Using Data Analysis to Examine Trends in Forest Fires

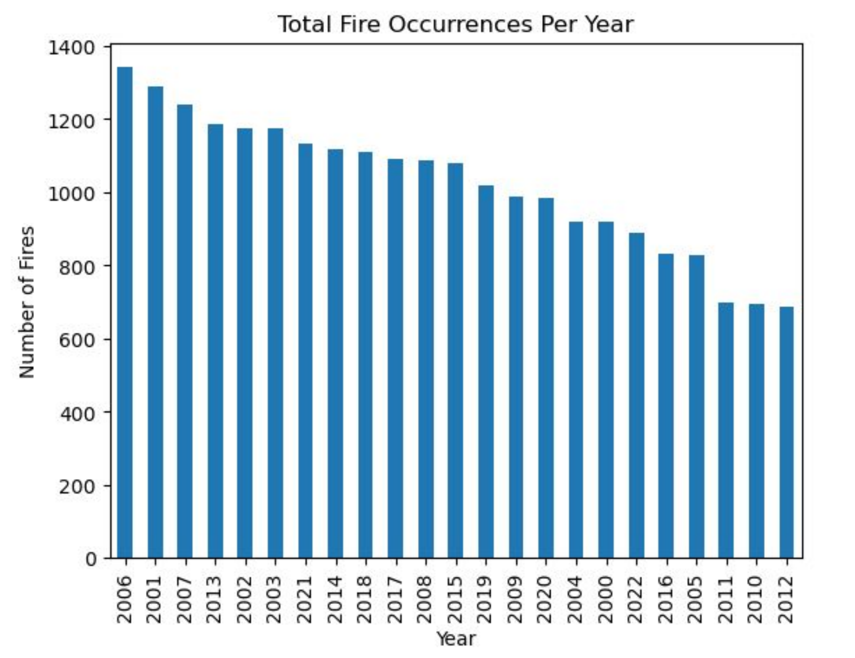
Winter ‘23, CS105

Group 13: Godfrey Lozada, Howie Nguyen, Tri Tran, Rovin Soriano, Ian lopez

GENERAL GRAPHS

Fire Occurrences Per Year

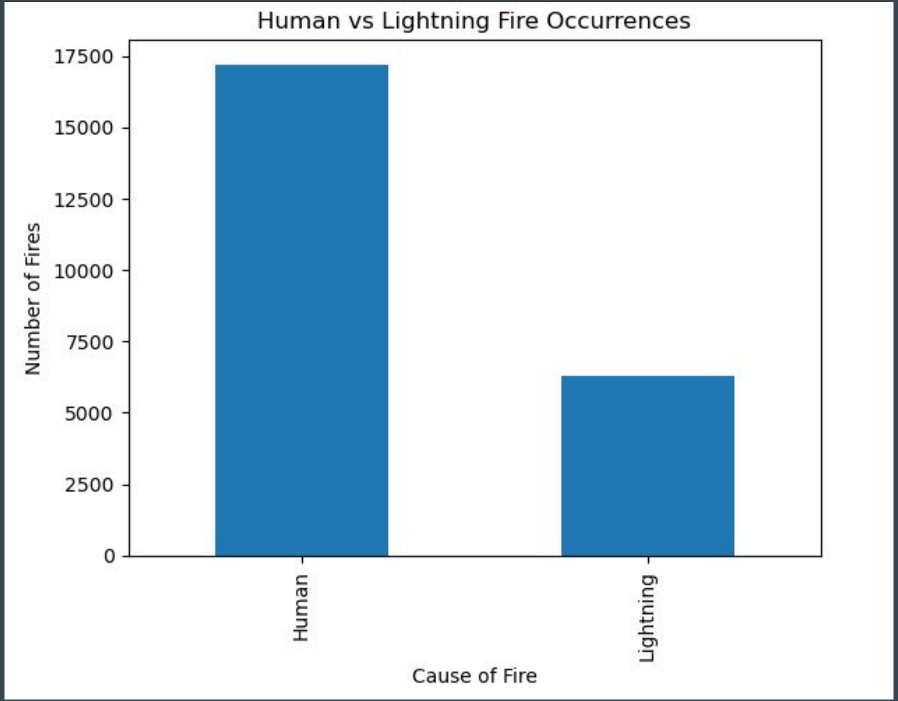
- Depicts frequency of fire incidents annually from 2006 to 2012
- 2006 and 2007 are in the top three for most fires
- Gradual decline ever since 2006



Human or Natural Cause?

