Predicting Forest Fires in California

Using Historical Weather and Fire Data

By: Adam Swan, Jeff Warchall, Omar Younis

Image Source: Noah Berger - Associated Press

Problem Statement:

To help the California Department of Forestry and Fire Protection allocate resources, can we predict the likelihood of fires utilizing historical weather and wildfire data?

Is it desirable to accurately predict as many fires as possible, or:

To minimize false alarms?

Data Sources

- Geographic and wildfire data from the State of the California
  - 4,279 fires from July 1, 2008 to December 31, 2020

- Weather data from World Weather Online
  - Historical weather records from every day in the same range
  - Aggregated to a monthly basis

Web Scraping - Weather

- World Weather Online:
    - Includes temperature, precipitation, and wind data for:
    - Every weather station, for:
    - Every day:
    - Since July 1, 2008

- Data obtained from every county in California
    - Aggregated by month

Web Scraping - Fires

- The available fire data includes:
  - Date of fire
  - Number of acres burned
  - Cause of fire

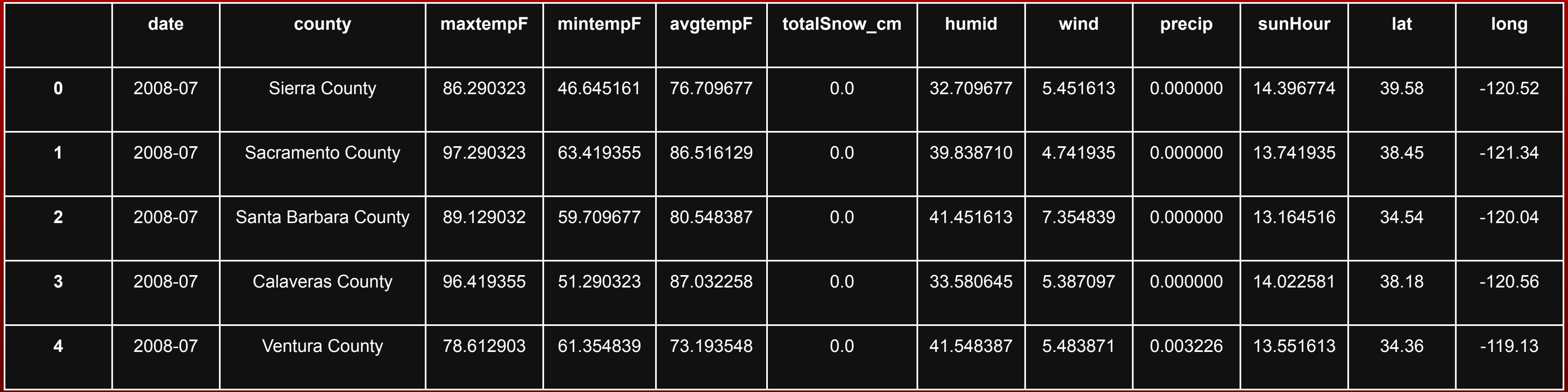
- Data extends back to 1952, but truncated at 2008 based on available weather data

Preprocessing

- Preprocessing was extensive for this project:
  - Multi-indexing
  - Monthly aggregation
  - Quarterly aggregation
  - Table joins
  - Multiple simultaneous fires
  - Data imputing
  - Time series (?)

Preprocessing

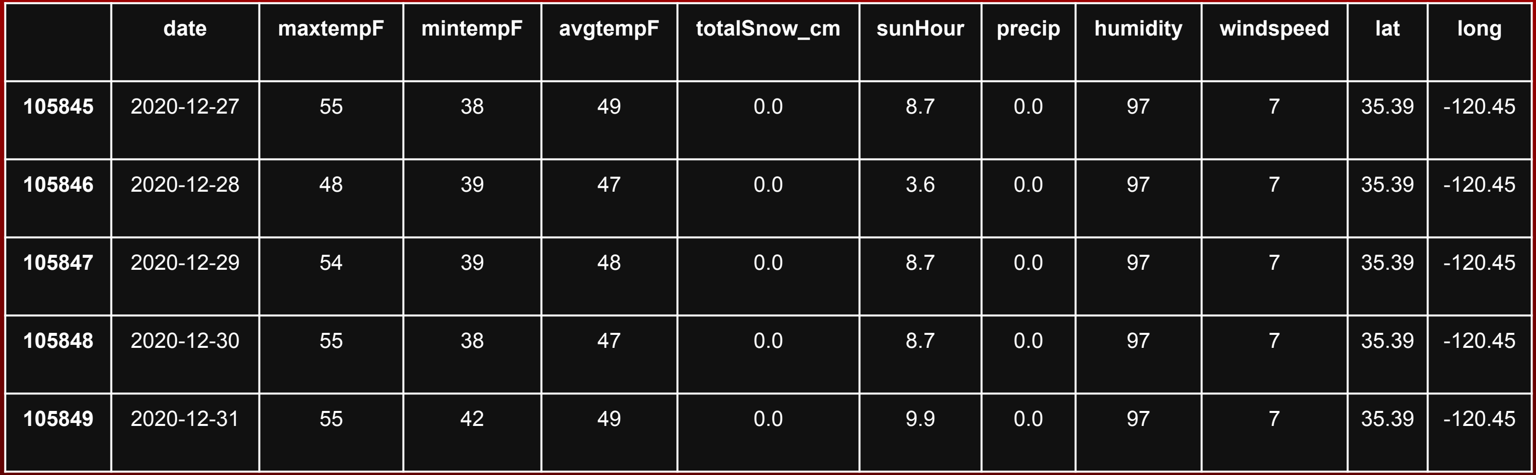
- Multi-Indexing



	date	county	maxtempF	mintempF	avgtempF	totalSnow_cm	humid	wind	precip	sunHour	lat	long
0	2008-07	Sierra County	86.290323	46.645161	76.709677	0.0	32.709677	5.451613	0.000000	14.396774	39.58	-120.52
1	2008-07	Sacramento County	97.290323	63.419355	86.516129	0.0	39.838710	4.741935	0.000000	13.741935	38.45	-121.34
2	2008-07	Santa Barbara County	89.129032	59.709677	80.548387	0.0	41.451613	7.354839	0.000000	13.164516	34.54	-120.04
3	2008-07	Calaveras County	96.419355	51.290323	87.032258	0.0	33.580645	5.387097	0.000000	14.022581	38.18	-120.56
4	2008-07	Ventura County	78.612903	61.354839	73.193548	0.0	41.548387	5.483871	0.003226	13.551613	34.36	-119.13

