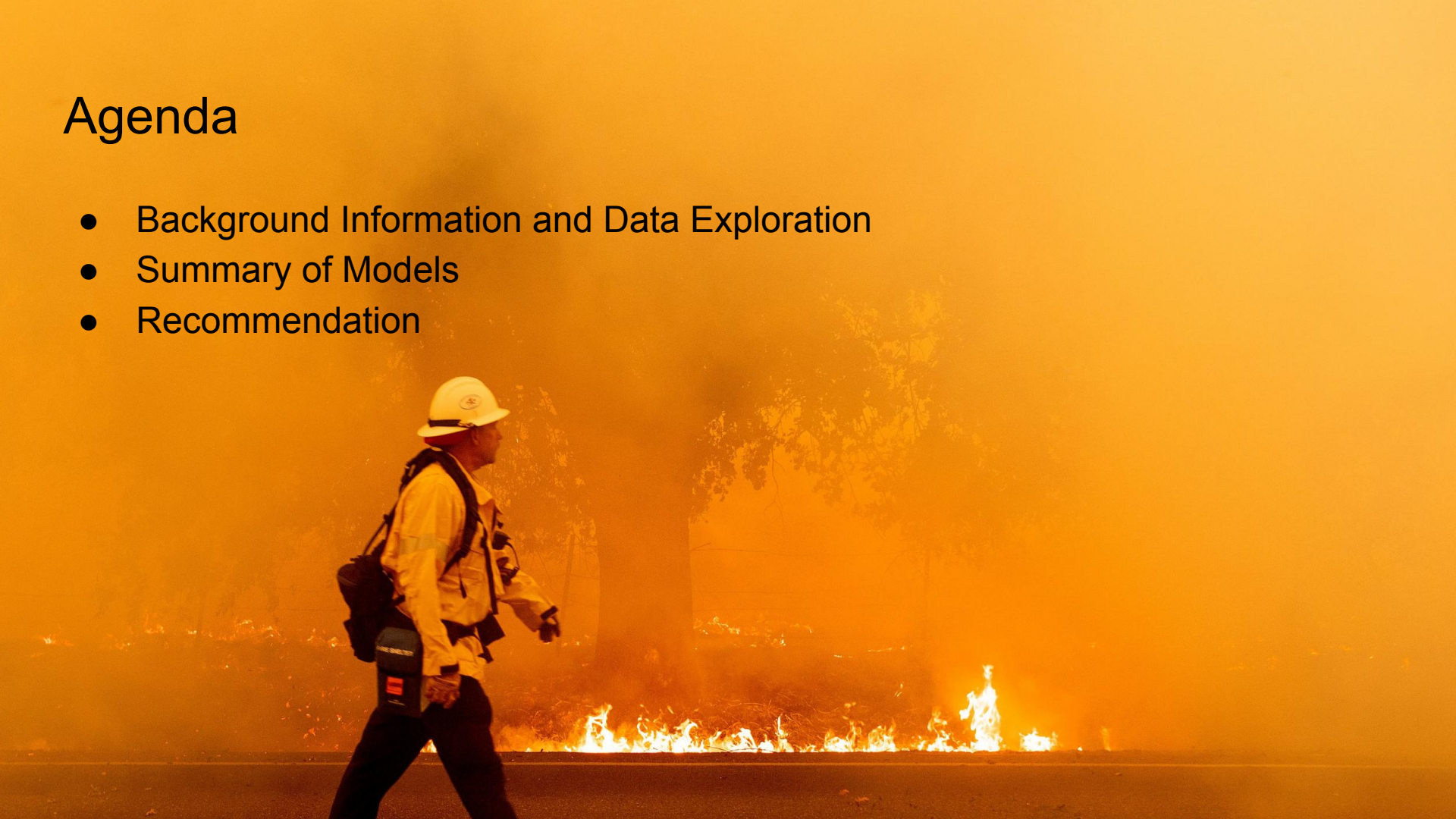
Predicting the Cause of Wildfires in California

Agenda

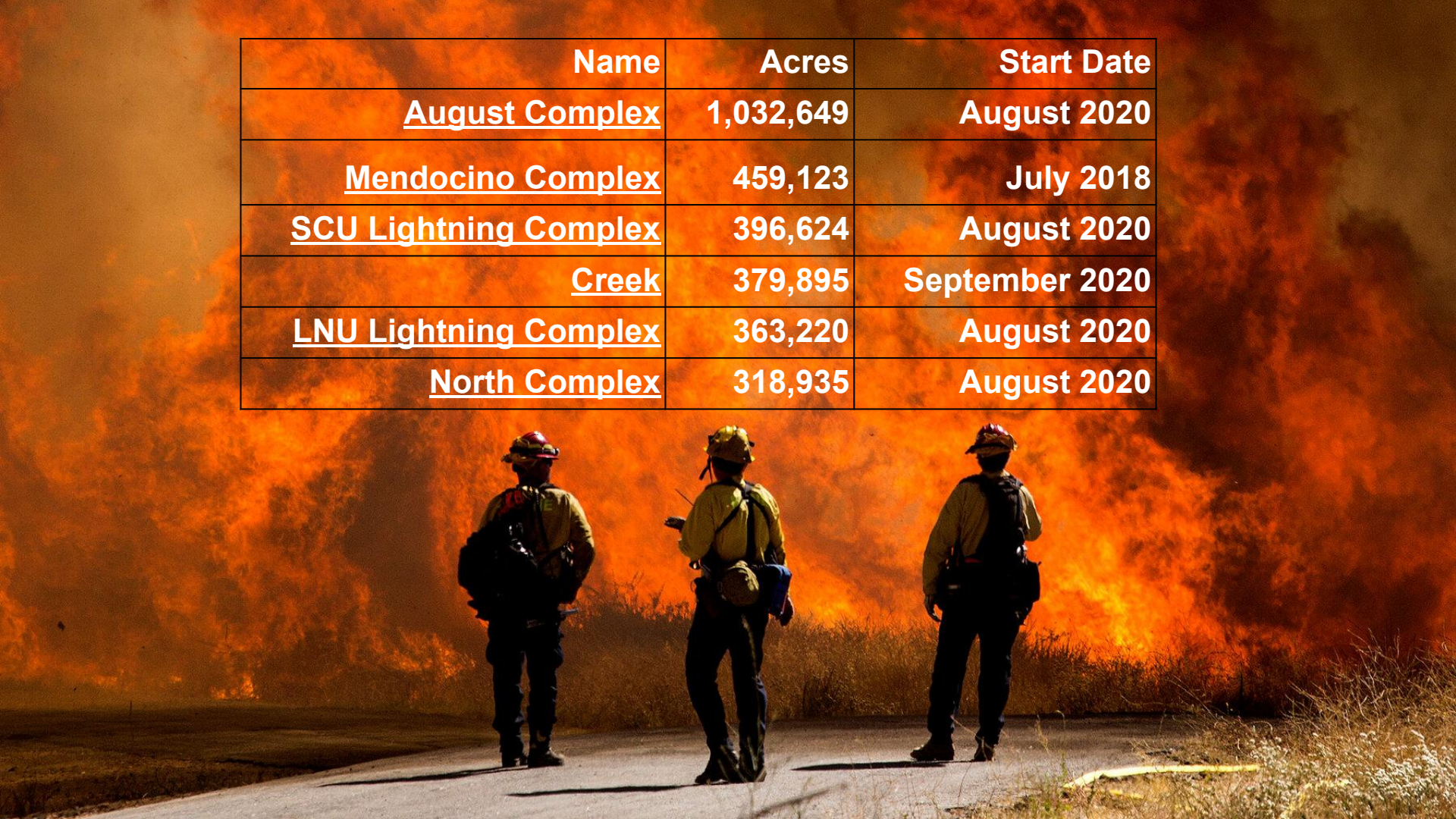
- Background Information and Data Exploration
- Summary of Models
- Recommendation

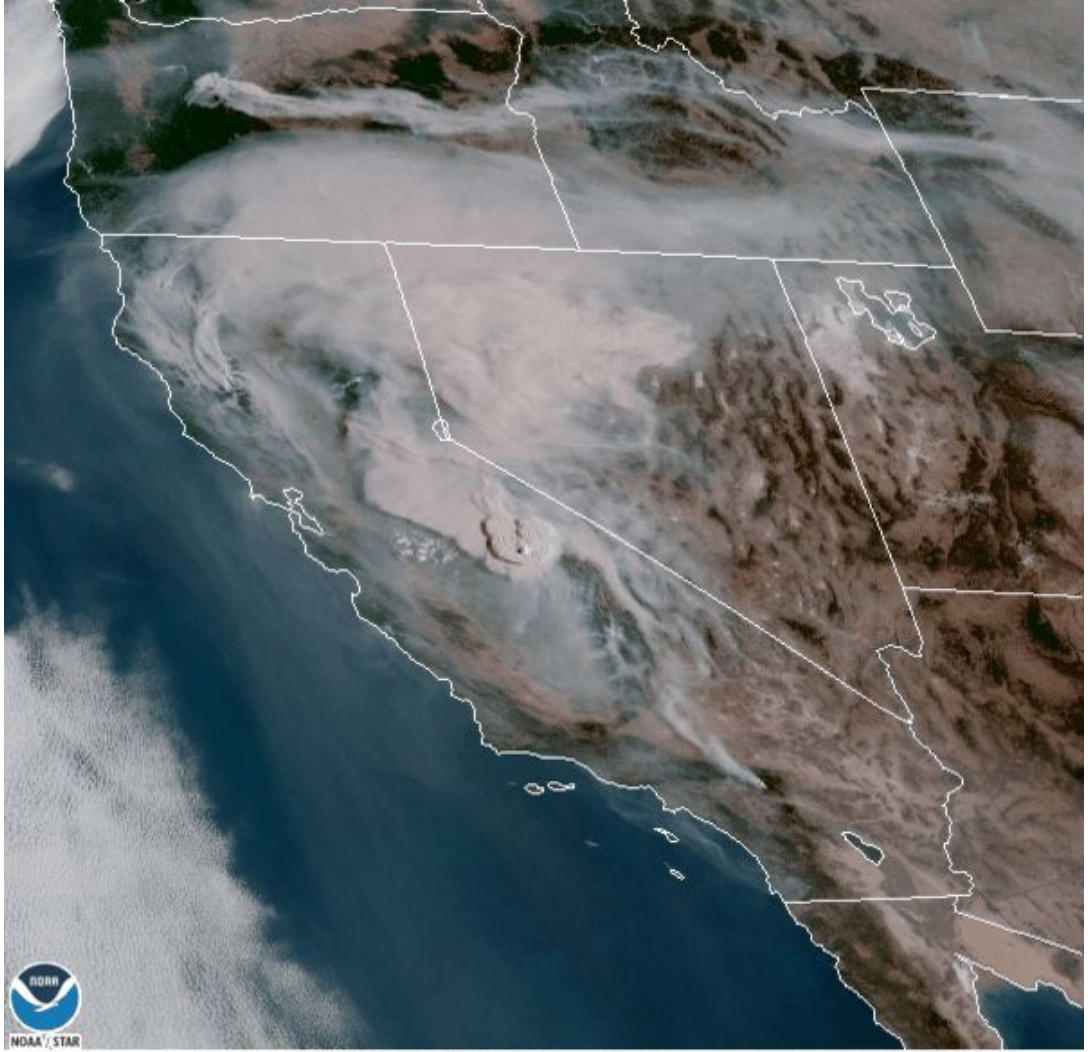



SENIOR CENTER
WEAR A MASK
WASH YOUR HANDS
SOCIAL DISTANCE
STAY SAFE
COME JOIN US


CAL FIRE
D2564
1512693

Name	Acres	Start Date
August Complex	1,032,649	August 2020
Mendocino Complex	459,123	July 2018
SCU Lightning Complex	396,624	August 2020
Creek	379,895	September 2020
LNU Lightning Complex	363,220	August 2020
North Complex	318,935	August 2020





06 Sep 2020 15:46Z NESDIS/STAR GOES-West GEOCOLOR


Slater Fire
Red Salmon Complex FIre
August Complex
North Complex Fire
Glass Fire
San Francisco
Creek Fire
SQF Complex Fire
N
100 km


Eureka
Redding
Carr fire
Susanville
Chico
Reno
Willits
Yuba City
Carson City
Mendocino Complex
Sacramento
Hawthorne
Santa Rosa
Donnell fire
Stockton
Yosemite National Park
San Francisco
Ferguson fire
San José
Merced
Santa Cruz
Fresno
California
Salinas
Visalia
Junipero Serra Pk.