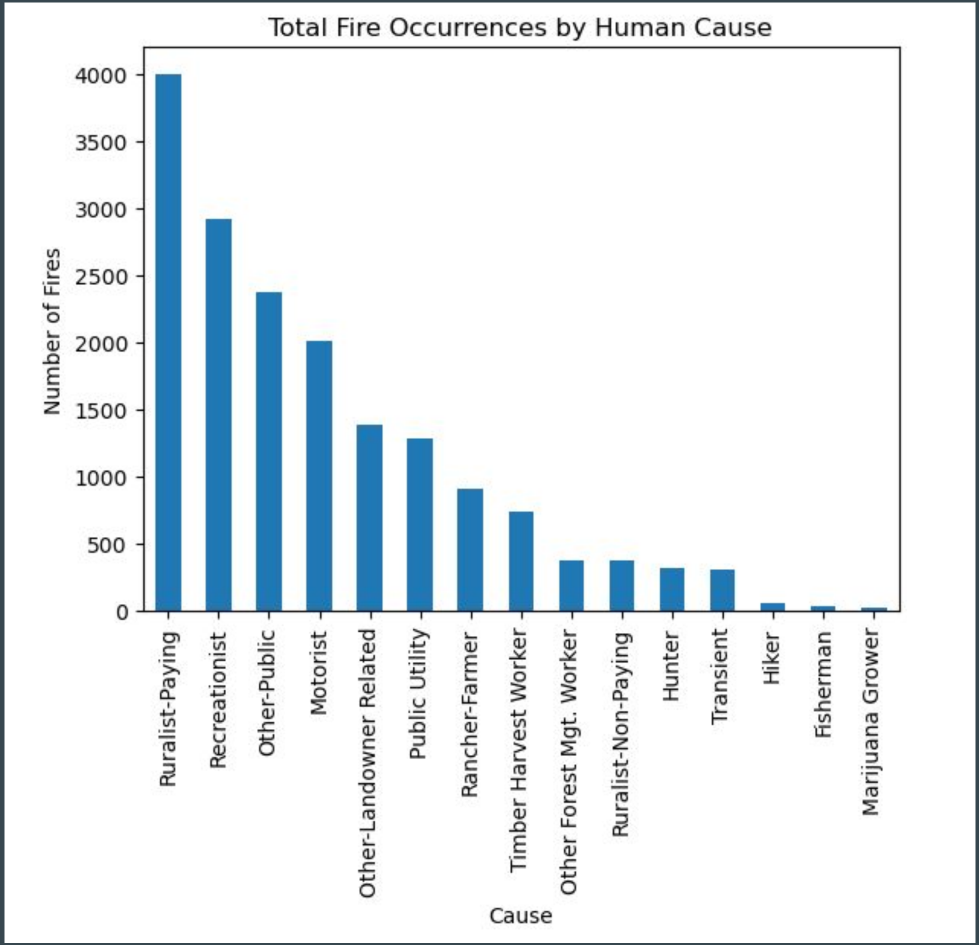
- Depicts incidence of fires based on human or lightning

- Humans cause over three times more fires than lightning strikes

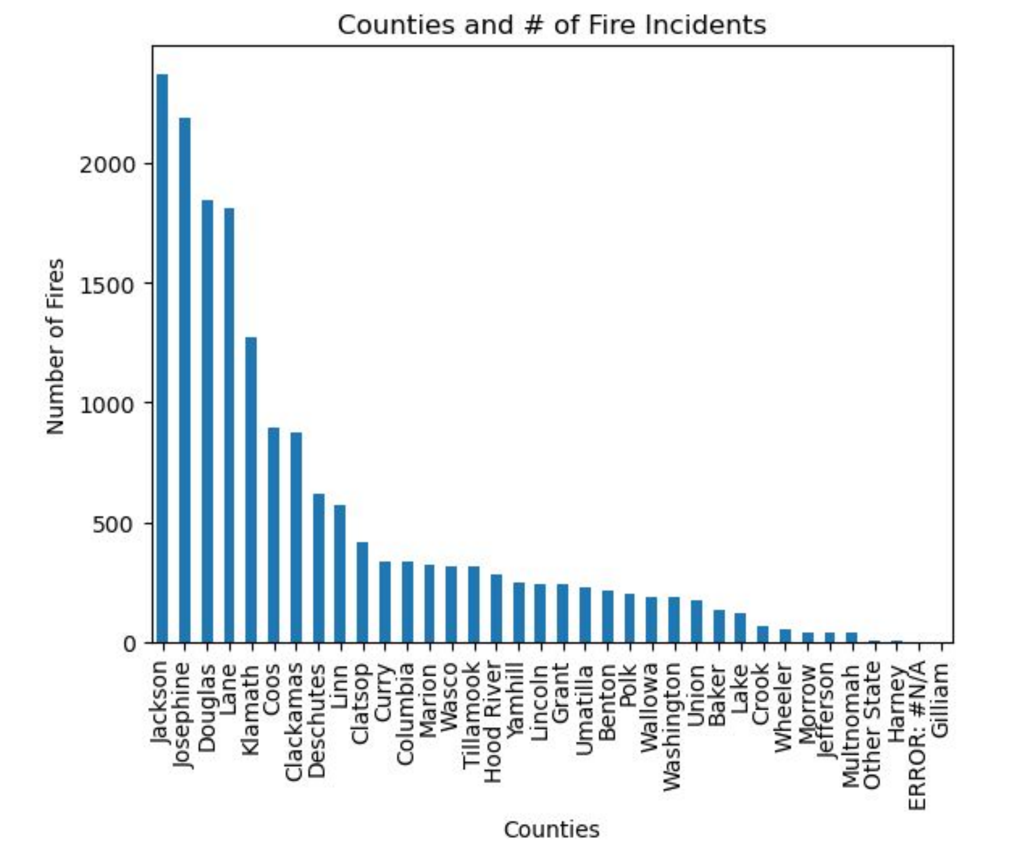


Human Cause Breakdown

- Illustrates the methods in which a fire is caused by humans.
- Ruralist-Paying and Recreationist cause the most fires.
- Fisherman and Marijuana Grower cause the least fires.



