
US Wildfires
An analysis of Data collected over 23 years

Topic Overview

- Background
- Problem
- Dataset
- Analysis and Result
- Machine Learning
- Conclusion
- Recommendation

Background

- Wildfires create economic and ecological damage while endangering human lives.
- Fast detection is a key element for controlling such a phenomenon.
- Each year millions of forest hectares are destroyed all around the world.
- USA is fighting wildfire problems for decades.
- 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.

Problems

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 help of growing technology? Has the size of fire decreased over 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?

Dataset

- This data publication contains a spatial database of wildfires that occurred in the United States from 1992 to 2015.

- It includes 1.88 million geo-referenced wildfire records, representing a total of 140 million acres burned during the 24 years.

- 14 relevant features and 1838525 entries were focused.

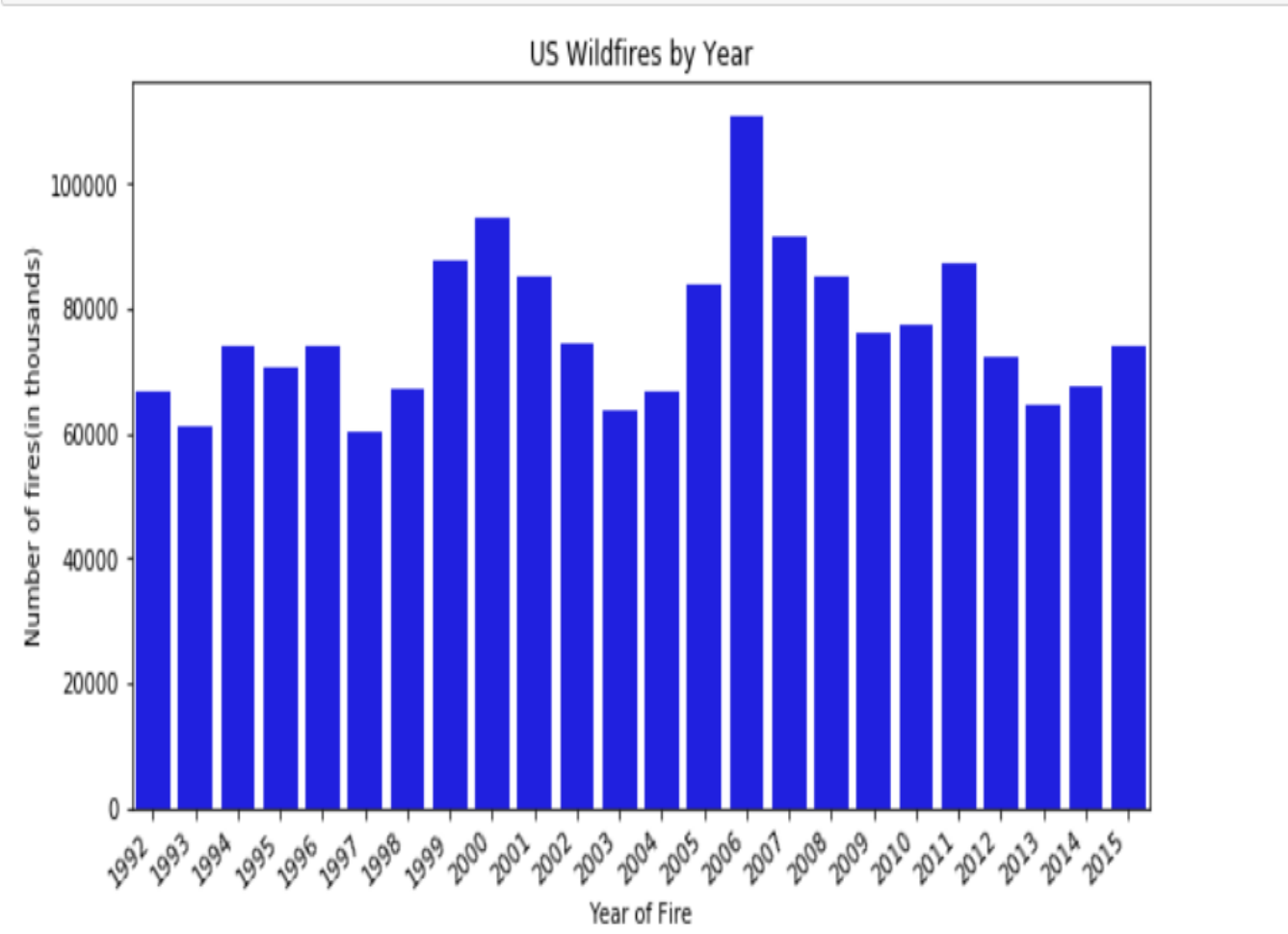
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1838525 entries, 0 to 636207
Data columns (total 14 columns):
GLOBAL_UNIQUE_ID    int64
DISCOVERY_DATE      object
FIRE_YEAR           int64
STAT_CAUSE_DESCR    object
FIRE_SIZE           float64
FIRE_SIZE_CLASS     object
STATE               object
STATE_NAME          object
COUNTY_NAME         object
LATITUDE            float64
LONGITUDE           float64
DISCOVERY_DOY       int64
OWNER_DESCR         object
FIPS_CODE           float64
```

Data Wrangling

- Dropped columns in SQL table "fires” which were related to Source of data and agency information which collected or prepared reports on Data.
- Dropped columns with more than 50% of its total data with null values.
- Filled the missing values of COUNTY_NAME column using 'LATITUDE' and 'LONGITUDE' columns.
- Converted "DISCOVERY_DATE“ column to a readable date format from Julian date.
- Created a column FIPS_CODE with FIPS code value for each County Name.
- Checked if there were any outliers with respect to the FIRE_YEAR column.

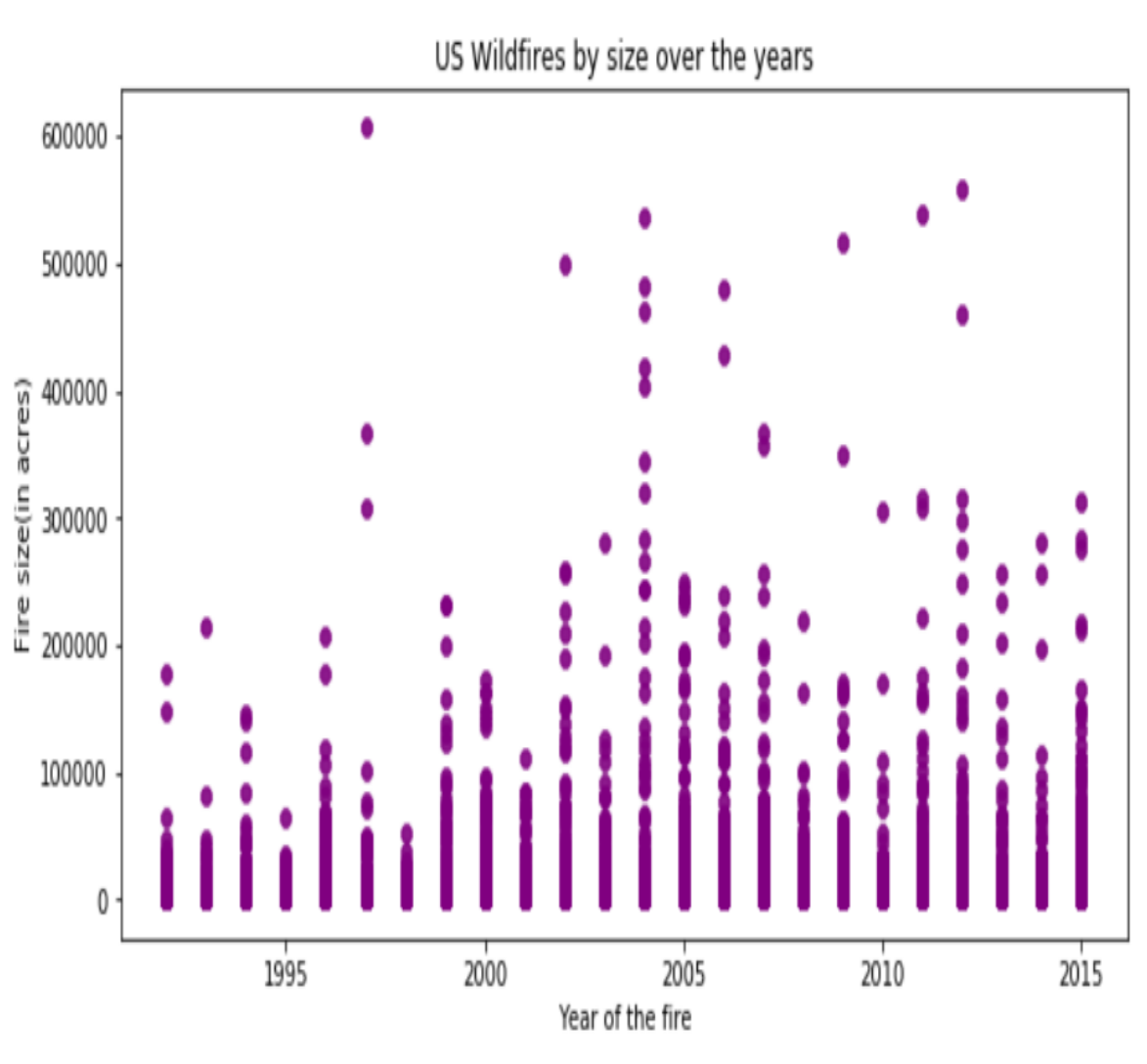
Number of fires over the period of 1992-2015

- 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 time periods, there is no continuous upward trend over the period.