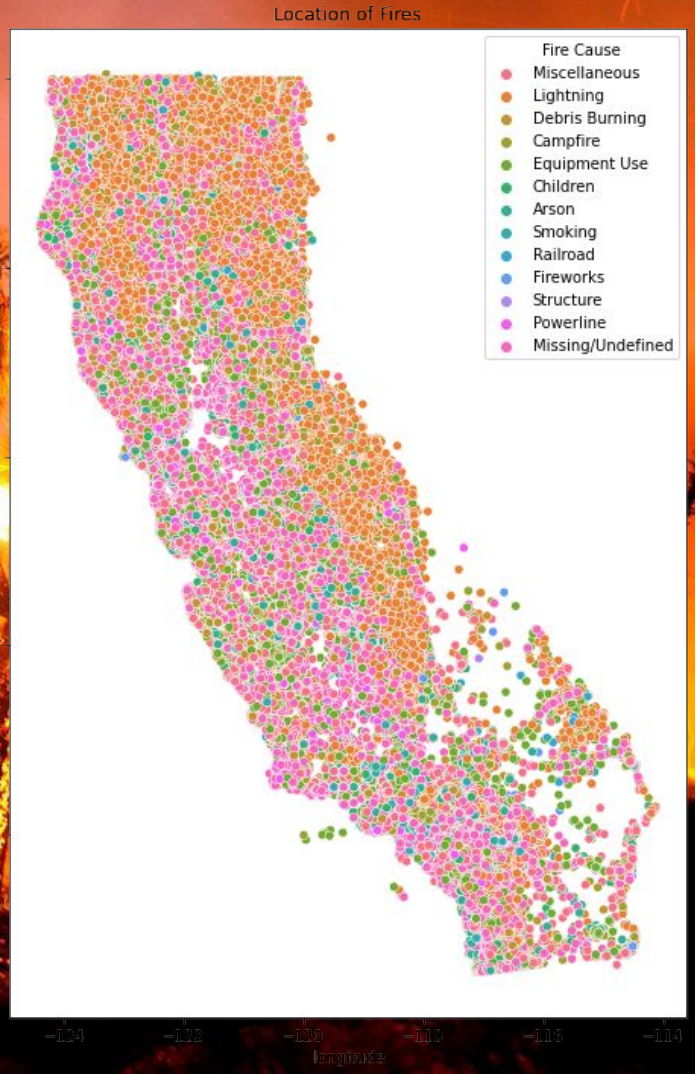
Exploratory Data Analysis

The Dataset

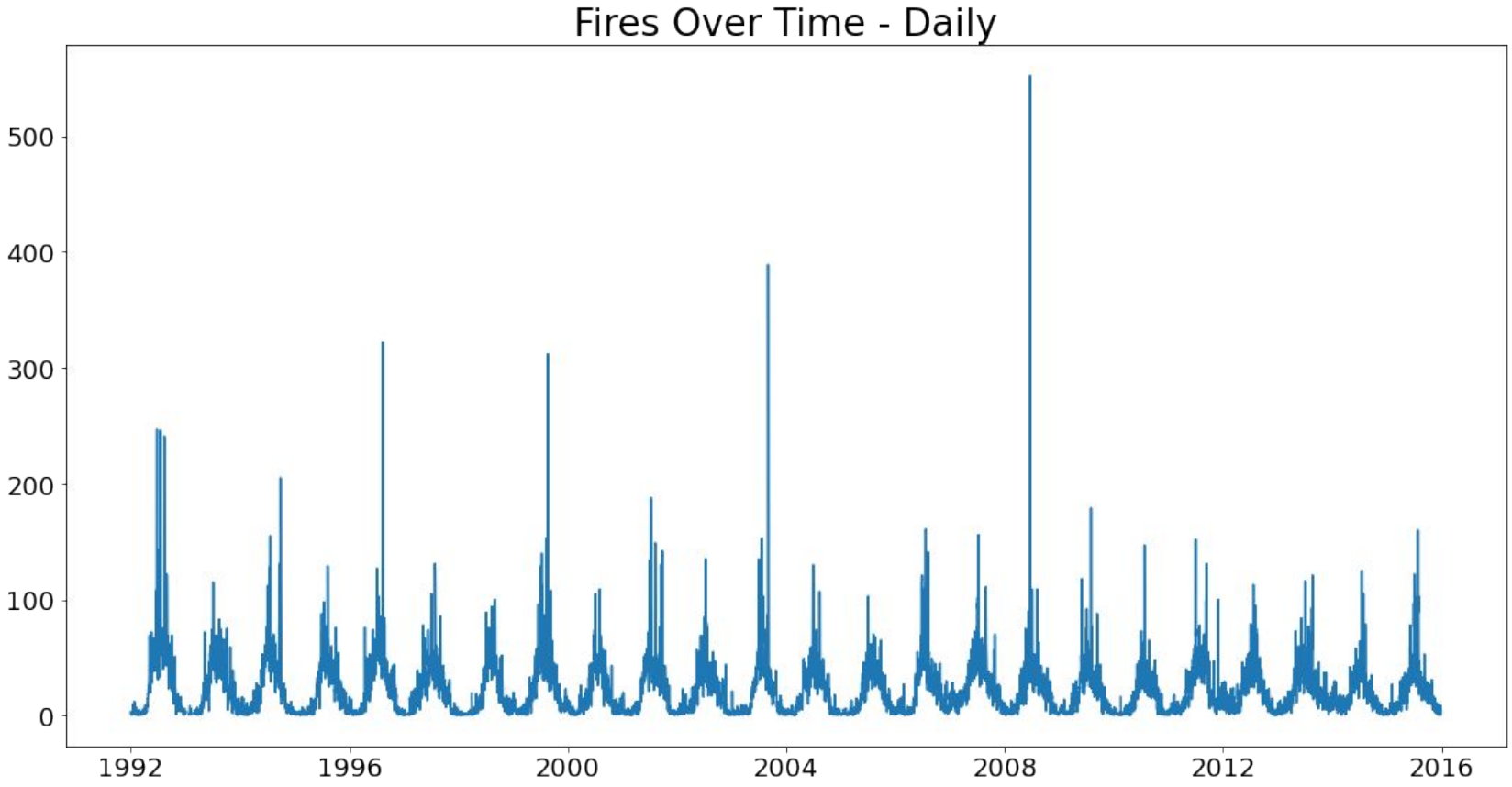
189,550 fires

January 1, 1992 - December 31, 2015

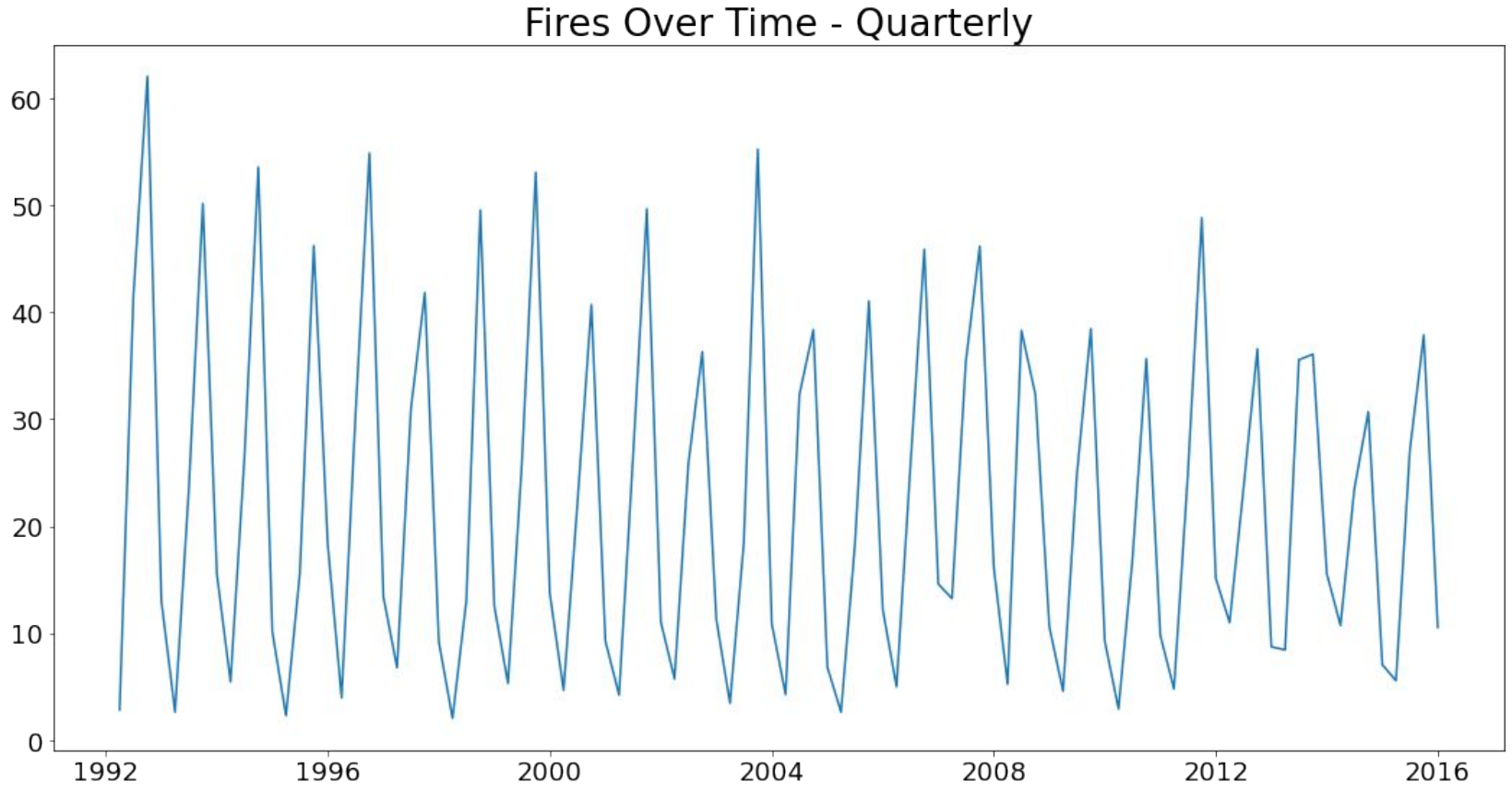
Features: fire year, discovery day, fire size, latitude, and longitude.

