Capstone Project #1 –
Identifying counties that are the most fire-prone and predicting the cause of a fire wildfire.

Problem:

Forest fires are a major environmental issue, creating economic and ecological damage while endangering human lives. Fast detection is a key element for controlling such a phenomenon. Despite an increase in state expenses to control this disaster, each year millions of forest hectares are destroyed all around the world. In California alone, we witnessed some of the deadliest and most destructive fires in state history. Billions of dollars have been spent by the various agencies to control and extinguish the fires. Wildfires are costly events, in so many ways [1]. The objective of this potential project would be to identify Which state and counties are the most fire-prone and to predict the cause of a wildfire.

Clients:

The clients would be agencies responsible for controlling and extinguishing the fires including the Fire department. This model enables the appropriate organizations to take preventative action, such as cutting firebreaks, as well as informing planning and preparedness activities, such as where to store fire retardant. It can be used in ensuring that the front-line firefights are deployed to the right locations to have the maximum impact, while simultaneously minimizing the risk to their safety[2].

Data: 1.88 Million US Wildfires (Kaggle1)

This data publication contains a spatial database of wildfires that occurred in the United States from 1992 to 2015. It is the third update of a publication originally generated to support the national Fire Program Analysis (FPA) system. The wildfire records were acquired from the reporting systems of federal, state, and local fire organizations. The following core data elements were required for records to be included in this data publication: discovery date, final fire size, and a point location at least as precise as the Public Land Survey System (PLSS) section (1-square mile grid). The data were transformed to conform, when possible, to the data standards of the National Wildfire Coordinating Group (NWCG). Basic error-checking was performed, and redundant records were identified and removed, to the degree possible. The resulting product, 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 during the 24 years.[3]

To Clean the Data obtained from Kaggle I applied Data Wrangling techniques. As the dataset was in the SQL file format, I did some cleaning activity in SQLite. There were many columns in SQL table "fires" which were related to the Source of data and agency information that collected or prepared reports on Data. I omitted these columns and imported only relevant data to the "US_Wildfire_data.csv" file. Also while importing I filled in all the null/blank values with "Nan" value.

By looking at the info of the data frame I observed many columns with more than 50% of its total data with null values. As those columns were not helping the data set I omitted those columns. As County_Name is the column which I am interested in, I checked if there are any Nan values in that column and found many 'Nan' values. I Filled the missing values of the COUNTY_NAME column using 'LATITUDE' and 'LONGITUDE' columns. I Retrieved CountyName by Passing Latitude and Longitude Value as tuple to the "reverse_geocoder.search" method as a parameter.

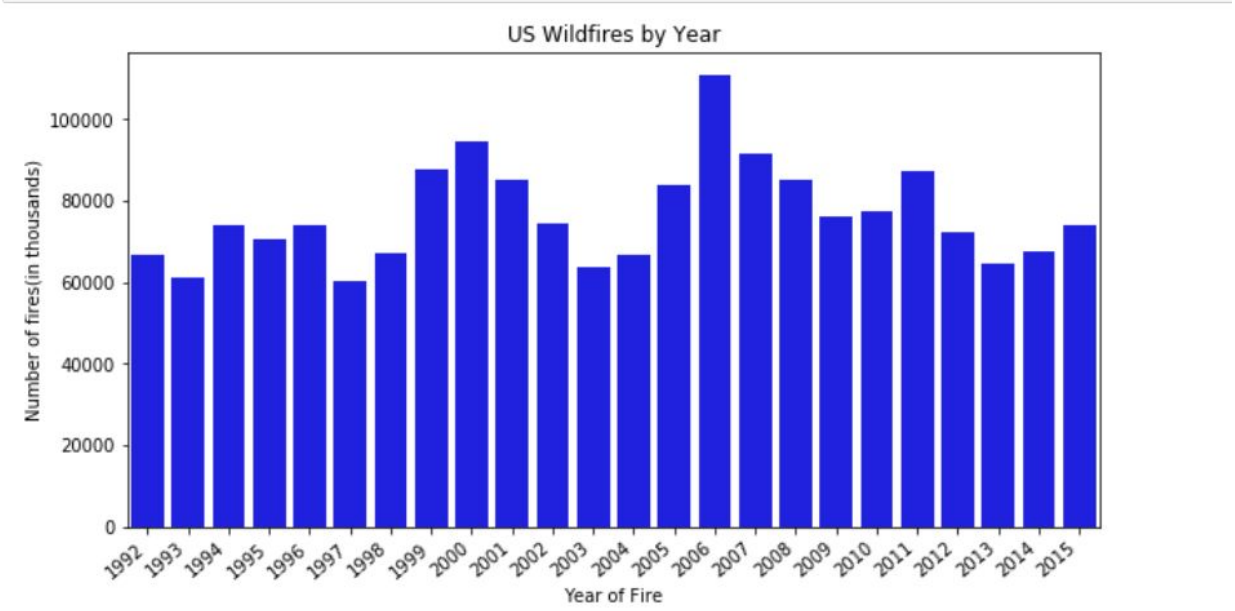
At first look, I observed a few of the county names were not matching with expected states in the "noCountyNameDF" DataFrame.(Note: When latitude and longitude lies in the US state border this discrepancy of value is observed). To get rows with the mismatching State value, a new column is mapped with State name and State Code as a dictionary with the former being key. Though there were no 'Nan' values anymore, column "DISCOVERY_DATE" was in data type float64 and not in Datetime format as expected. Hence I converted it to a readable date format from Julian date. I also Created a column FIPS_CODE with FIPS code value for each County Name.

inally, I checked if there were any outliers concerning the FIRE_YEAR column and found there were no outliers that needed to be discarded.

By exploring the cleaned US Wildfires data we would try to answer below questions

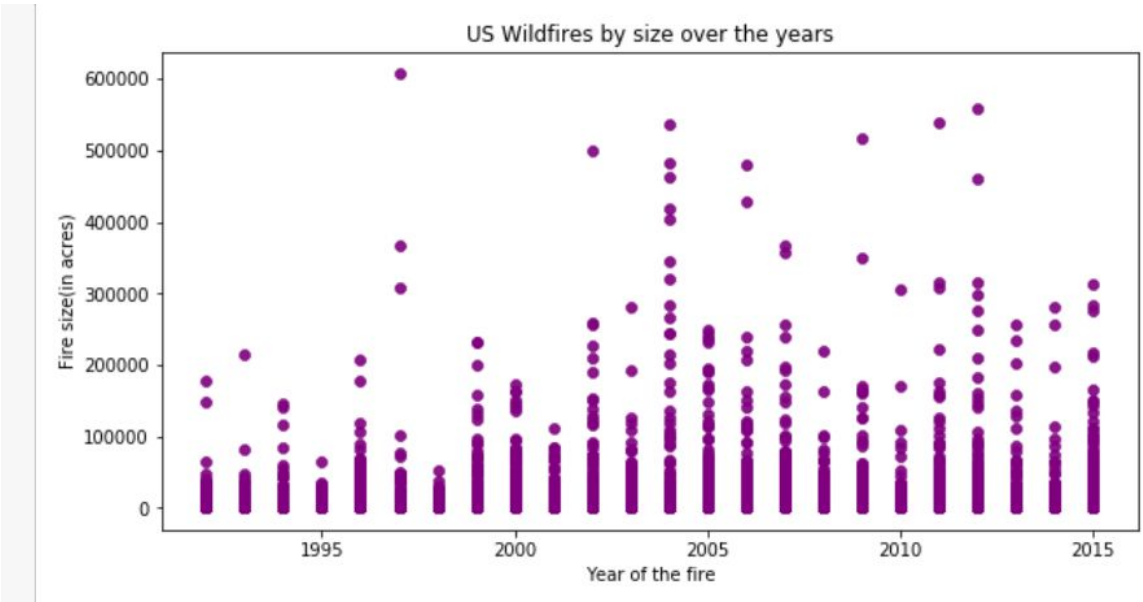
1. Is Global warming affecting the number of fires? Has the number of fires increased over the period 1992-2015?
2. Are we able to limit fire spread with the help of growing technology? Has the size of fire decreased over the years?
3. What causes the most fires? Which causes are associated with larger wildfires?
4. Which state in the USA is most affected by WildFires? Which county is more prone or less prone to WildFire?
5. What is causing more fires in each state?

Let us see trends in the number of fires throughout 1992-2015.