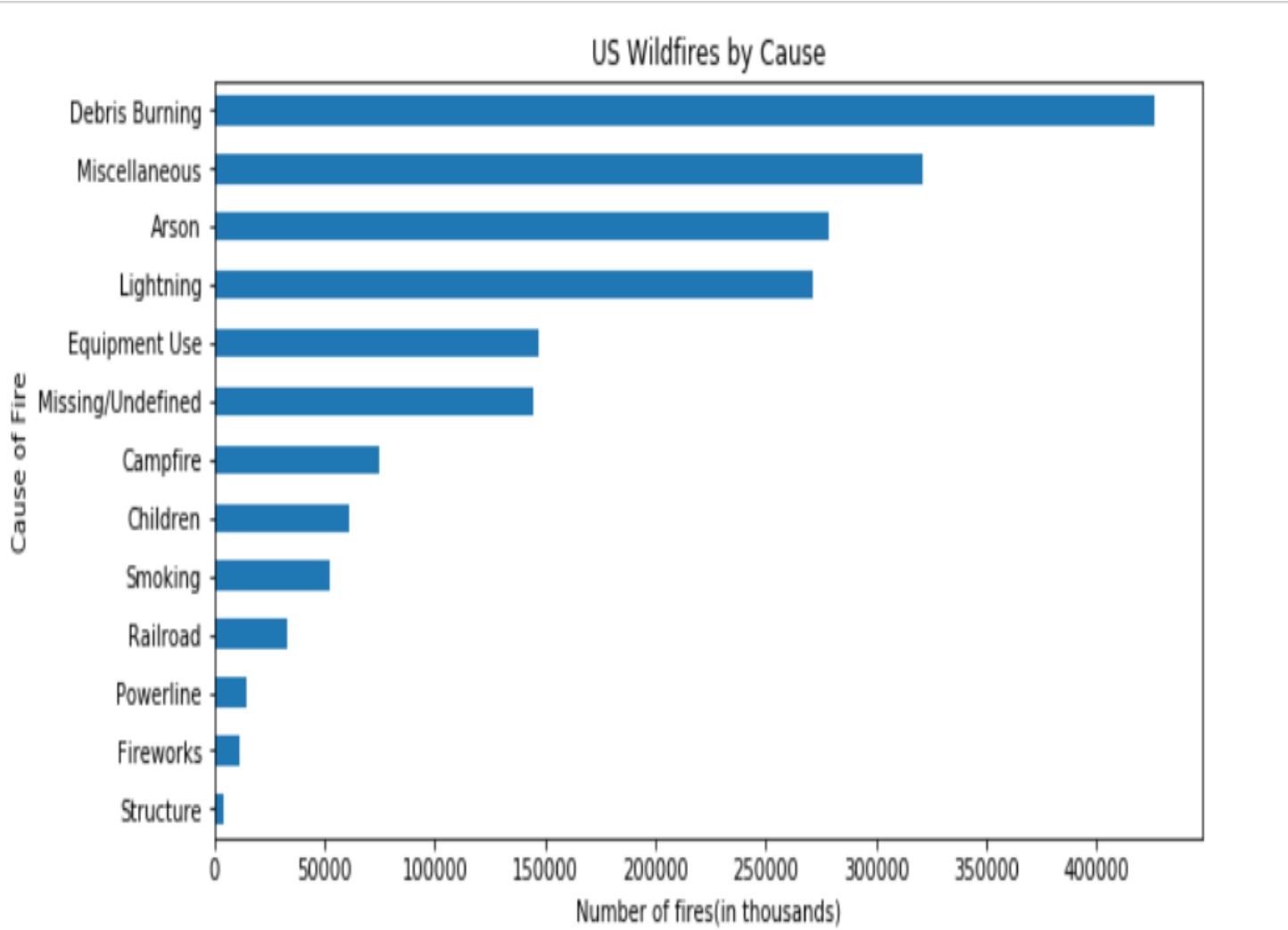
Wildfire size over the years

- There is no decrease or increase of fire size from 1992-2015.
- Every year has recorded fire of different sizes.



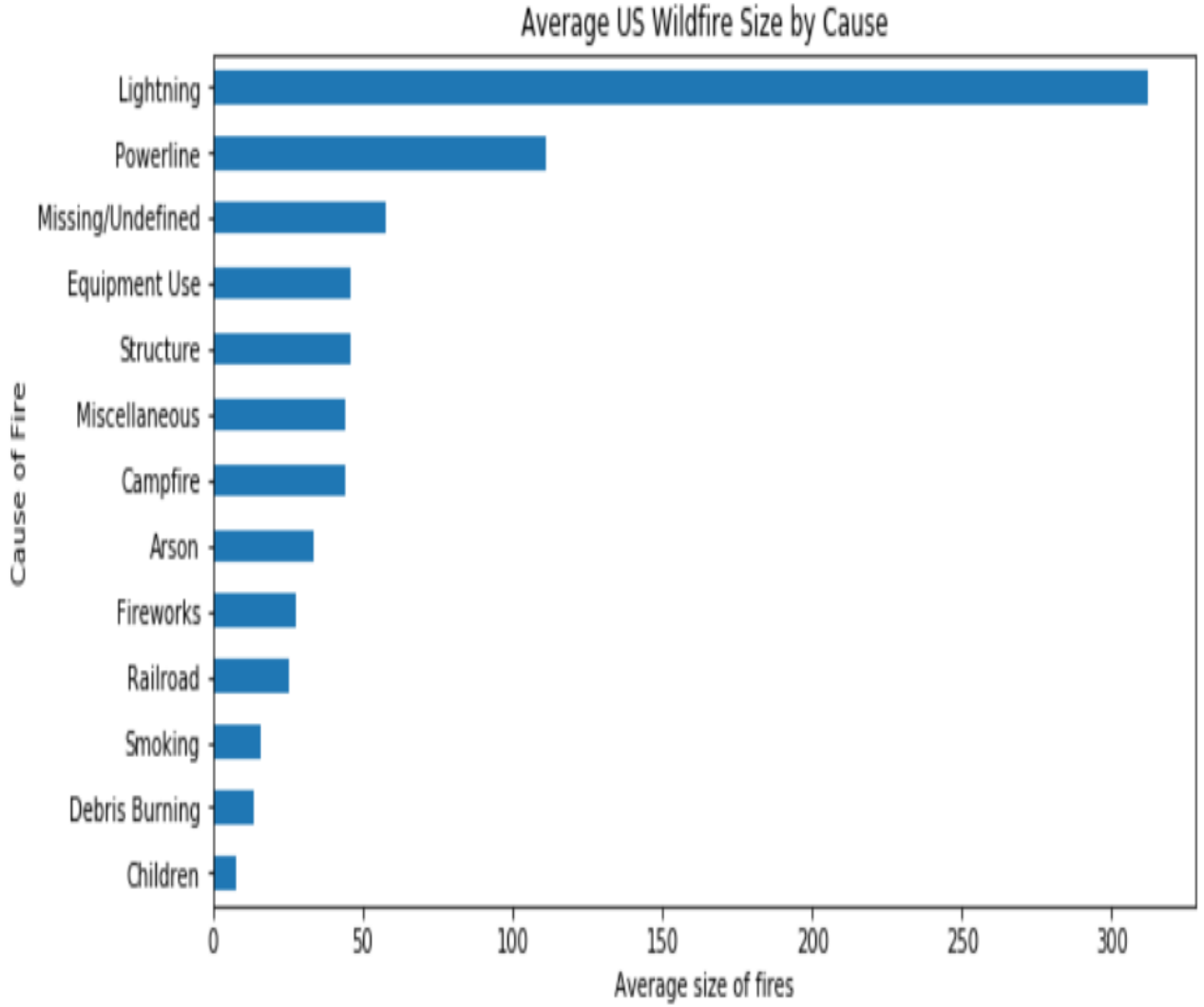
What causes the most fires?

- Trash burning was the largest cause of wildfire by a significant margin.
- Slightly more fires were started by arson than by lightning
- Interesting to note that one reason is just ‘Children’.



Relationship between cause and fire size?