Preprocessing

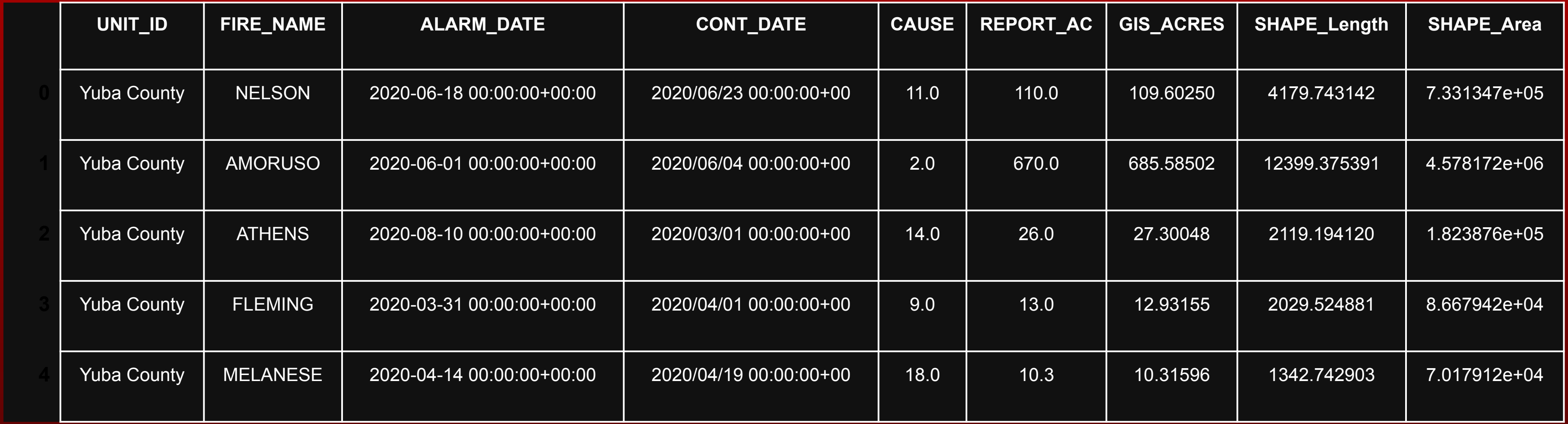
- Aggregation
  - Data collected form every day
  - Looking for monthly / quarterly data



	date	maxtempF	mintempF	avgtempF	totalSnow_cm	sunHour	precip	humidity	windspeed	lat	long
105845	2020-12-27	55	38	49	0.0	8.7	0.0	97	7	35.39	-120.45
105846	2020-12-28	48	39	47	0.0	3.6	0.0	97	7	35.39	-120.45
105847	2020-12-29	54	39	48	0.0	8.7	0.0	97	7	35.39	-120.45
105848	2020-12-30	55	38	47	0.0	8.7	0.0	97	7	35.39	-120.45
105849	2020-12-31	55	42	49	0.0	9.9	0.0	97	7	35.39	-120.45

Preprocessing

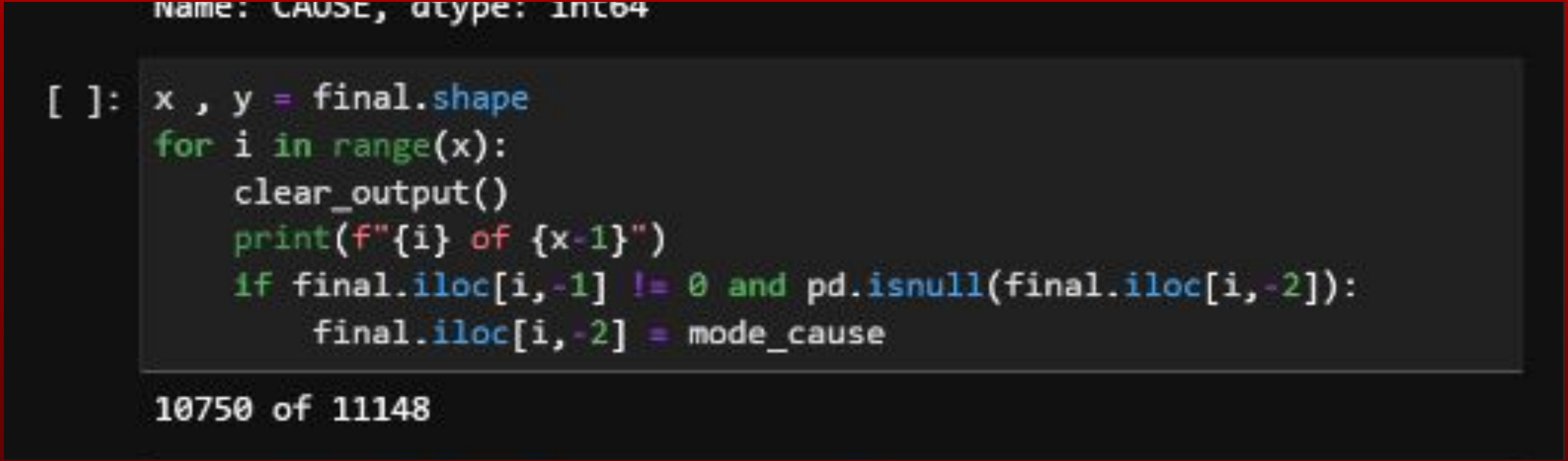
- Joining simultaneous records
  - More than 1 fire in a county in a month



	UNIT_ID	FIRE_NAME	ALARM_DATE	CONT_DATE	CAUSE	REPORT_AC	GIS_ACRES	SHAPE_Length	SHAPE_Area
0	Yuba County	NELSON	2020-06-18 00:00:00+00:00	2020/06/23 00:00:00+00	11.0	110.0	109.60250	4179.743142	7.331347e+05
1	Yuba County	AMORUSO	2020-06-01 00:00:00+00:00	2020/06/04 00:00:00+00	2.0	670.0	685.58502	12399.375391	4.578172e+06
2	Yuba County	ATHENS	2020-08-10 00:00:00+00:00	2020/03/01 00:00:00+00	14.0	26.0	27.30048	2119.194120	1.823876e+05
3	Yuba County	FLEMING	2020-03-31 00:00:00+00:00	2020/04/01 00:00:00+00	9.0	13.0	12.93155	2029.524881	8.667942e+04
4	Yuba County	MELANESE	2020-04-14 00:00:00+00:00	2020/04/19 00:00:00+00	18.0	10.3	10.31596	1342.742903	7.017912e+04

Preprocessing

- Data imputation
    - Months when there is no fire
    - Cases when a fire's cause is missing



```
Name: CAUSE, dtype: int64

[ ]: x , y = final.shape
     for i in range(x):
         clear_output()
         print(f"{i} of {x-1}")
         if final.iloc[i,-1] != 0 and pd.isnull(final.iloc[i,-2]):
             final.iloc[i,-2] = mode_cause

     10750 of 11148
```