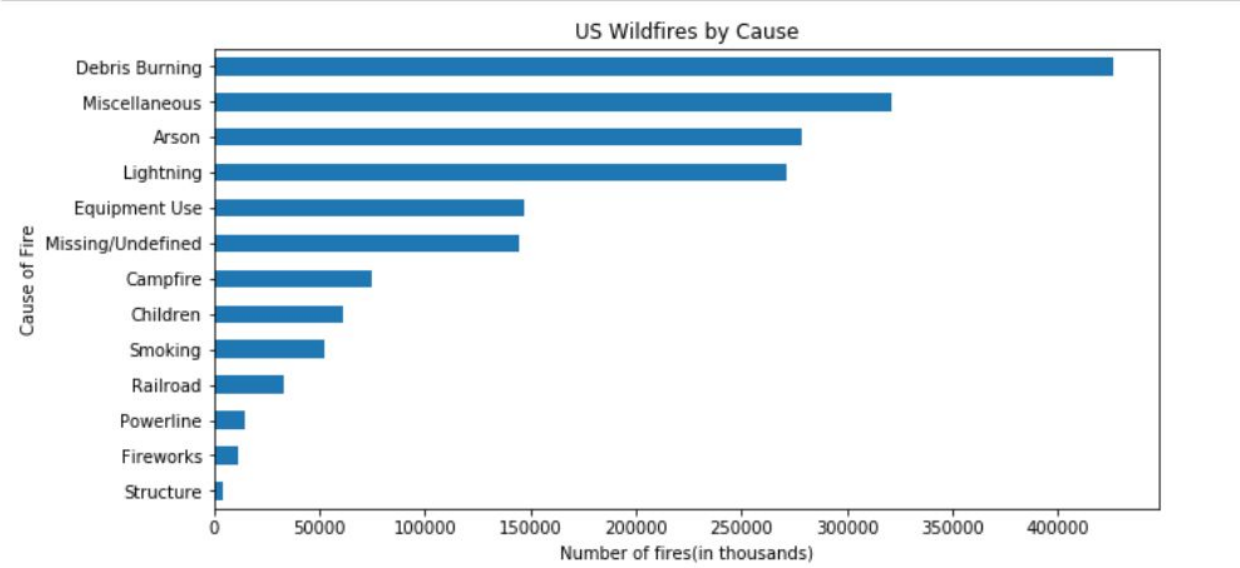
The number of fires per year ran between 60,000 and 100,000 from 1992 to 2015. There was a spike in fires in 2006. Though we can see a small upward trend at certain periods, there is no continuous upward trend over the period.

Now to see if concerned departments can restrict the fire size with the help of modern technology over the years, plotting fire year against fire size.

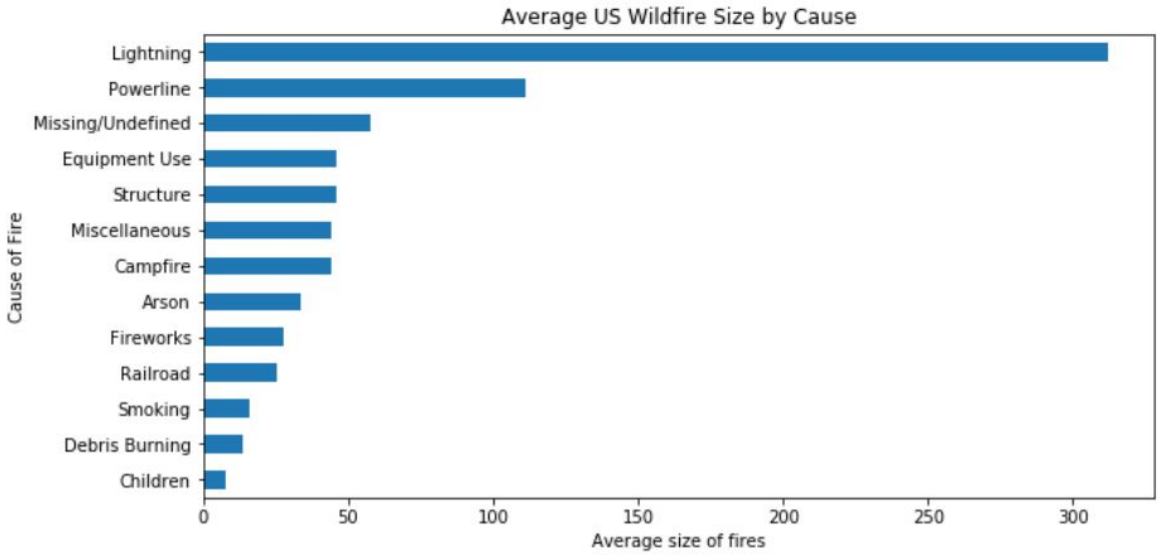


As we can see, there is no decrease or increase in fire size from 1992-2015. Every year has recorded fires of different sizes.

Now that we have seen that number of fires or its size has not changed over the years. Let us look at the cause of these Wildfires. What causes the most fires? Which causes are associated with larger and longer-burning wildfires?

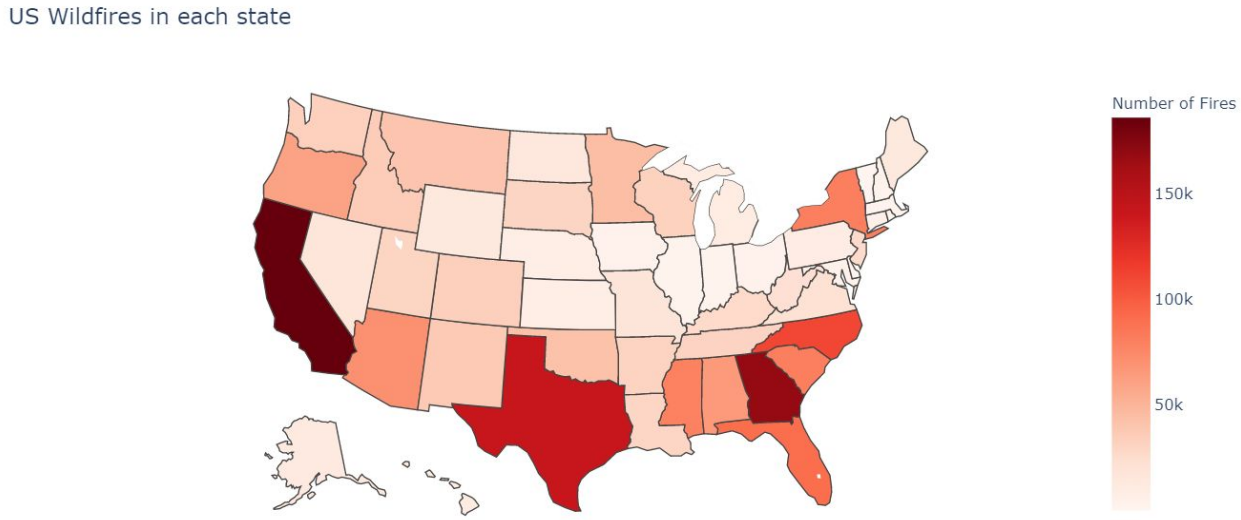


Trash burning was the largest cause of wildfire by a significant margin. Interestingly, slightly more fires were started by arson than by lightning. I am surprised that one of the causes is just ‘children’.
This leads to a question. Is there a relationship between cause and fire size?



We can observe that fires started by electricity are the most damaging. Though the number of fires caused by debris burning is large in number, the average size of the fire caused by it is very low.