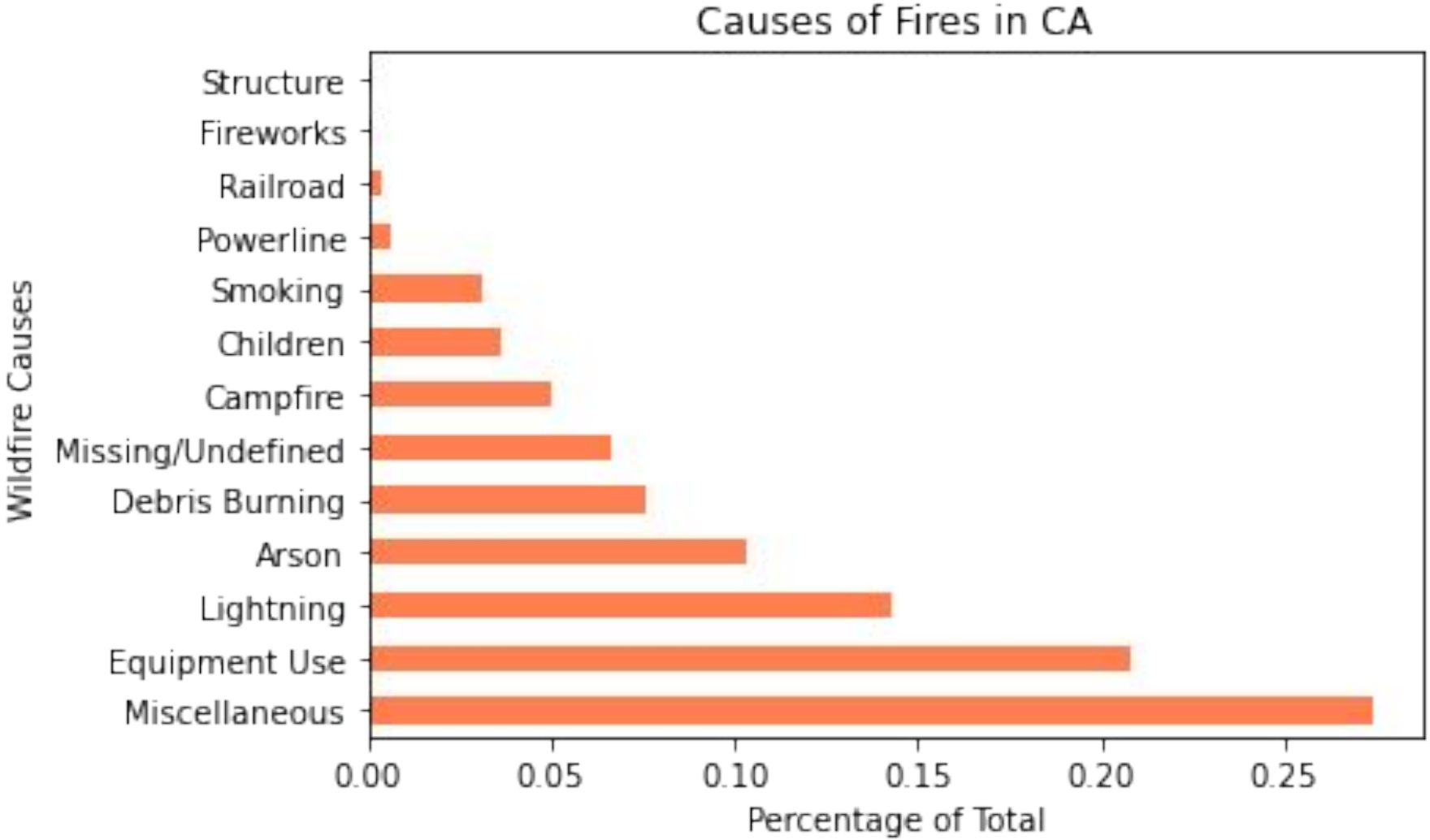


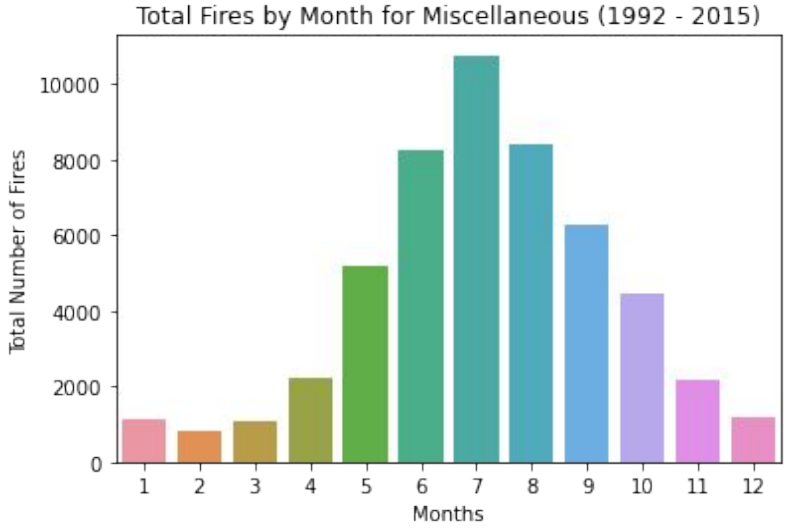
Seasonality

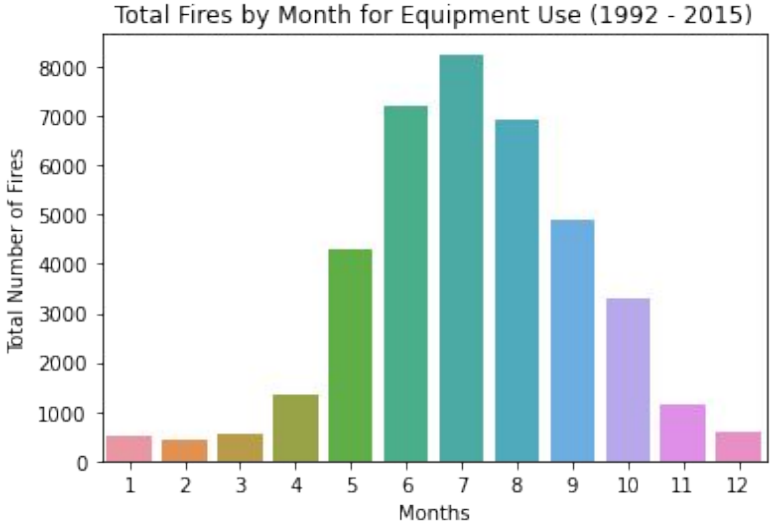


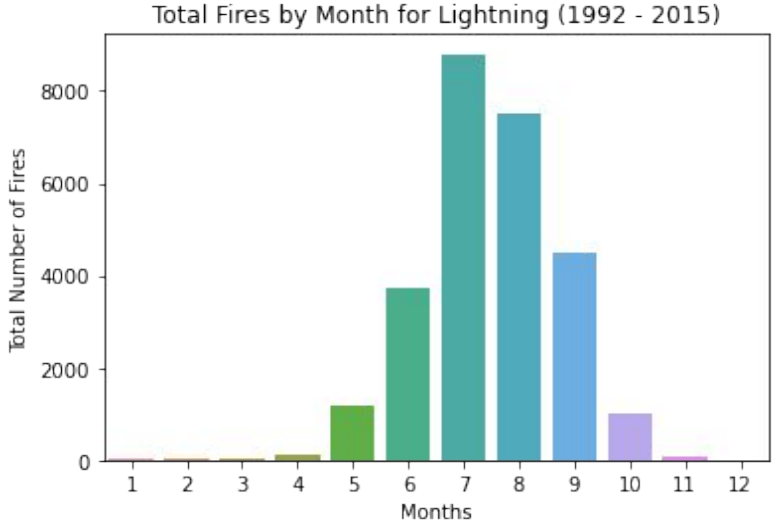
Seasonality




Causes of Fires in CA
Structure
Fireworks
Railroad
Powerline
Smoking
Children
Campfire
Missing/Undefined
Debris Burning
Arson
Lightning
Equipment Use
Miscellaneous
Wildfire Causes
0.00
0.05
0.10
0.15
0.20
0.25
Percentage of Total

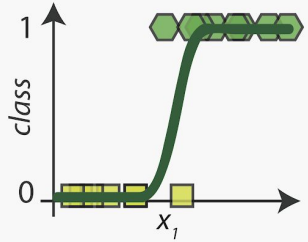

Total Fires by Month for Miscellaneous (1992 - 2015)
Total Number of Fires
0
2000
4000
6000
8000
10000
1
2
3
4
5
6
7
8
9
10
11
12
Months


Total Fires by Month for Equipment Use (1992 - 2015)
Total Number of Fires
0
1000
2000
3000
4000
5000
6000
7000
8000
1
2
3
4
5
6
7
8
9
10
11
12
Months

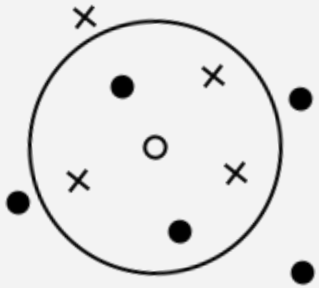

Total Fires by Month for Lightning (1992 - 2015)
Total Number of Fires
0
2000
4000
6000
8000
1
2
3
4
5
6
7
8
9
10
11
12
Months

Classification Models

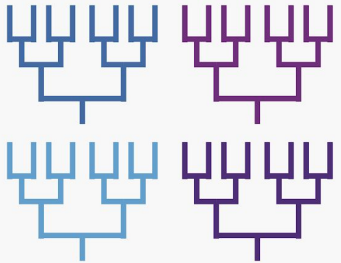
Logistic regression



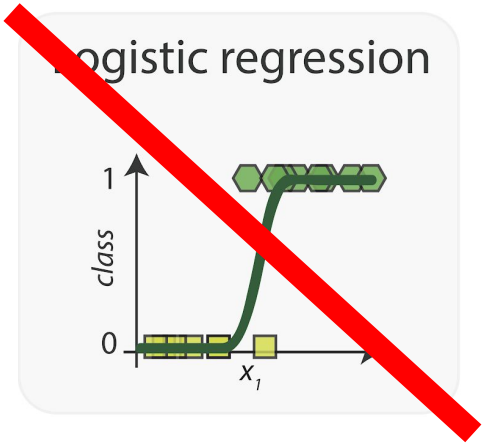
KNN



Random forests

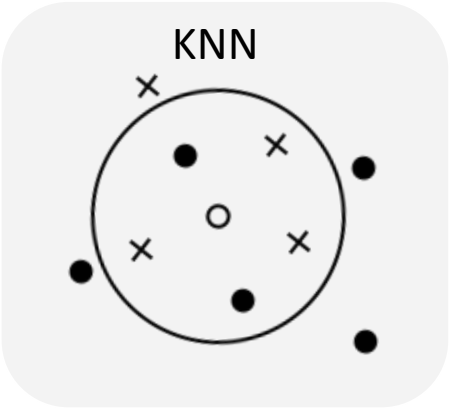


Classification Models



Accuracy:

Natural - 44%
Accidental - 66%
Malicious - 0%
Other - 47%



Accuracy:

Natural - 81%
Accidental - 62%
Malicious - 31%
Other - 42%

Accuracy:

Natural - 0%
Accidental - 0%
Malicious - 0%
Other - 100%

Recommendation

- KNN Model
- Strengths - lightning strikes
- Weaknesses - arson

Next Steps

Incorporate weather data

Incorporate newer fires

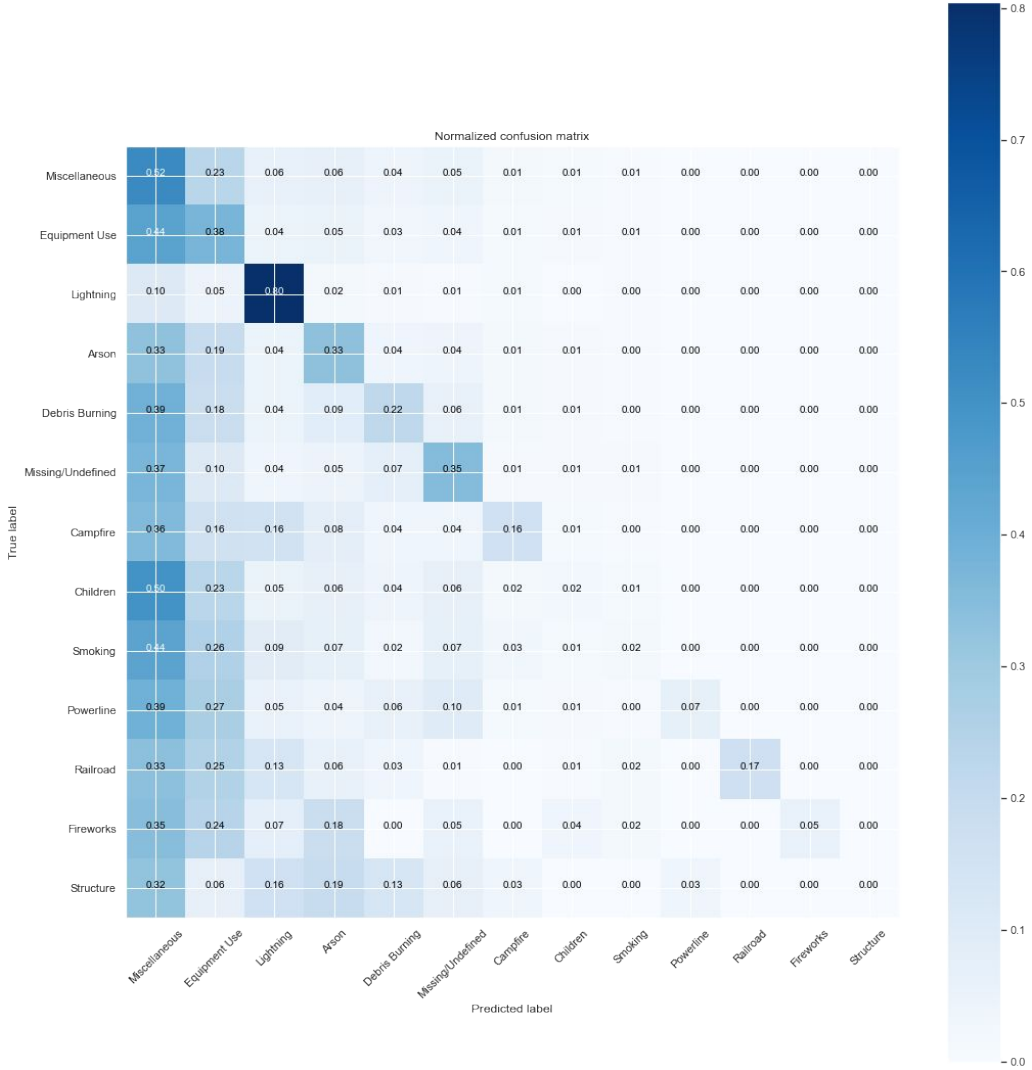
Forest conditions

Try a neural network

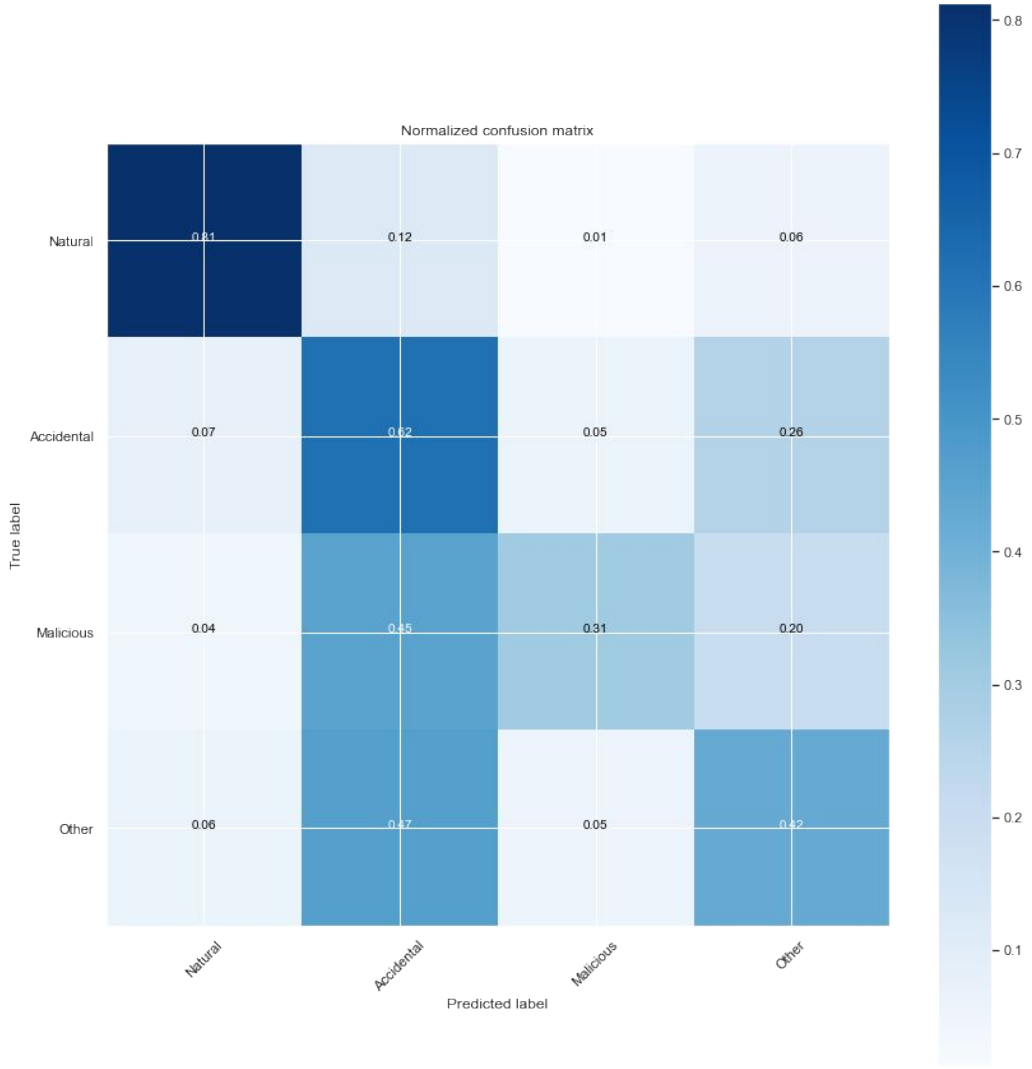
More Information for a Technical Audience

Modeling

13 Classes


Normalized confusion matrix
Miscellaneous 0.52 0.23 0.06 0.06 0.04 0.05 0.01 0.01 0.01 0.00 0.00 0.00 0.00
Equipment Use 0.44 0.38 0.04 0.05 0.03 0.04 0.01 0.01 0.01 0.00 0.00 0.00 0.00
Lightning 0.10 0.05 0.80 0.02 0.01 0.01 0.01 0.00 0.00 0.00 0.00 0.00 0.00
Arson 0.33 0.19 0.04 0.33 0.04 0.04 0.01 0.01 0.00 0.00 0.00 0.00 0.00
Debris Burning 0.39 0.18 0.04 0.09 0.22 0.06 0.01 0.01 0.00 0.00 0.00 0.00 0.00
Missing/Undefined 0.37 0.10 0.04 0.05 0.07 0.35 0.01 0.01 0.01 0.00 0.00 0.00 0.00
Campfire 0.36 0.16 0.16 0.08 0.04 0.04 0.16 0.01 0.00 0.00 0.00 0.00 0.00
Children 0.50 0.23 0.05 0.06 0.04 0.06 0.02 0.02 0.01 0.00 0.00 0.00 0.00
Smoking 0.44 0.26 0.09 0.07 0.02 0.07 0.03 0.01 0.02 0.00 0.00 0.00 0.00
Powerline 0.39 0.27 0.05 0.04 0.06 0.10 0.01 0.01 0.00 0.07 0.00 0.00 0.00
Railroad 0.33 0.25 0.13 0.06 0.03 0.01 0.00 0.01 0.02 0.00 0.17 0.00 0.00
Fireworks 0.35 0.24 0.07 0.18 0.00 0.05 0.00 0.04 0.02 0.00 0.00 0.05 0.00
Structure 0.32 0.06 0.16 0.19 0.13 0.06 0.03 0.00 0.00 0.03 0.00 0.00 0.00
True label
Miscellaneous
Equipment Use
Lightning
Arson
Debris Burning
Missing/Undefined
Campfire
Children
Smoking
Powerline
Railroad
Fireworks
Structure
Predicted label
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

4 Classes


Normalized confusion matrix
Natural
Accidental
Malicious
Other
True label
0.81
0.12
0.01
0.06
0.07
0.62
0.05
0.26
0.04
0.45
0.31
0.20
0.06
0.67
0.05
0.42
Natural
Accidental
Malicious
Other
Predicted label
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