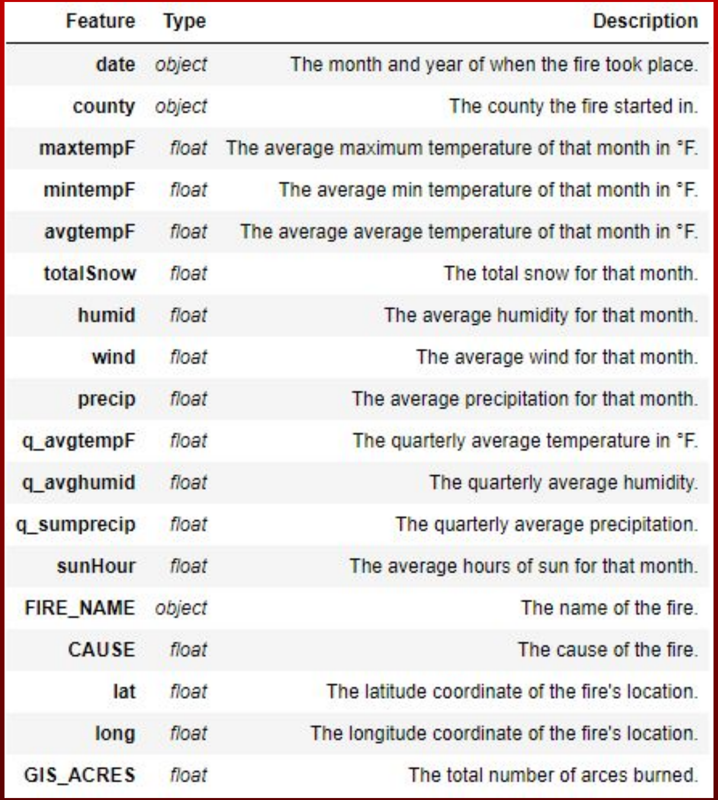
Preprocessing

- Is this a time series?
    - Ultimately, no; but:
    - Time series tools were useful in feature engineering

Final Dataset

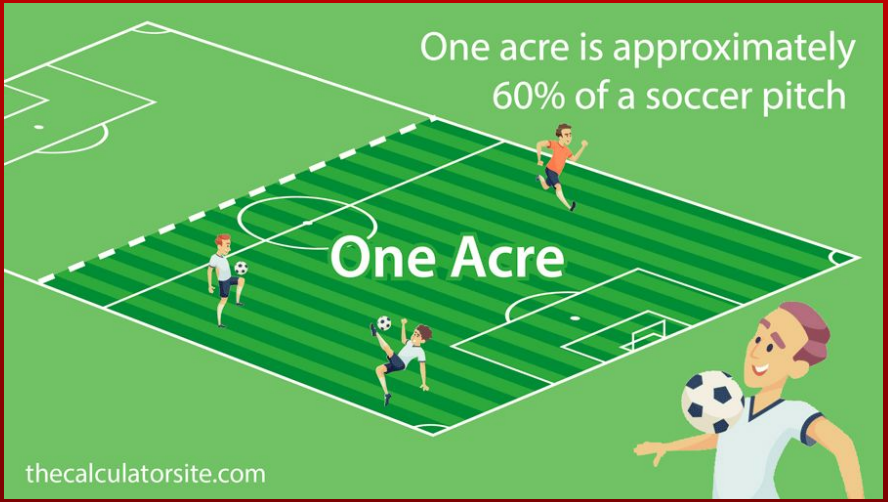
- The final dataset passed to the visualization department contained:
    - 10,988 records
    - 4,297 of which had fires
    - Features for:
        - Monthly weather averages
        - Quarterly weather averages
        - Quarterly cumulative precipitation
        - Acres burned
        - Cause of fire
        - Geo Location

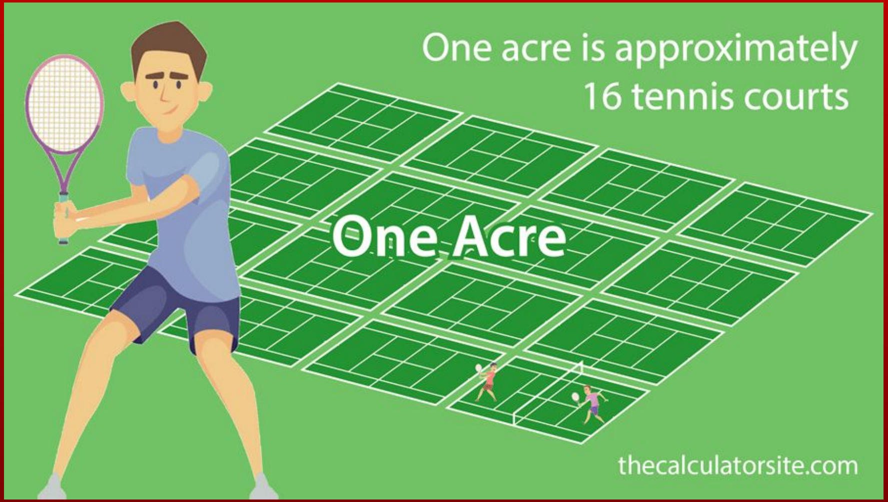
Data Dictionary



Feature	Type	Description
date	*object*	The month and year of when the fire took place.
county	*object*	The county the fire started in.
maxtempF	*float*	The average maximum temperature of that month in °F.
mintempF	*float*	The average min temperature of that month in °F.
avgtempF	*float*	The average average temperature of that month in °F.
totalSnow	*float*	The total snow for that month.
humid	*float*	The average humidity for that month.
wind	*float*	The average wind for that month.
precip	*float*	The average precipitation for that month.
q_avgtempF	*float*	The quarterly average temperature in °F.
q_avghumid	*float*	The quarterly average humidity.
q_sumprecip	*float*	The quarterly average precipitation.
sunHour	*float*	The average hours of sun for that month.
FIRE_NAME	*object*	The name of the fire.
CAUSE	*float*	The cause of the fire.
lat	*float*	The latitude coordinate of the fire's location.
long	*float*	The longitude coordinate of the fire's location.
GIS_ACRES	*float*	The total number of arces burned.

How Big is an Acre?