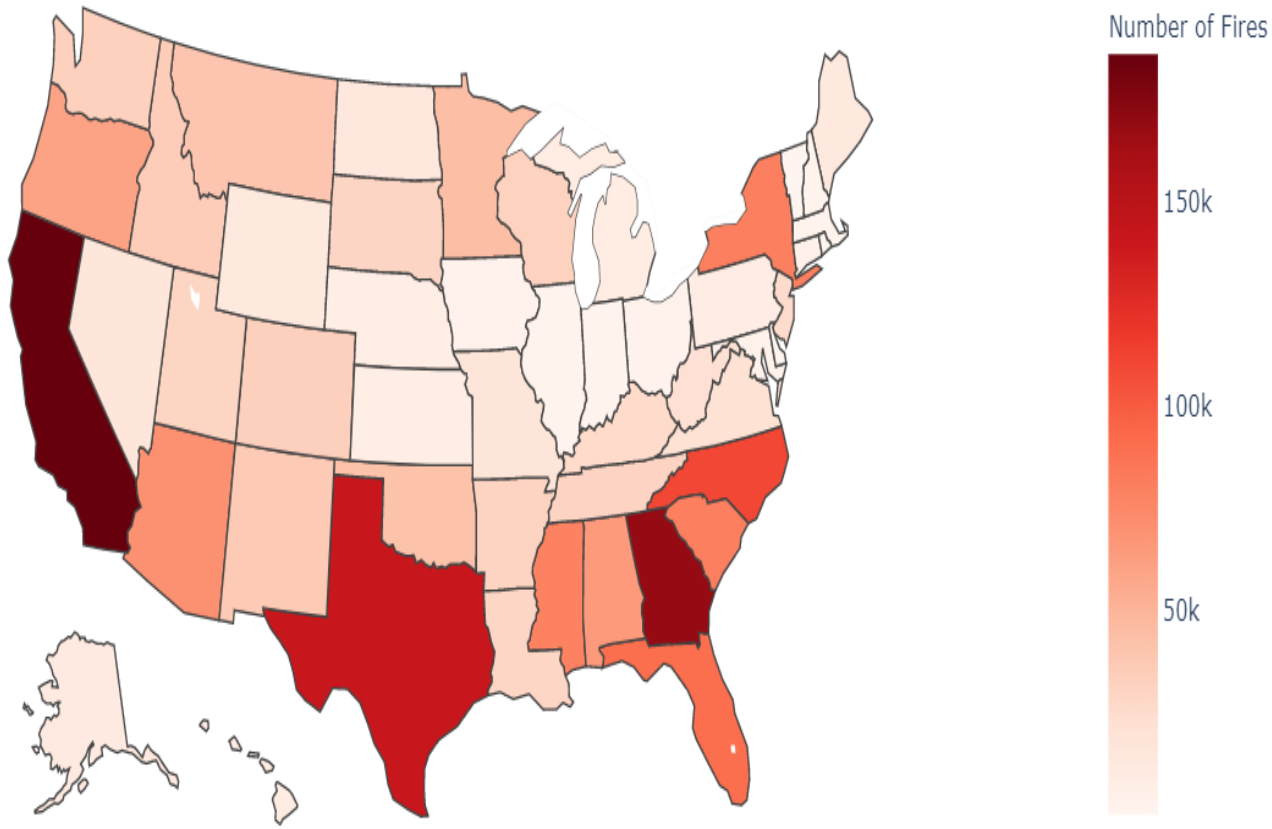
- We can observe that fires started by ‘Lightning’ are the most damaging.
- Though the number of fires caused by debris burning is large in number, the average size of the fire caused by it is very low.



Wildfires in each state of USA

- California and Texas have the most wildfires.

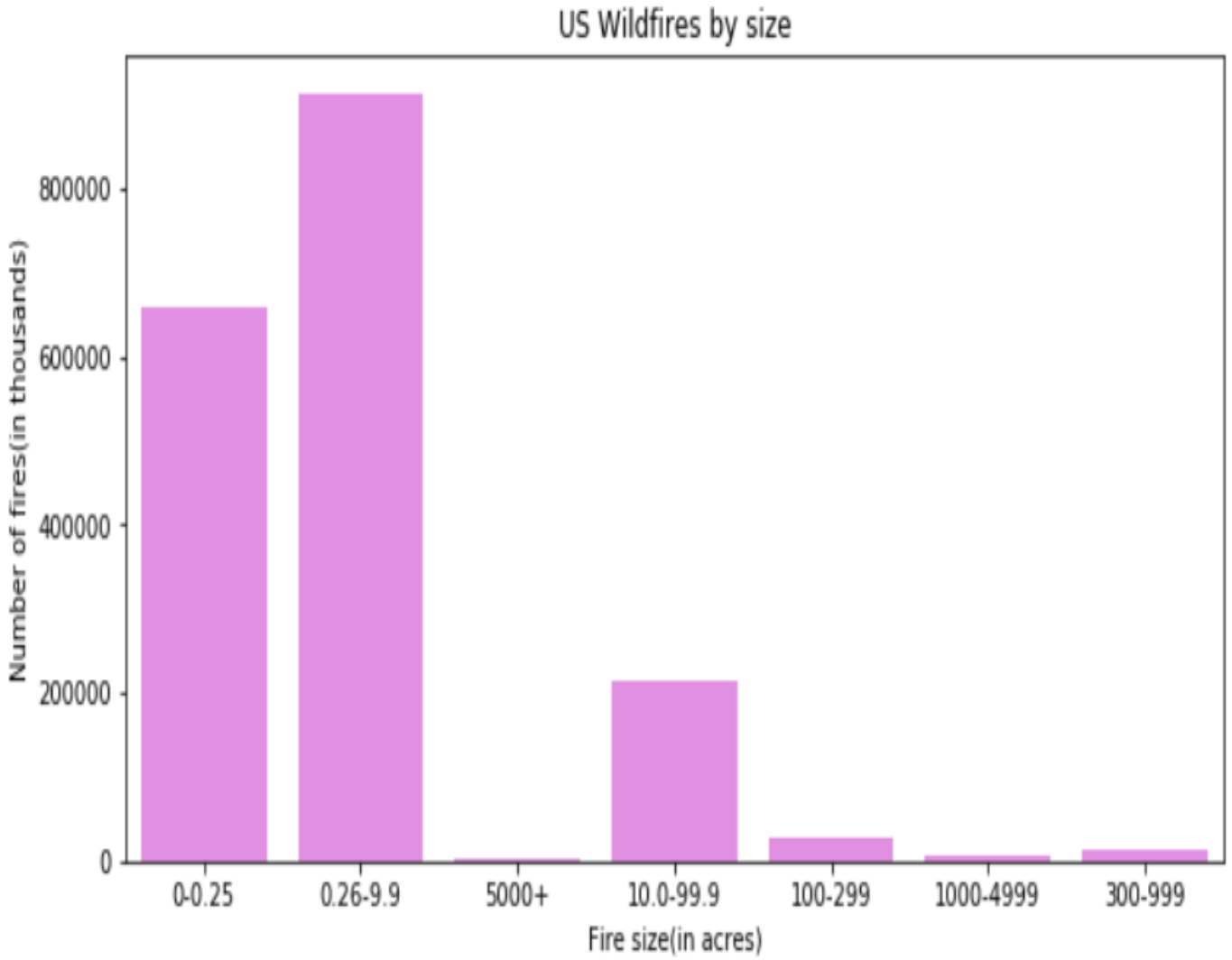
- Surprisingly, Georgia which is much smaller state than Texas has larger number of wildfires than Texas.



Wildfires by size

- Over 800k of total fire incidents are in between 0.26-0.99 acres within the final fire perimeter expenditures.

- The number of wildfires reported which have caused damage to more than 1000 acres are comparatively less

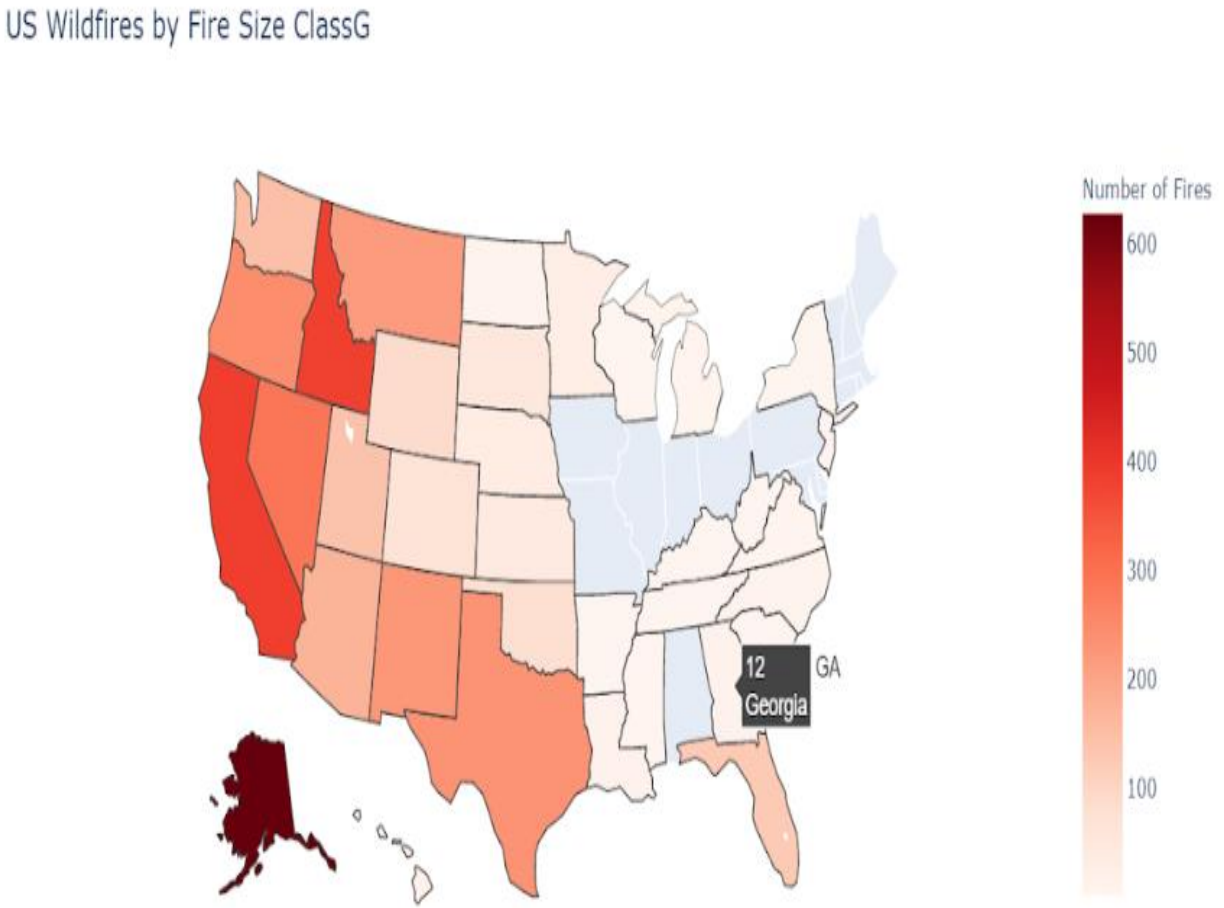


Wildfires in each state with respect to fire size

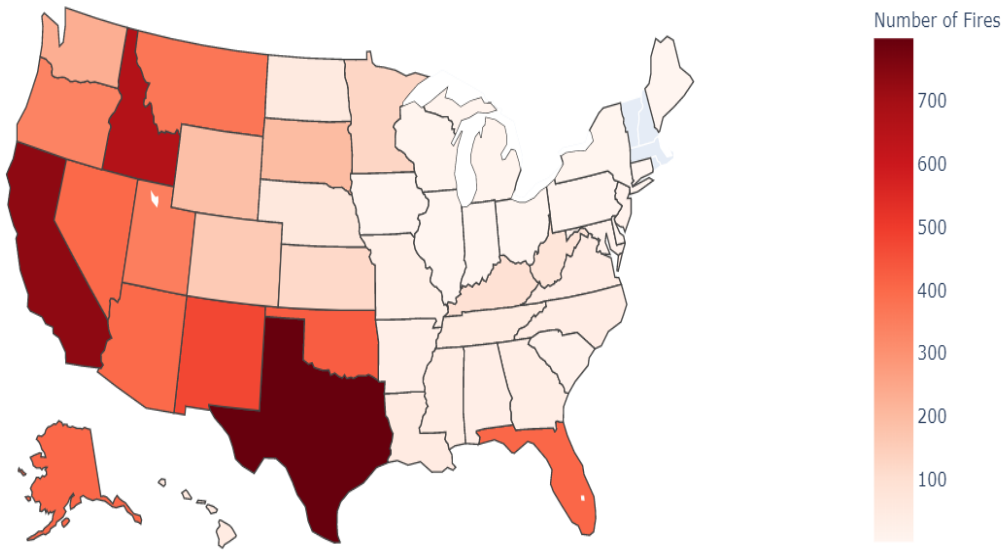
- From Dataset we know that
  - **Size 'A' represents '0-0.25' acres**