Counties Prone to Fire Incidents

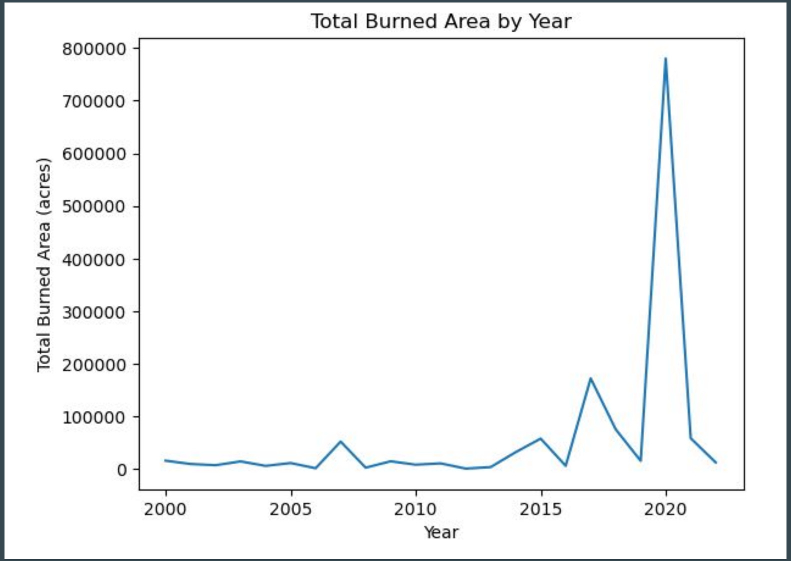


- Illustration of the number of fires according to county in Oregon
- Jackson county and Josephine county hold the two highest number of fires.
- Harney county and Gilliam county each garner less than 10 fires

Yearly Fire Spread

- Depicts the area burned by fires according to year (from 2000 to 2021).

- Consistent correlation from 2000 to 2015



- Surge in area burned between 2016 and 2017

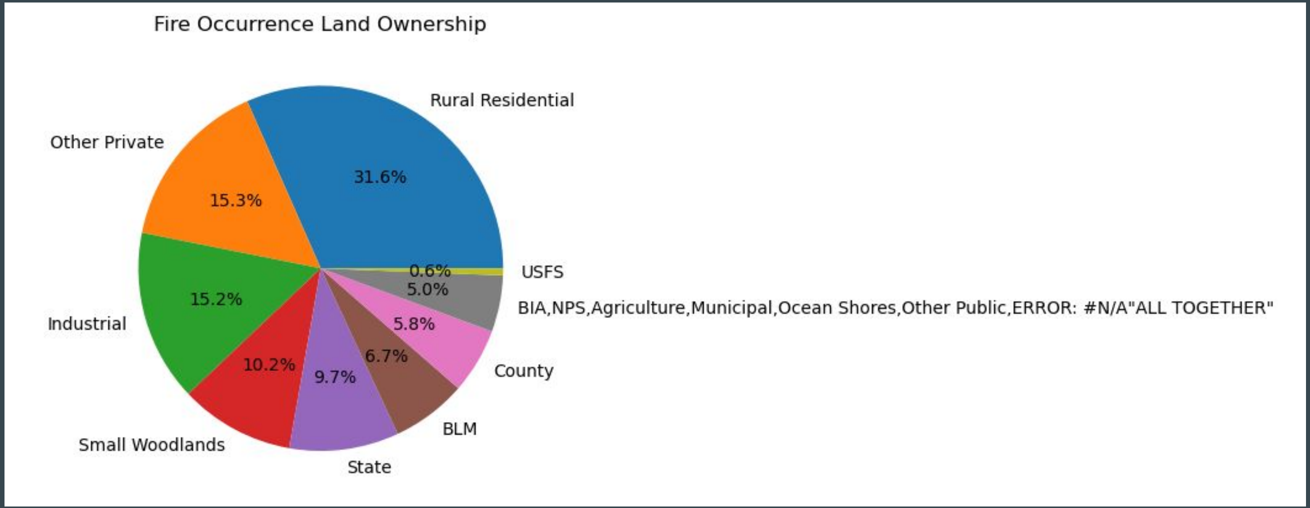
- Deviation from trend with sharp increase in 2020

In What Property Does a Fire occur?

- Illustrates the distribution of ownership of the land where fires occurred.

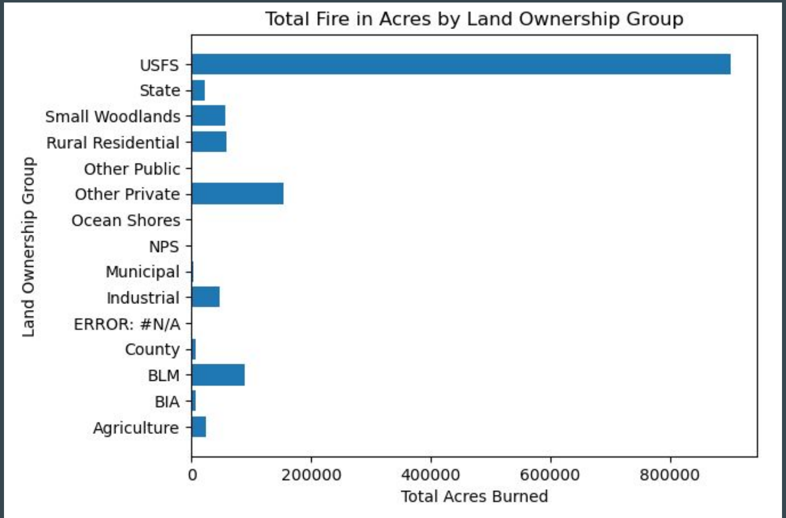
- Lowest percentage of fire occurrence is relegated to the United States Forest Service (USFS)

- Highest percentage of fires occur in Rural Residency



Ownership of Land Burnt

- Portrays the total area burnt according to land ownership groups.



