Applied Competitive Lab Final Project – Wildfires in the USA

Group 4

Dor Tal 205501380

Liav Aharon 318492089

David Guedalia 208505883

Assaf Gozlan 318642287

Table of Contents

Introduction

In this paper, we portray our workflow working on our project. We shall present our design choices in each stage of the process, as well as address other, somewhat less successful solutions we tried. In this paper we thoroughly walk through the main stages in of our workflow which are:

- Exploratory data analysis - understanding the data via visualizations, while exploring possible relations between the features in the data.
- Pre-Processing – Handling problems encountered in the data, such as missing values, and adjusting some of the features to fit our predictive model.
- Feature Engineering – Adjusting existing features (mostly categorical) to fit the model, while trying to "keep" as much information as possible stored in them. This section includes addition of features, such as aggregative features as well.
- Model Selection – experimenting several classification models for comparison, as well as hyper-parameter optimizations.
- Performance Evaluation – reviewing the classification performance of the model.

Objective: Given a test set of wildfires in the USA, **predicting the cause of each fire in the data**.

Exploratory data analysis

In this section we portray our analysis of the provided data and present some of the conclusions we derived based on our analysis:

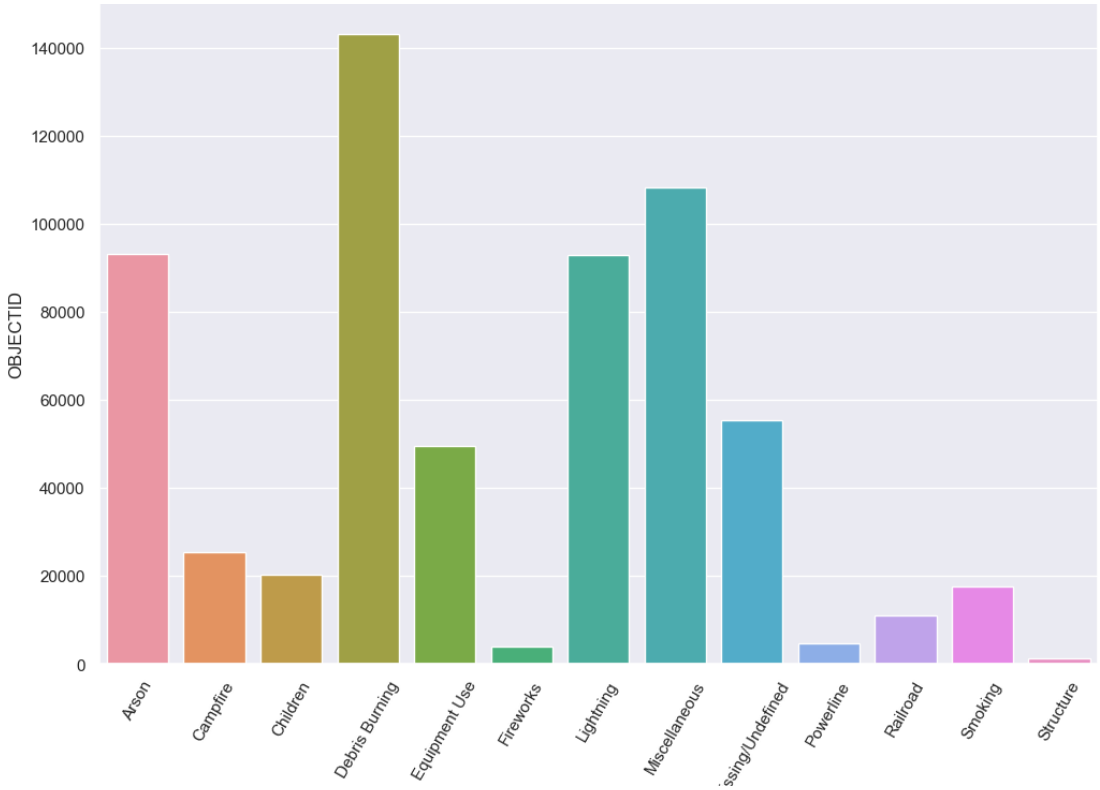
Fire Causes Analysis

First, we would like to analyze our target column, we are particularly interested in examining it's values distribution and relations with other features in the data.

- The frequency of each fire cause:

```
STAT_CAUSE_DESCR
Arson                93304
Campfire             25367
Children             20354
Debris Burning      143074
Equipment Use        49423
Fireworks             3865
Lightning            93057
Miscellaneous       108372
Missing/Undefined    55397
Powerline             4733
Railroad             11053
Smoking              17571
Structure             1252
Name: OBJECTID, dtype: int64
```

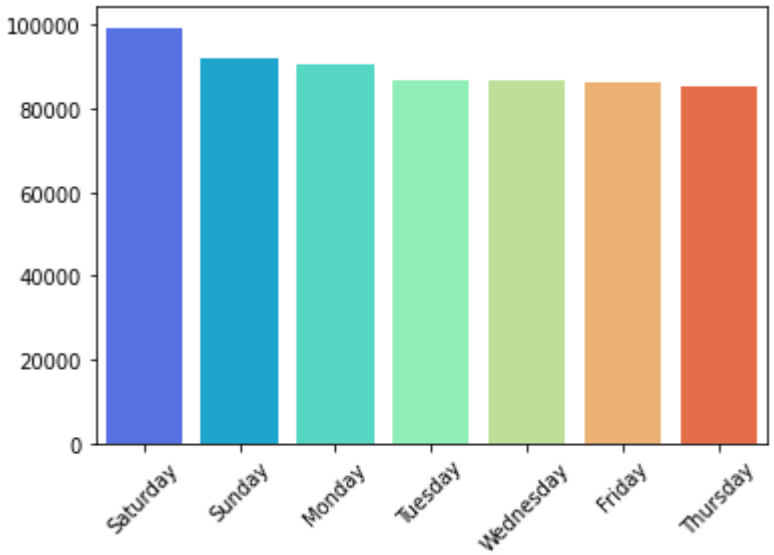
- A plot of the fires number as function of the reasons:



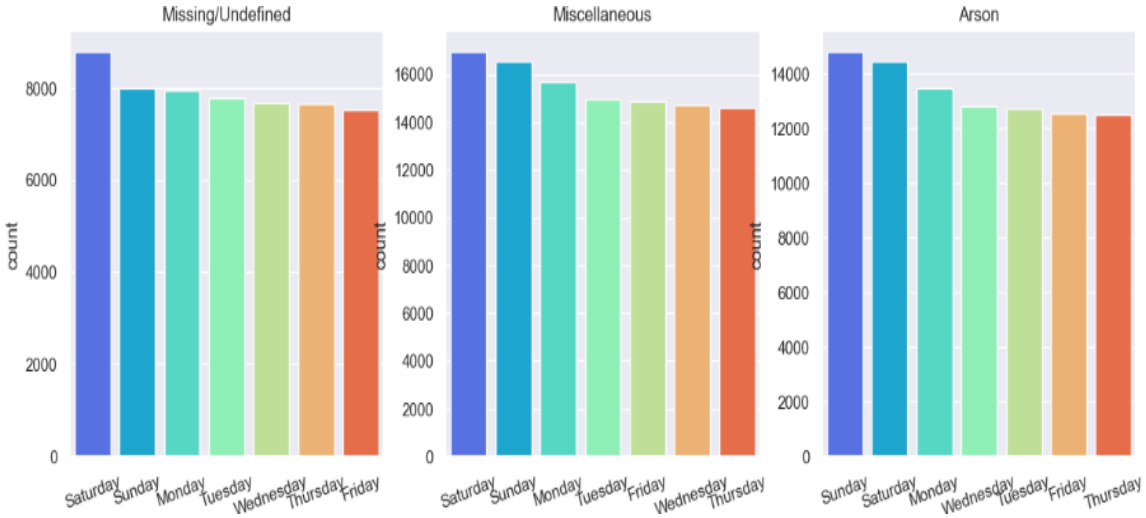
Fire Time Analysis

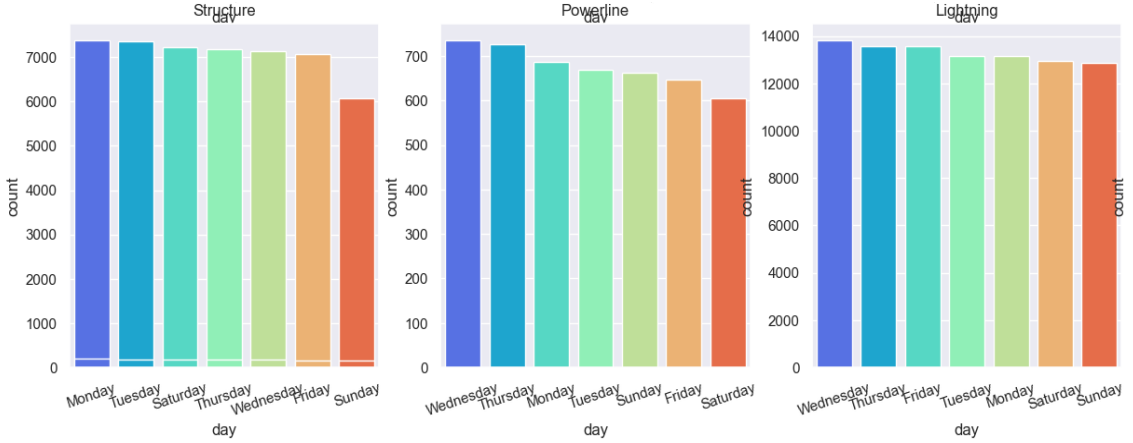
Let us explore the relation between the time timing and frequency of fires:

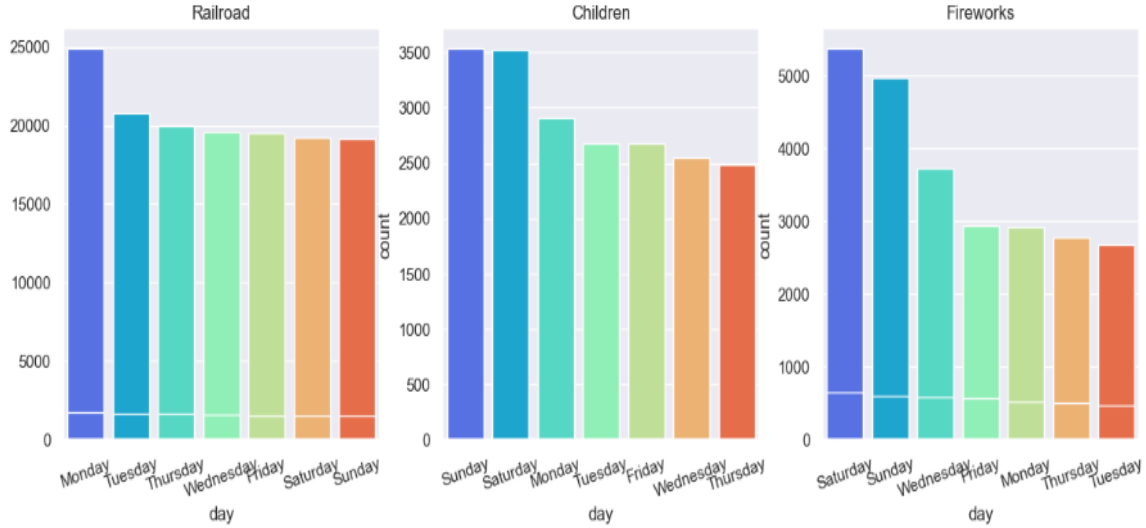
- A plot of the fires number as function of the **days of week**:



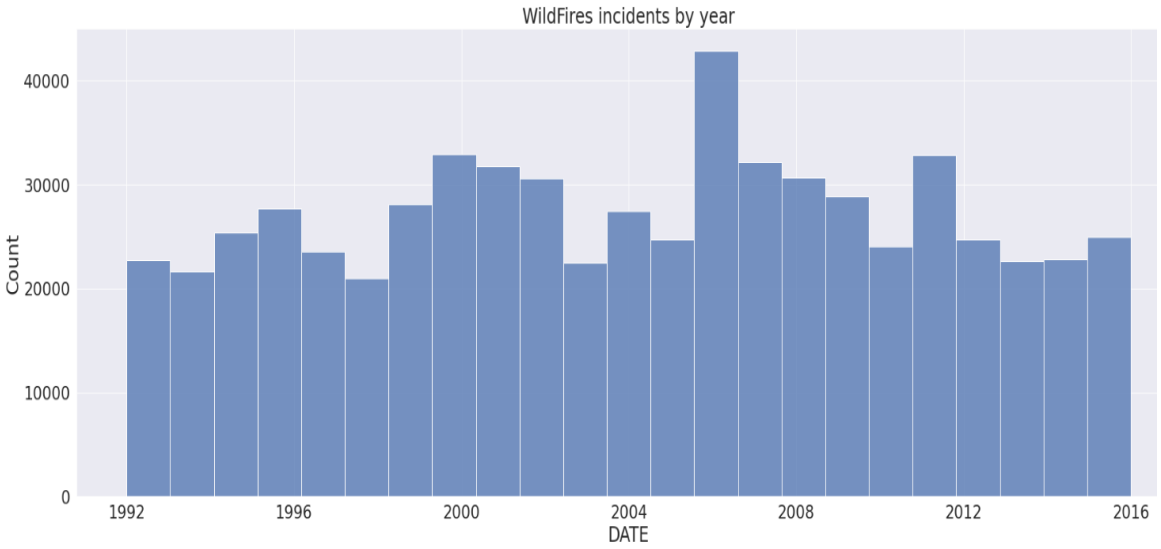
- From the plot, it can be inferred that fires are common **in the weekend.**
- Plots of the **fires number of each cause as function of the days of week**:


Missing/Undefined
count
Saturday
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Miscellaneous
count
Saturday
Sunday
Monday
Tuesday
Friday
Wednesday
Thursday
Arson
count
Sunday
Saturday
Monday
Wednesday
Tuesday
Friday
Thursday


Structure
day
count
Monday
Tuesday
Saturday
Thursday
Wednesday
Friday
Sunday
day
Powerline
day
count
Wednesday
Thursday
Monday
Tuesday
Sunday
Friday
Saturday
day
Lightning
day
count
Wednesday
Thursday
Friday
Tuesday
Monday
Saturday
Sunday
day


Railroad
count
Monday
Tuesday
Thursday
Wednesday
Friday
Saturday
Sunday
day
Children
count
Sunday
Saturday
Monday
Tuesday
Friday
Wednesday
Thursday
day
Fireworks
count
Saturday
Sunday
Wednesday
Friday
Monday
Thursday
Tuesday
day

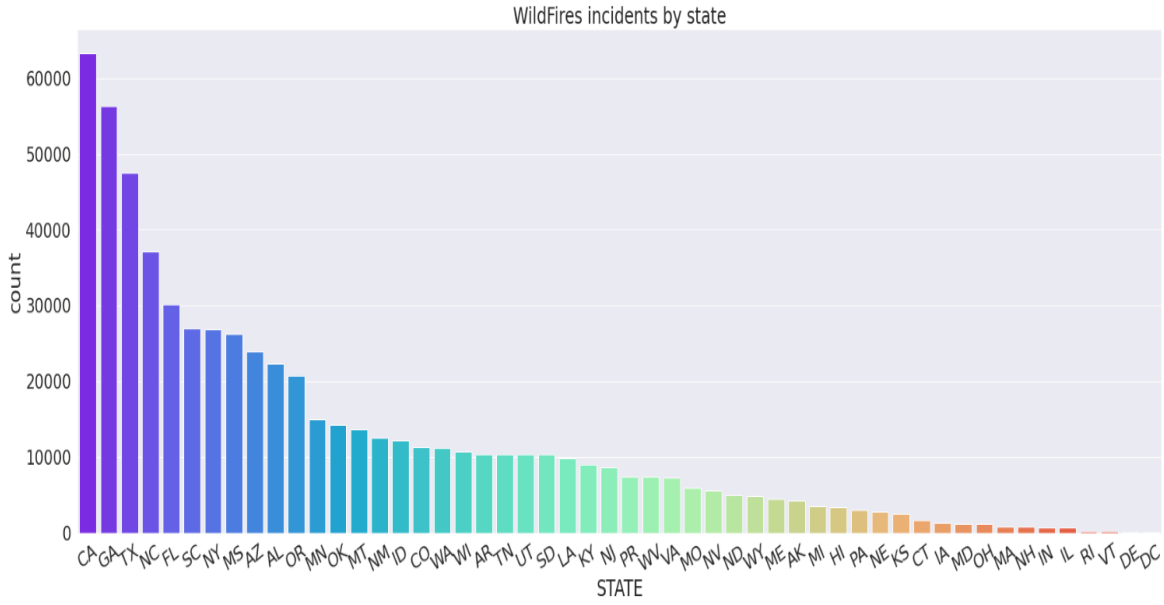
- According to the graphs above, for most of the causes, fires occurred during the weekend. However, note that the causes: Powerline and Structure appear to be common in the middle of the week.
- A graph of **wild fires incidents as a function of years**:



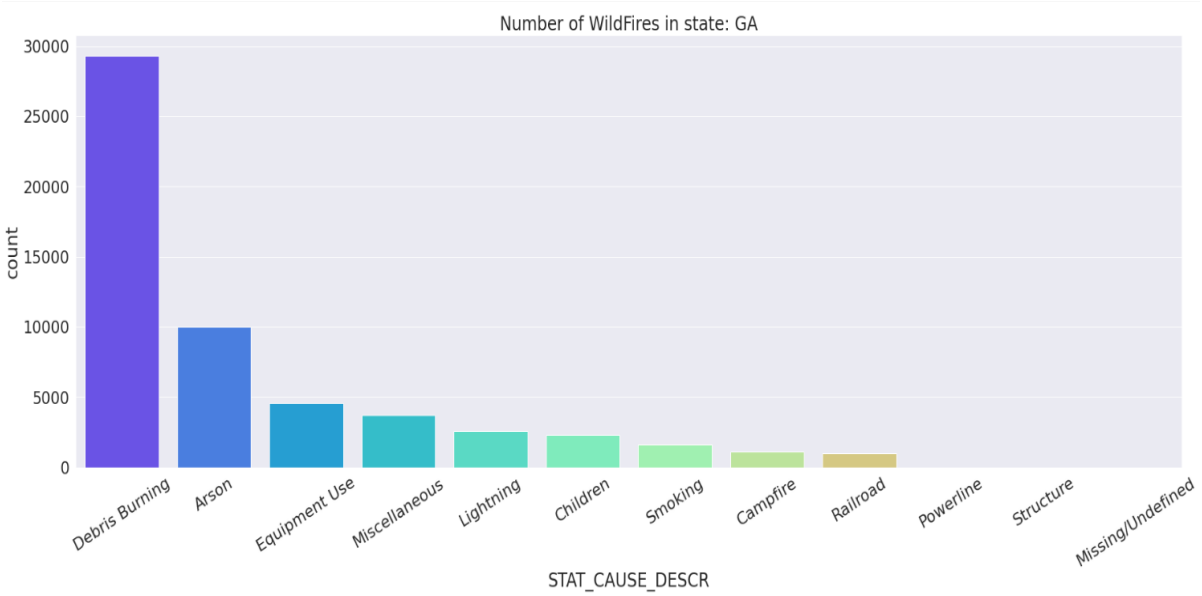
Fire Location Analysis

Now, we proceed to examine the relation between the location of a fire to other features. We start by slicing the data by states and visualizing the results.