  - **Size 'B' represents '0.26-9.9' acres**
  - **Size 'C' represents '10.0-99.9' acres**
  - **Size 'D' represents '100-299' acres**
  - **Size 'E' represents '300-999' acres**
  - **Size 'F' represents '000-4999' acres**
  - **Size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

- Georgia has 12 fires over the span of two decades which are of size ‘5000”’ acres

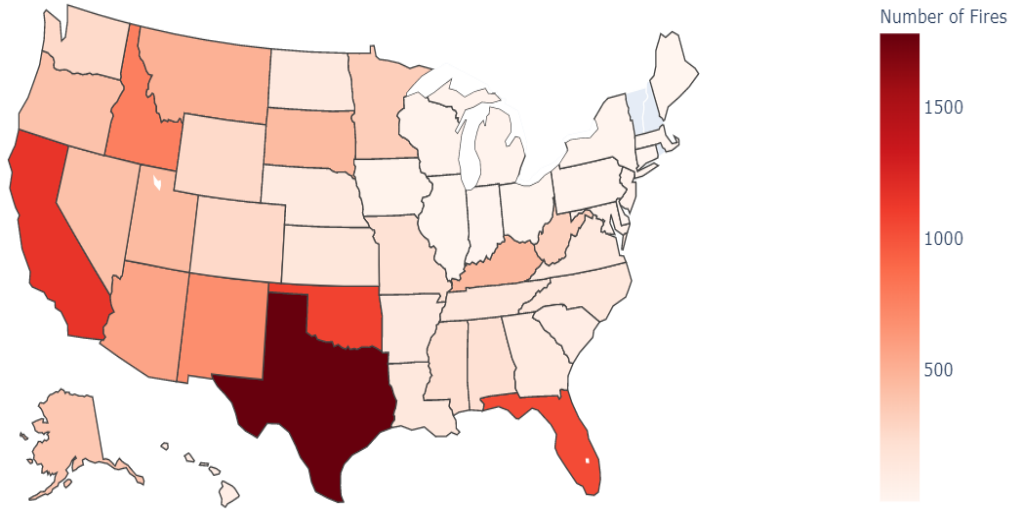
- Alaska has the large number of fires of largest size



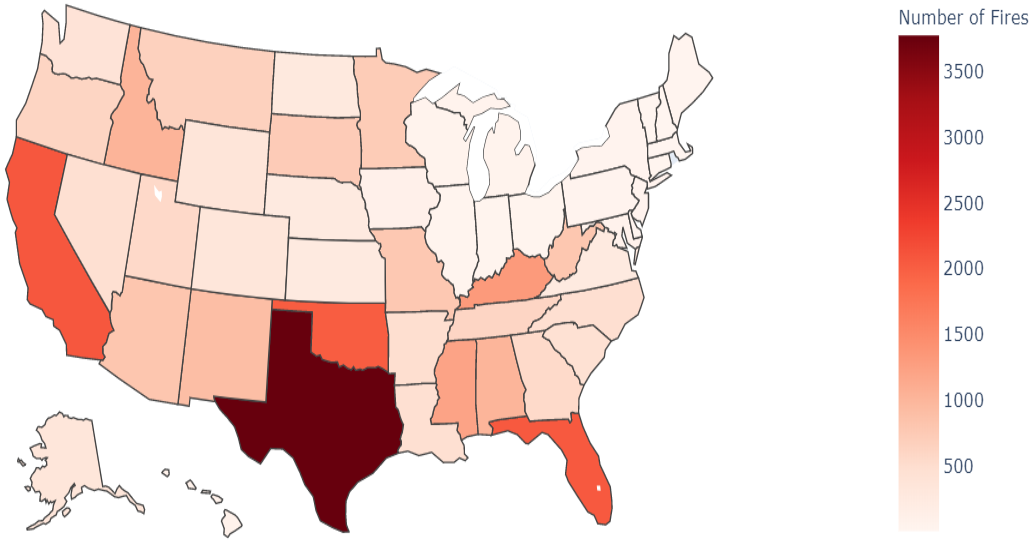
US Wildfires by Fire Size ClassF



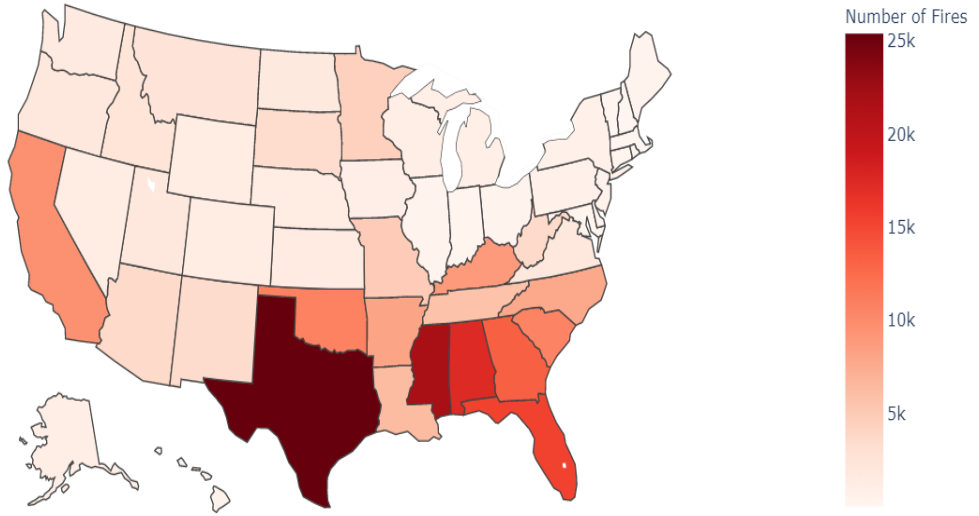
US Wildfires by Fire Size ClassE



US Wildfires by Fire Size ClassD



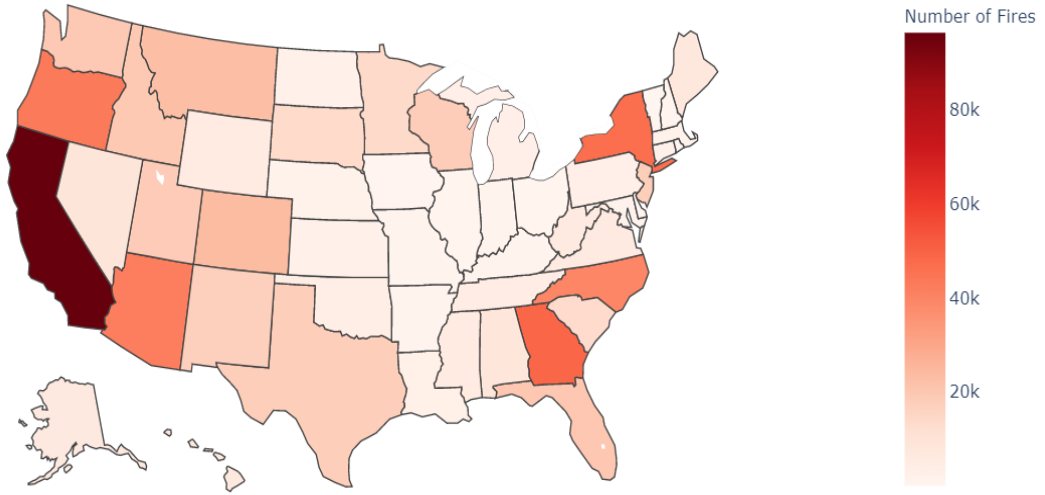
US Wildfires by Fire Size ClassC



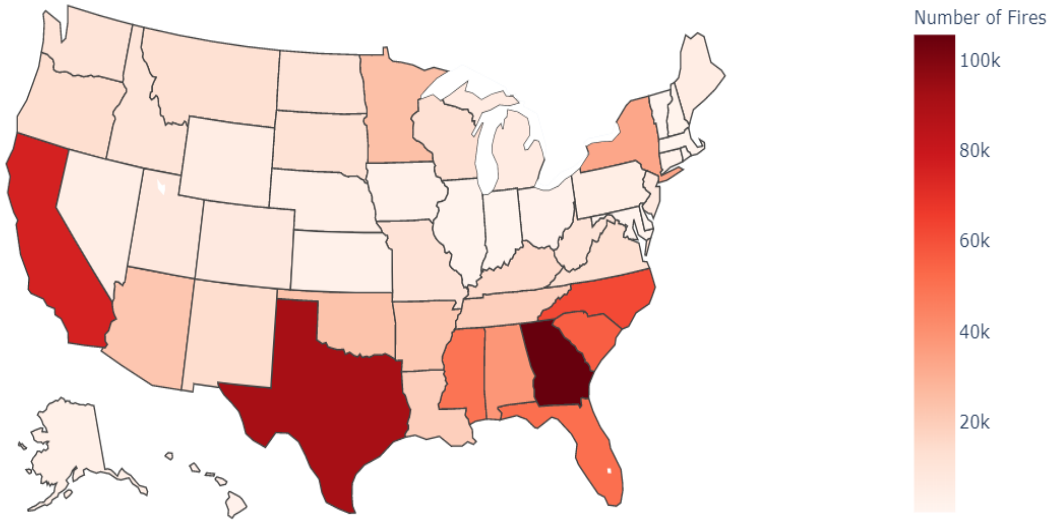
- After looking at State maps by Fire size class, we can note that Georgia has most fires with size under 100 acres.