Comparing All Three Models

Model	Baseline Score	Accuracy Score - Train	Accuracy Score - Test	CrossValScore - Train	CrossValScore - Test
Logistic Regression - 13 Classes	0.274	0.340	0.342	0.340	0.344
Logistic Regression - 4 Classes	0.413	0.500	0.500	0.500	0.495
KNN - 13 Classes	0.274	0.600	0.420	0.340	0.344
KNN - 4 Classes	0.413	0.683	0.546	0.497	0.495
Random Forest - 13 Classes	0.274	0.999	0.489	0.483	0.457
Random Forest - 4 Classes	0.413	0.999	0.614	0.607	0.582

Unsupervised Learning Experiments

Clustering with K Means

Methodology

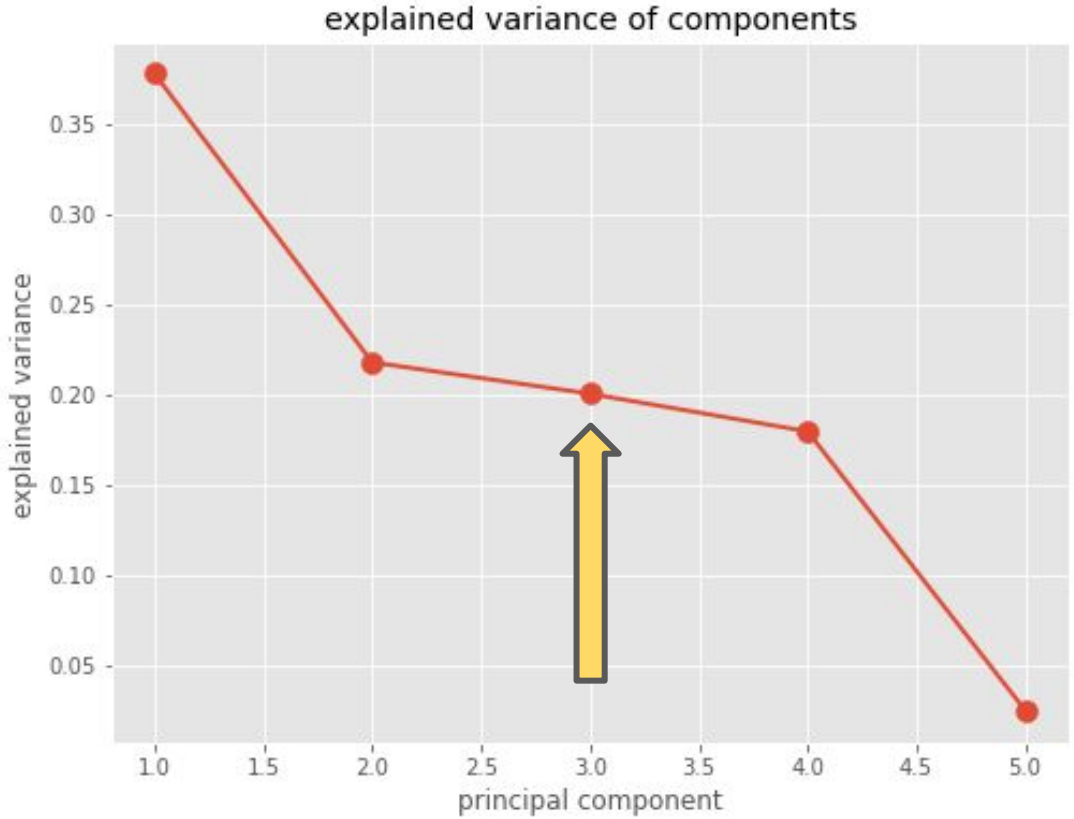
My methodology includes the following steps:

- Standardize the data
- Consolidate the features using principal component analysis
- Determine the optimal number of K clusters using the elbow method
- Instantiate a K Means model
- Observe the results with a scatter plot
- Evaluate the result with a cluster report

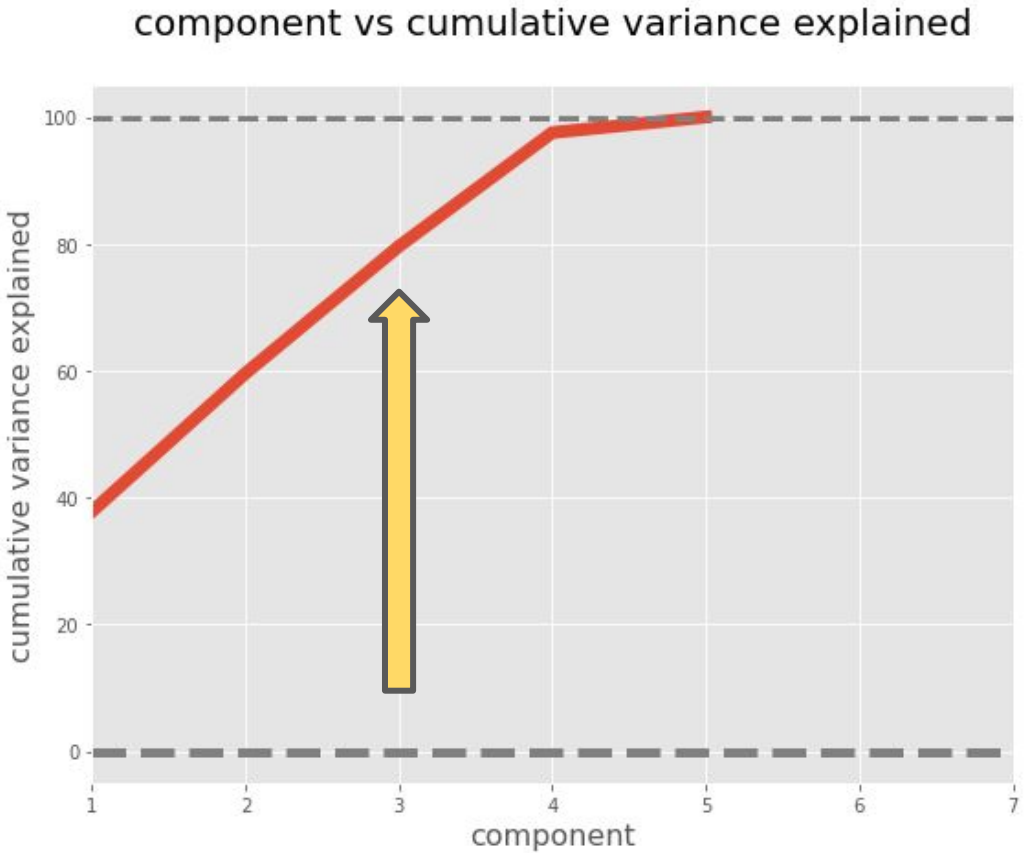
Model Summary

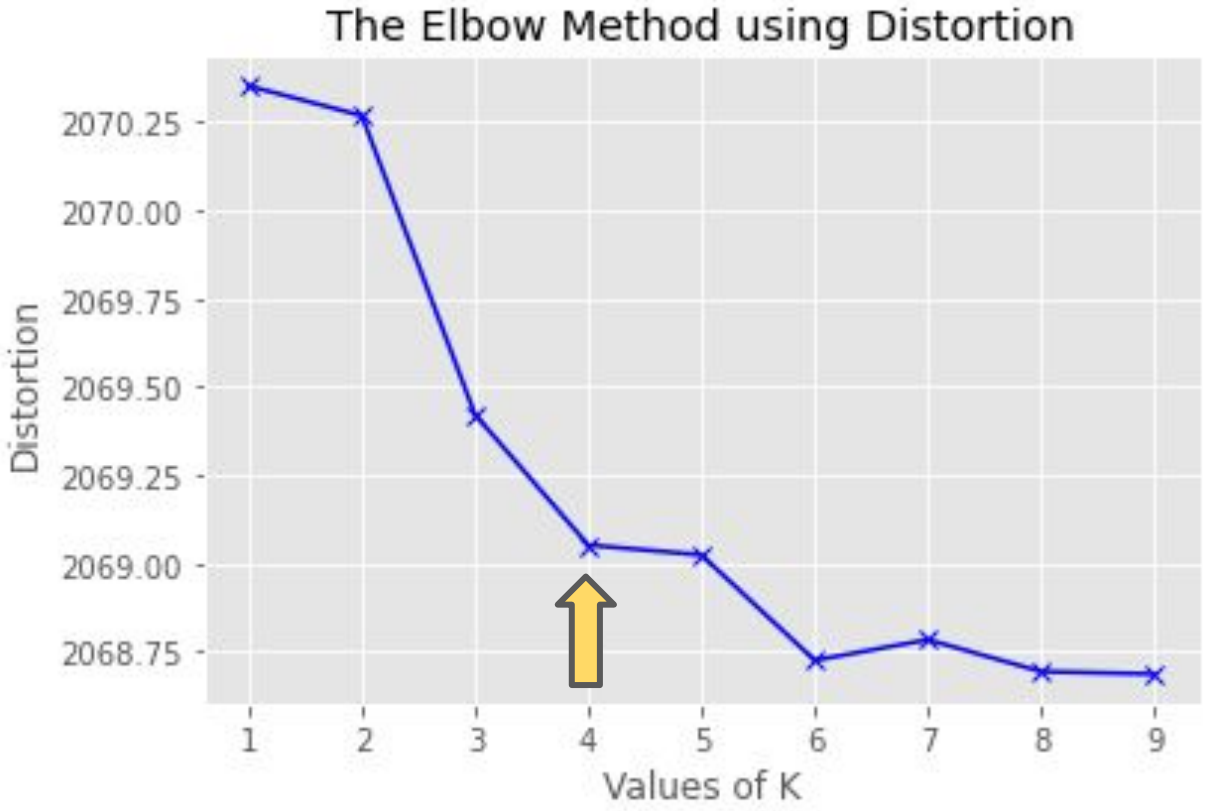
Clustering Model Version	# of PCA Features	Variance Explained by PCA	# of K CLusters	Silhoutee Score
A	2	60%	6	0.358
B	0	n/a	6	0.232
C	3	80%	4	0.366