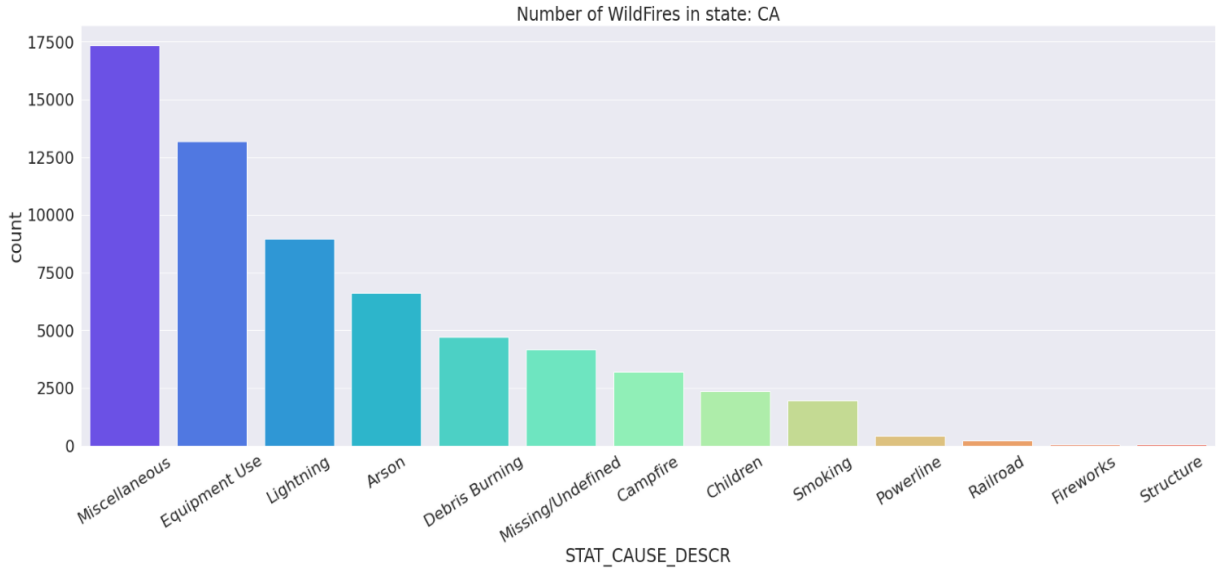
- The following is a bar-plot of **wildfire incidents count as a function of States**:



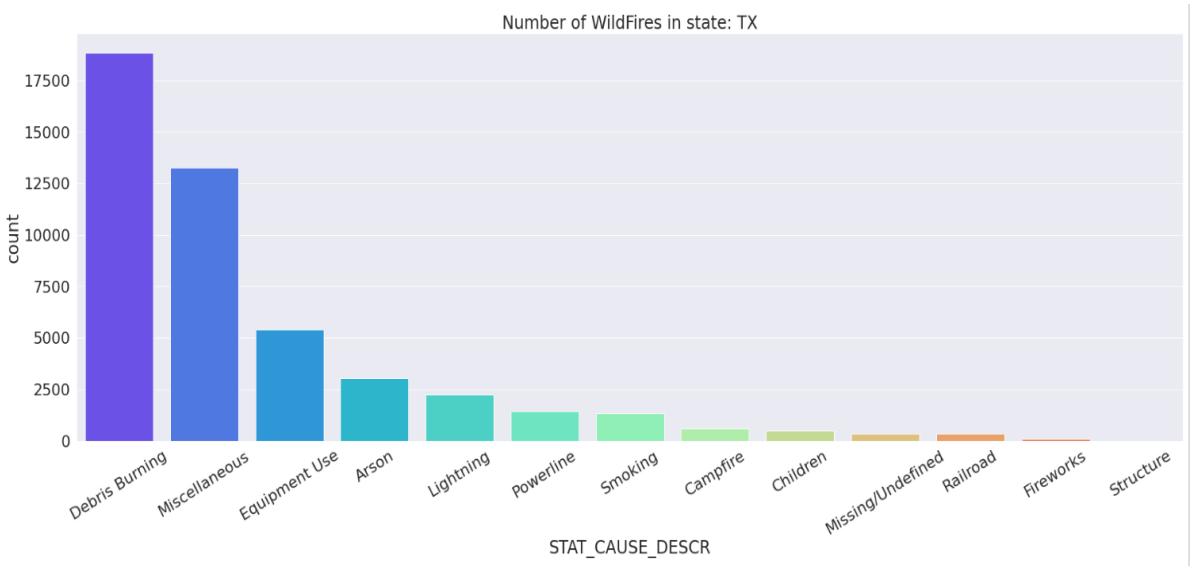
- Bar-plot of **wildfire incidents count as a function of the causes in the state of Georgia:**



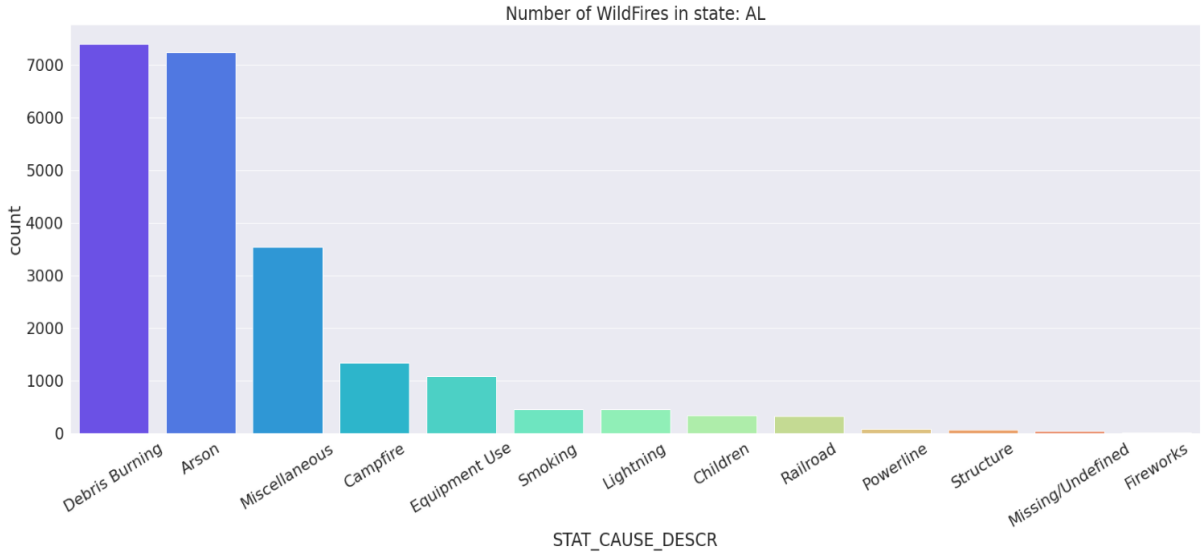
- Bar-plot of **wildfire incidents count as a function of the causes in the state of California:**



- Bar-plot of **wildfire incidents count as a function of the causes in the state of Texas:**



- For the state of Alabama:

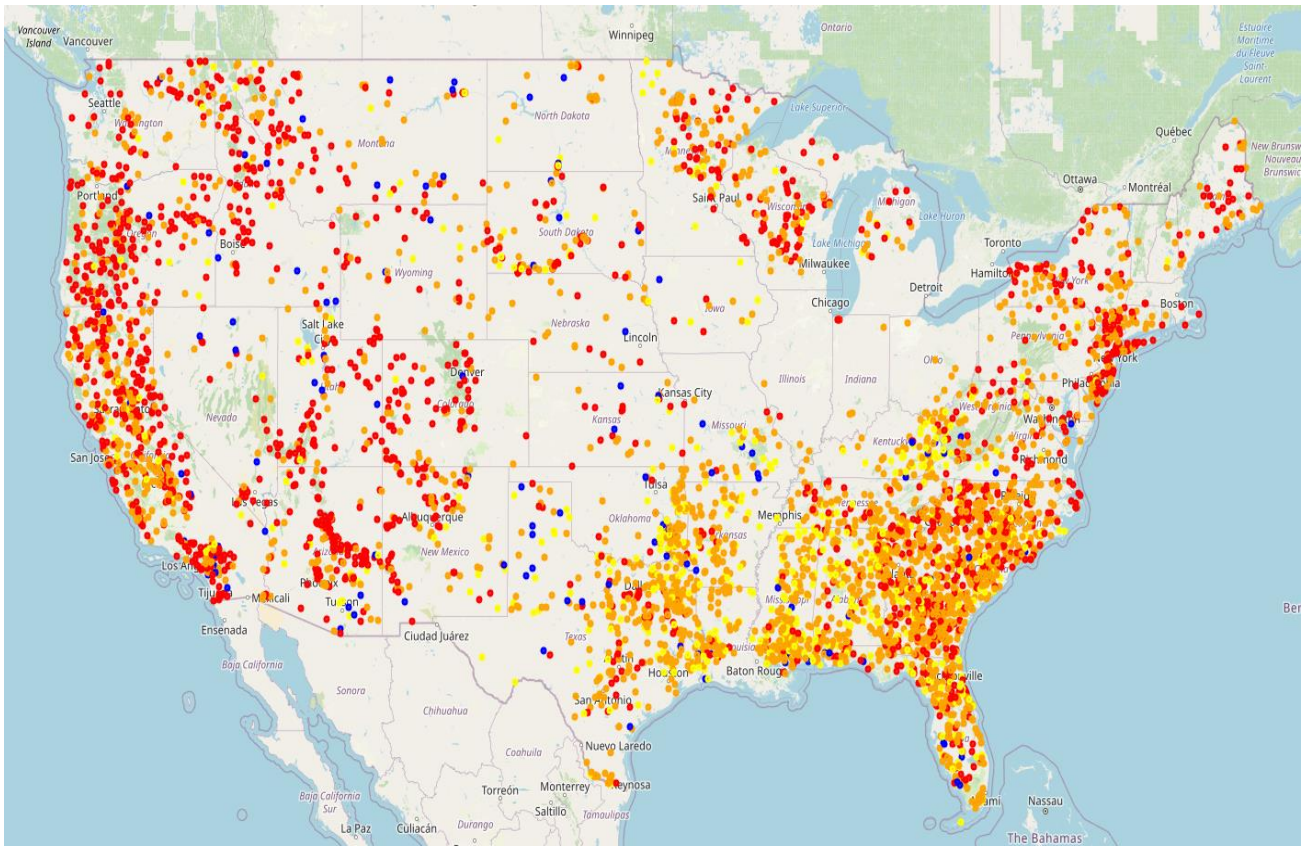


As can be seen in the above plots, the causes distribution slightly shifts as we examine different states. This behavior teaches us that location is likely to be important for our prediction purposes. So, we proceed to examine this relation, but now, we take on a slightly more advanced approach.

The following visualizations demonstrate the impact of wildfire location on it's size and cause.

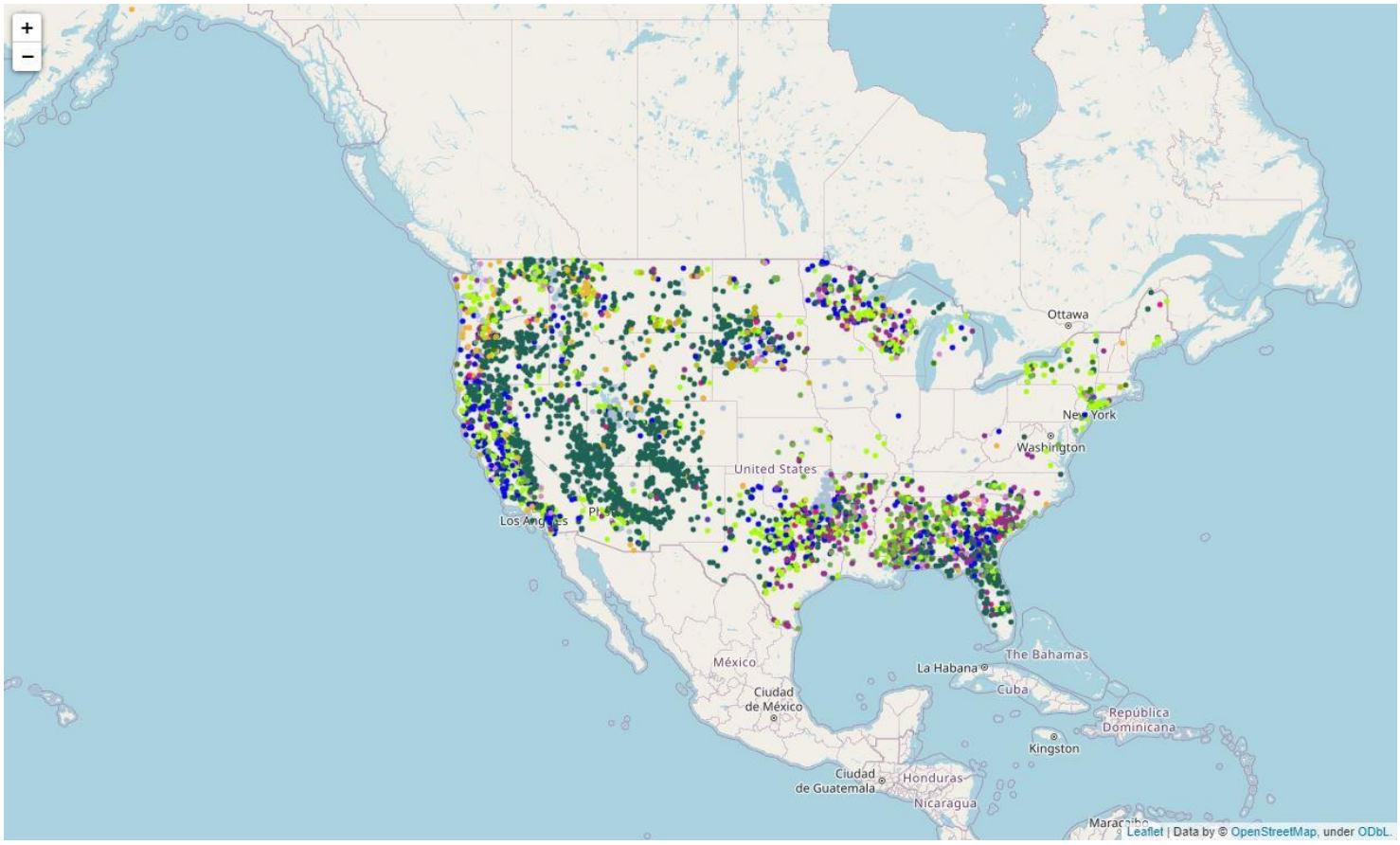
The next map portrays 10,000 random data samples, such that each sample (fire) in the data is colored based on its 'FIRE_SIZE_CLASS' value. where:

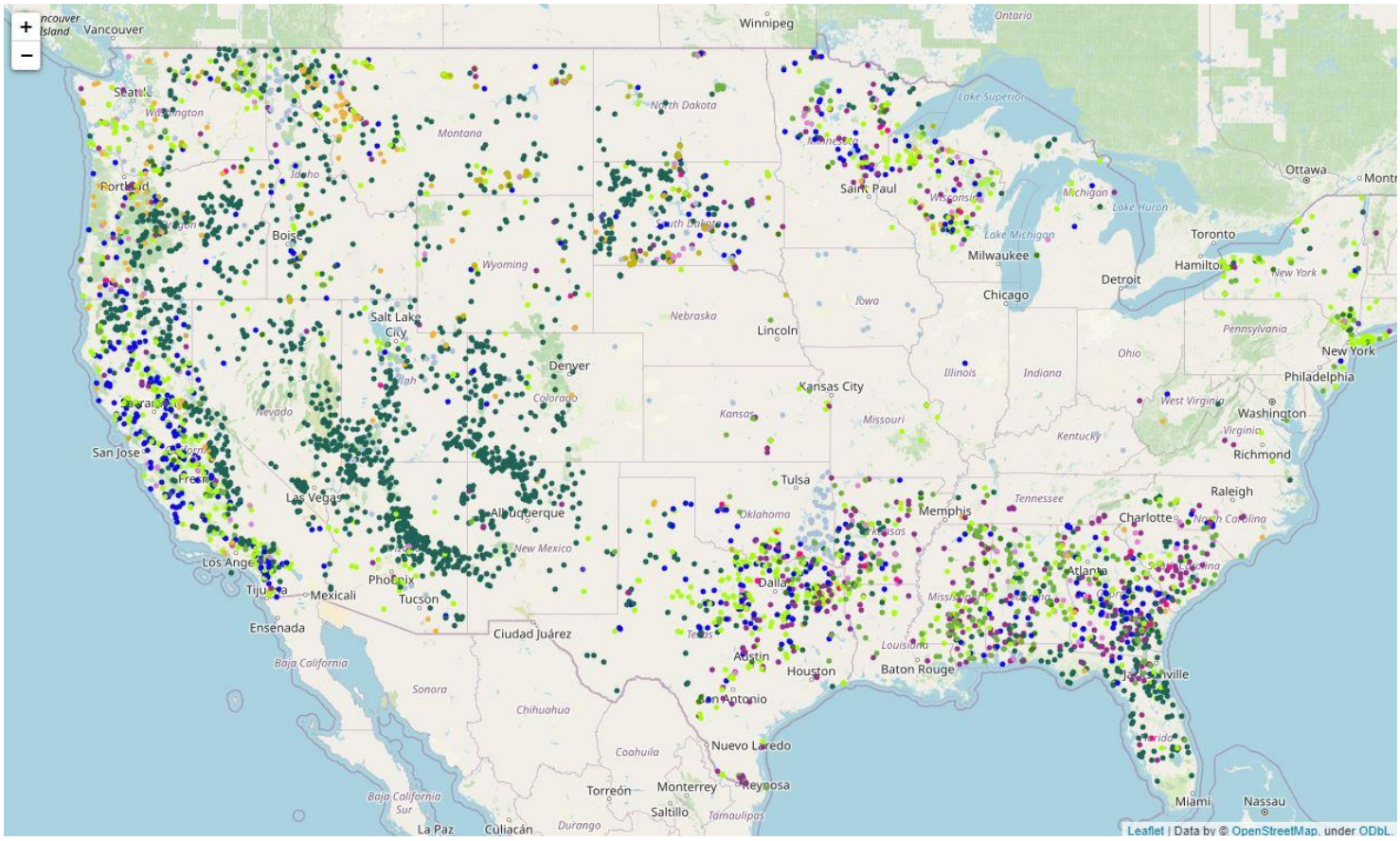
- 'A' = RED
- 'B' = ORANGE
- 'C' = YELLOW
- 'D' and above = BLUE