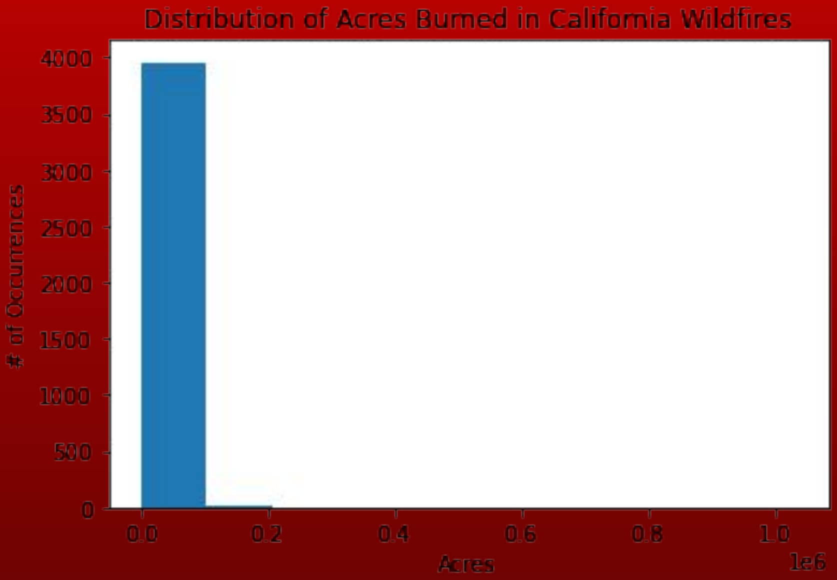


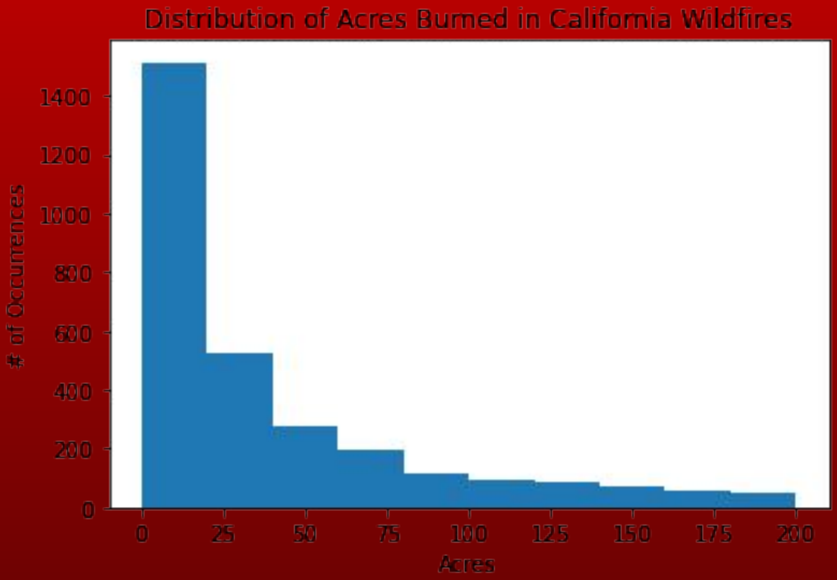


43,560 sq/ft or .001562 sq/mi

Image source - www.thecalculatorsite.com

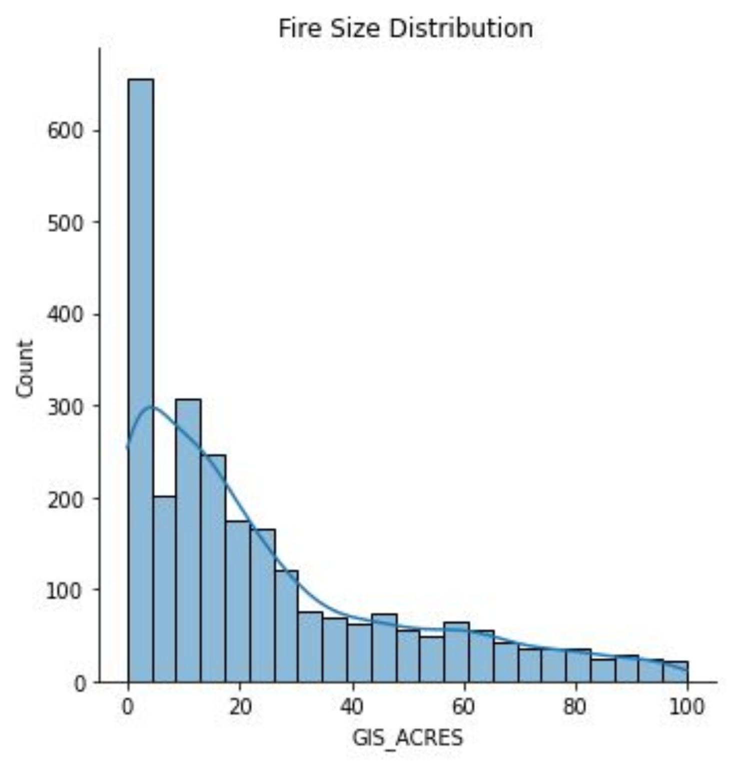
Fire Size Distribution:





61% of 4,279 fires between 2008 - 2020 were smaller than 100 acres - roughly .15 sq/mi

Fire Size Distribution (Continued)



- Break Down Version.

Largest Fire:

Year - 2020
Cause - Lightning
Damage - 1.03 million acres (1,562 sq/mi), nearly the size of Harris county, Texas

