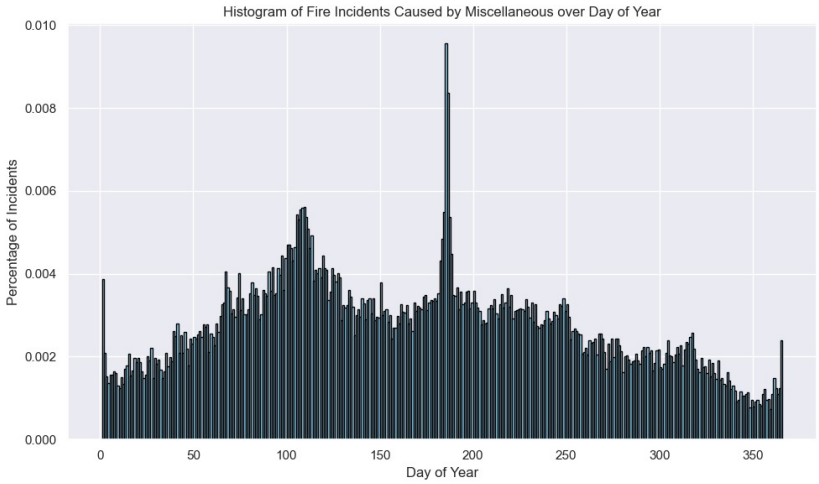

Figure 27: Fire incidents caused by miscellaneous over DOY

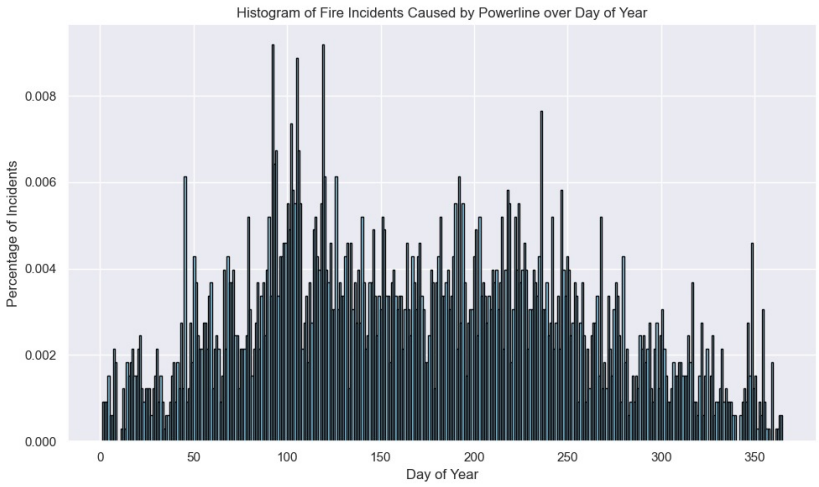

Figure 28: Fire incidents caused by powerline over DOY

2. Dataset Feature Legend

- `FOD ID` Global unique identifier.
- `FPA ID` Unique identifier that contains information necessary to track back to the original record in the source dataset.
- `SOURCESYSTEMTYPE` Type of source database or system that the record was drawn from (federal, nonfederal, or interagency).
- `SOURCESYSTEM` Name of or other identifier for source database or system that the record was drawn from. See Table 1 in Short (2014), or .pdf, for a list of sources and their identifier.
- `NWCGREPORTINGAGENCY` Active National Wildlife Coordinating Group (NWCG) Unit Identifier for the agency preparing the fire report (BIA = Bureau of Indian Affairs, BLM = Bureau of Land Management, BOR = Bureau of Reclamation, DOD = Department of Defense, DOE = 1 Department of Energy, FS = Forest Service, FWS = Fish and Wildlife Service, IA = Interagency Organization, NPS = National Park Service, ST/C&L = State, County, or Local Organization, and TRIBE = Tribal Organization).
- `NWCGREPORTINGUNIT ID` Active NWCG Unit Identifier for the unit preparing the fire report.
- `NWCGREPORTINGUNIT NAME` Active NWCG Unit Name for the unit preparing the fire report. SOURCEREPORTINGUNIT = Code for the agency unit preparing the fire report, based on code/name in the source dataset.
- `SOURCEREPORTINGUNIT NAME` Name of reporting agency unit preparing the fire report, based on code/name in the source dataset.
- `LOCALFIREREPORT ID` Number or code that uniquely identifies an incident report for a particular reporting unit and a particular calendar year.
- `LOCALINCIDENTID` Number or code that uniquely identifies an incident for a particular local fire management organization within a particular calendar year.
- `FIRE CODE` Code used within the interagency wildland fire community to track and compile cost information for emergency fire suppression (https://www.firecode.gov/).
- `FIRE NAME` Name of the incident, from the fire report (primary) or ICS-209 report (secondary).
- `ICS209INCIDENT NUMBER` Incident (event) identifier, from the ICS209 report.
- `ICS209NAME` Name of the incident, from the ICS-209 report.
- `MTBS ID` Incident identifier, from the MTBS perimeter dataset.