For our next visualization, we decided to explore the geographic distribution of the causes. To do so, we chose to visualize and compare 2 months from the same year and examine the causes distributions on the map. Specifically, we chose march and July of 2006, as they were the months with the highest sample count.

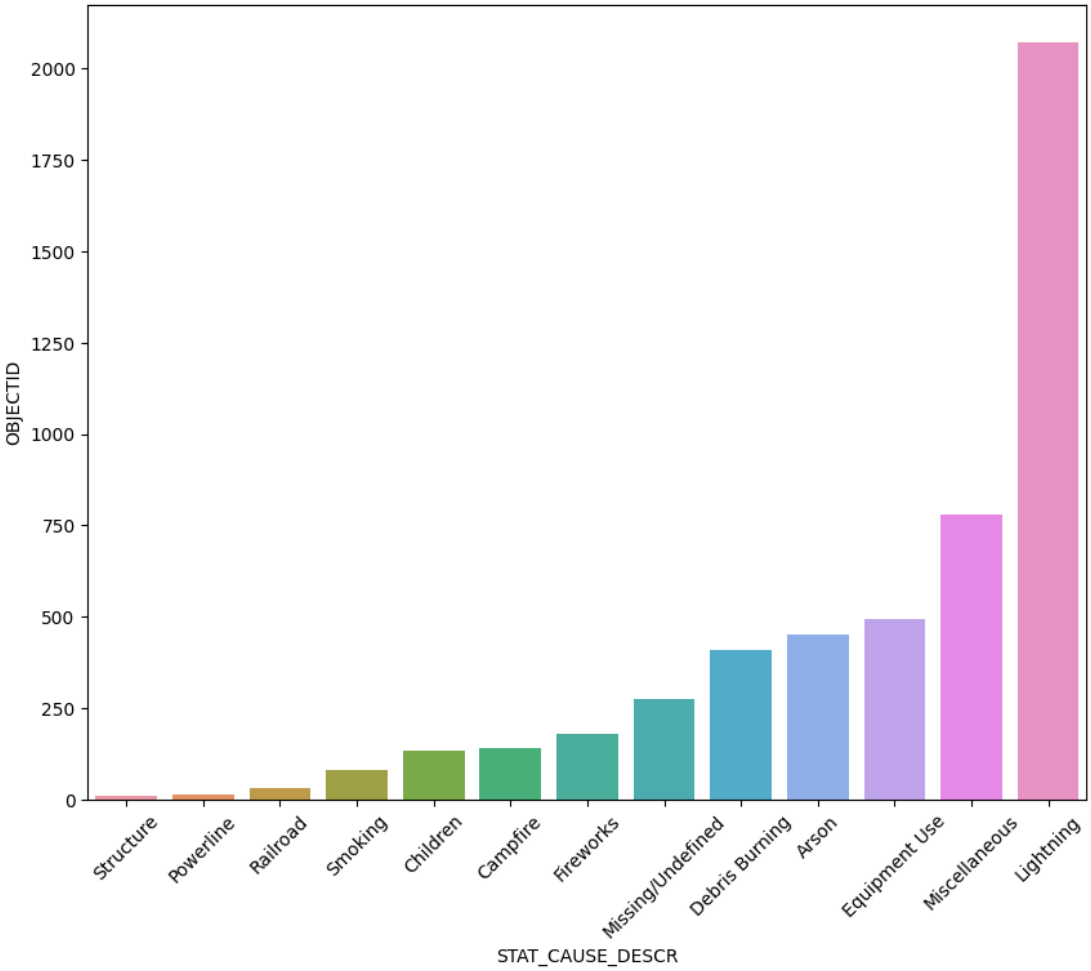
The following two maps show the causes distribution for July 2006 (color legend is in the next page):


Ottawa
New York
Washington
United States
Los Angeles
México
Ciudad de México
La Habana
The Bahamas
Cuba
República Dominicana
Kingston
Ciudad de Guatemala
Honduras
Nicaragua
Leaflet | Data by © OpenStreetMap, under ODbL.

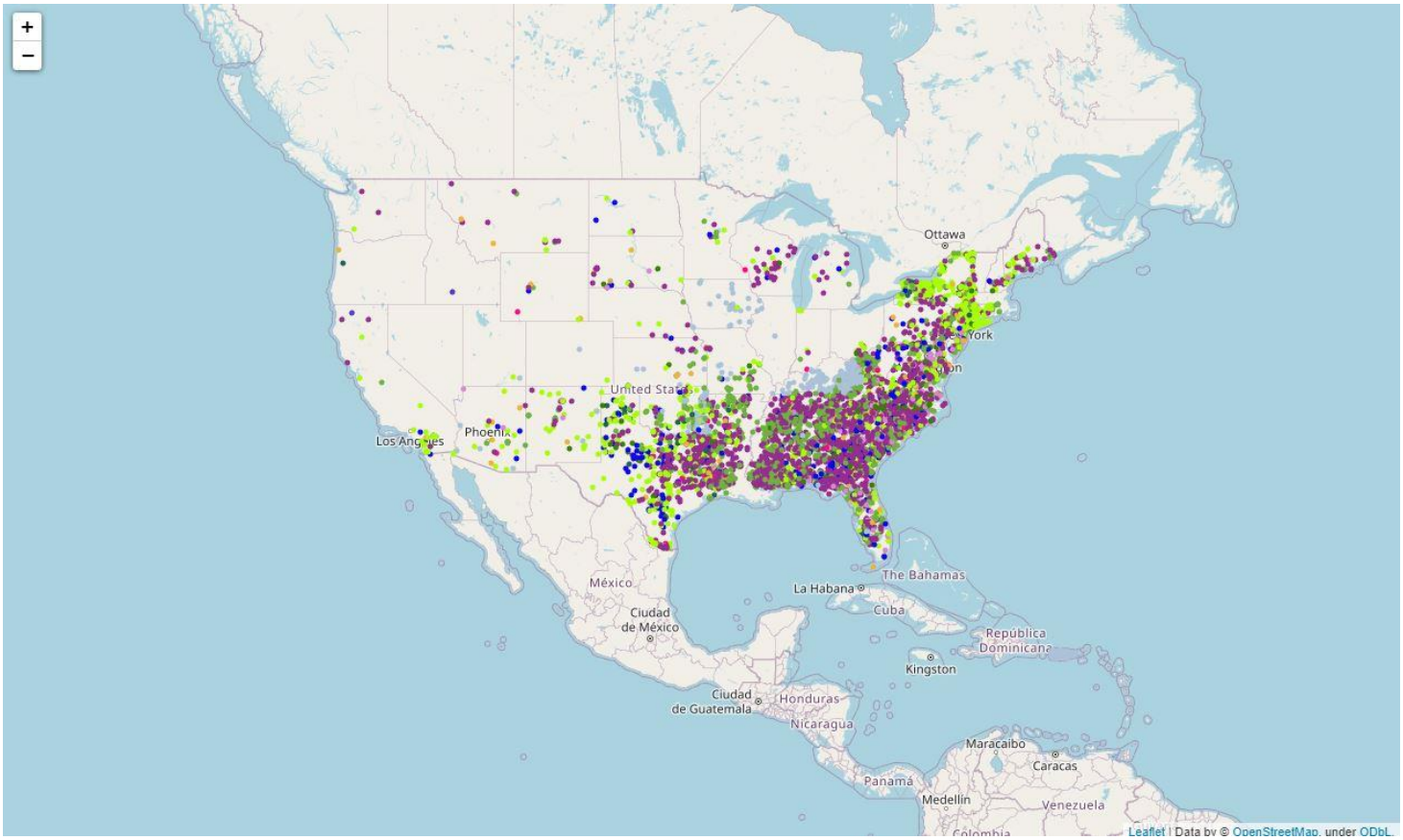

Vancouver
Winnipeg
Ontario
Seattle
Portland
Boise
Salt Lake City
Denver
Las Vegas
San Jose
Los Angeles
Phoenix
Tucson
Albuquerque
Mexicali
Tijuana
Ensenada
Ciudad Juárez
Baja California
Sonora
Chihuahua
Coahuila
Nuevo Laredo
Torreón
Monterrey
Reynosa
Saltillo
Tamaulipas
La Paz
Culiacán
Durango
Baja California Sur
Lincoln
Kansas City
Omaha
Saint Paul
Minneapolis
Milwaukee
Chicago
Detroit
Toronto
Hamilton
Ottawa
Philadelphia
New York
Washington
Richmond
Raleigh
Charlotte
Memphis
Atlanta
Tulsa
Dallas
Austin
Houston
Baton Rouge
San Antonio
Miami
Nassau
Leaflet | Data by © OpenStreetMap, under ODbL.