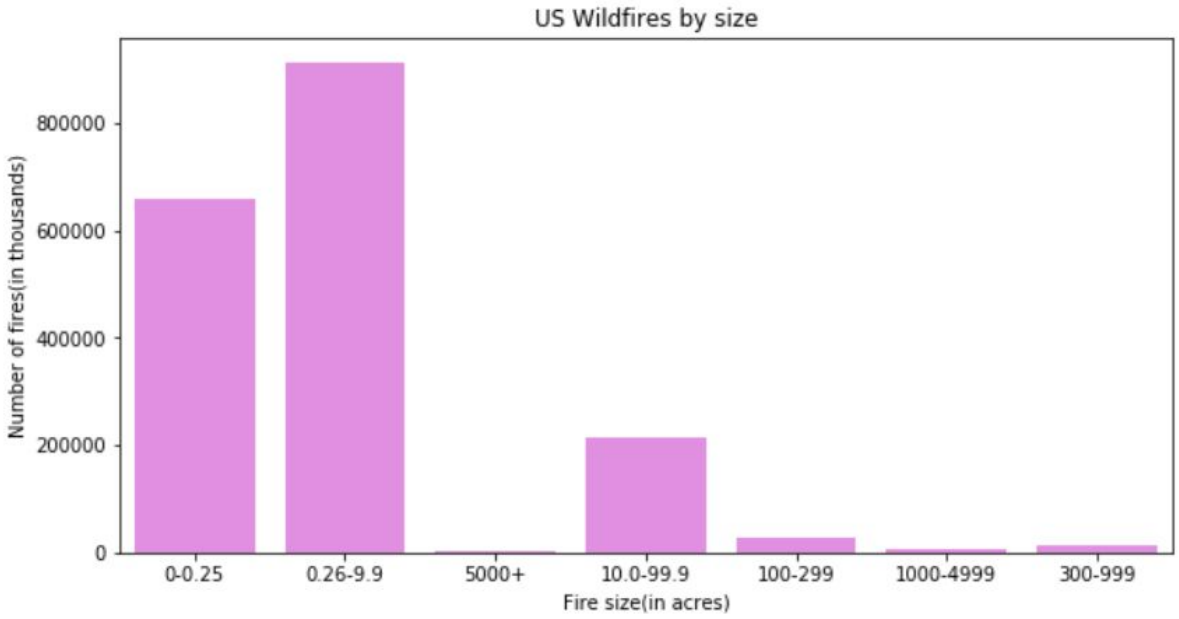
To answer questions like Which State in the USA is most affected by WildFires? Which county is more prone or less prone to WildFire? Let's compare the number of wildfires by state.



As we might expect California and Texas have the most wildfires due to sheer size and climate. However, it is surprising to see Georgia with so many fires.

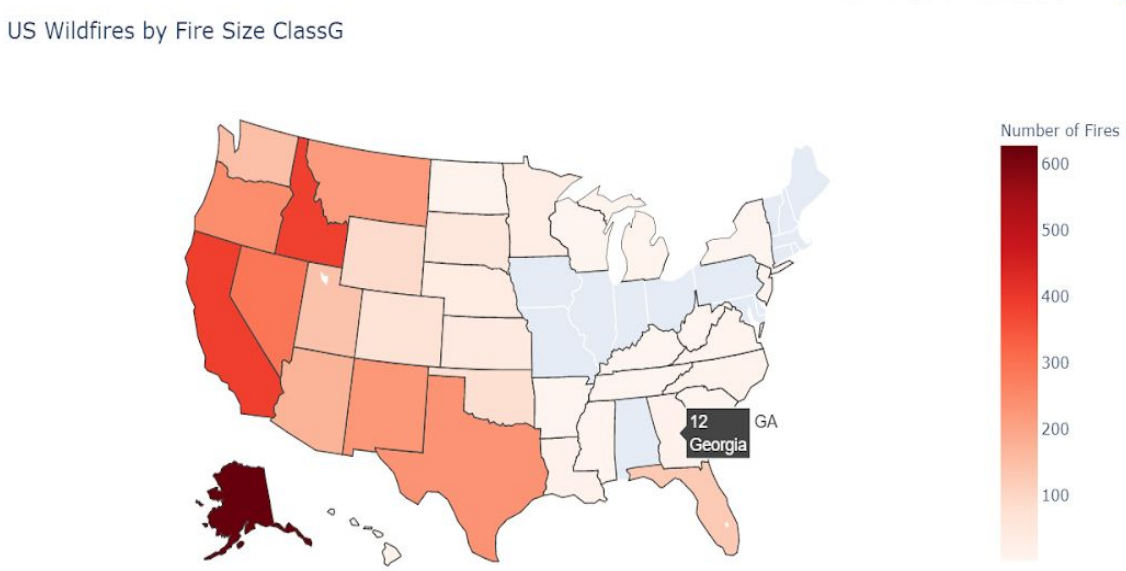
Though Georgia reported more fires than other bigger states, I am curious about the Fire size of fires reported in Georgia.

From DataSet we know that **Size 'A' represents '0-0.25' acres,Size 'B' represents '0.26-9.9' acres,Size 'C' represents '10.0-99.9' acres,Size 'D' represents '100-299' acres,Size 'E' represents '300-999' acres,Size 'F' represents '000-4999' acres and Size 'G' represents '5000+' acres.**



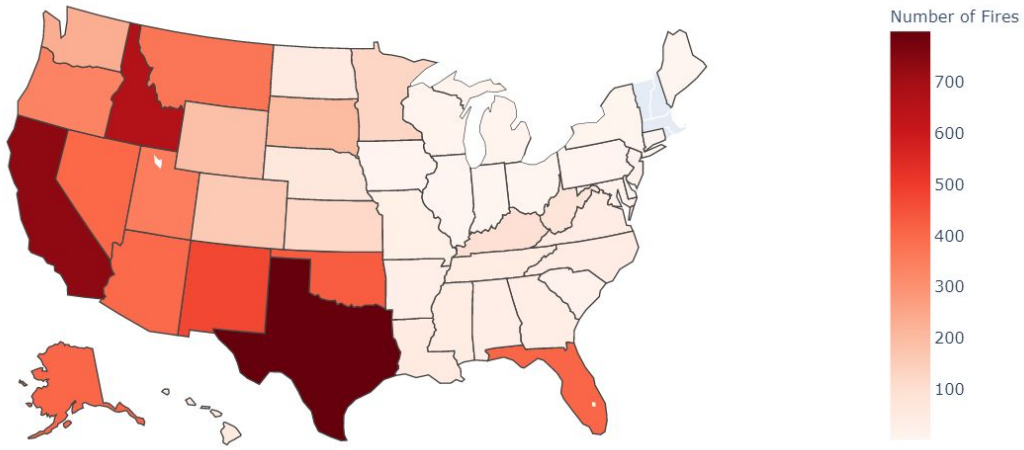
We can see in the above graph that over 800k of total fire incidents are in between 0.26-0.99 acres within the final fire perimeter expenditures.

To know more about this let us compare the number of fires per state against fire size reported.

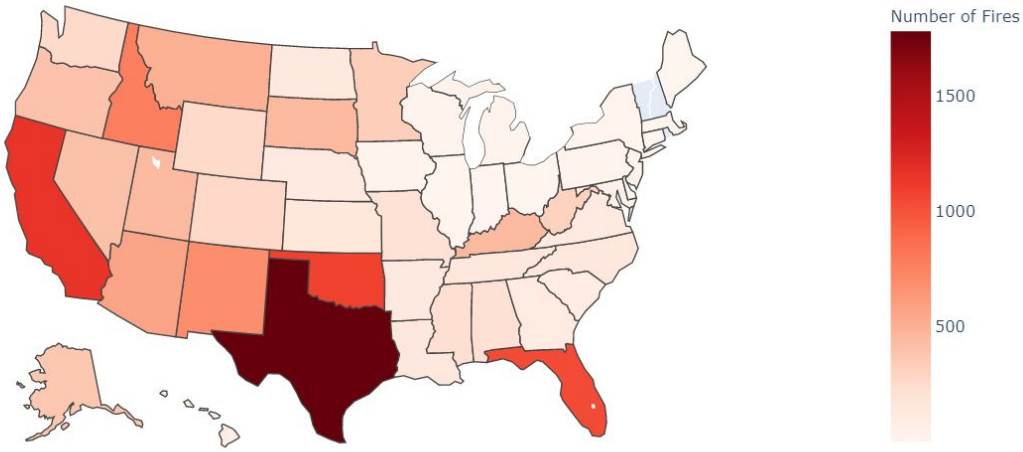


We can see that Georgia has 12 fires over two decades which are of size ‘5000”’ acres and Alaska has the most number of fires of the largest size.