- California and Texas have reported fires of all sizes in more numbers over two decades.

US Wildfires by Fire Size ClassA

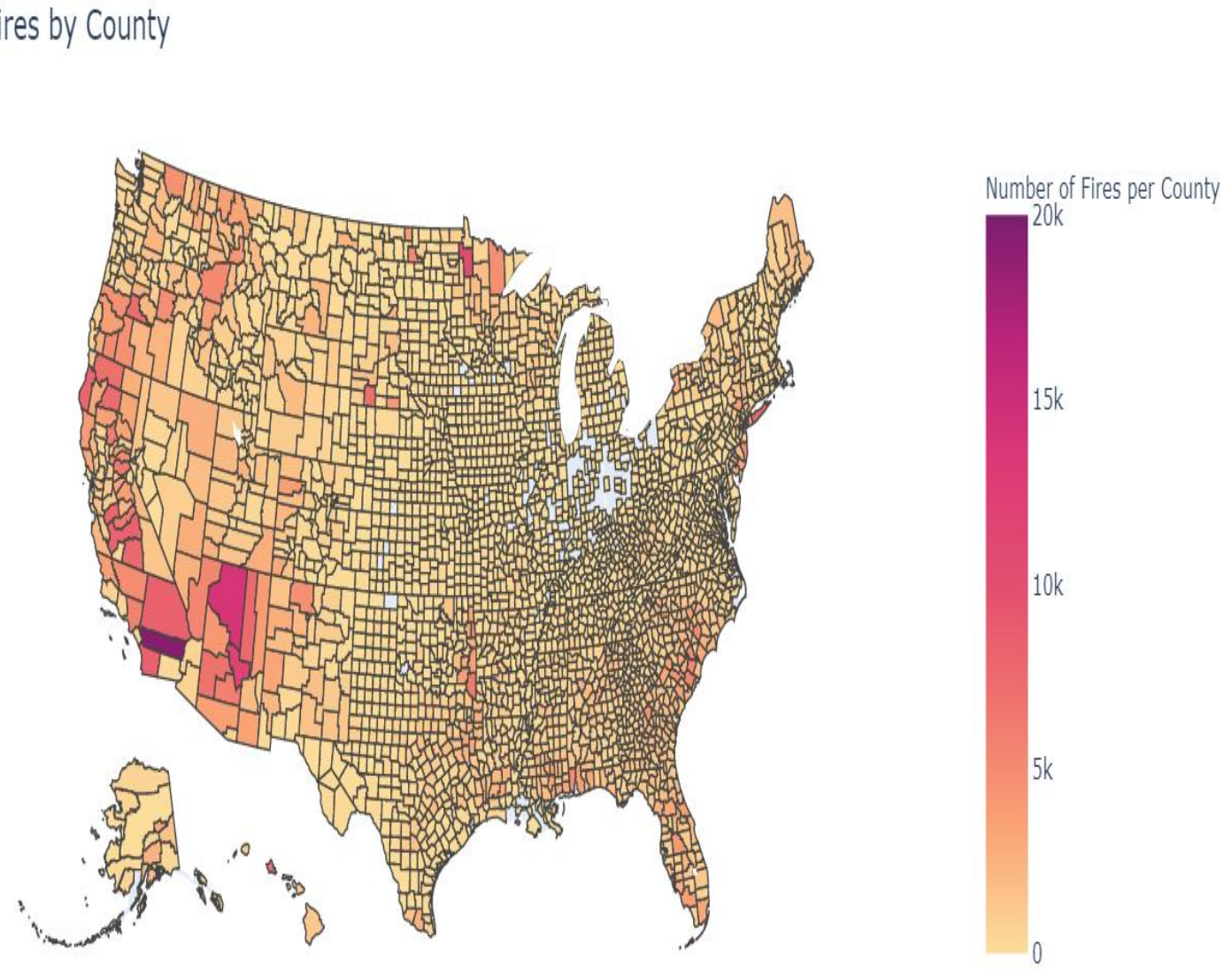


US Wildfires by Fire Size ClassB



Which County is more prone to Wildfires in USA?

- “Riverside County” in California has the most numbers of wildfires.
- Below counties comes next in the top list.
  - “Coconino” and “Gila” in Arizona
  - “Beltrami” in Minnesota
  - “San Bernardino” in California
  - “San Diego” in California
- Number of fires are more in counties of California, Arizona, Minnesota, and New York.



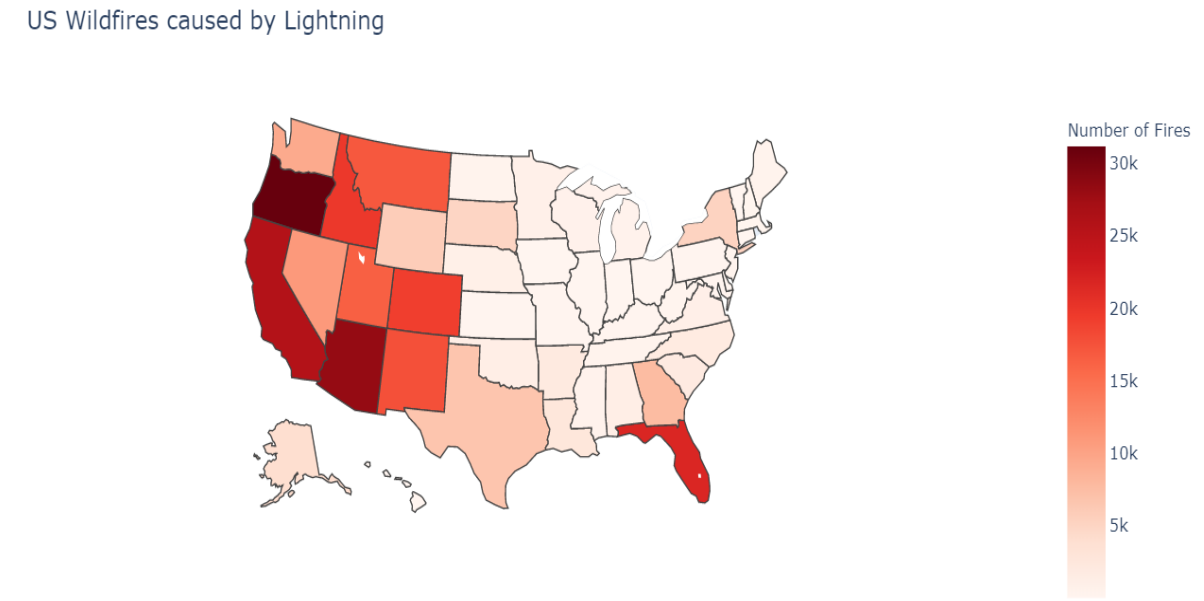
Which counties are most affected in terms of acres of land burnt due to wildfire?