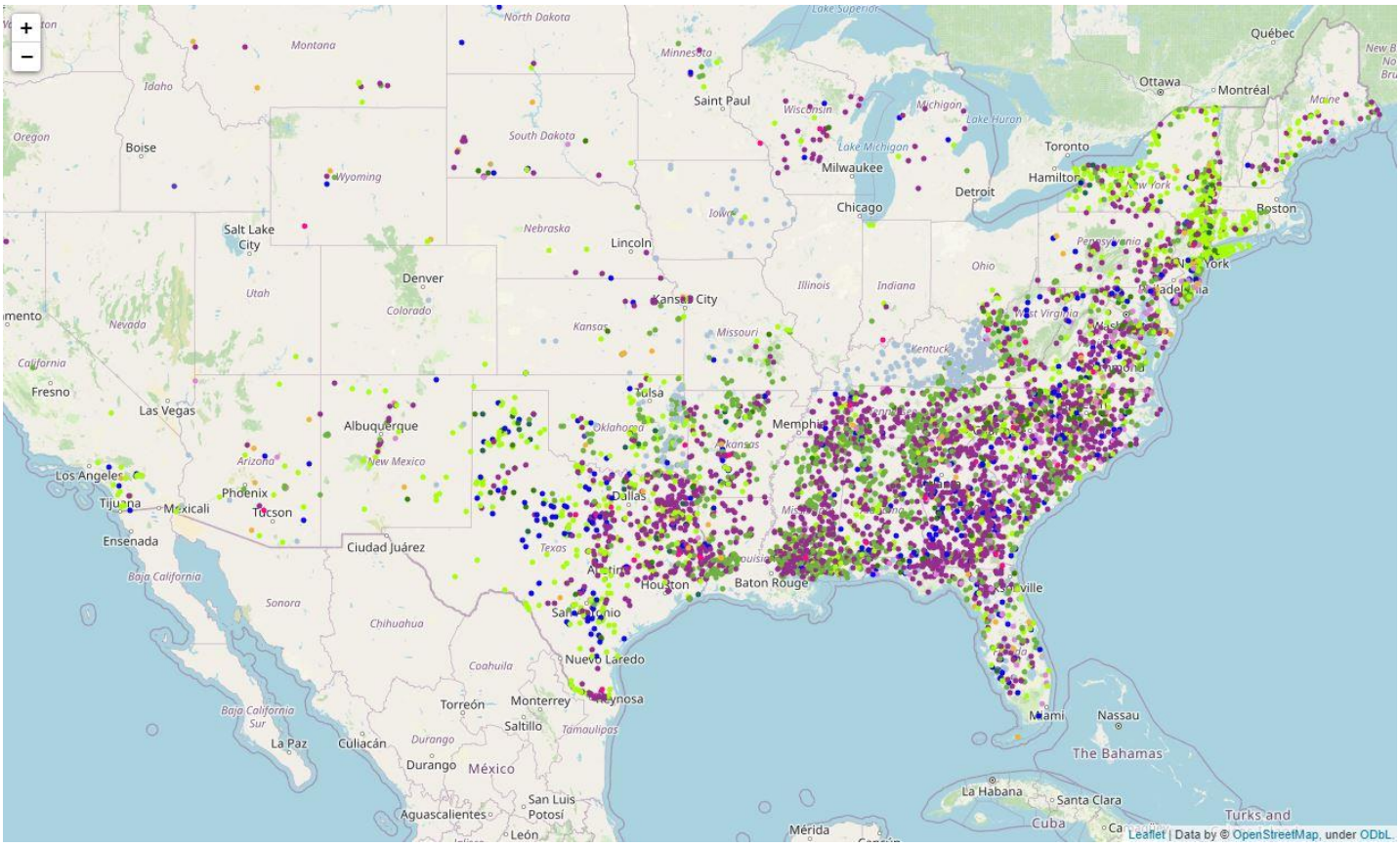
- Color Legend:

Bar-plot of the causes count in July 2006:



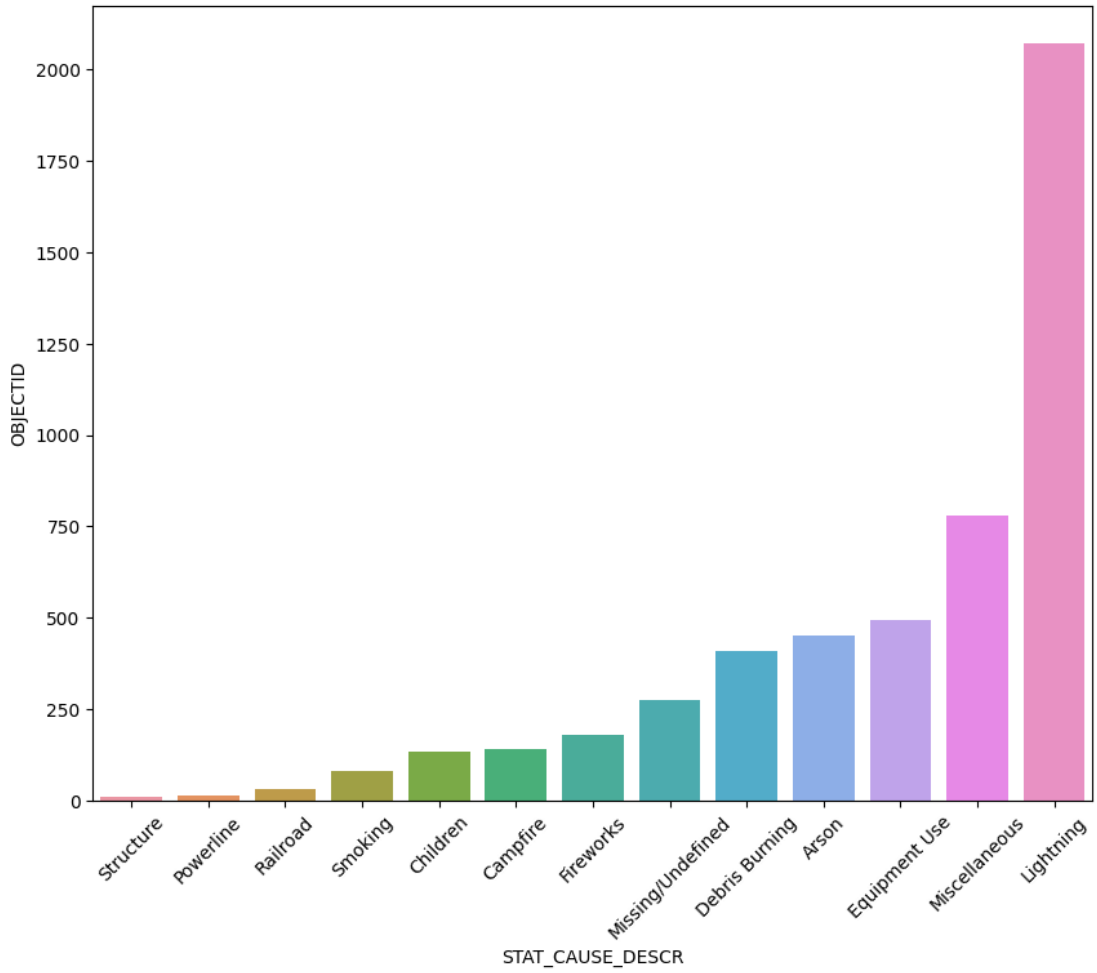
Similar maps for March 2006





Color Legend:

Bar-plot of the causes count in March 2006:



Pre – Processing

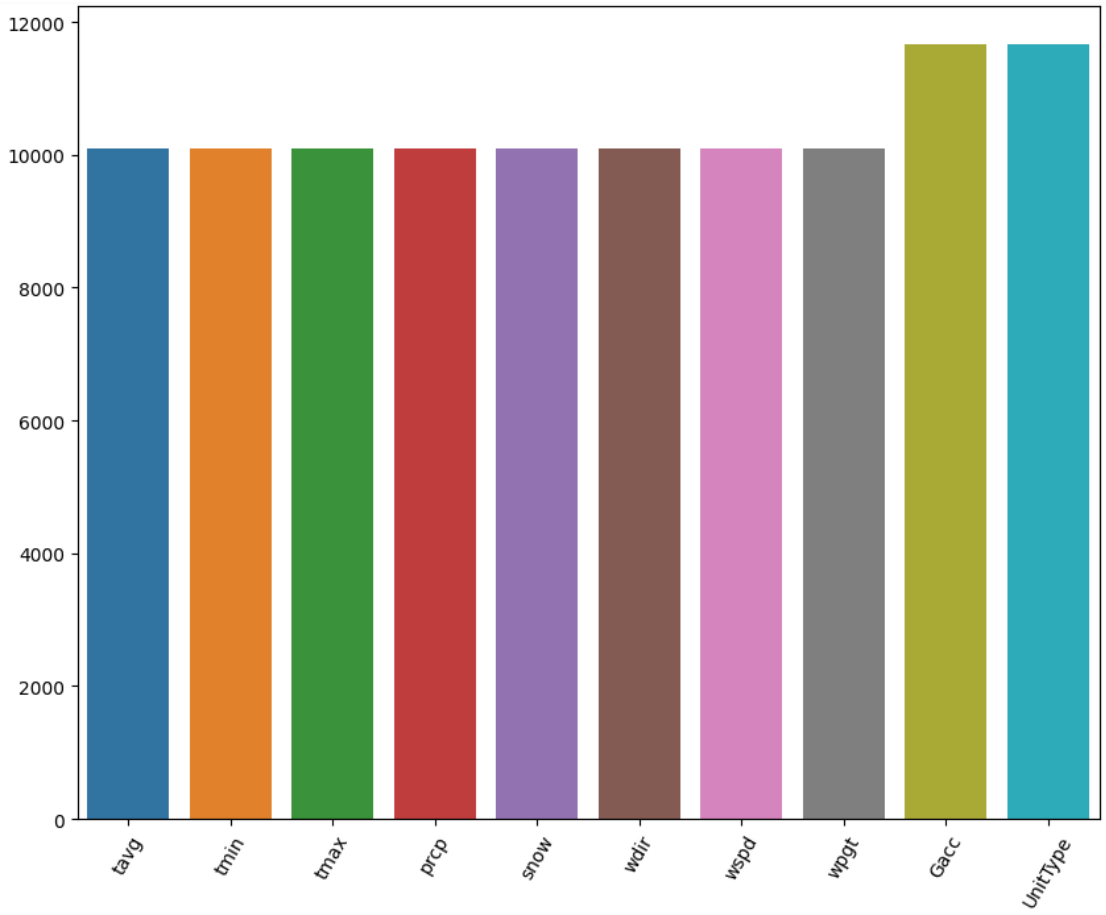
Integration of External Data

- We decided to integrate the existing data with two outer data frames: **weather data** and **parks data**.
- **Weather data** – data containing details about the weather in the USA in the relevant time. As far as we were concerned it was reasonable wildfires would happen in hot weather, and lightings would happen in cold weather. Thus, we thought it would be a good idea to add this kind of data frame.

- **Note about stage 2**: for some days, the requested data was missing. In such cases, we used linear interpolation (built in pandas function) to fill measurements data. Moreover, for some stations, the requested data was not available at all - in such case, the station was removed from consideration.
- **Parks data** – data containing number of national parks in each state of the USA. We thought it is likely in states with lots of parks, there would be more fires in comparison to stated with little amount of parks. As a result, we decided to add this data frame.

Handling Missing Values

- **Note:** some of the procedures in the pre-processing stage were done **before** splitting the data into train and test sets. We only performed the "one-hot" encoding in such manner, as we knew that the encoded values are of a finite, pre-known, set of values. For instance, we knew all the possible values for the "state" feature, so there's effectively no harm in that.
- A graph of **number of missing values as function of the features**:



- As visible in the above plot, the majority of the missing values were in the **'Gacc'** and **'UnitType'** columns. While the weather columns had small, insignificant number of missing values. Therefore, we decided to impute the missing values for 'Gacc' and 'UnitType' and drop the samples with missing weather data(**'wpgt', 'wspd', 'wdir', 'snow', 'prcp', 'tmax', 'tmin' and 'tvag'**).
- **Note:** the imputation was done separately on the train and test sets to avoid information leakage. Hence, as a first step - we split the data to train and test sets. In a later stage, when we process the Categorical data using one-hot encoding we would want to use the full data to avoid feature mismatch between the train and test sets. Having said that, we wish to stress out that the train and test sets are determined only once and only in this section. Even if we used the full data in later stages, we did so in a leakage-safe manner, where the train and test sets contained the same samples throughout the entire process.

Aggregative Features

We decided to add the following features, using aggregation:

1. Average number of acres (FIRE_SIZE) consumed by wildfire per **Gacc**, per year, named: "**fire_size_mean**". As far as we were concerned, it is likely for wildfires to occur in a location where a large number of acres are being consumed by wildfire in a year. So, this was our way to insert this sort of information to the data.
2. Number of fires per Gacc per year, named: "**fire_gacc_count**". It is possible that some locations have higher likelihood for wildfires over others. So, to 'capture' this sort of information, we measured the fire frequencies in different 'Gacc' areas.
3. Number of fires per reporting unit, named: "**fire_unit_count**". The purpose of this feature is to capture the impact of the reporting unit on the possible cause.
4. Average temperature per Gacc per year, named: "**gcc_mean_year_temp**". The purpose of this feature is to capture climate differences in different geographic ares.

Handling Categorical Data

- We split the categorical data to **serial data** and **non – serial data.**
  - **The serial features are**: 'SOURCE_SYSTEM_TYPE', 'UnitType'.
  - **The non serial features are**: 'OWNER_CODE', 'Gacc', 'Agency', 'STATE'.
- **Serial Features:** We decided to apply the **following order** on 'SOURCE_SYSTEM_TYPE': NONFED > FED > INTERAGCY. Moreover, we decided to apply the **following order** on 'UnitType': Tribe > US County/Local > US Federal > Interagency.
- **Non – Serial Features:** We used **one-hot encoding** as learned in class.

Adding 'prior' knowledge to the causes distribution

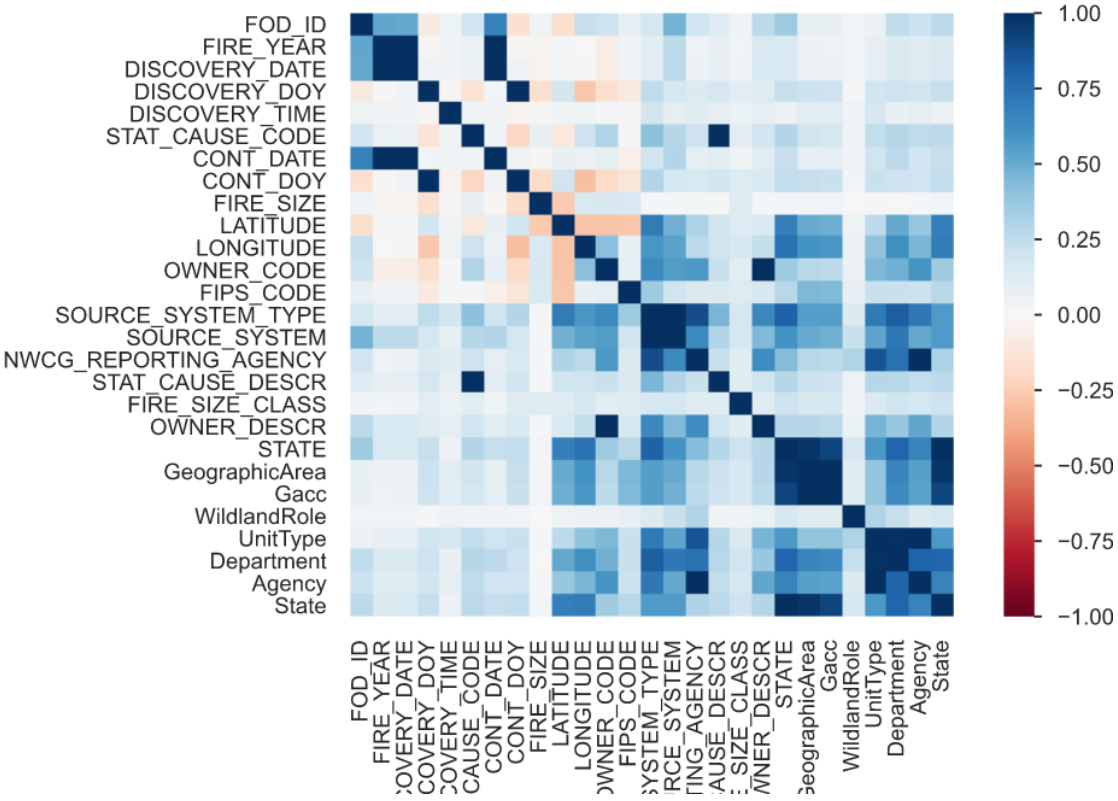
- **Note:** The imputation and addition of the prior and aggregative features are done after splitting train and test.
- In this section, we used the fact that our data is quite large to obtain some information regarding the causes' distribution. We then plug this information to the model to, hopefully, boost the classification performance
- The following process was implemented in a separate script in order to achieve what was mentioned in the previous point:

  1. First, we split the data into train and test sets.

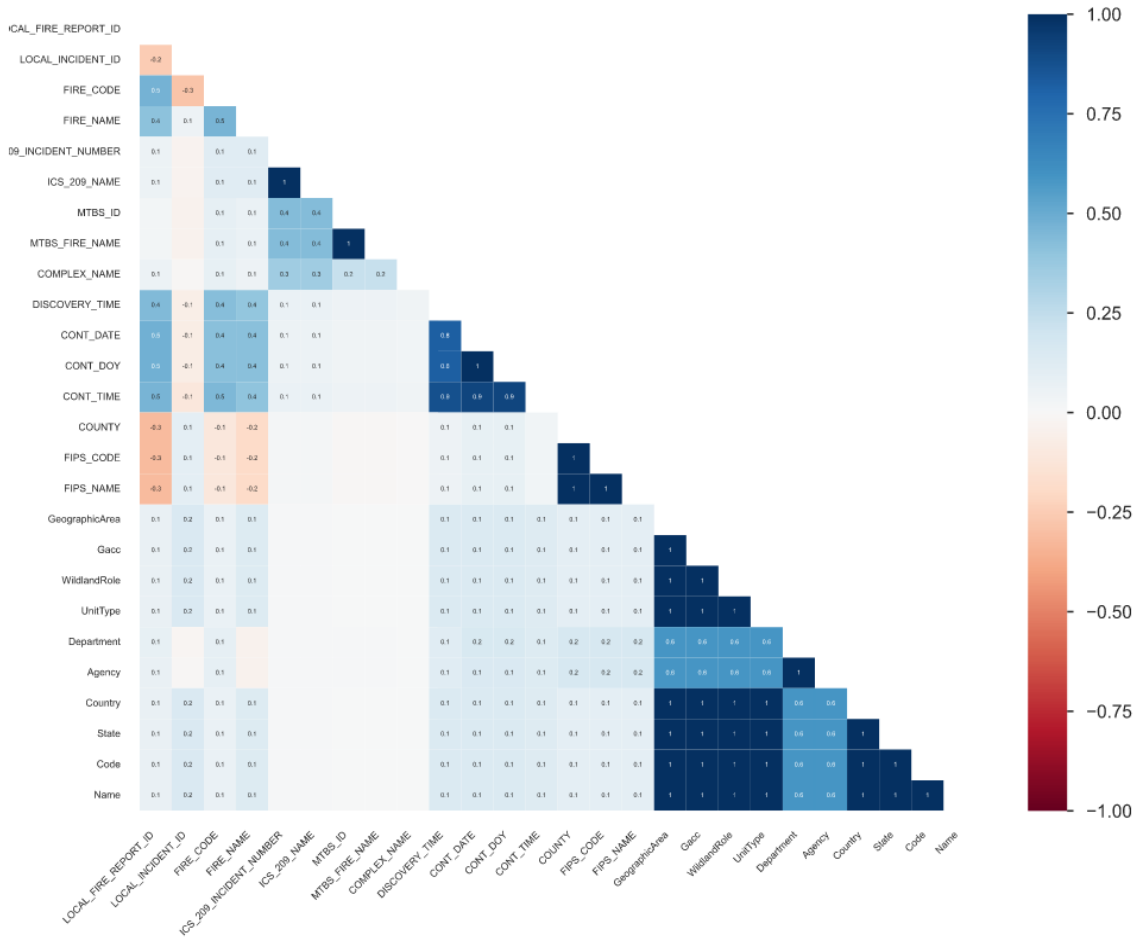
2. Afterwards, we took a small slice piece of train set (about 30,000 samples) and dropped it altogether to avoid inserting bias to the training stage.
3. On the sliced data, we computed the frequency ratio of each cause for each (unit, year) pair.
4. After computing the mentioned values, we plugged them into the train and test sets.

Features Selection

- We wanted to check correlation between features in order to understand which features to choose for our model.
- Correlation Heatmap:



- It can be seen the features: 'Agency', 'Department', 'Unit_Type' are correlated. Thus we decided not include all of them in our model.
- Furthermore, it can be seen 'Fire_Year' and 'Discovery_Date' are correlated, and 'SOURCE_SYSTEM_TYPE', 'SOURCE_SYSTEM' are correlated.
- Missing values Correlation Heatmap:



- It can be seen the features: 'GeographicArea', 'Gacc','WildLandRole' and 'Unit_Type' are missing values - correlated.
- Moreover, it can be seen 'Country', 'State' and 'Code' are missing values – correlated.

Model Selection

- After experimenting with different models, we decided to use **Random Forest Classifier** to predict the cause of the fires.
- Throughout the process, we experimented with linear models as well. However, we decided to abandon them in favor of tree-based models.
- The main reason for this choice is our strong belief that the wildfires causes are strongly impacted by weather conditions, and season changes. These factors are inherently cyclic; hence, it is safe to assume that linear models will not out-perform the tree-based models.
- The secondary reason for this choice of ours is runtime considerations. As we experimented with different models, we learned that the training time required for linear models is significantly higher than what is required for tree-based models.
- So, having said all that, we decided to focus and compare tree-based models.
- Before settling on the random forest classifier, we decided to test another classification model: **XGB Classifier**. Here we present it's performance:

	precision	recall	f1-score	support
Arson	0.55	0.49	0.52	23400
Campfire	0.50	0.32	0.39	6662
Children	0.40	0.11	0.18	5220
Debris Burning	0.51	0.74	0.60	36307
Equipment Use	0.42	0.25	0.32	12434
Fireworks	0.56	0.46	0.51	1013
Lightning	0.74	0.84	0.79	24277
Miscellaneous	0.52	0.52	0.52	27309
Missing/Undefined	0.87	0.89	0.88	11414
Powerline	0.48	0.14	0.22	1268
Railroad	0.48	0.40	0.44	2765
Smoking	0.28	0.02	0.04	4462
Structure	0.66	0.24	0.36	397
accuracy			0.58	156928
macro avg	0.54	0.42	0.44	156928
weighted avg	0.56	0.58	0.56	156928

- We learned that the XGB Classifier yields the same performance as the RF model at best, so we decided to go with the RF model as it is simpler and thus takes less time to train.
- Note that before between the models, we performed Bayesian HPO on both of them.
- Finally, we present the classification results of our RF model on the test set:

	precision	recall	f1-score	support
Arson	0.52	0.50	0.51	22811
Campfire	0.48	0.33	0.39	6396
Children	0.30	0.15	0.20	4987
Debris Burning	0.52	0.67	0.58	35596
Equipment Use	0.35	0.25	0.29	11495
Fireworks	0.58	0.44	0.50	994
Lightning	0.74	0.85	0.79	23591
Miscellaneous	0.51	0.50	0.51	26174
Missing/Undefined	0.91	0.90	0.91	13708
Powerline	0.34	0.14	0.20	1252
Railroad	0.50	0.42	0.46	2580
Smoking	0.18	0.06	0.09	4278
Structure	0.55	0.21	0.30	387
accuracy			0.57	154249
macro avg	0.50	0.42	0.44	154249
weighted avg	0.56	0.57	0.56	154249

For the train set:

	precision	recall	f1-score	support
Arson	0.82	0.79	0.80	65903
Campfire	0.85	0.59	0.70	18641
Children	0.88	0.35	0.50	14565
Debris Burning	0.71	0.91	0.80	101964
Equipment Use	0.81	0.62	0.70	35013
Fireworks	0.87	0.62	0.72	2836
Lightning	0.84	0.93	0.88	67654
Miscellaneous	0.76	0.80	0.78	76917
Missing/Undefined	0.95	0.97	0.96	39442
Powerline	0.95	0.31	0.47	3530
Railroad	0.72	0.57	0.64	7797
Smoking	0.93	0.18	0.30	12468
Structure	0.97	0.31	0.47	1082
accuracy			0.79	447812
macro avg	0.85	0.61	0.67	447812
weighted avg	0.80	0.79	0.78	447812

We can infer by the performance on the train set that our model hasn't reached over-fitting, and thus is more likely to better generalize.

Hyper - parameters tuning

- We used Bayesian HPO in order to optimize the following hyper parameters: "estimators","min_samples_leaf" and "min_samples_split".
- Visualization of the result we got, after hyper - parameters tuning:

	precision	recall	f1-score	support
Arson	0.55	0.50	0.52	23400
Campfire	0.54	0.31	0.39	6662
Children	0.41	0.11	0.17	5220
Debris Burning	0.52	0.72	0.60	36307
Equipment Use	0.41	0.25	0.31	12434
Fireworks	0.60	0.36	0.45	1013
Lightning	0.73	0.85	0.79	24277
Miscellaneous	0.52	0.54	0.53	27309
Missing/Undefined	0.88	0.88	0.88	11414
Powerline	0.58	0.09	0.16	1268
Railroad	0.49	0.40	0.44	2765
Smoking	0.31	0.03	0.05	4462
Structure	0.75	0.19	0.30	397
accuracy			0.58	156928
macro avg	0.56	0.40	0.43	156928
weighted avg	0.57	0.58	0.56	156928