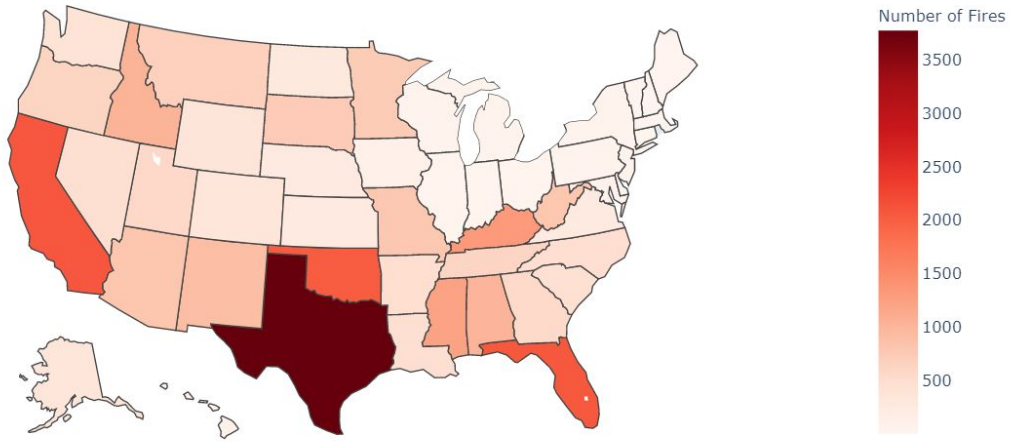
US Wildfires by Fire Size ClassF



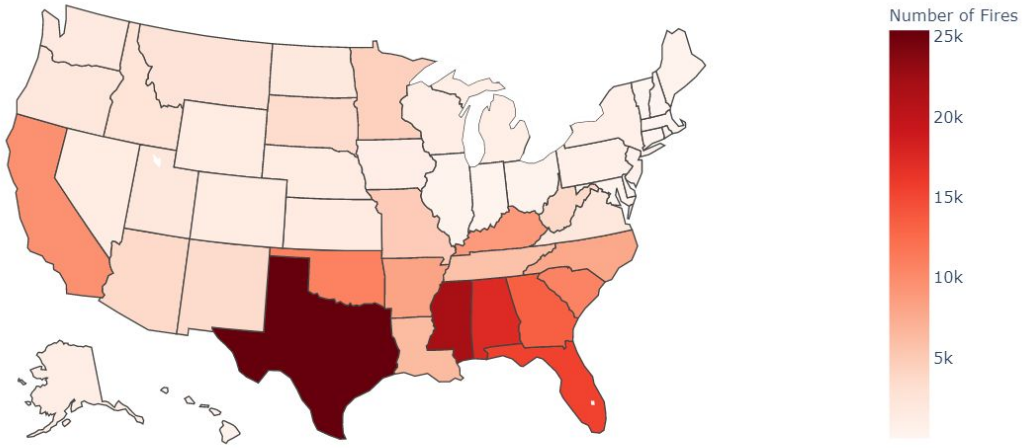
US Wildfires by Fire Size ClassE



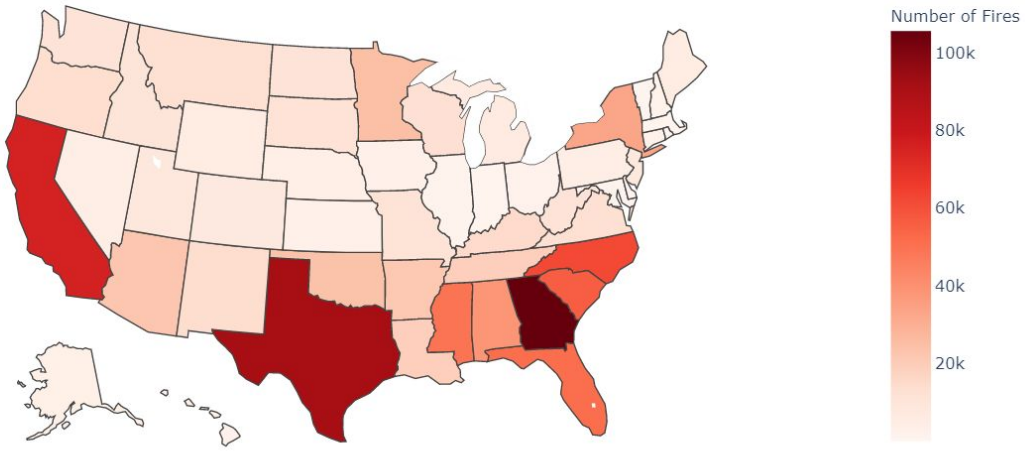
US Wildfires by Fire Size ClassD



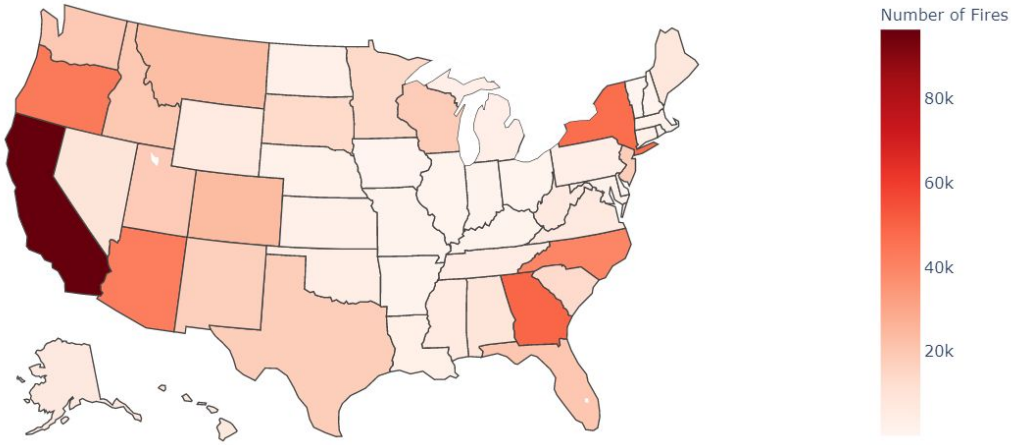
US Wildfires by Fire Size ClassC



US Wildfires by Fire Size ClassB



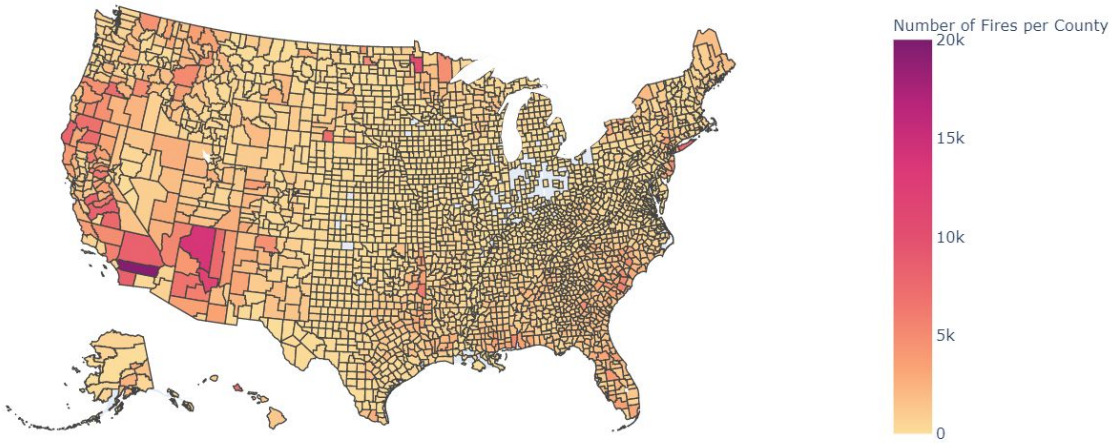
US Wildfires by Fire Size ClassA



After looking at State maps by Fire size class, we can note that Georgia has most fires with size under 100 acres. However, California and Texas have reported fires of all sizes in more numbers over two decades.

Now let's look at one of the main questions we are trying to answer. Which County is more prone to wildfire.

US Wildfires by County



	FIPS_CODE	County_Count	COUNTY_NAME	STATE_NAME
0	6065.0	19398	Riverside	California
1	4005.0	13864	Coconino	Arizona
2	4007.0	11662	Gila	Arizona
3	27007.0	10518	Beltrami	Minnesota
4	6071.0	8301	San Bernardino	California
5	6073.0	8271	San Diego	California
6	6019.0	8229	Fresno	California
7	6023.0	7963	Humboldt	California
8	6047.0	7894	Merced	California
9	36103.0	7663	Suffolk	New York

From the above map, we can know that “Riverside County” in California has more fires. .“Coconino” and “Gila” in Arizona, “Beltrami” in Minnesota, “San Bernardino” in California, and “San Diego” in California come next in the top list of counties. Also, we can see that counties like “Hamilton-Indiana” have only 1 case throughout 1992-2015. We can see that Riverside County of California has the highest number of fire incidents. However, we cannot assume that more fires have caused more damage. Let us see which counties are most affected in terms of acres of land burnt due to wildfire.