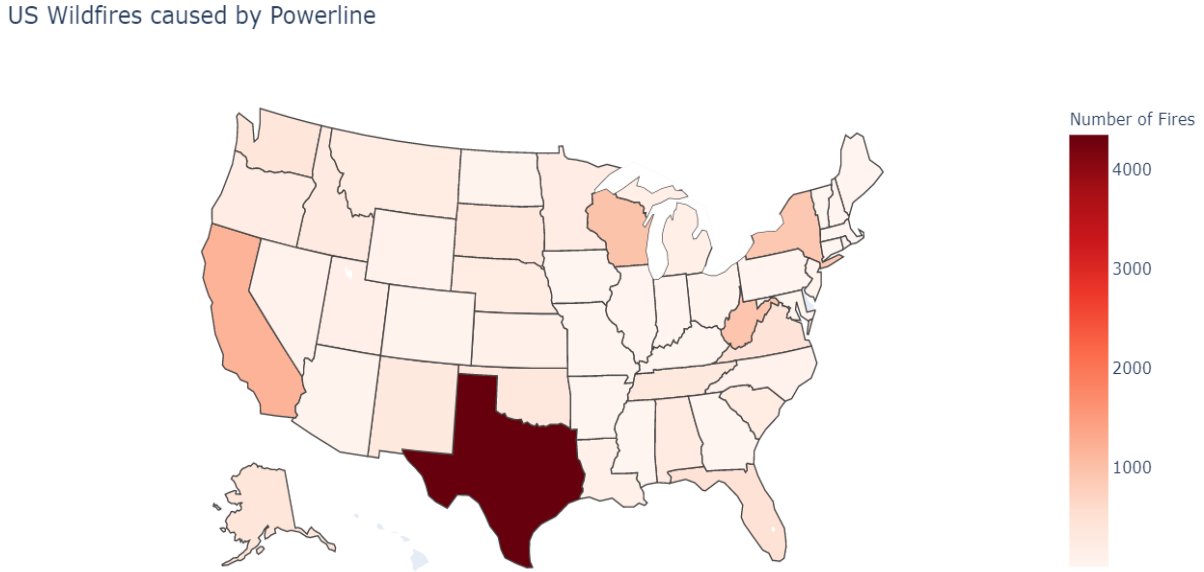
- Yukon-Koyukuk is the most affected county over the period.
- More destructive fires are reported in counties of Alaska, Nebraska, and Oklahoma.
- Riverside County of California reported 19398 fire incidents over 13 years with the average fire size being approximately 31 acres.
- Yukon-Koyukuk county of Alaska reported only 51 fire incidents over the same period but with an average fire size of approximately 35000 acres.

FIRE_SIZE_mean	FIRE_SIZE_min	FIRE_SIZE_max	COUNTY_NAME	STATE_NAME
34400.542941	0.10	312918.3	Yukon-Koyukuk	Alaska
13467.000000	1.00	40000.0	Stanton	Nebraska
9360.666667	604.00	23488.0	Woodward	Oklahoma
9083.589286	0.25	70000.0	Barber	Kansas
8983.950000	0.10	53640.0	Pondera	Montana
7770.115425	0.10	517078.0	Denali Borough	Alaska
7342.829783	0.10	606945.0	Dillingham Census Area	Alaska
6050.744675	0.10	308120.0	Bethel Census Area	Alaska
6007.200000	1.00	162625.0	Kent	Rhode Island
5000.000000	5000.00	5000.0	Ellis	Texas

Wildfires in each state with respect to Cause of Fire

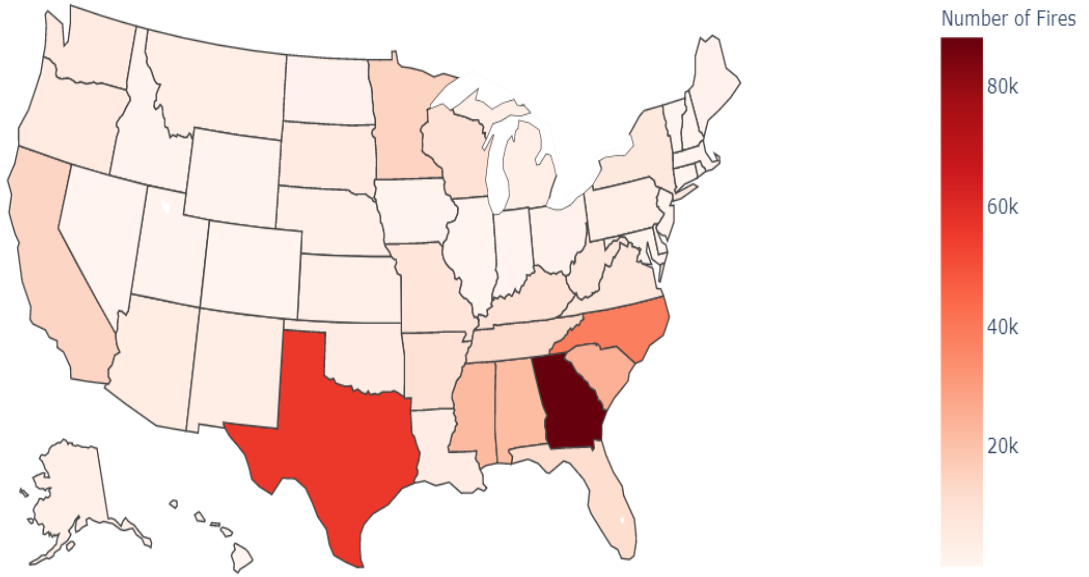
- The west states of USA has recorded most wildfires caused by “Lightning”.
- Florida has recorded 21.8k wildfire caused by lightning
- “Powerline” has been stated as cause for over4k wildfires in Texas alone.



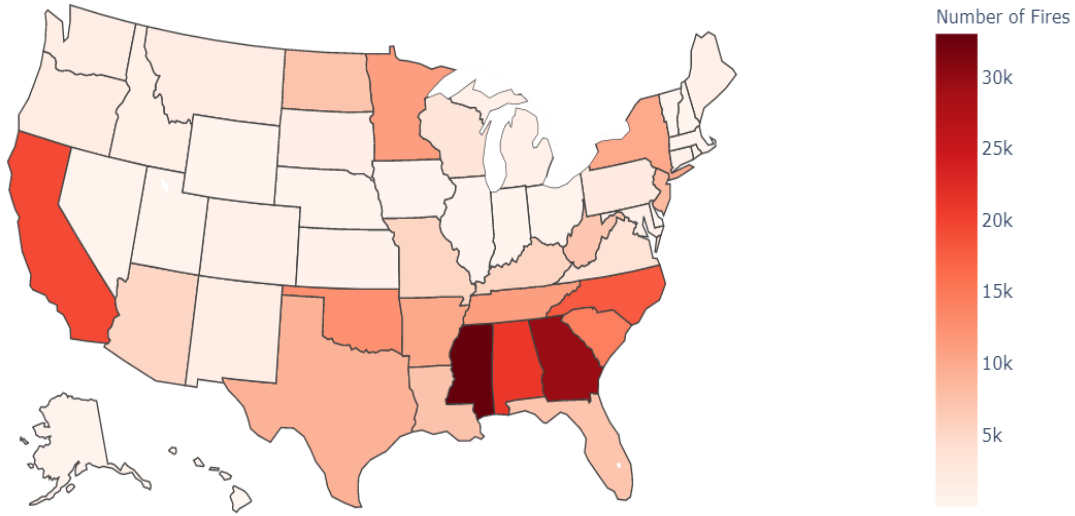


- Trash burning is the cause for almost all the cases in Georgia.
- Mississippi has 30k wildfires caused by trash burning and arson alone.