- USFS holds the highest total burned area.

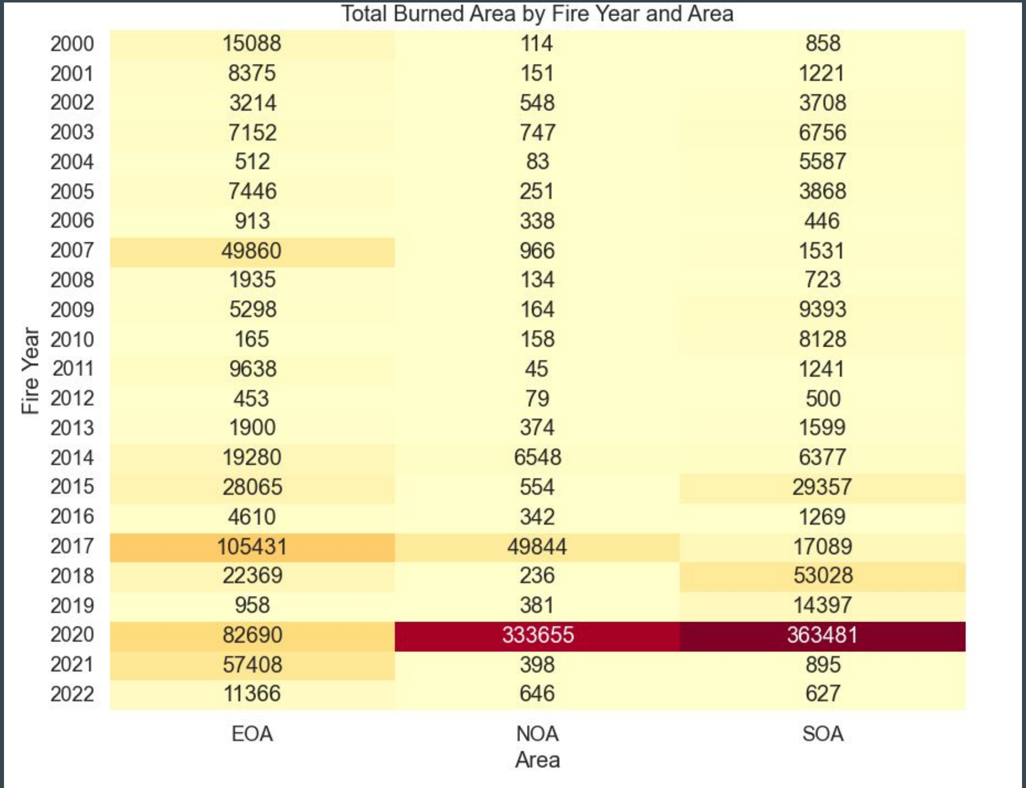
- Other Public, Ocean Shores, and NPS (National Park Service) do not have any burnt land in acres.

Yearly Burnt Acres Per Area

- Describes total area burnt according to year and specific area of Oregon.

- Eastern Oregon Area (EOA), Northern Oregon Area (NOA), Southern Oregon Area (SOA)

- Each area in each year suffered relatively less area burnt than in 2020 with the exception of EOA in 2017



- However, spikes in area burnt is increasingly more common to see in later years

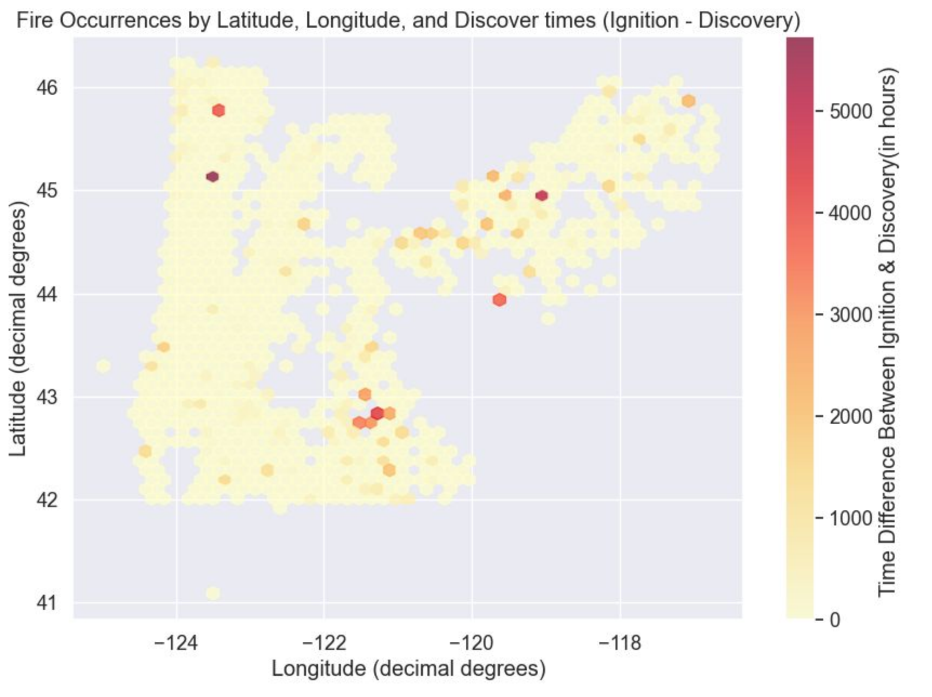
Fire Discovery Time Based on Location Coordinates