	FIPS_CODE	FIRE_SIZE_mean	FIRE_SIZE_min	FIRE_SIZE_max	COUNTY_NAME	STATE_NAME
91	2290.0	34400.542941	0.10	312918.3	Yukon-Koyukuk	Alaska
910	20187.0	13467.000000	1.00	40000.0	Stanton	Nebraska
2081	40153.0	9360.666667	604.00	23488.0	Woodward	Oklahoma
823	20007.0	9083.589286	0.25	70000.0	Barber	Kansas
1540	30073.0	8983.950000	0.10	53640.0	Pondera	Montana
71	2068.0	7770.115425	0.10	517078.0	Denali Borough	Alaska
72	2070.0	7342.829783	0.10	606945.0	Dillingham Census Area	Alaska
70	2050.0	6050.744675	0.10	308120.0	Bethel Census Area	Alaska
2527	48263.0	6007.200000	1.00	162625.0	Kent	Rhode Island
2028	40045.0	5000.000000	5000.00	5000.0	Ellis	Texas

Now the picture of counties affected by wildfire has completely changed. Yukon-Koyukuk is the most affected county over the period. Also, it is interesting to note that though the number of fires was more in counties of California, Arizona, Minnesota, and New York, more destructive fires are reported in counties of Alaska, Nebraska, and Oklahoma.

```
grouped_countyFireData[grouped_countyFireData['COUNTY_NAME'] == 'Riverside']
```

	FIPS_CODE	FIRE_SIZE_mean	FIRE_SIZE_min	FIRE_SIZE_max	COUNTY_NAME	STATE_NAME
214	6065.0	30.995669	0.01	40200.0	Riverside	California

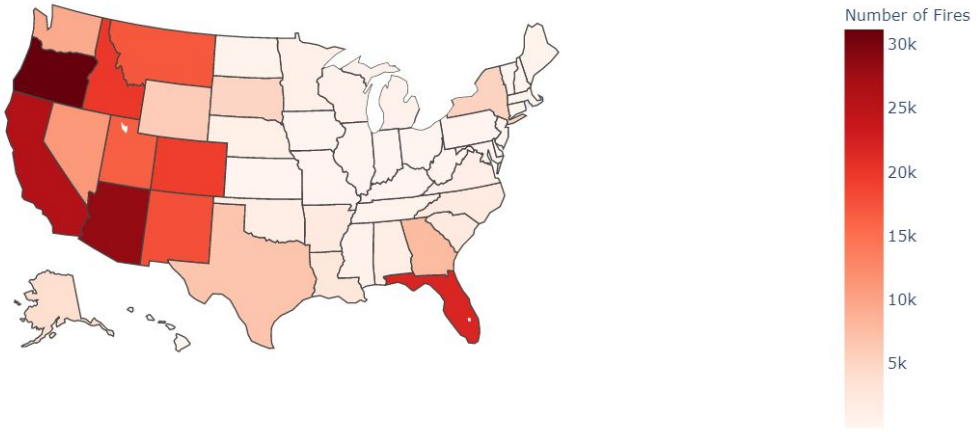
```
countyFireData[countyFireData['COUNTY_NAME']=='Yukon-Koyukuk']
```

	FIPS_CODE	County_Count	COUNTY_NAME	STATE_NAME
2228	2290.0	51	Yukon-Koyukuk	Alaska

Riverside County of California reported 19398 fire incidents over 13 years with the average fire size being approximately 31 acres. However, Yukon-Koyukuk county of Alaska reported only 51 fire incidents over the same period but with an average fire size of approximately 35000 acres.

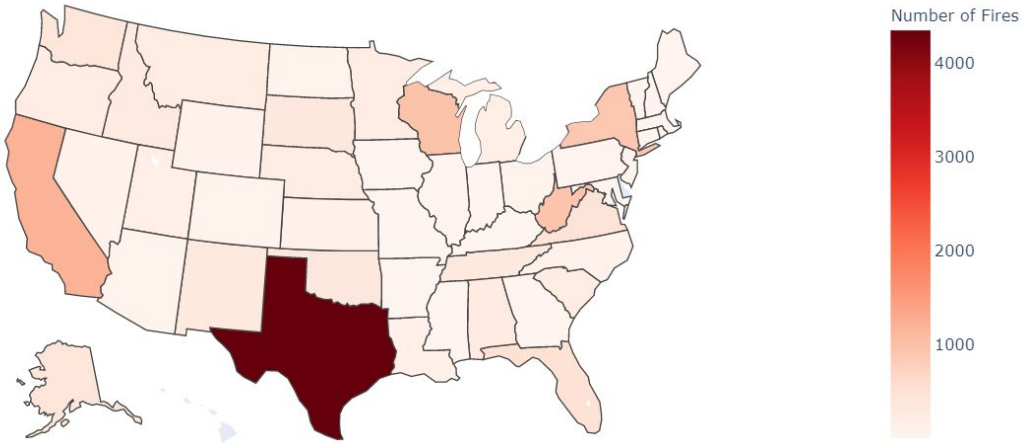
Though we see the number of fires per state and county in the above maps, it is not clear what is causing these fires in any given state. If we know the reason behind these wildfires, State/County departments can take appropriate precautionary actions depending on the most likely reason for the fire in their state/county.

US Wildfires caused by Lightning

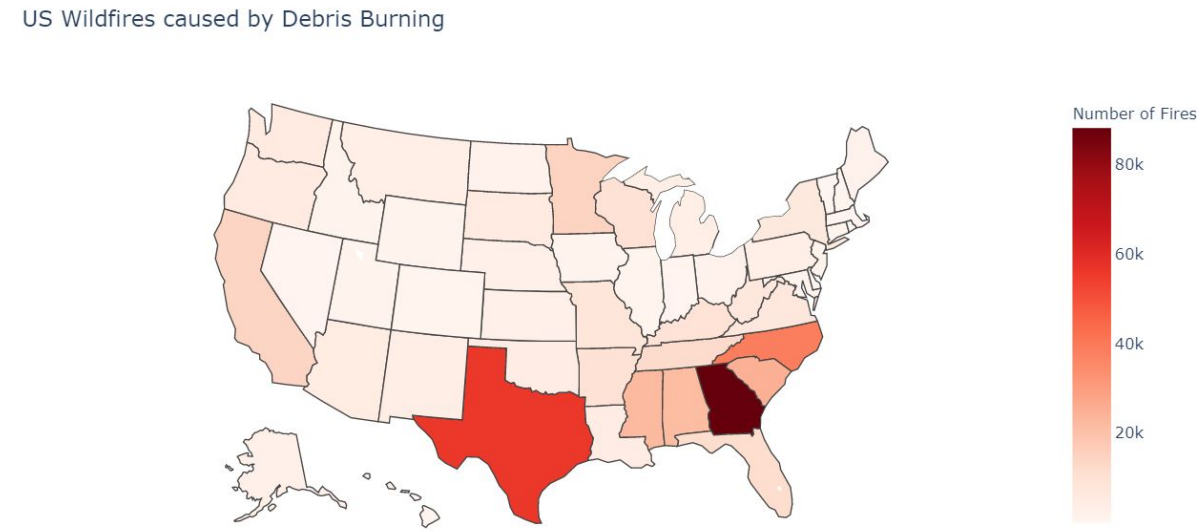


As we saw before Lightning is causing more destructive fires. This can be seen in most of the western states. We can note that other states of the USA have also suffered from wildfires caused by lightning.

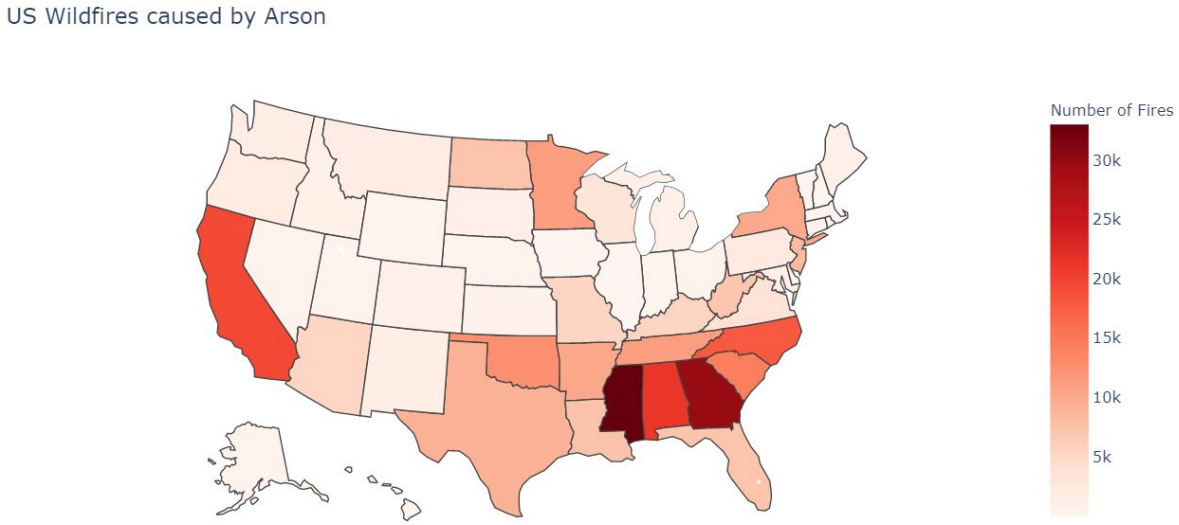
US Wildfires caused by Powerline



However, Powerline has caused around 4K wildfires in Texas. This is something that needs to be addressed at priority by the Texas government. Controlling ‘Lightning’ might not be under human control, But fixing the powerline problem definitely is.