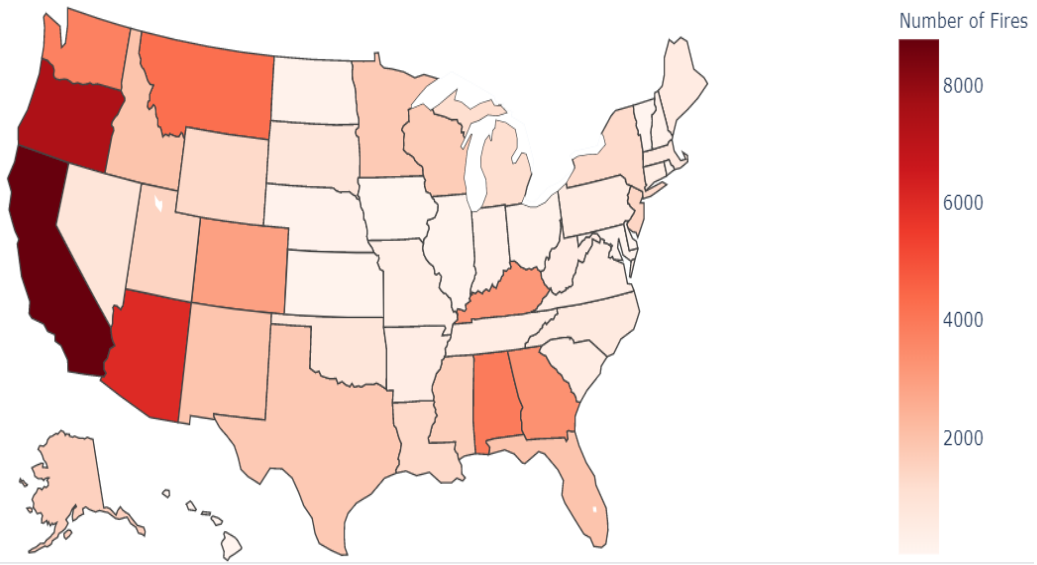
US Wildfires caused by Debris Burning



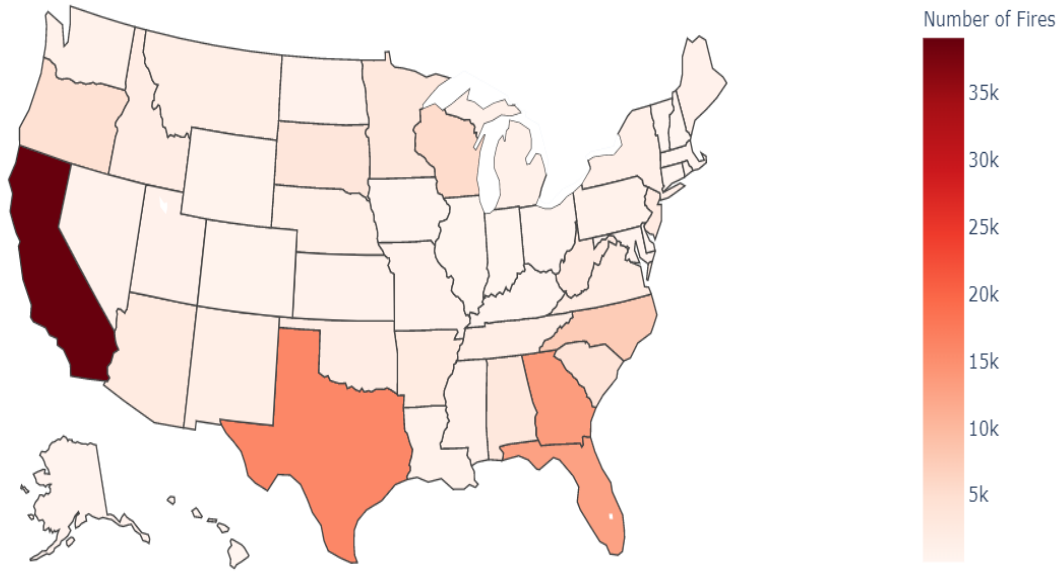
US Wildfires caused by Arson



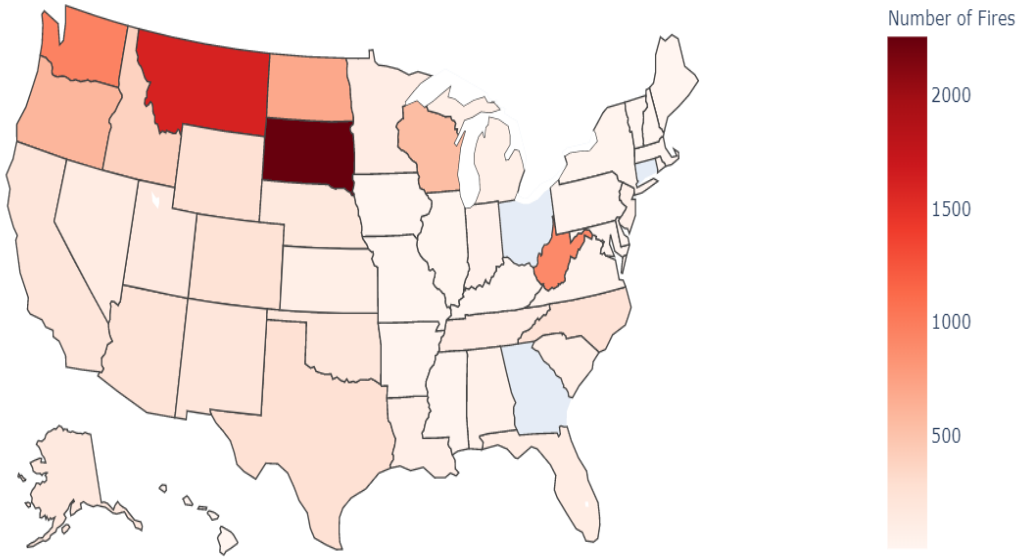
US Wildfires caused by Campfire



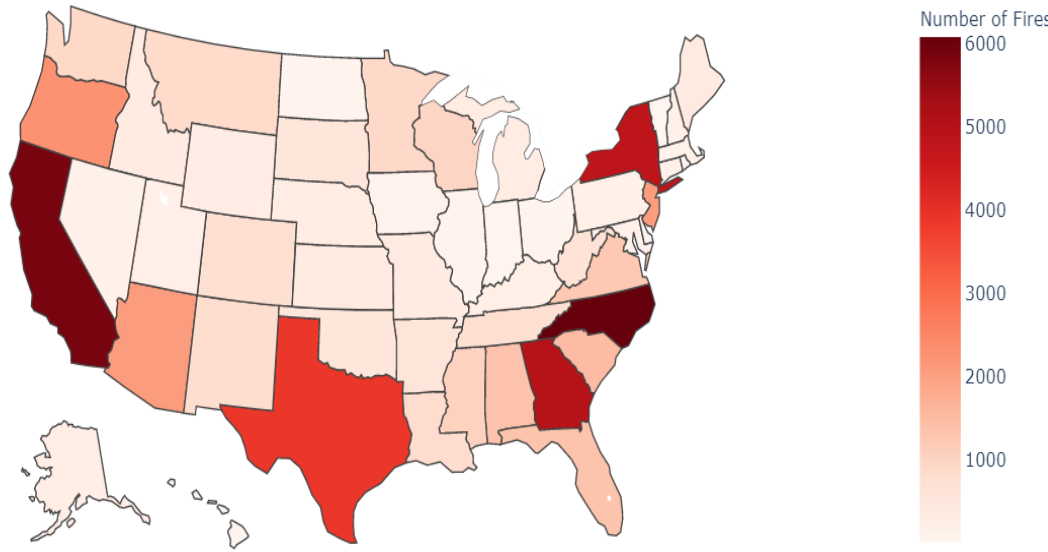
US Wildfires caused by Equipment Use



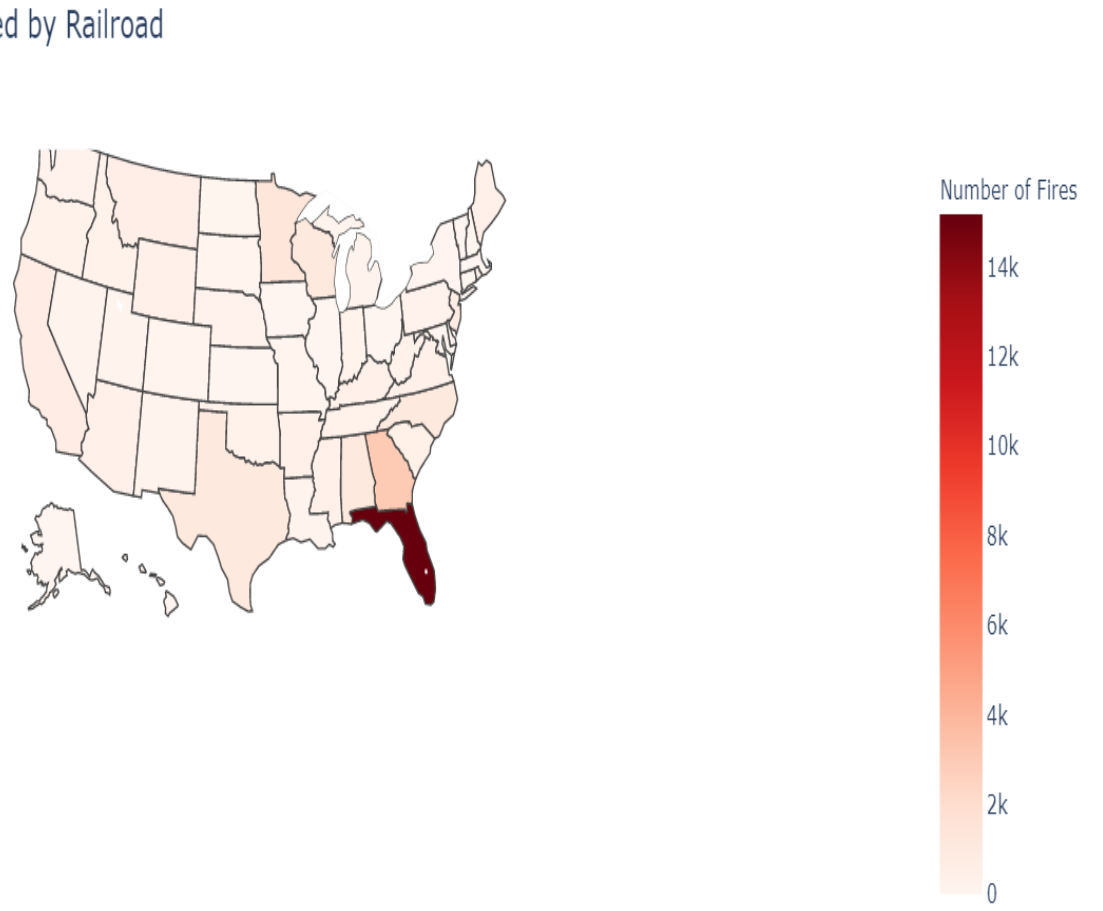
US Wildfires caused by Fireworks



US Wildfires caused by Smoking



- Railroads are causing about 14K wildfires in Florida.
- only state majorly affected by Railroad caused wildfires.