- Describes the discovery times of a fire according to their location coordinates

- Discovery time = time of ignition - time of discovery

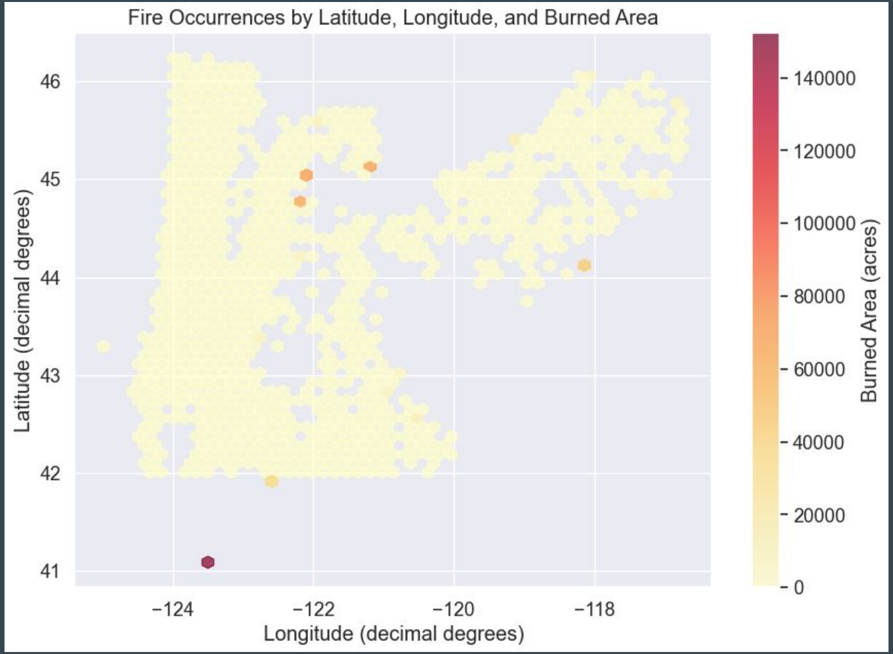
- The fires in Central Southern and Central part of Oregon have higher discovery time.

- Additionally, extreme discovery times take place in the Northwestern part of Oregon



Area Burnt Based on Location Coordinates

- Portrays the area burnt by a fire according to their location coordinates



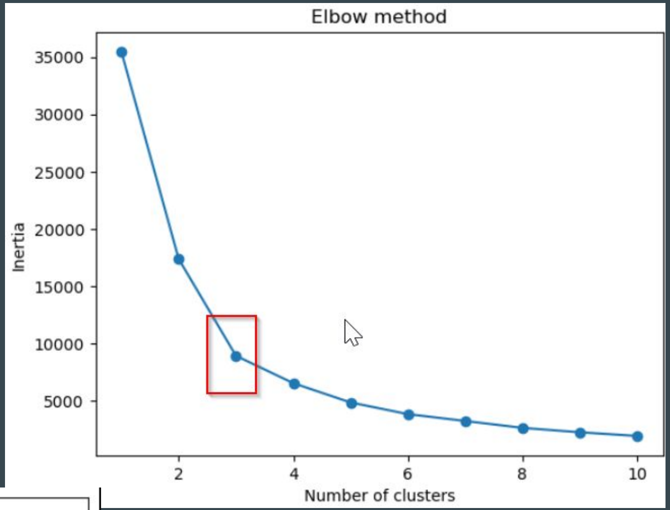
- An overwhelming majority of fires in Oregon burn less than 20000 acres

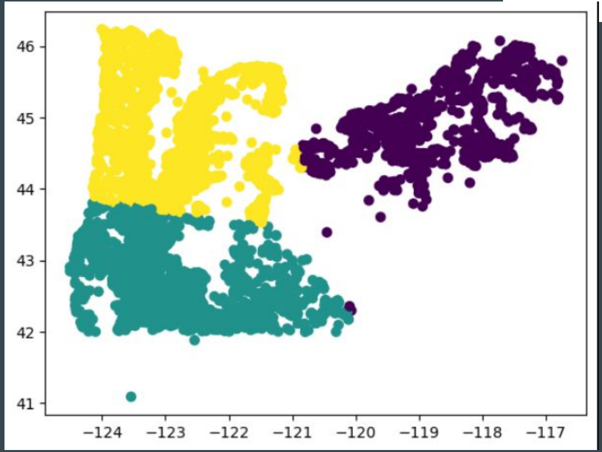
- However, in the Northern part of Oregon, fires have a tendency to burn between 60000 and 80000 acres