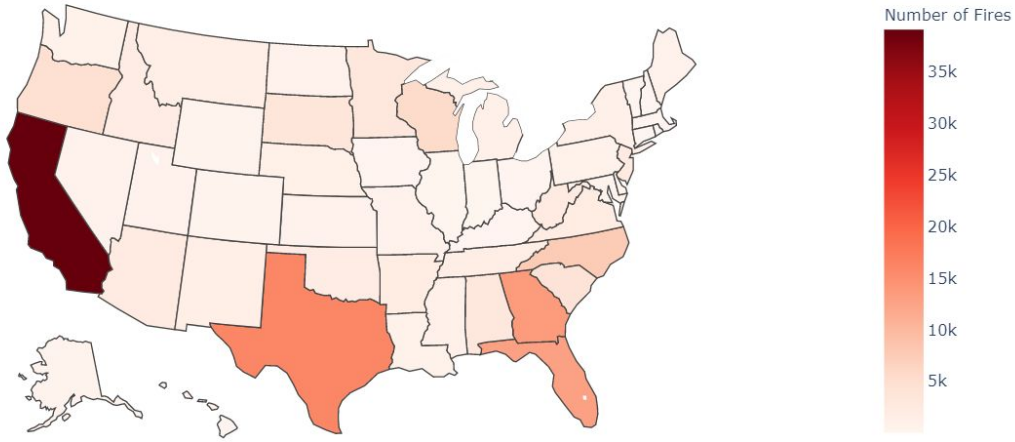


Looks like trash burning is the cause for almost all the cases in Georgia.

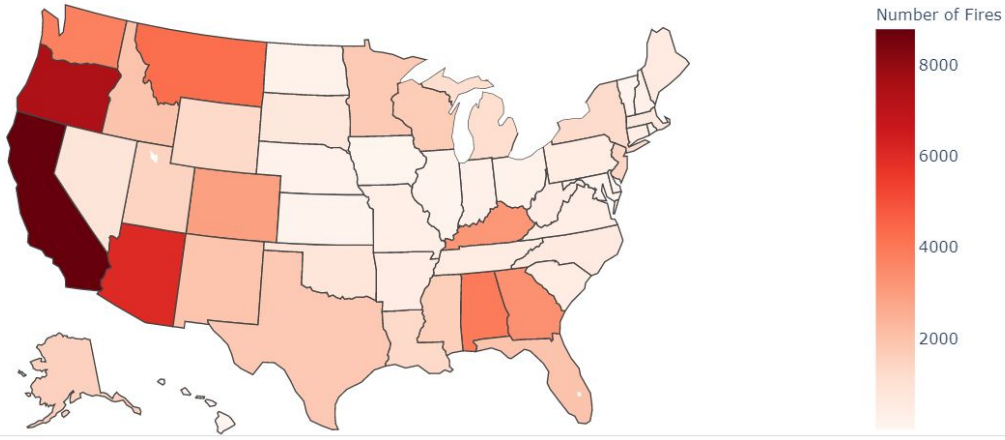


Mississippi has 30k wildfires caused by trash burning and arson alone.

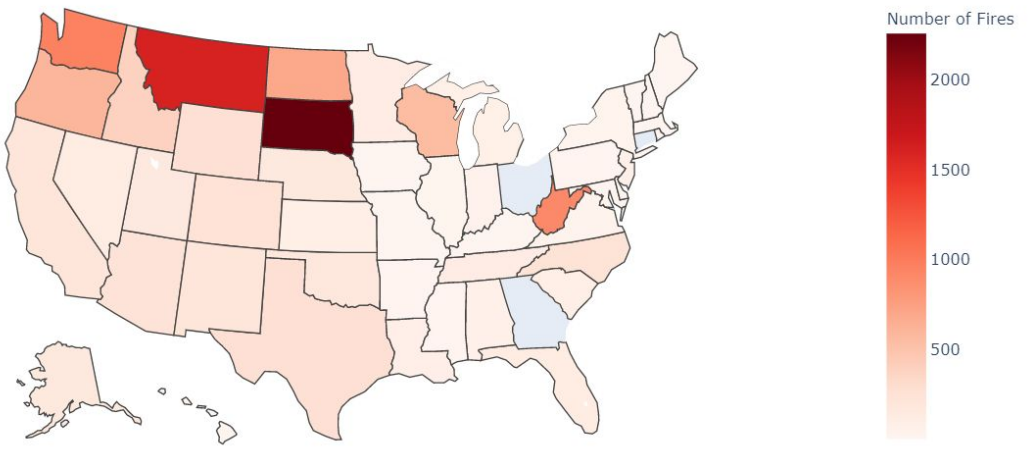
US Wildfires caused by Equipment Use



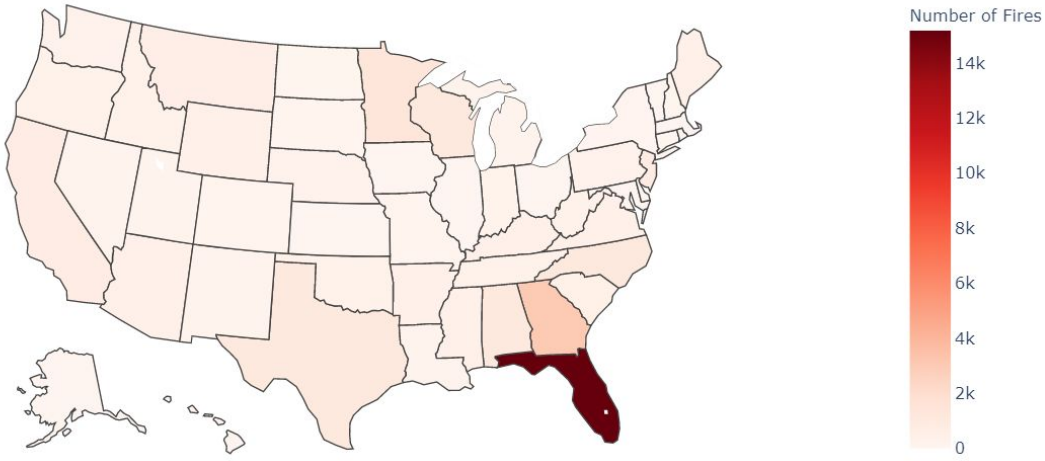
US Wildfires caused by Campfire



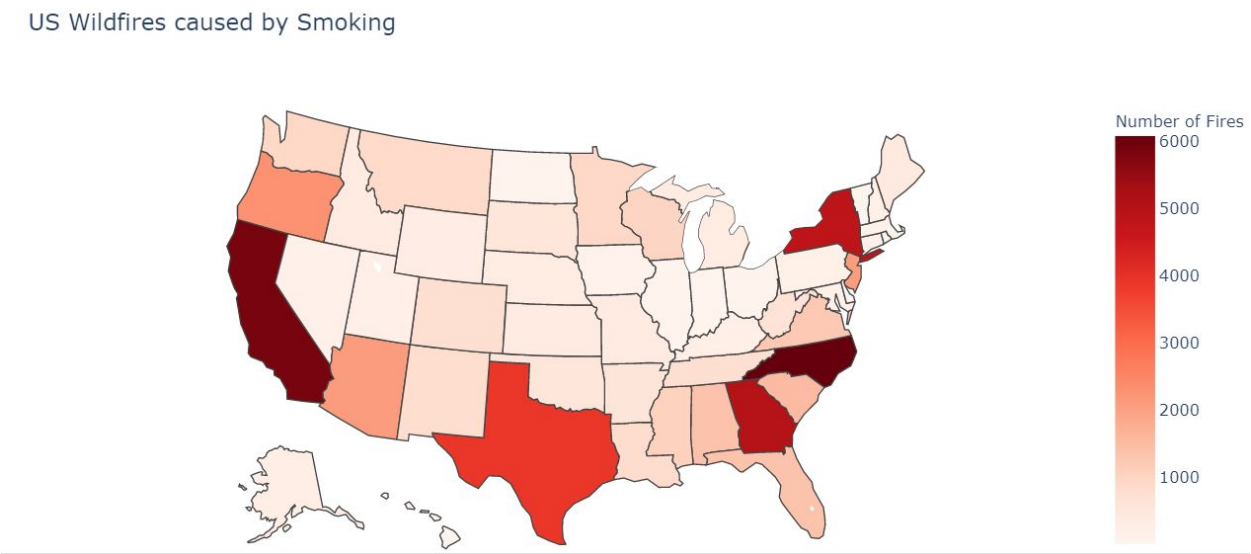
US Wildfires caused by Fireworks



US Wildfires caused by Railroad



Railroads are causing about 14K wildfires in Florida. Interestingly, it is most likely the only state affected by Railroad caused wildfires.