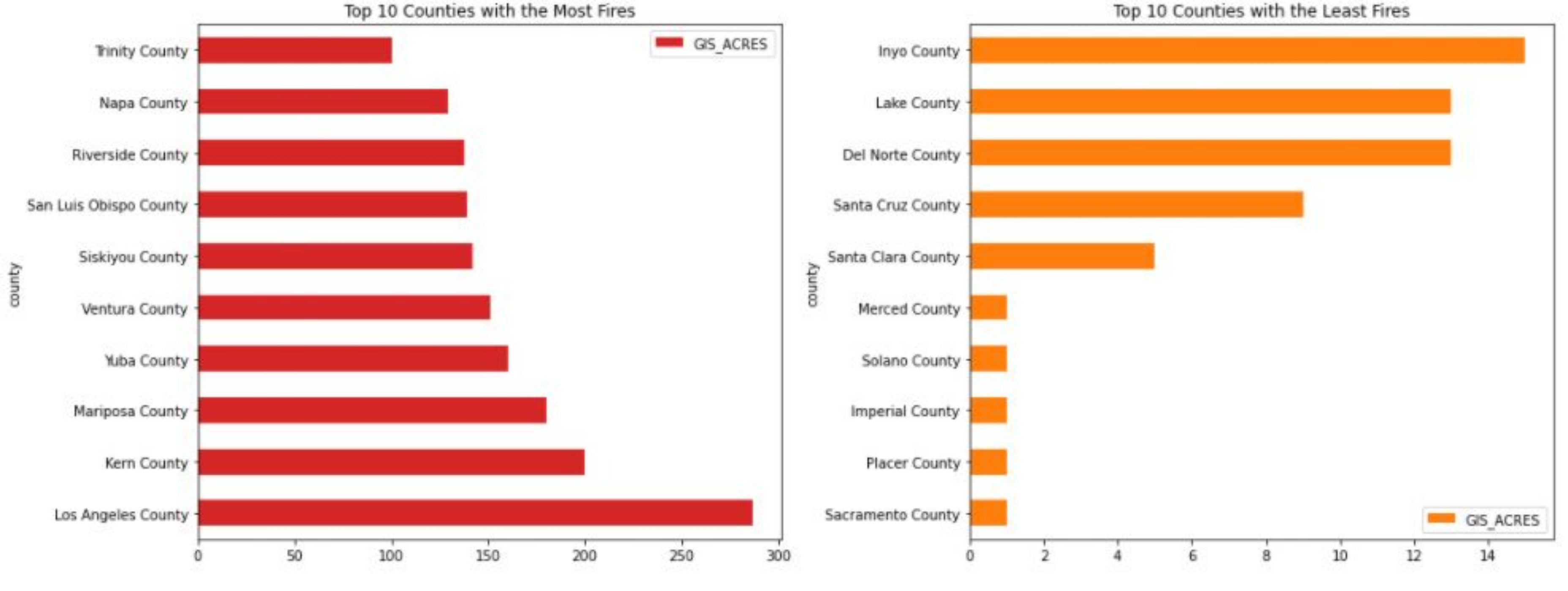
MONTGOMERY
LIBERTY
CHAMBERS
GALVESTON
BRAZORIA
FORT BEND
WALLER
Waller
Hockley
Hufsmith
Tomball
Spring
Klein
Kingwood
Atascocita
Huffman
Humble
George Bush Intercontinental Airport
Lake Houston
Cypress
Aldine
Jersey Village
Crosby
Sheldon
SHELDON WILDLIFE MGMT AREA
SHELDON LAKE STATE PARK
Barrett
Highlands
McNair
HOUSTON
Katy
Barker
Addicks
Addicks Reservoir
Barker Reservoir
Spring Valley
Hilshire Vil.
Hedwig Vil.
Hunters Creek Vil.
Piney Point Vil.
Bunker Hill Vil.
Jacinto City
Channelview
Galena Park
Baytown
BATTLESHIP TEXAS
SAN JACINTO MONUMENT
ATKINSON ISLAND WMA
Morgans Point
Alief
Bellaire
West University Place
Southside Place
Pasadena
Deer Park
South Houston
La Porte
Stafford
Missouri City
William P. Hobby Airport
Shoreacres
Taylor Lake Village
Johnson Space Center
El Lago
Seabrook
Pearland
Friendswood
Webster
Nassau Bay
League City
Galveston Bay
Houston Ship Channel
Clear Lake
Clear Creek
Spring Creek
Cypress Creek
Buffalo Bayou
Greens Bayou
Cedar Bayou
San Jacinto River
West Fork San Jacinto
East Fork San Jacinto
Free Ferry
0 8 MILES
Harris County © Texas Almanac

Image source - www.hsaonline.org

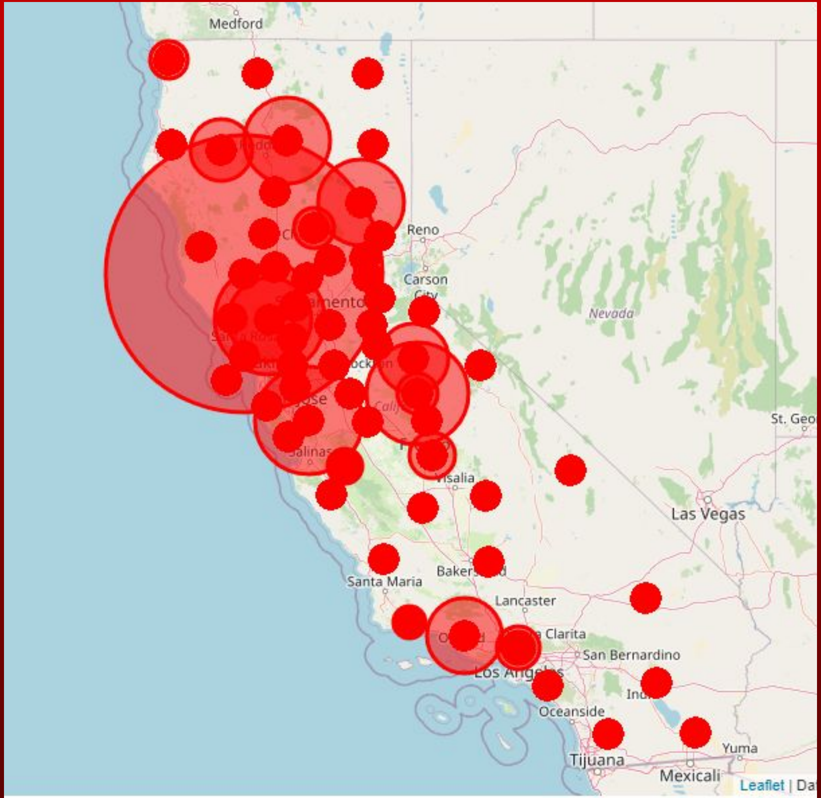
Top 8 Causes of CA Wildfires:

Rank	Cause	# of Occurrences
1	Unknown	1,208
2	Lightning	832
3	Equipment use	439
4	Vehicle	262
5	Powerline	207
6	Arson	184
7	Debris	153
8	Campfire	96