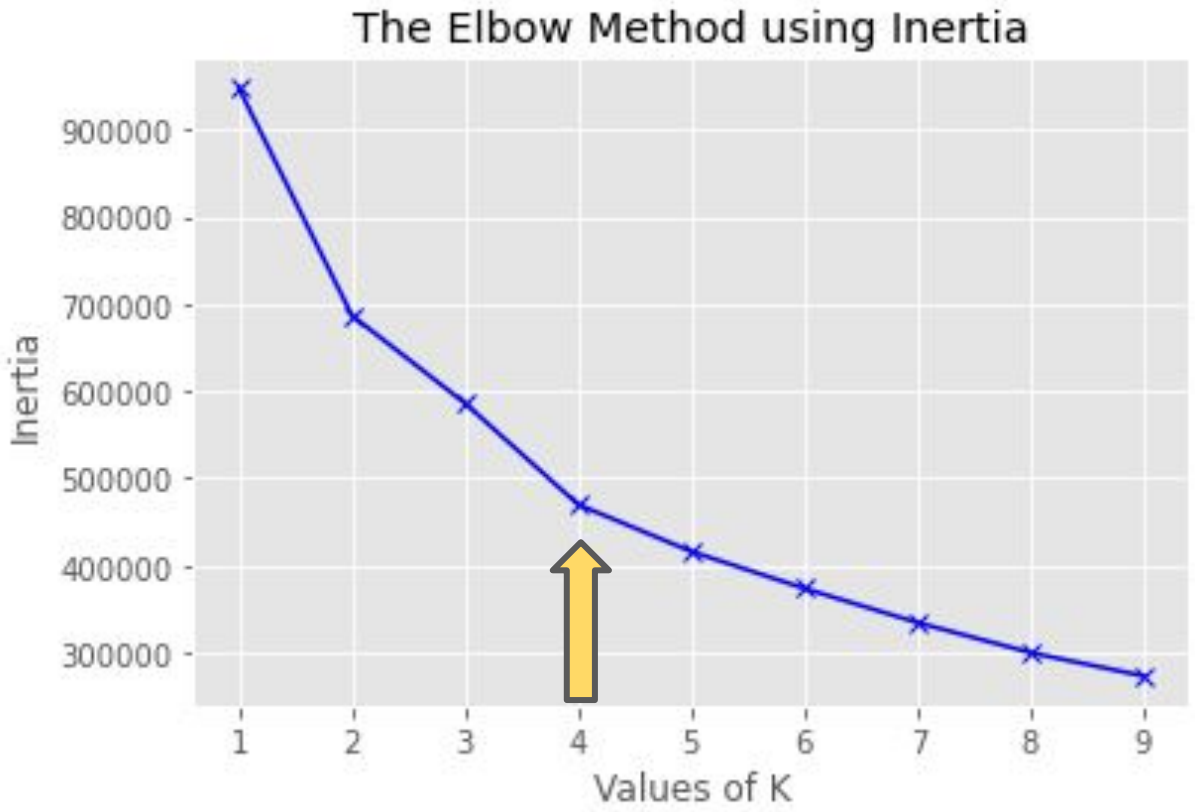
Model C - Principal Component Analysis



Model C - Principal Component Analysis




The Elbow Method using Distortion
Distortion
2070.25
2070.00
2069.75
2069.50
2069.25
2069.00
2068.75
1
2
3
4
5
6
7
8
9
Values of K

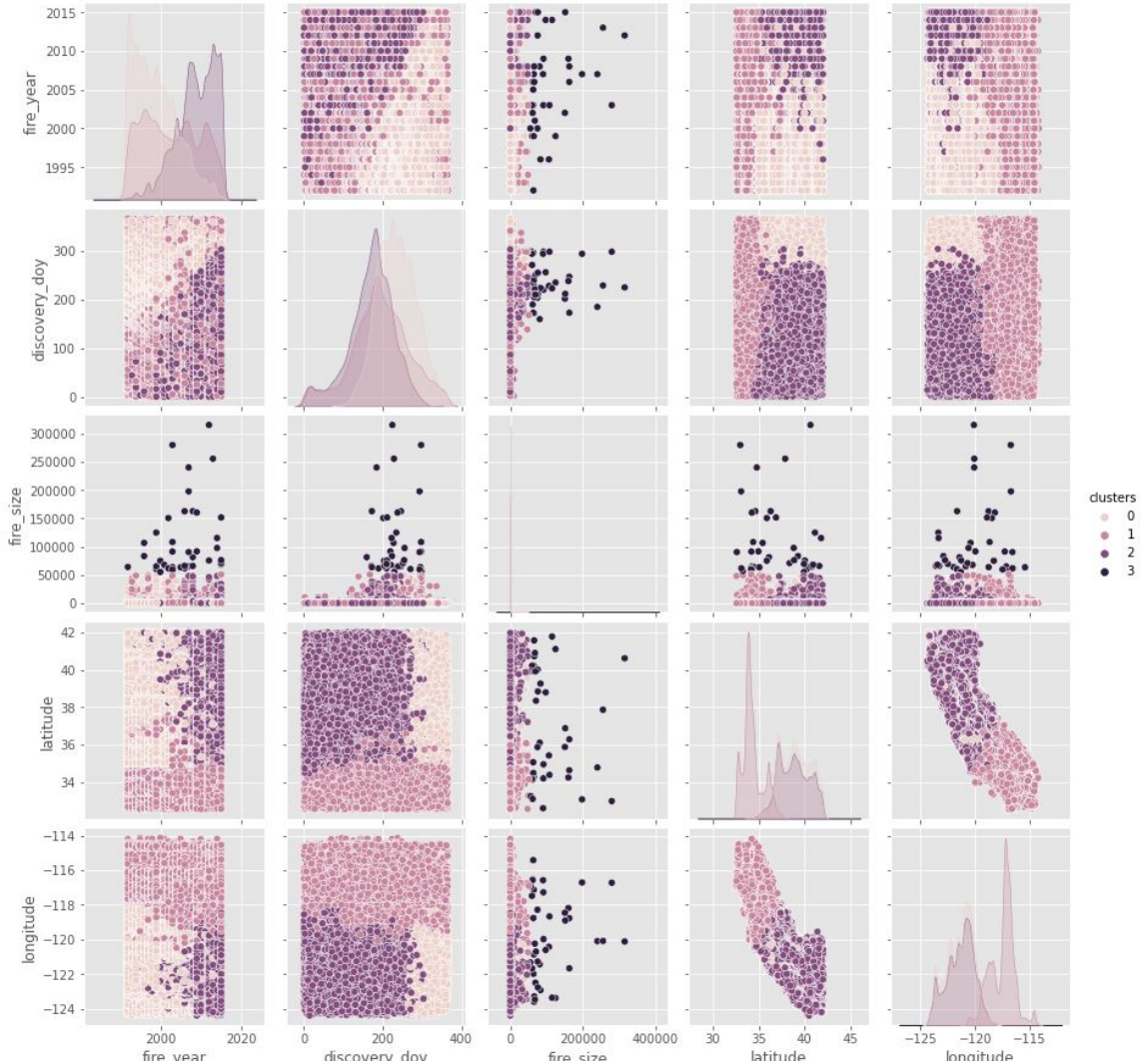

The Elbow Method using Inertia
Inertia
900000
800000
700000
600000
500000
400000
300000
1
2
3
4
5
6
7
8
9
Values of K

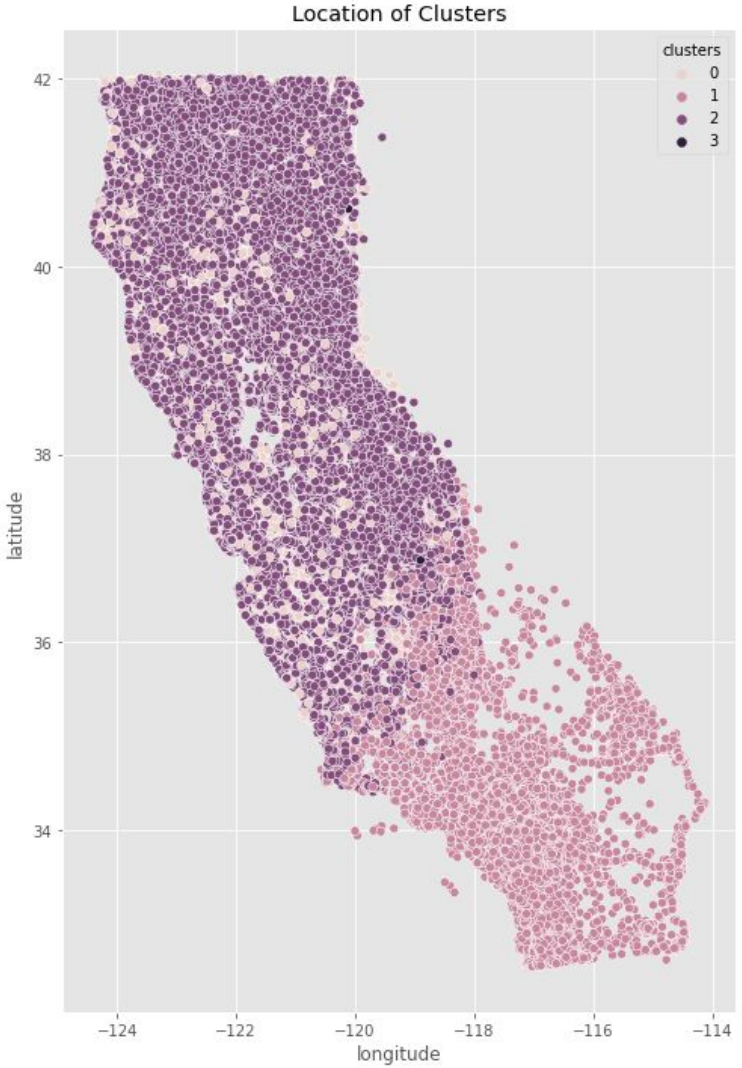
Summary of Clusters

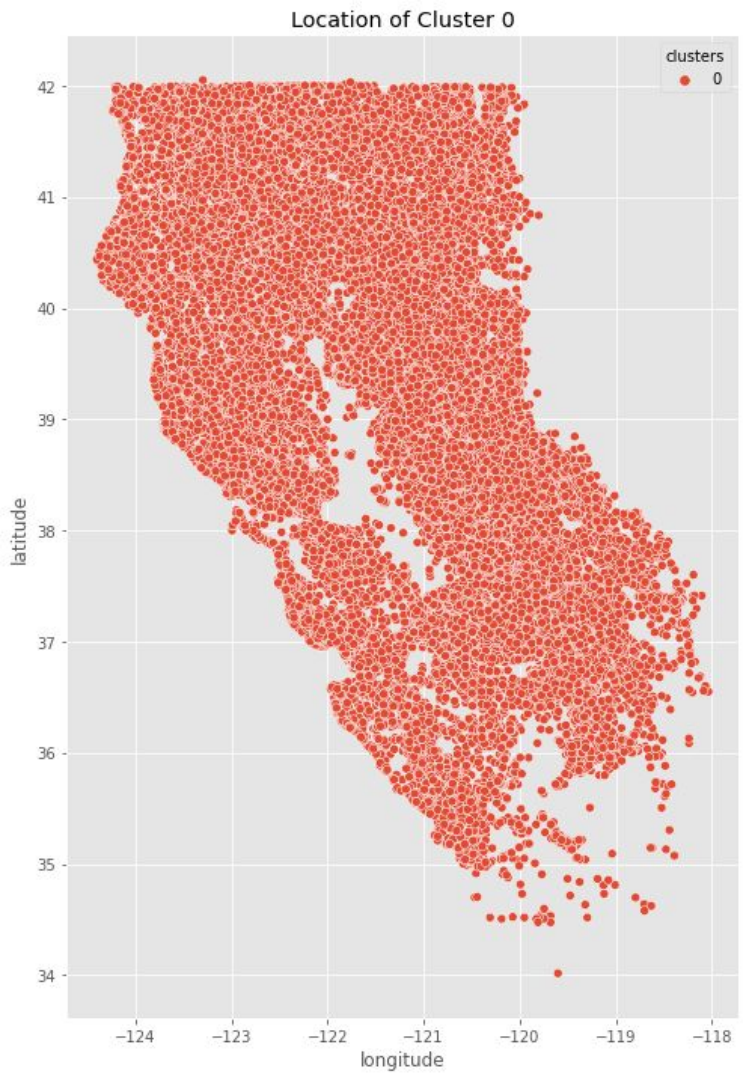
Cluster Summary:

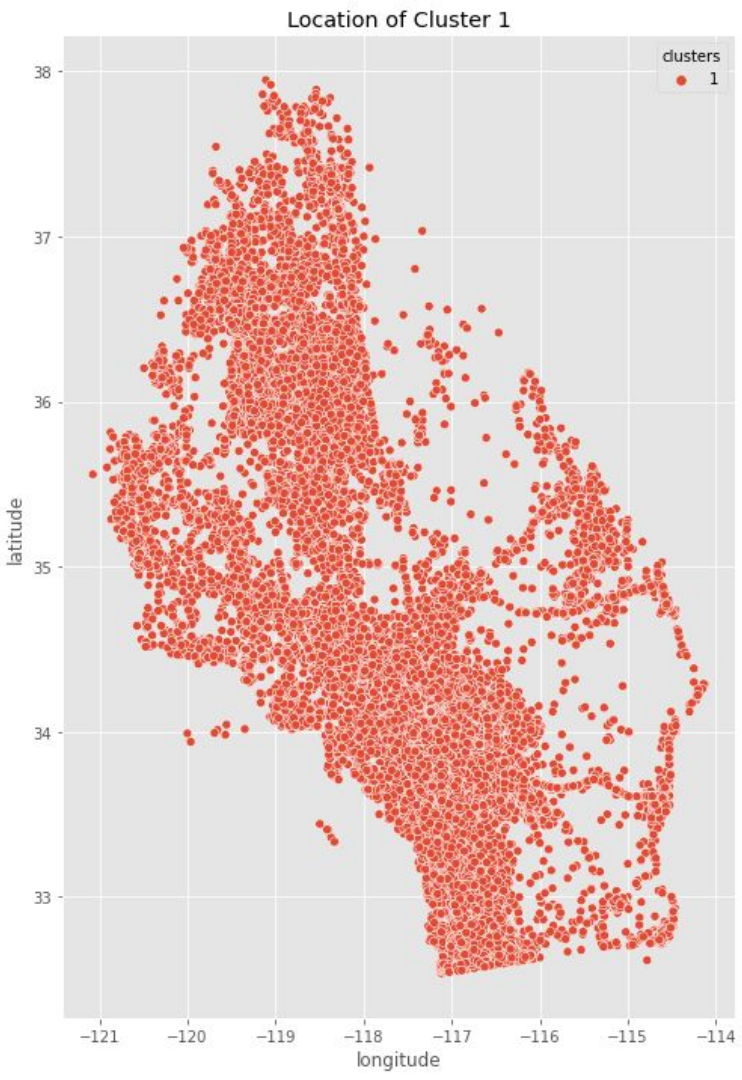
- Cluster 0: These are fires before 2003 and discovered in the second half of their year. They are primarily located in the north and west parts of the state.
- Cluster 1: These fires are all located in southern California. Location is its only attribute.
- Cluster 2: This is a very small cluster of only 40 fires that did not fit with the other clusters. They represent the largest fires.
- Cluster 3: These are primarily fires located in northern California that were discovered earlier in the year.

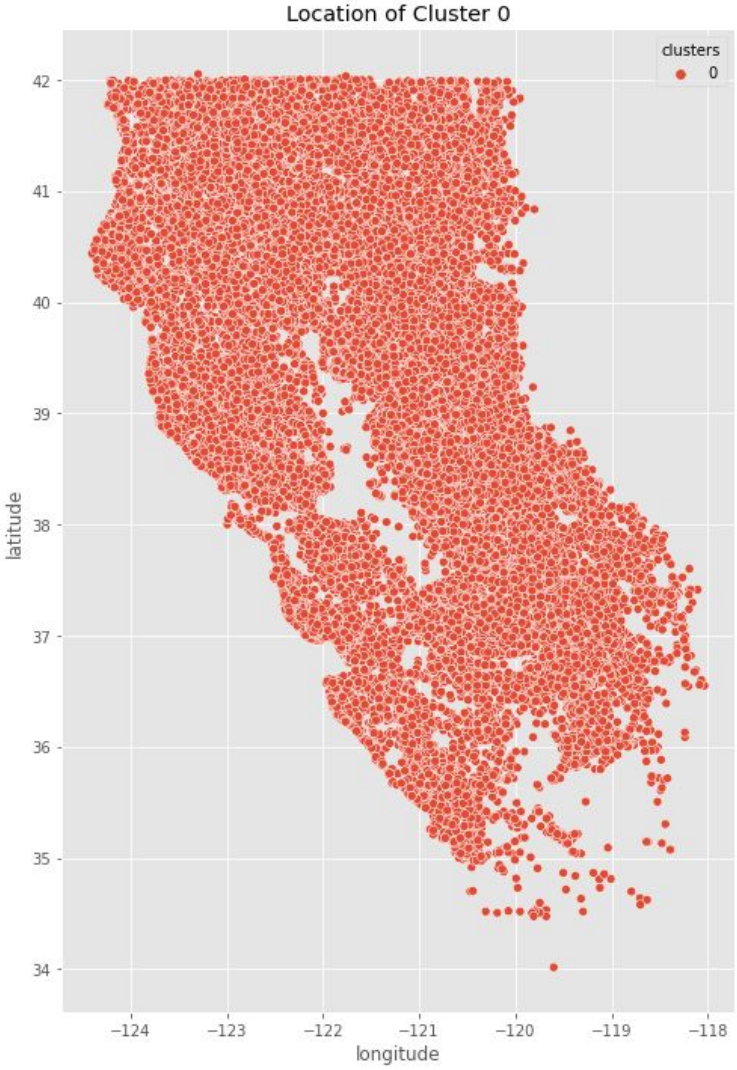
Pairplot Analysis

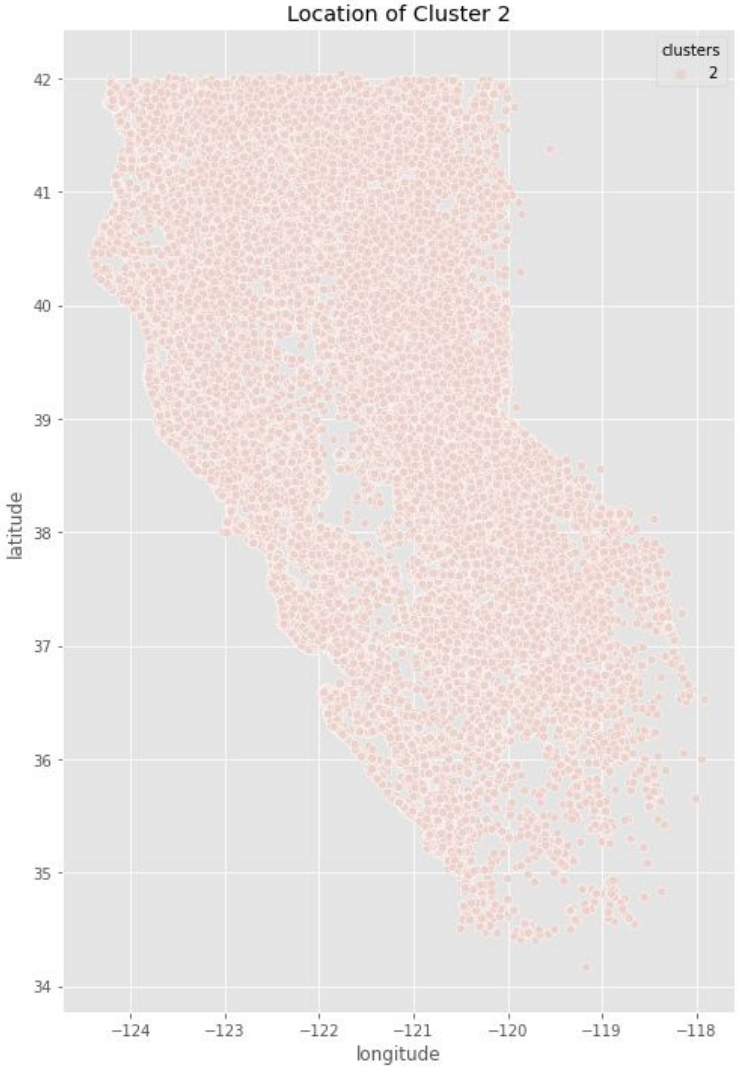


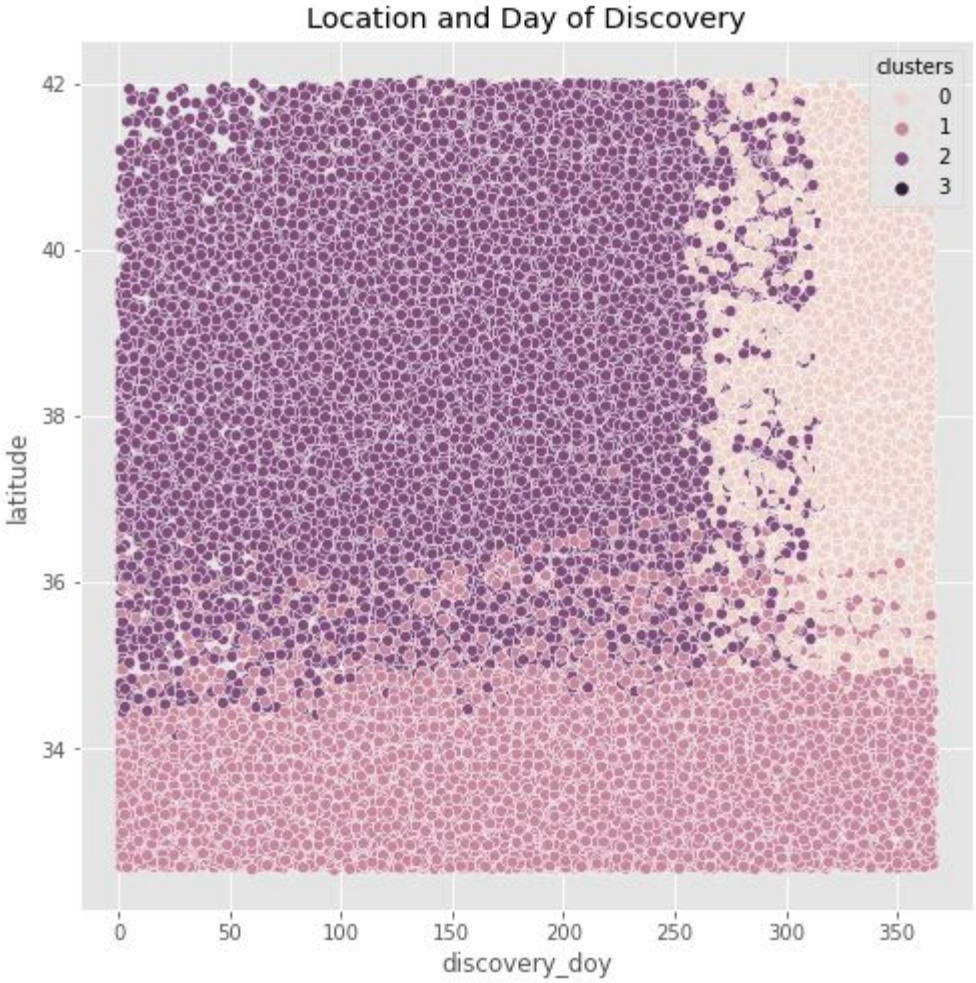

Location of Clusters
clusters
0
1
2
3
42
40
38
36
34
latitude
−124
−122
−120
−118
−116
−114
longitude