We have seen what has caused more fires over the years. Also, we did not see any direct increase or decrease in the number of fires over the years. Now the question I am trying to answer is "Is the cause of wildfire-related to season?" To answer this question I performed a chi-square test.

The following are the null and alternative hypothesis I have chosen.

H0(null hypothesis): The features Cause of Fire and Season are independent (which means they are not associated).
H1(alternate hypothesis): Cause of Fire and Season are not independent (which means they are associated).
Firstly, I created a Contingency table using the features in question. Also, I choose significance level **α = 0.05**

SEASON	Autumn	Spring	Summer	Winter
STAT_CAUSE_DESCR				
Arson	56818	118176	48426	54785
Campfire	18560	21397	27054	7778
Children	9320	26595	15606	9225
Debris Burning	73292	189394	60094	103616
Equipment Use	31556	41534	53747	19785
Fireworks	834	1076	9037	387
Lightning	31007	24724	213776	1640
Miscellaneous	63120	109267	99036	49440
Missing/Undefined	22874	50117	43742	27831
Powerline	2801	5092	4644	1847
Railroad	4152	14749	6906	7467
Smoking	10984	18179	15002	8274
Structure	833	1373	984	572

To perform chi-square tests we need **Chi-Square Statistic value, p-value, degree of freedom, and expected value.**

- The mathematical formula to create **Expected Frequency(value)** Is :
  E = (row total * column total)/grand total

- The formula to calculate **Chi-square value or X ²** is:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

  Where X is not the English alphabet we know but the 22nd Greek alphabet Chi.

  And X ² is the summation of the squared difference between Observed and Expected frequencies divided by the Expected frequency for all the cells

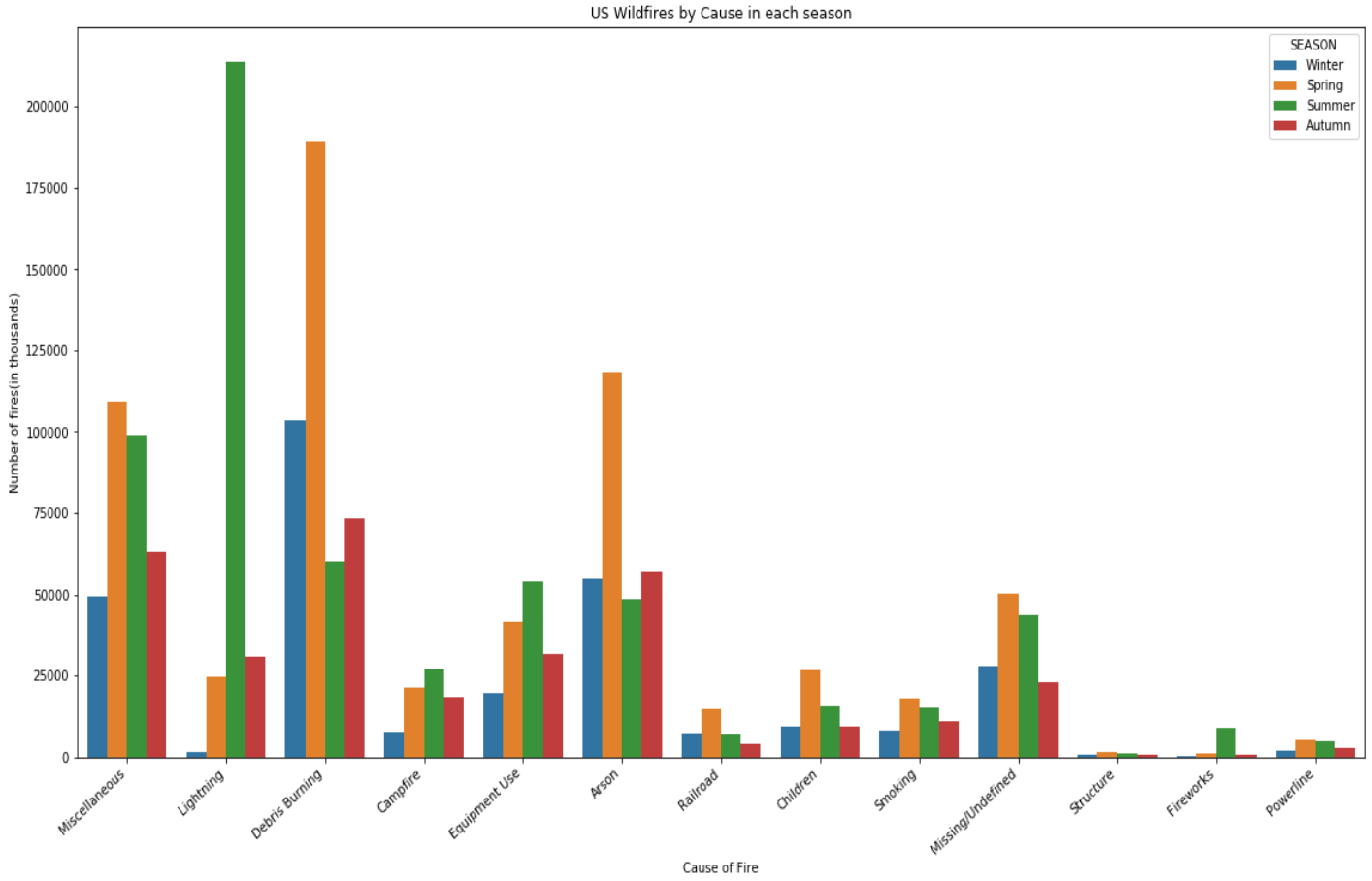
- The **degrees of freedom** can be calculated as:
  df = (total_rows - 1) * (total_cols - 1)

we make a decision based on the p-value. If our p-value is less than the significance value we reject the Null Hypothesis and if our p-value is greater than the significance value we do not reject it.

p-value=0.000000, significance=0.05

At 0.05 level of significance, we reject the null hypothesis and accept H1.
They are not independent.

As "Cause of Fire" and "Season" are not independent, let us see visualized data for the same.



We already knew that wildfires caused by "Lightning" were more damaging. From the above observation, we can also note that Lightning during Summer causes most wildfires.

We saw that there is a visible relationship between the States and the Cause of fire when we visualized data. However, to establish a statistical relationship between these two variables I did apply a chi-square test on STATE_NAME and Cause of Fire.

After performing chi-square tests on these two features and choosing a significance level as 0.05, we can conclude that STATE_NAME and Cause of Fire are associated with each other.

I want to point out that in both tests performed as described above, the p-value calculated is exactly 0.0. We have to treat this value with caution due to the large sample size.
In our statistical analysis, we saw what causes more fire across the states of the USA. Now let us find out if we can predict the cause of Fire in the future given the location, month, and year of fire reported. I did not consider 'fire size' as a feature because the goal is to predict the cause of wildfire at its initial phase where fire_size is small.

While selecting the required features I dropped the ‘Discoved_Date’ column by keeping only month data. First I will convert all non-numeric features like the cause of the fire and state name to numeric values, this is necessary for machine learning using *preprocessing.LabelEncoder.fit_transform* method.

The goal is to predict the cause of the fire = STAT_CAUSE_DESCR. Since I'll be using supervised learning I created training and test datasets. I choose features as below and I have split the data in 30% for testing, 70% for training.

Features/Predicted Variable = FIRE_YEAR,STATE,LONGITUDE,LATITUDE,MONTH
Target Variable = STAT_CAUSE_DESCR

I have tried applying Decision Tree Algorithm and Gradient Boosting Decision Tree Algorithm to the data. The prediction score is approximately 49% and 17% respectively. I have used the Random Forest Algorithm here as I think that fits my dataset better than the Decision tree classifier. The prediction score of this model is approximately 57%, with Training data accuracy 74%. We can clearly see that this model is overfitting.