Counties with Most and Least Fires

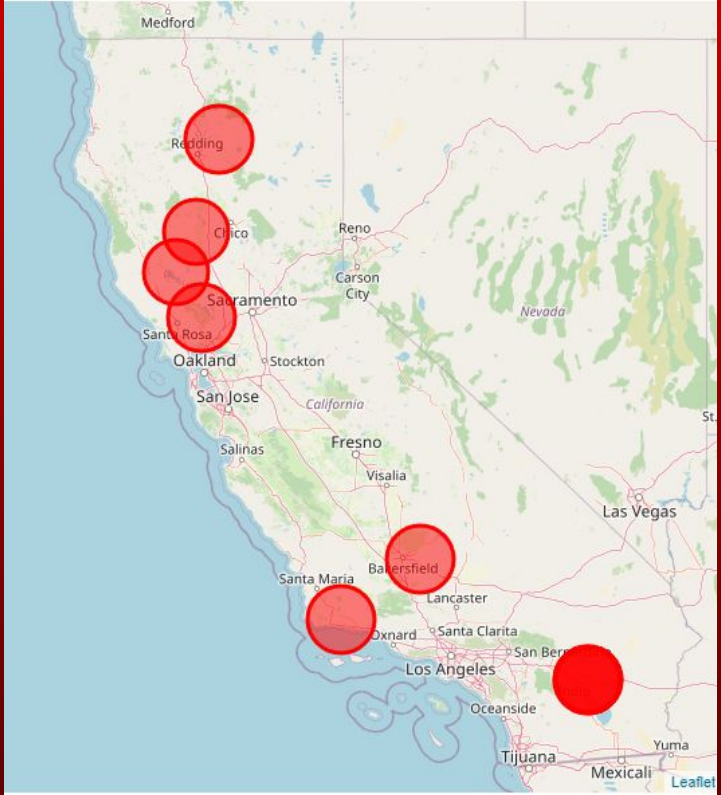


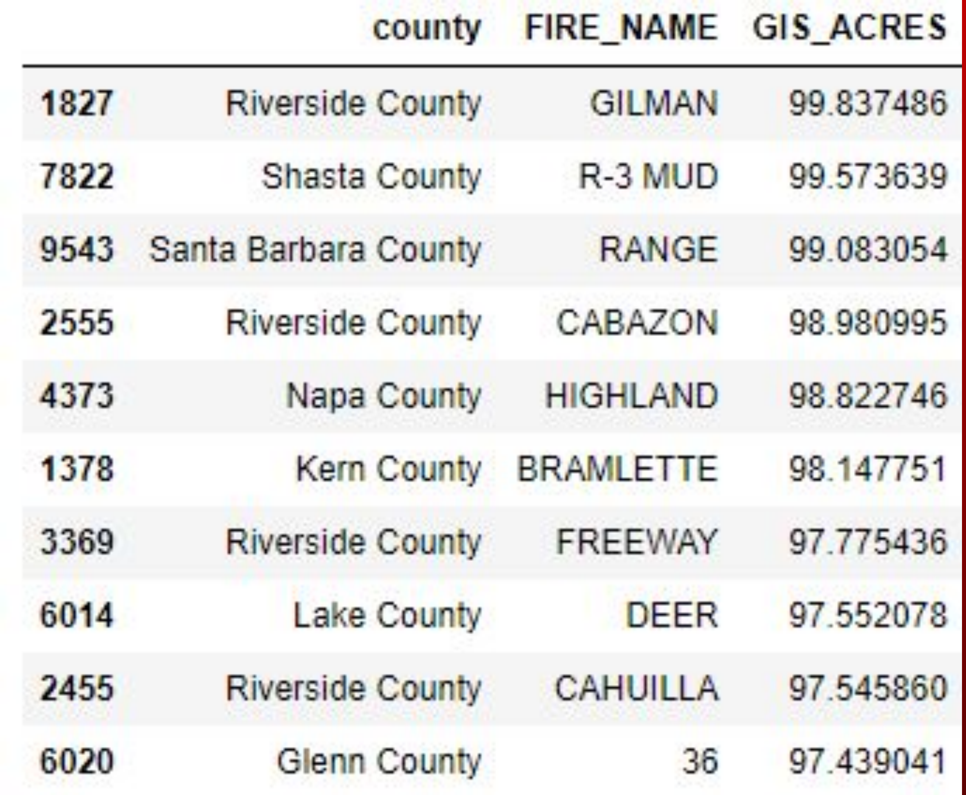
Interactive Fire Map



- Interactive .html file.
- Fire Size (marker size)
- All Fires (not limited to sizes less than 100 acres).
- Most fires in Northern California.

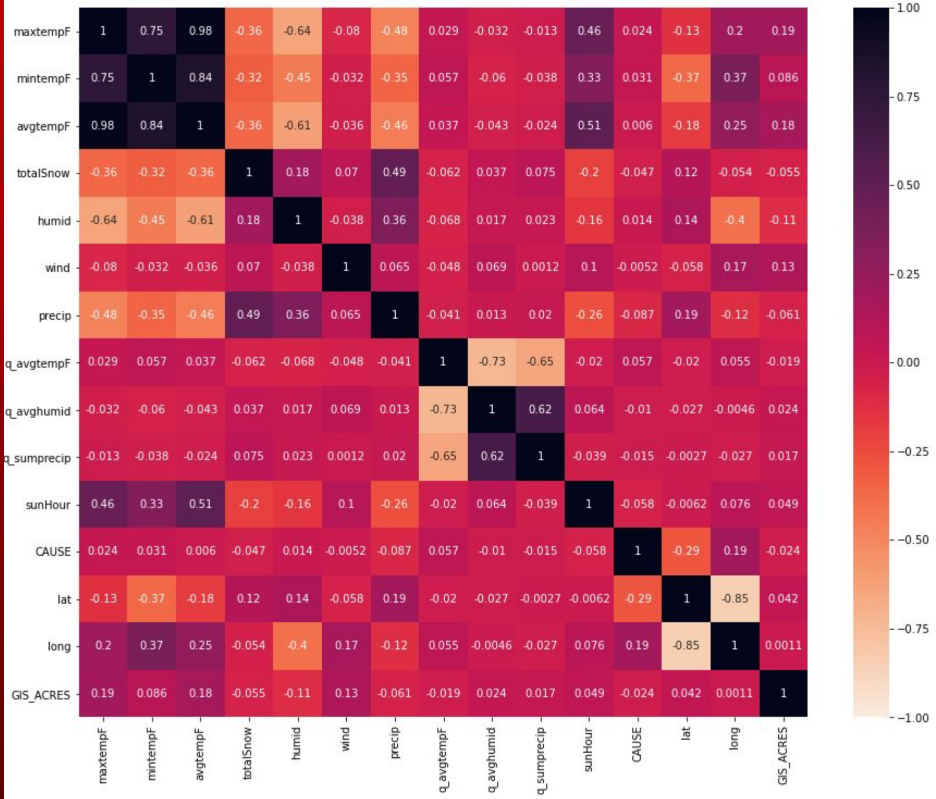
Top 10 Under 100





	county	FIRE_NAME	GIS_ACRES
1827	Riverside County	GILMAN	99.837486
7822	Shasta County	R-3 MUD	99.573639
9543	Santa Barbara County	RANGE	99.083054
2555	Riverside County	CABAZON	98.980995
4373	Napa County	HIGHLAND	98.822746
1378	Kern County	BRAMLETTE	98.147751
3369	Riverside County	FREEWAY	97.775436
6014	Lake County	DEER	97.552078
2455	Riverside County	CAHUILLA	97.545860
6020	Glenn County	36	97.439041

Heatmap (All)