SPLITTING OREGON TO FURTHER ANALYZE DATA

Splitting Oregon: K-Means Clustering

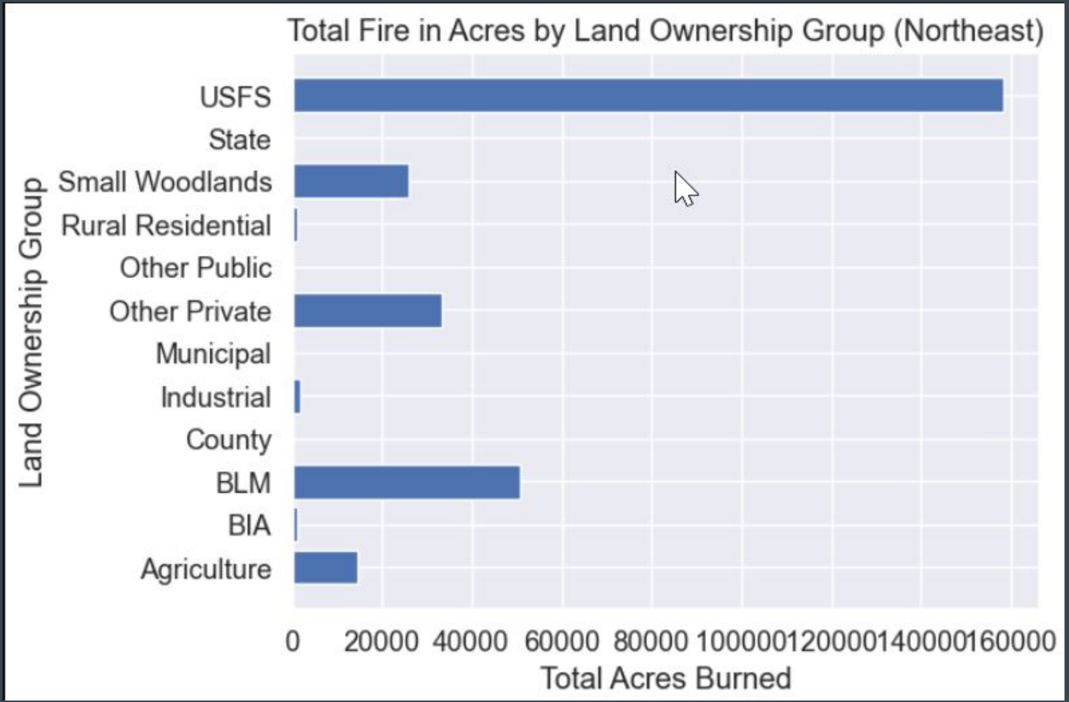
- Ran K-Means Clustering to get distinctive sub-groups
  - Ideal value was 3
- Fit model to a scatter plot. Get following regions
  - Northwest
  - Southwest
  - Northeast.