An overfit model may look impressive on the training set but will be useless in a real application. Therefore, I will try to tune the hyperparameter. The standard procedure for hyperparameter optimization accounts for overfitting through cross-validation. Using Scikit-Learn's RandomizedSearchCV method, we can define a grid of hyperparameter ranges and randomly sampled from the grid, performing K-Fold CV with each combination of values.

So far I have used a Random Forest classifier with the default parameters other than for n_estimator value which I have chosen to be 60. A higher n_estimator may yield better results but I am running into timeout/memory error with higher values. To use RandomizedSearchCV, I have created a parameter grid. I tried using a wide range of values but due to limited RAM access narrowed the range of values.

Using the below random grid parameters to search for best hyperparameters.
{'bootstrap': [True],
 'max_depth': [90],
 'max_features': ['auto', 'sqrt'],
 'min_samples_leaf': [2, 4],
 'min_samples_split': [5, 10],
 'n_estimators': [60]}

Created a Random Forest Classifier model and tried to tune it by passing it to RandomizedSearchCV method along with n_iter=100, which controls the number of different combinations to try, and cv=3 which is the number of folds to use for cross validation. More iterations will cover a wider search space and more cv folds reduces the chances of overfitting, but raising each has increased the run time and at times lead to memory error.

To determine if a random search yielded a better model, I again created a RandomForest model with the best parameters obtained using the RandomizedSearchCV method.
{'bootstrap': True,
 'max_depth': 90,
 'max_features': 'sqrt',
 'min_samples_leaf': 4,
 'min_samples_split': 5,
 'n_estimators': 60}

We achieved an unspectacular improvement in accuracy of 1% after using bet parameters found using random search.

There are a lot of different Cause classes, I wanted to put some of these together and have just 3 classes for the cause of fires. Then test again to see if the score improves.

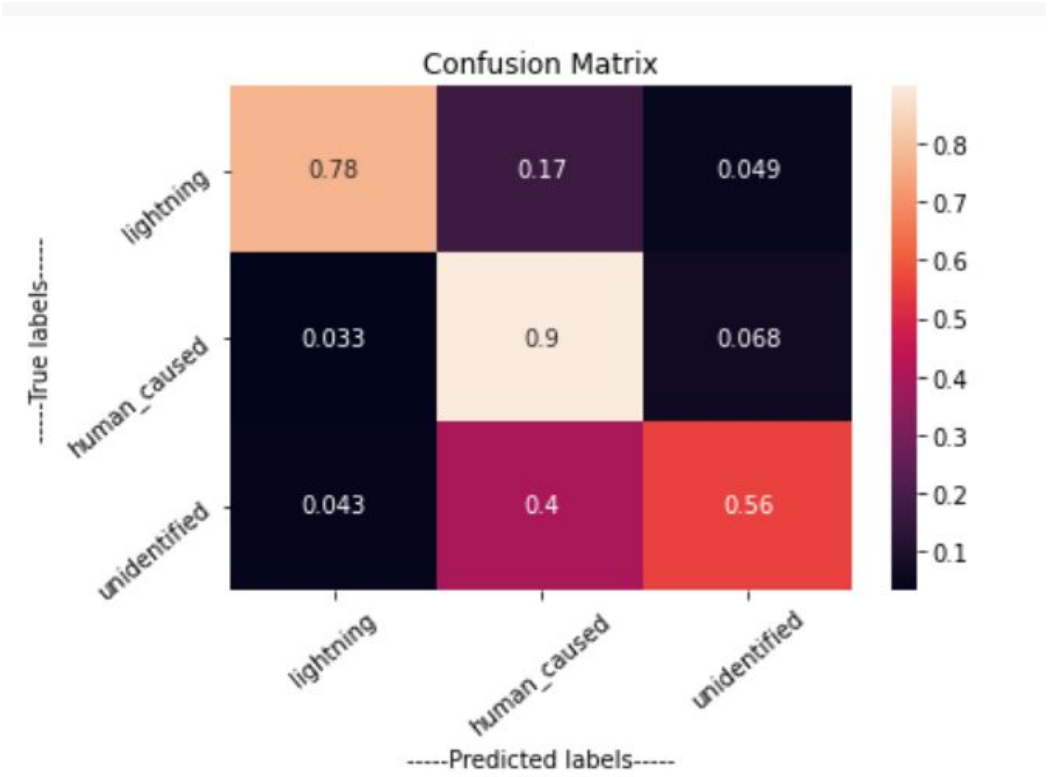
The 3 classes are: **lightning, human_caused and Unidentified/other**

where **lightning** = ['Lightning']

human_caused = ['Arson','Fireworks','Powerline','Railroad','Smoking','Children','Campfire','Equipment Use','Debris Burning','Structure']

Unidentified/other = ['Missing/Undefined','Miscellaneous']

I've now replaced STAT_CAUSE_DESCR with LABEL. So now I tried to predict LABEL. Also, I will be passing the best parameter values obtained by the RandomSearchCV method. Reducing the number of categories improved the prediction score significantly (from around 58% to 80%). Prediction score is a good metric to measure success but we can look at other metrics like confusion matrix.



The random forest algorithm did well with the first two labels: lightning(78%) and human_caused (90%) but did not do as well with the 'unidentified' label. It labeled 40% of unidentified labels as human_caused fire.

Given that it is easier to make accurate predictions if the number of classes is reduced, I wanted to look at just one state at a time and build a system that can predict cause in that State. First I wanted to see How well the model predicts only for California.

The generic model accuracy(80%) was significantly more compared to the model per state(66%). Thus I decided to not go ahead with this approach.

I wanted to check if taking a random sample of the dataset and applying a Random Forest Classifier Algorithm on this sample would yield a more efficient model. Thus I took 60% of the dataset randomly and built a classifier. The model using the subset of the dataset did not yield an any better result. The accuracy is exactly the same as the model built using a full dataset.

Summary: Through this project not only we saw which State and county face more wildfires but also which state and county have faced the most destructive wildfires. Counties like Riverside County, San Bernardino, and San Diego in California, Coconino, and Gila in Arizona, Beltrami in Minnesota have more number of fires. Yukon-Koyukuk, Stanton, and other counties in Alaska have fires that have destroyed over thousands of acres. Though states like California and Texas report more numbers of fires, it seems like they are successful at controlling the larger damage. Concerned government departments in bigger states like California and Texas might be better prepared or equipped as they expect wildfires many times in a year. However, smaller states like Alaska, Georgia, and Rhode Island should be prepared with the right equipment and safety measures to prevent/fight wildfires which have recklessly caused harm to the states in the past.

We have seen that most wildfires across the USA are caused by debris burning and arson. However, fires started by electricity(caused by lightning and Powerline) are the most damaging. We also saw that Lighting during summer causes the most dangerous fires. By educating people about local Ordinance regarding trash burning, being careful when having a campfire, using fireworks or fire pits, and implementing strict laws against arsonists we can prevent or at least reduce the number of fires caused by human negligence. Parks in California post the risks of forest fires on a daily basis. Other states and counties should implement this too. If people are aware of the risks, they can prevent doing any activities that could end up causing a wildfire.

Using the Random forest classifier model trained in this project, we can predict the cause of these wildfires, at least to an accuracy of 58% or better. Reducing the number of labels(Fire Cause classes) significantly improves the prediction score to 80% for the random forest algorithm. But with further tuning or a different algorithm, it may be possible to reach a better score.

PowerPoint presentation of the above report can be found in the link below.

https://github.com/lasyabheemendra/Sprigboard-DatascienceProjects/blob/master/Capstone1_US-Wildfire-Prediction/Final%20Project%20Report/US_WildFires_Project.pptx

Below is the link to the Github repository of code files of this project.
https://github.com/lasyabheemendra/Sprigboard-DatascienceProjects/tree/master/Capstone1_US-Wildfire-Prediction

References: The sources are listed in the order in which they are cited in the report, as in the following book/article/website.

[1] Paulo Cortez and An´ıbal Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. Department of Information Systems/R&D Algoritmi Centre, University of Minho, 4800-058 Guimar˜aes, Portugal, pcortez@dsi.uminho.pt
WWW home page: http://www.dsi.uminho.pt/˜pcortez

[2] The Role of Data Scientists in Helping to Fight Wildfires.*Rick Hutley is the Program Director and Clinical Professor of Analytics at the University of the Pacific.*
https://www.pacific.edu/academics/schools-and-colleges/school-of-engineering-and-computer-science/about-the-school/academics-/graduate-programs/ms-in-data-science/rick-wildfire-story-npr.html

[3] Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPAFOD20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive.
https://doi.org/10.2737/RDS-2013-0009.4
https://www.kaggle.com/rtatman/188-million-us-wildfires