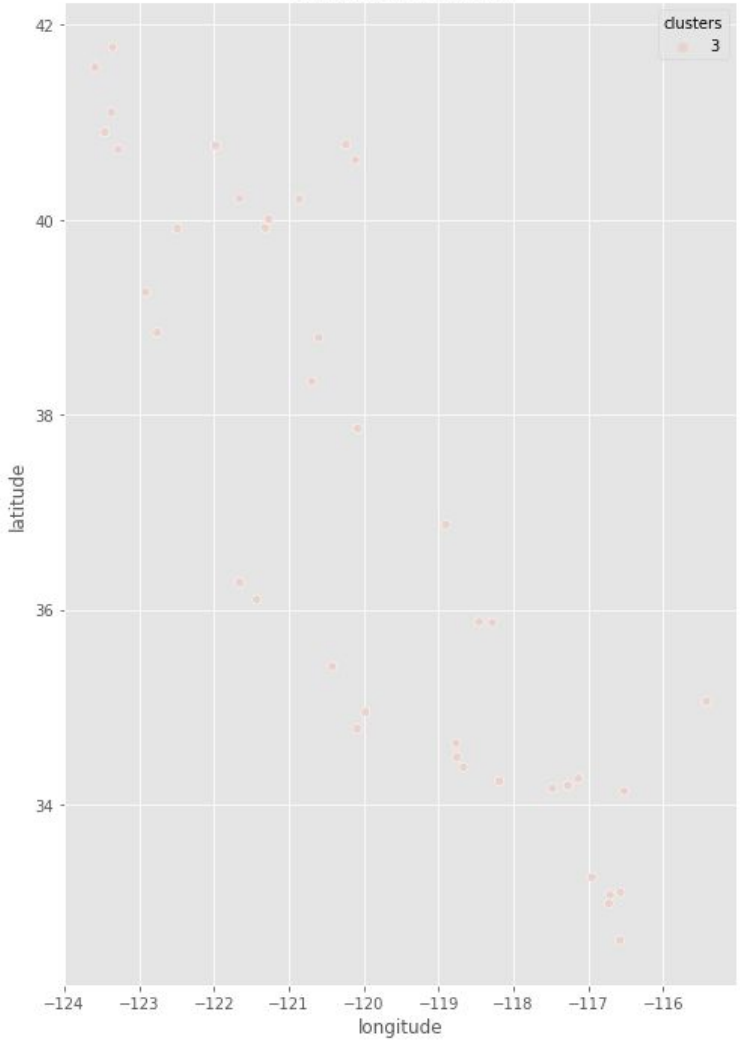

Location of Cluster 0
clusters
0
latitude
longitude
42
41
40
39
38
37
36
35
34
−124
−123
−122
−121
−120
−119
−118

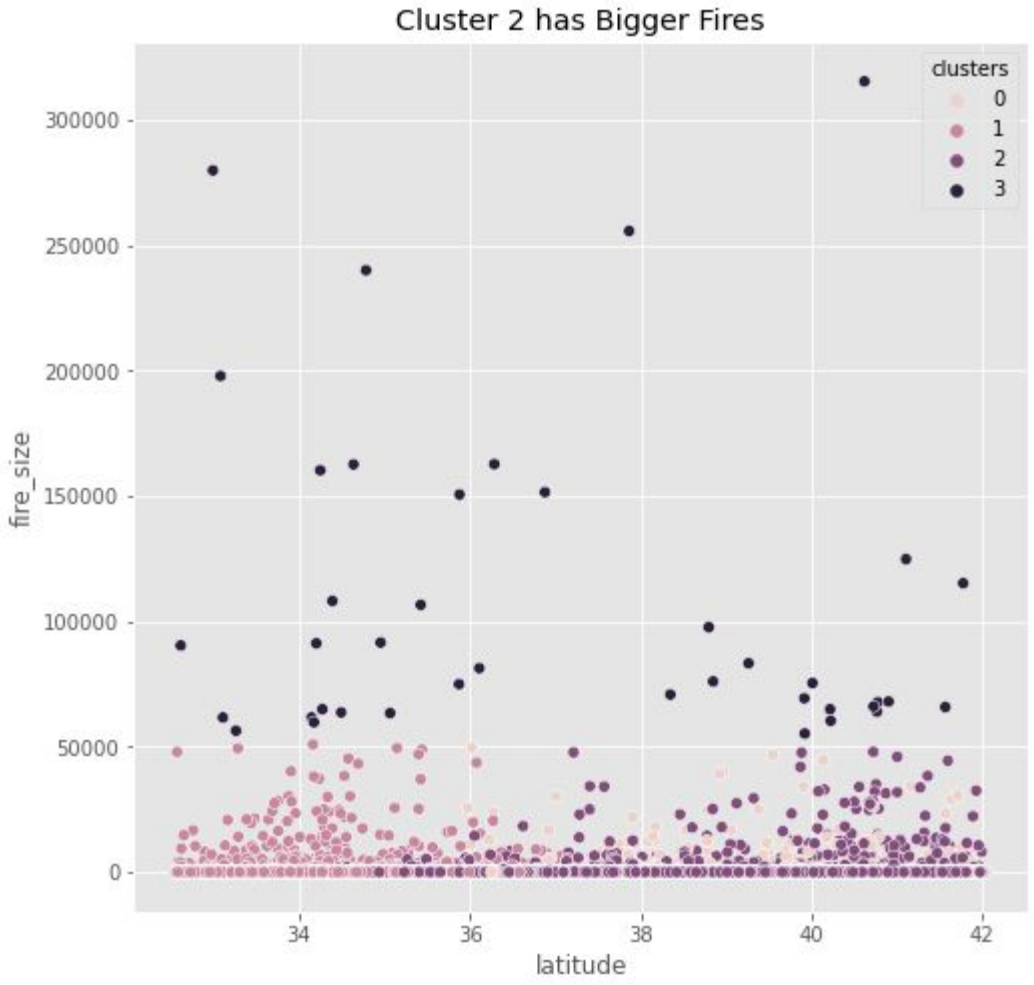

Location of Cluster 1
clusters
1
latitude
longitude
38
37
36
35
34
33
−121
−120
−119
−118
−117
−116
−115
−114

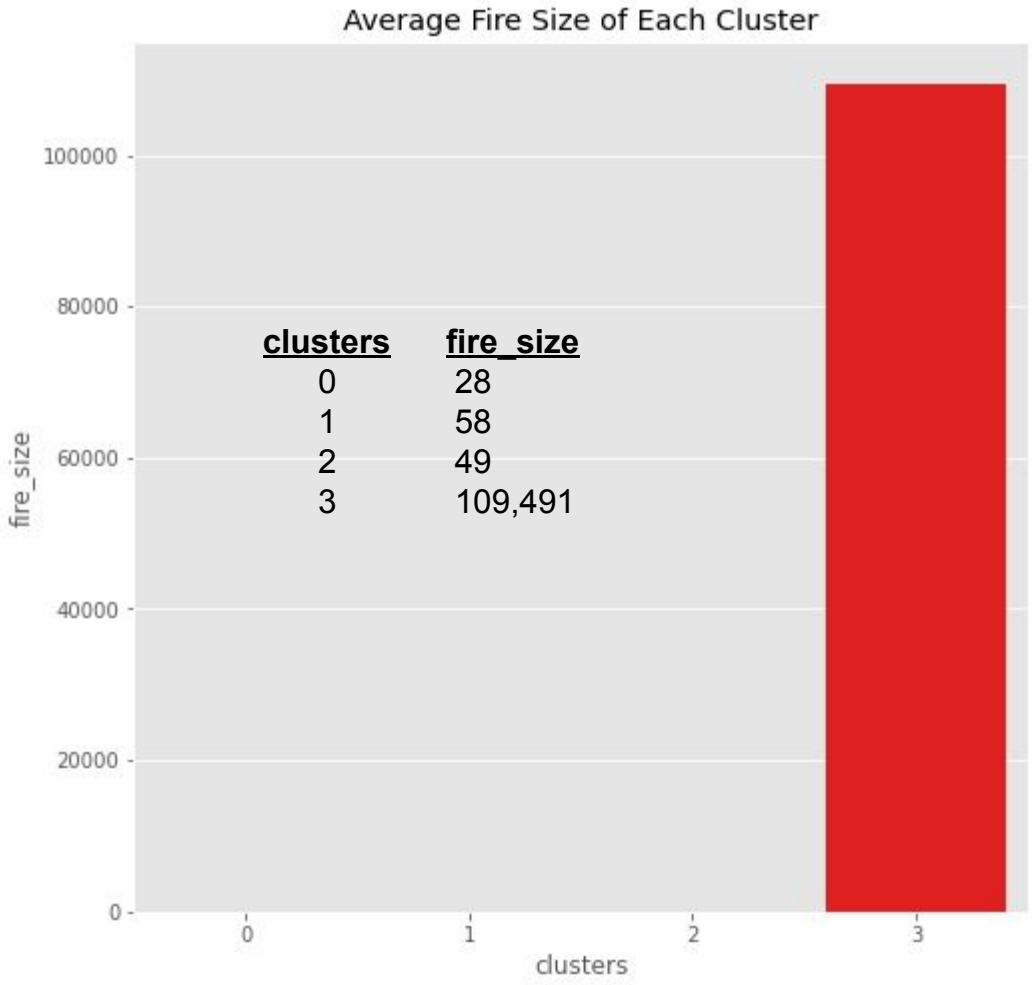

Location of Cluster 0
clusters
0
latitude
longitude
42
41
40
39
38
37
36
35
34
−124
−123
−122
−121
−120
−119
−118


Location of Cluster 2
clusters
2
latitude
longitude
42
41
40
39
38
37
36
35
34
−124
−123
−122
−121
−120
−119
−118


Location and Day of Discovery
clusters
0
1
2
3
42
40
38
36
34
latitude
0
50
100
150
200
250
300
350
discovery_doy


Location of Cluster 3
clusters
3
42
40
38
36
34
latitude
−124
−123
−122
−121
−120
−119
−118
−117
−116
longitude


Cluster 2 has Bigger Fires
clusters
0
1
2
3
300000
250000
200000
150000
100000
50000
0
fire_size
34
36
38
40
42
latitude



clusters	fire_size
0	28
1	58
2	49
3	109,491

Interactive Maps and Articles

Why Does California Have So Many Wildfires?

https://archive.is/sOF86#selection-1023.1-1035.239

California Fire Map & Tracker

https://www.sfchronicle.com/projects/california-fire-map/2020-august-complex-fire

California wildfires map

https://www.latimes.com/wildfires-map/?fire=mm-49