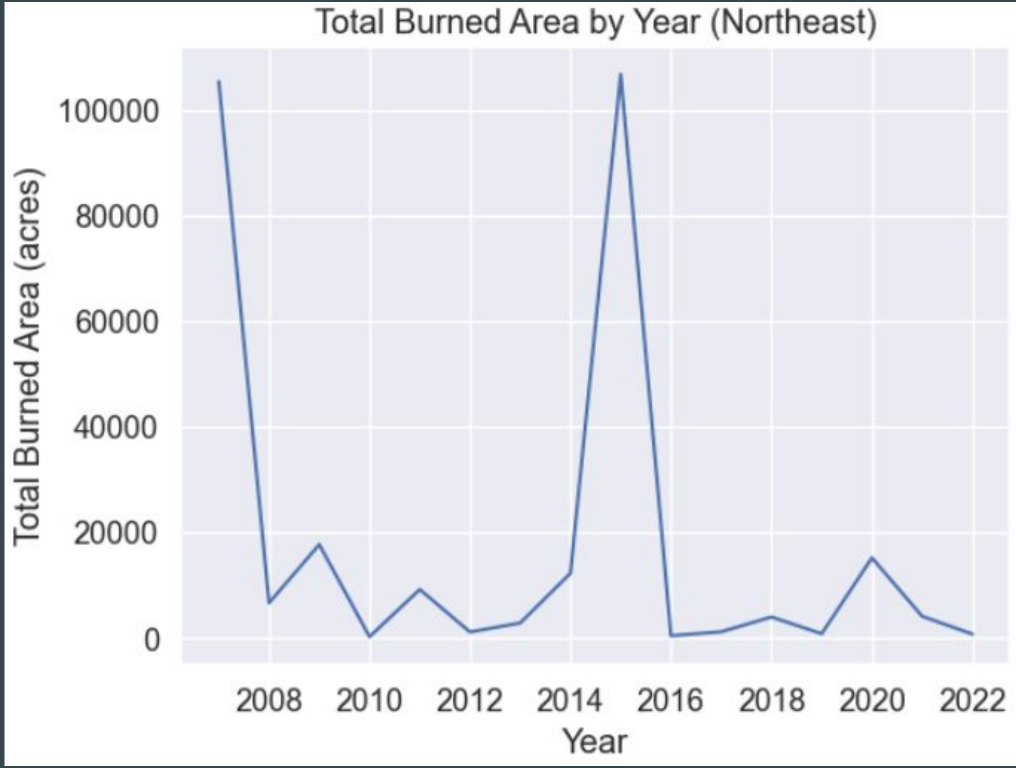


CLUSTERS

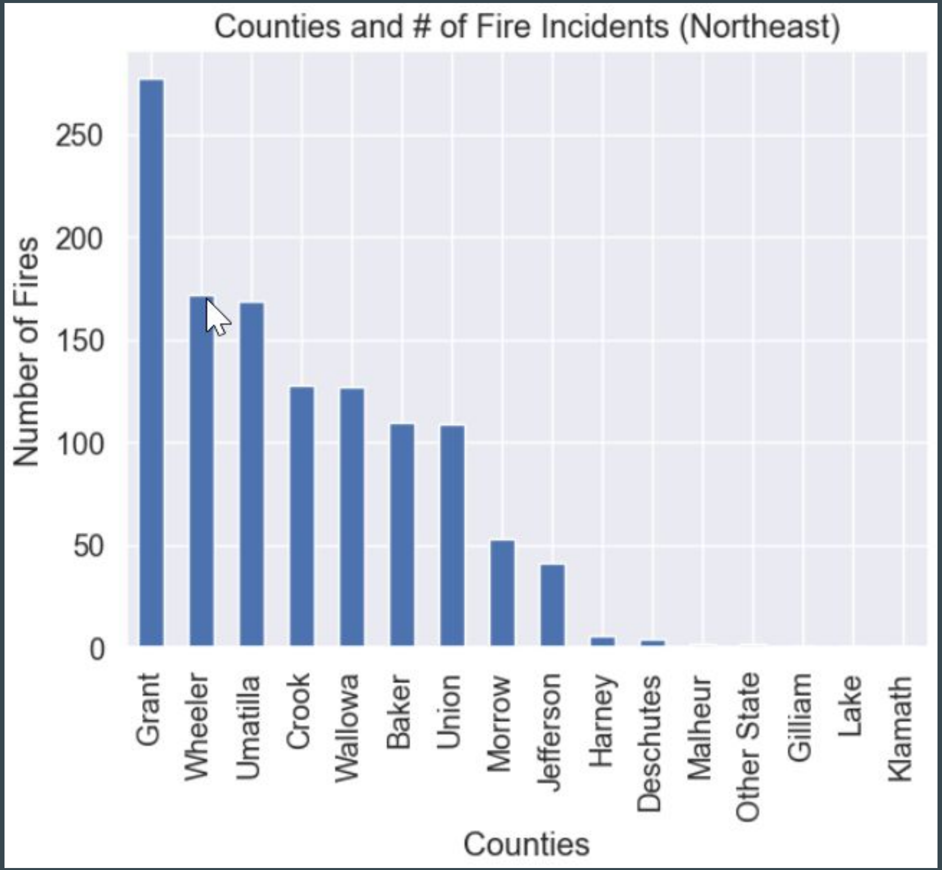
Cluster 1 Northeast Bar Graph



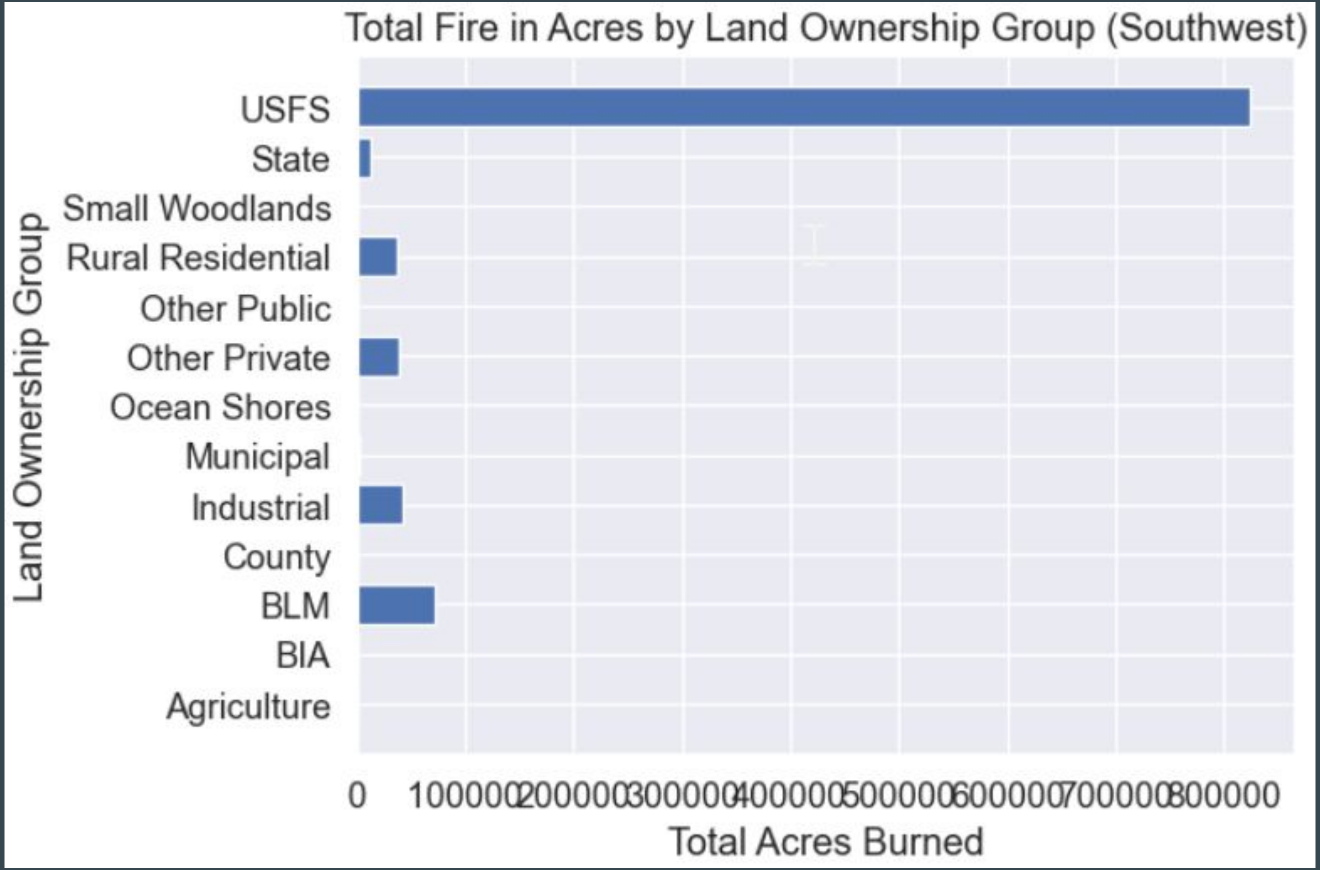