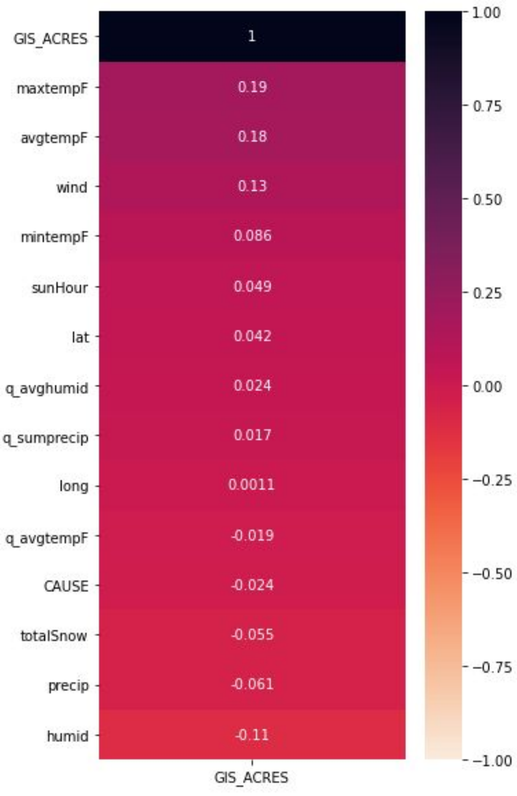


	maxtempF	mintempF	avgtempF	totalSnow	humid	wind	precip	q_avgtempF	q_avghumid	q_sumprecip	sunHour	CAUSE	lat	long	GIS_ACRES
maxtempF	1	0.75	0.98	-0.36	-0.64	-0.08	-0.48	0.029	-0.032	-0.013	0.46	0.024	-0.13	0.2	0.19
mintempF	0.75	1	0.84	-0.32	-0.45	-0.032	-0.35	0.057	-0.06	-0.038	0.33	0.031	-0.37	0.37	0.086
avgtempF	0.98	0.84	1	-0.36	-0.61	-0.036	-0.46	0.037	-0.043	-0.024	0.51	0.006	-0.18	0.25	0.18
totalSnow	-0.36	-0.32	-0.36	1	0.18	0.07	0.49	-0.062	0.037	0.075	-0.2	-0.047	0.12	-0.054	-0.055
humid	-0.64	-0.45	-0.61	0.18	1	-0.038	0.36	-0.068	0.017	0.023	-0.16	0.014	0.14	-0.4	-0.11
wind	-0.08	-0.032	-0.036	0.07	-0.038	1	0.065	-0.048	0.069	0.0012	0.1	-0.0052	-0.058	0.17	0.13
precip	-0.48	-0.35	-0.46	0.49	0.36	0.065	1	-0.041	0.013	0.02	-0.26	-0.087	0.19	-0.12	-0.061
q_avgtempF	0.029	0.057	0.037	-0.062	-0.068	-0.048	-0.041	1	-0.73	-0.65	-0.02	0.057	-0.02	0.055	-0.019
q_avghumid	-0.032	-0.06	-0.043	0.037	0.017	0.069	0.013	-0.73	1	0.62	0.064	-0.01	-0.027	-0.0046	0.024
q_sumprecip	-0.013	-0.038	-0.024	0.075	0.023	0.0012	0.02	-0.65	0.62	1	-0.039	-0.015	-0.0027	-0.027	0.017
sunHour	0.46	0.33	0.51	-0.2	-0.16	0.1	-0.26	-0.02	0.064	-0.039	1	-0.058	-0.0062	0.076	0.049
CAUSE	0.024	0.031	0.006	-0.047	0.014	-0.0052	-0.087	0.057	-0.01	-0.015	-0.058	1	-0.29	0.19	-0.024
lat	-0.13	-0.37	-0.18	0.12	0.14	-0.058	0.19	-0.02	-0.027	-0.0027	-0.0062	-0.29	1	-0.85	0.042
long	0.2	0.37	0.25	-0.054	-0.4	0.17	-0.12	0.055	-0.0046	-0.027	0.076	0.19	-0.85	1	0.0011
GIS_ACRES	0.19	0.086	0.18	-0.055	-0.11	0.13	-0.061	-0.019	0.024	0.017	0.049	-0.024	0.042	0.0011	1

- Significant correlation
  - |0.4| or greater

- No strong correlation between variables
  - Dependent Features

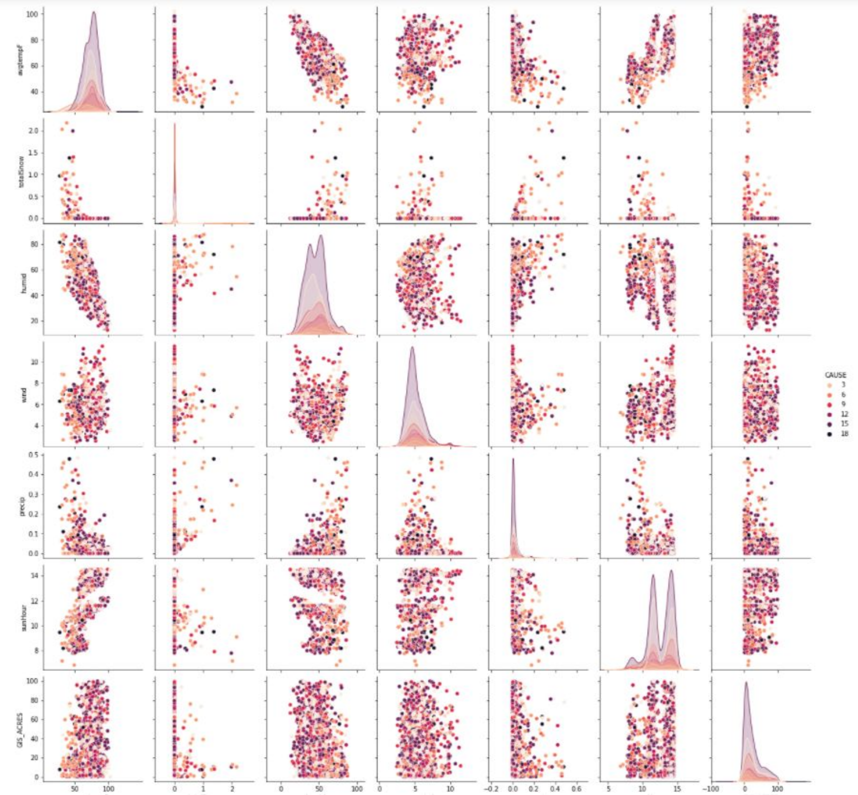
Heatmap (GIS_ACRES)