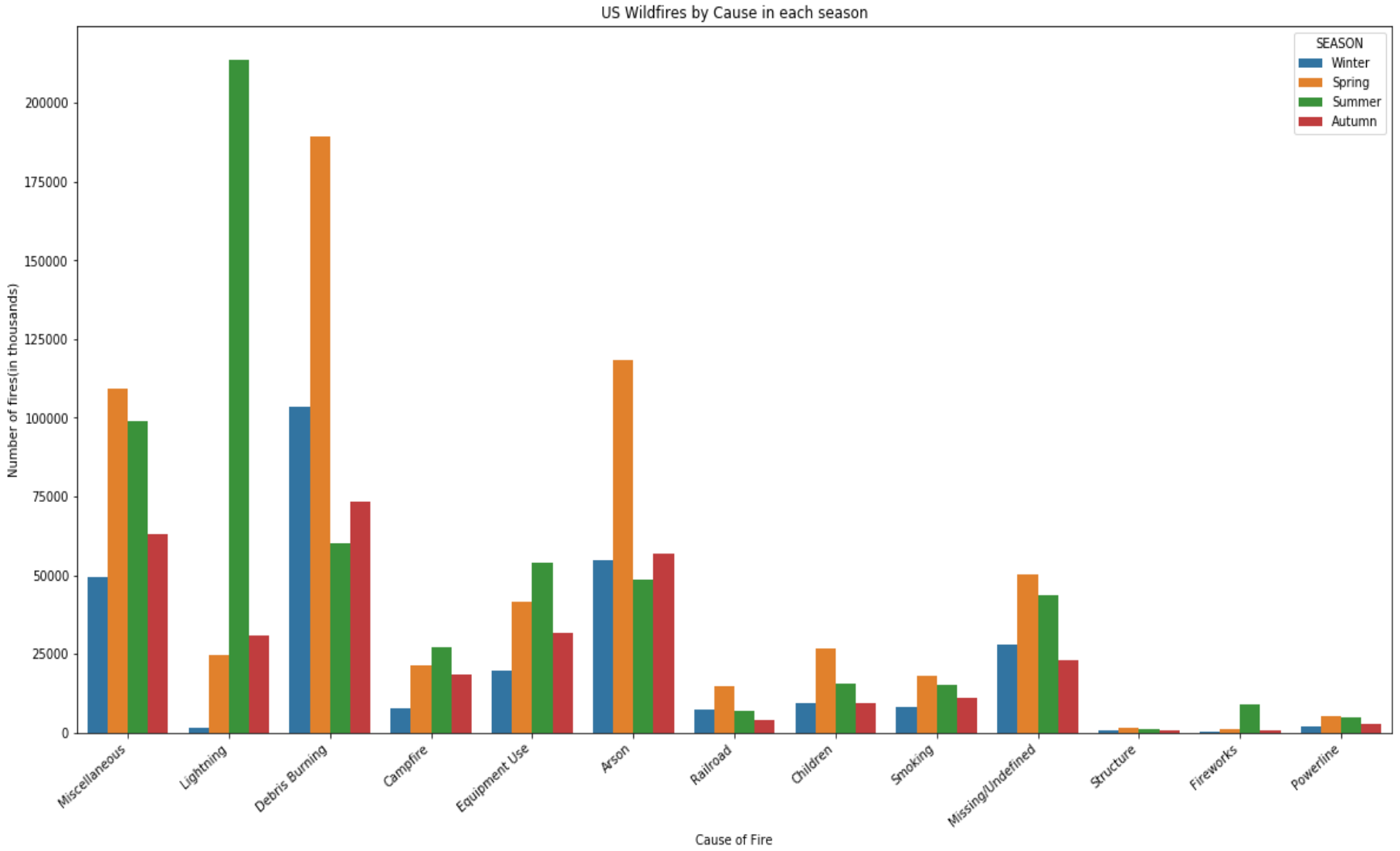
Is the cause of wildfire related to season?

- **chi-square Test**
  - 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).
  - **p-value=0.000000**
- **At 0.05 level of significance, we reject the null hypothesis and accept H1.**
  They are not independent.

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

- "Cause of Fire" and "Season" are not independent.

- we can also note that Lightning during Summer causes most wildfires.



Machine Learning-Predicting the cause of Fire

- Select features to build model
  - **Features/Predicted Variable = FIRE_YEAR,STATE,LONGITUDE,LATITUDE,MONTH**
  - **Target Variable = STAT_CAUSE_DESCR**

- I have used Supervised learning with training and test datasets

- Split the data in 30% for testing and 70% for training.

- Convert all non-numeric features like the cause of the fire and state name to numeric values.

- I have evaluated Decision Tree, Random Forest Classifier and Gradient Boosting Decision Tree models with prediction score 47%, 57% and 17% respectively.

Random Forest Classifier

- Prediction score of this model is approximately 57%, with Training data accuracy 74%.
- Model is overfitting . Overfit model may look impressive on the training set but will be useless in a real application.
- Tune the hyperparameter
- The standard procedure for hyperparameter optimization for overfitting is 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.

Hyperparameter Tuning

- 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.

- 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]}
```

Hyperparameter Tuning Contd..

- 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.
  - cv=3, which is the number of folds to use for cross validation.

- created a Random Forest 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 }
```

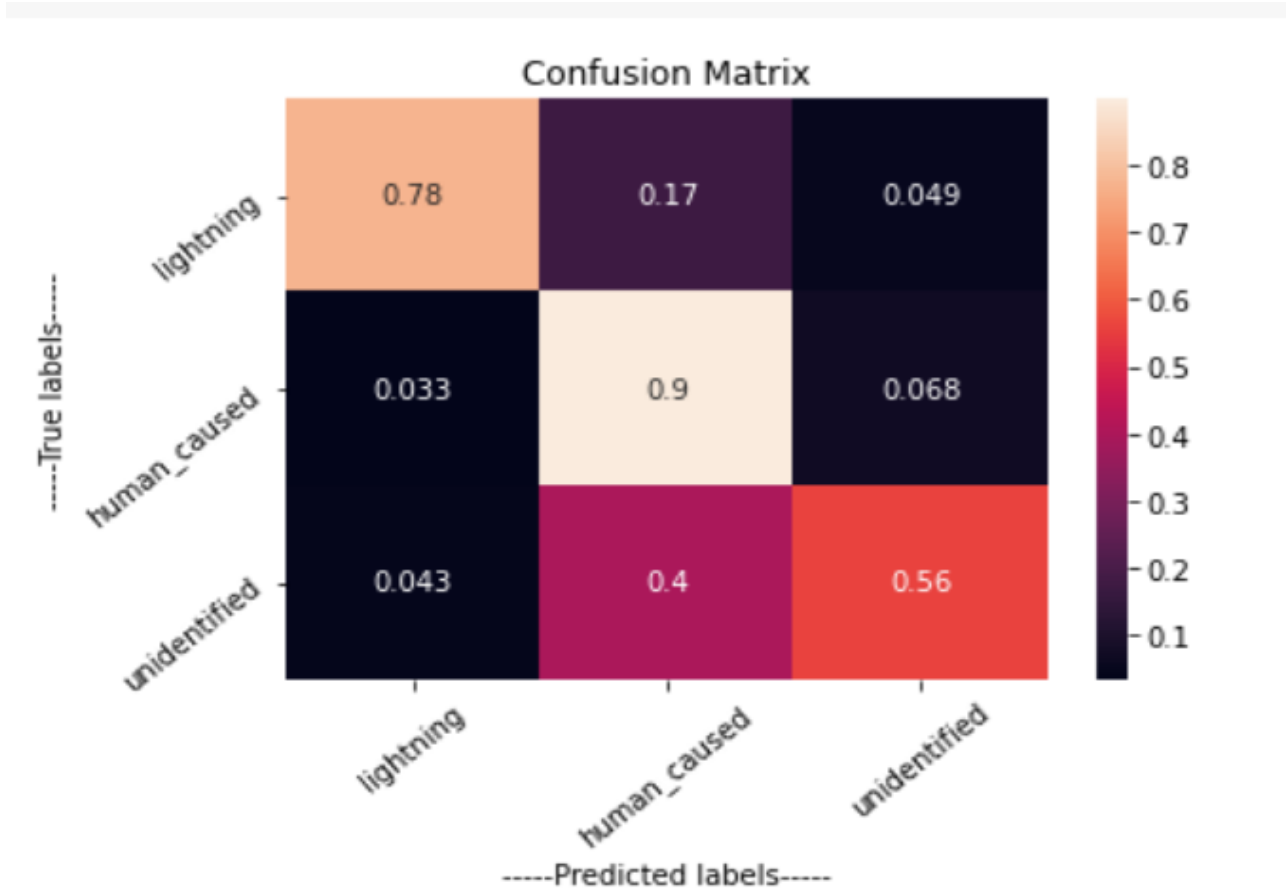
- We achieved an unspectacular improvement in accuracy of 1% after using bet parameters found using random search.

Feature Classification

- Classify cause of fire into 3 major classes and test if prediction score improves.
- The 3 classes are: **lightning, human_caused and Unidentified/other.**
  - **lightning** = ['Lightning’]
  - **human_caused** = ['Arson','Fireworks','Powerline','Railroad','Smoking','Children','Campfire','Equipment Use’, 'Debris Burning’, 'Structure’]
  - **Unidentified/other** = ['Missing/Undefined’, 'Miscellaneous’]
- Replaced STAT_CAUSE_DESCR with LABEL.
- 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%.

confusion matrix

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



Other approaches

- Created a Random Forest model to predict cause of fire only for California.
  - Prediction score was 66%

- Taking a random sample of the full dataset and applying a Random Forest Classifier Algorithm on this sample.
  - Prediction score was 57%

Conclusion

- Counties like Riverside County, San Bernardino, and San Diego in California, Coconino, and Gila in Arizona, Beltrami in Minnesota have a greater number of fires.
- Yukon-Koyukuk, Stanton, and other counties in Alaska have largest fires that have destroyed over thousands of acres.
- Most wildfires across the USA are caused by debris burning and arson.
- Fires started by electricity(caused by lightning and Powerline) are the most damaging.
- Lighting during summer causes the most dangerous fires.
- 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.

Recommendation

- 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.

- 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 daily. 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.

Thank You