- `MTBSFIRENAME` Name of the incident, from the MTBS perimeter dataset.
- `COMPLEX NAME` Name of the complex under which the fire was ultimately managed, when discernible.
- `FIRE YEAR` Calendar year in which the fire was discovered or confirmed to exist.
- `DISCOVERY DATE` Date on which the fire was discovered or confirmed to exist.
- `DISCOVERY DOY` Day of year on which the fire was discovered or confirmed to exist.
- `DISCOVERY TIME` Time of day that the fire was discovered or confirmed to exist.
- `STATCAUSECODE` Code for the (statistical) cause of the fire.
- `STATCAUSEDESCR` Description of the (statistical) cause of the fire.
- `CONT DATE` Date on which the fire was declared contained or otherwise controlled (mm/dd/yyyy where mm=month, dd=day, and yyy=year).
- `CONT DOY` Day of year on which the fire was declared contained or otherwise controlled.
- `CONT TIME` Time of day that the fire was declared contained or otherwise controlled (hhmm where hh=hour, m=minutes).
- `FIRE SIZE` Estimate of acres within the final perimeter of the fire.
- `FIRESIZECLASS` Code for fire size based on the number of acres within the final fire perimeter expenditures (A=greater than 0 but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, and G=5000+ acres).
- `LATITUDE` Latitude (NAD83) for point location of the fire (decimal degrees).
- `LONGITUDE` Longitude (NAD83) for point location of the fire (decimal degrees).
- `OWNER CODE` Code for primary owner or entity responsible for managing the land at the point of origin of the fire at the time of the incident.
- `OWNER DESCR` Name of primary owner or entity responsible for managing the land at the point of origin of the fire at the time of the incident.
- `STATE` Two-letter alphabetic code for the state in which the fire burned (or originated), based on the nominal designation in the fire report.

- `COUNTY` County, or equivalent, in which the fire burned (or originated), based on nominal designation in the fire report.
- `FIPS CODE` Three-digit code from the Federal Information Process Standards (FIPS) publication 6-4 for representation of counties and equivalent entities.
- `FIPS NAME` County name from the FIPS publication 6-4 for representation of counties and equivalent entities. 3
- `NWCGUnitIDActive20170109` Look-up table containing all NWCG identifiers for agency units that were active (i.e., valid) as of 9 January 2017, when the list was downloaded from https://www.nifc.blm.gov/unit id/Publish.html and used as the source of values available to populate the following fields in the Fires table: NWCGREPORTINGAGENCY, NWCGREPORTINGUNITID, and NWCGREPORTINGUNITNAME.
- `UnitId` NWCG Unit ID.
- `GeographicArea` Two-letter code for the geographic area in which the unit is located (NA=National, IN=International, AK=Alaska, CA=California, EA=Eastern Area, GB=Great Basin, NR=Northern Rockies, NW=Northwest, RM=Rocky Mountain, SA=Southern Area, and SW=Southwest).
- `Gacc` Seven or eight-letter code for the Geographic Area Coordination Center in which the unit is located or primarily affiliated with.
- `WildlandRole` Role of the unit within the wildland fire community.
- `UnitType` Type of unit (e.g., federal, state, local).
- `Department` Department (or state/territory) to which the unit belongs.
- `Agency` Agency or bureau to which the unit belongs.
- `Parent` Agency subgroup to which the unit belongs (A concatenation of State and Unit from this report - https://www.nifc.blm.gov/unit id/publish/UnitIdReport.rtf).
- `Country` Country in which the unit is located (e.g. US= United States).
- `State` Two-letter code for the state in which the unit is located (or primarily affiliated).
- `Code` Unit code (follows state code to create UnitId).
- `Name` Unit name.