- Significant correlation
  - |0.4| or greater

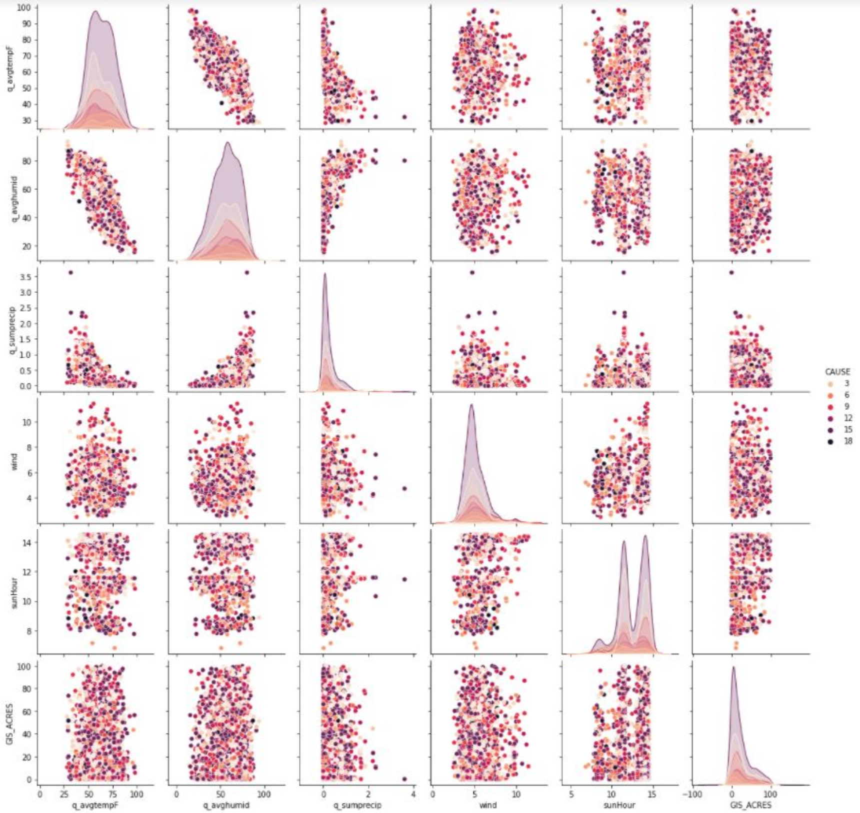
- No strong correlations

Pairplot Monthly



- No strong patterns in the scatter plots.
- Dependant Features.

Pairplot Monthly



- No strong patterns in the scatter plots.

- Dependant Features.

Modeling:

1) Logistic Regression
2) KNN Classifier
3) Random Forest Classifier
4) Voting Classifier

Model Performance:

Model Type	Accuracy
Logistic Regression	76%
KNN Classifier	85%
Random Forest Classifier	88%
Voting Classifier	87%

Logistic Regression Coefficients:

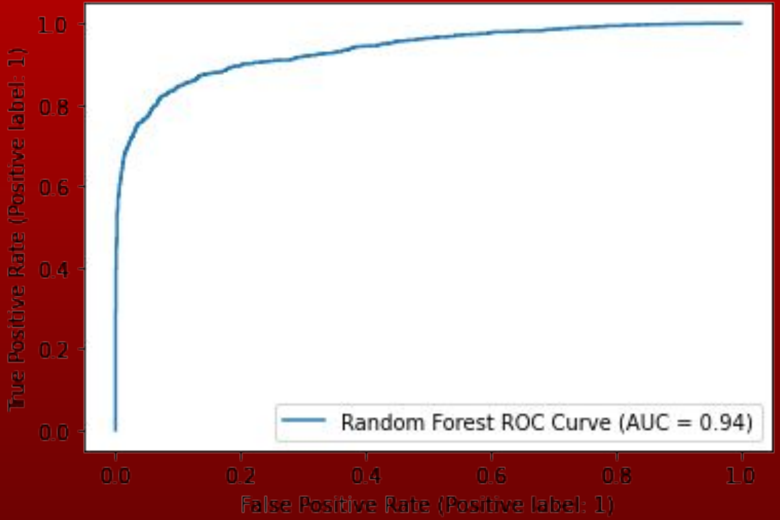
Variable:	Effect:
August	+108%
September	+69%
Hours of sun	+51%
November	+42%
July	+37%
December	+21%
Total snowfall	+2%

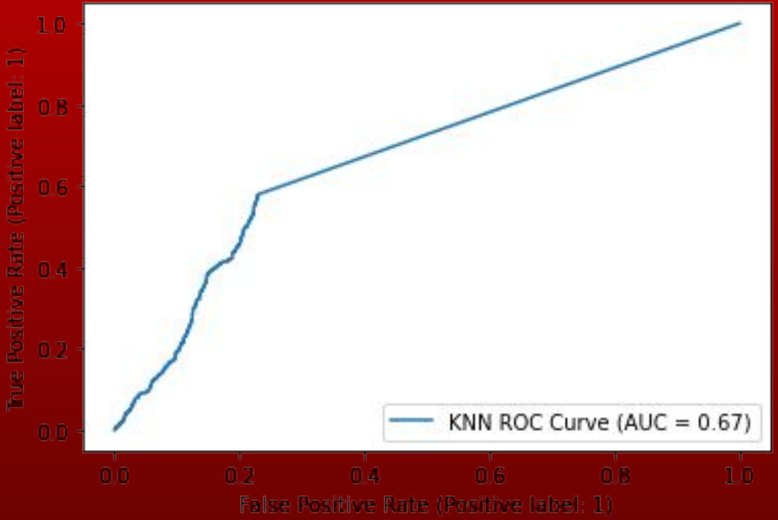
Variable:	Effect:
February	-64%
March	-50%
May	-49%
January	-49%
April	-44%
Wind	-28%
June	-22%

Random Forest Features:

Rank	Feature
1	Hours of sunlight
2	Average temp
3	Average humidity
4	Average wind speed
5	Average precipitation
6	Year
7	July
8	Monthly snowfall

Additional Metrics:





Additional Metrics cont:

Null model: 61% accuracy

Random Forest Classifier	
Metric	**Score**
AUC - ROC	87%
Recall	83%
Precision	84%
Accuracy	88%

Voting Classifier	
Metric	**Score**
AUC - ROC	N/A
Recall	86%
Precision	80%
Accuracy	87%

Conclusion :

Random Forest:

1) Higher accuracy
2) Higher Precision

Voting Classifier:

1) Lower precision
2) Higher recall

Recommendation:

Random Forest

If ample resources and funding are available, we recommend the Random Forest model as it positively identifies forest fires more often than our other models.

Cons - More false positives, possibly wasting resources.

Final Considerations:

Forest fires are:

- Unavoidable
- Healthy for a forest's life cycle
- Generally small when responded to quickly
- Increasing in danger and damage as humans populate and expand further out from urban and suburban areas

End

Thank you for your time, we will now take questions.