Cluster 1 Northeast Line Graph



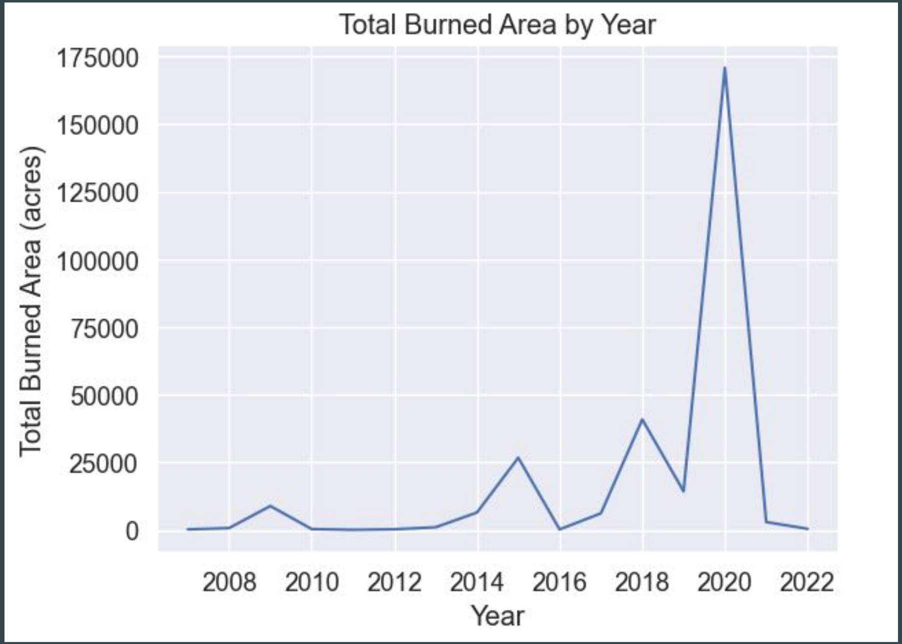
Cluster 1 Northeast Bar Graph



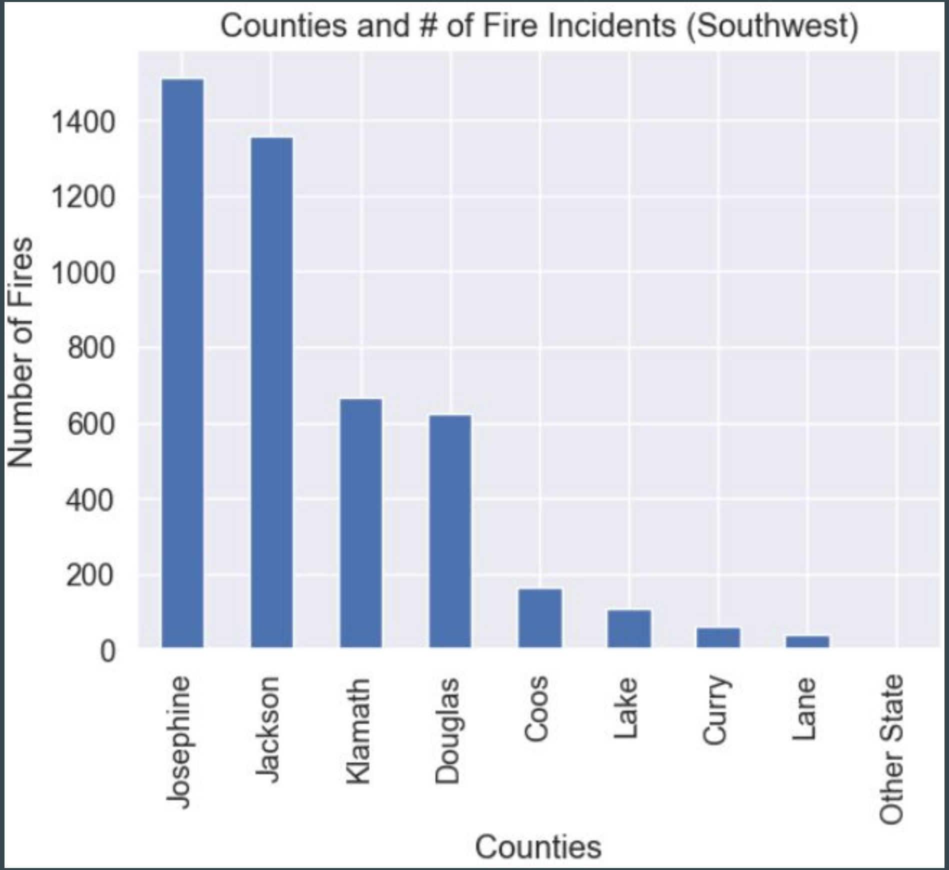
Cluster 2 Southwest Bar Chart



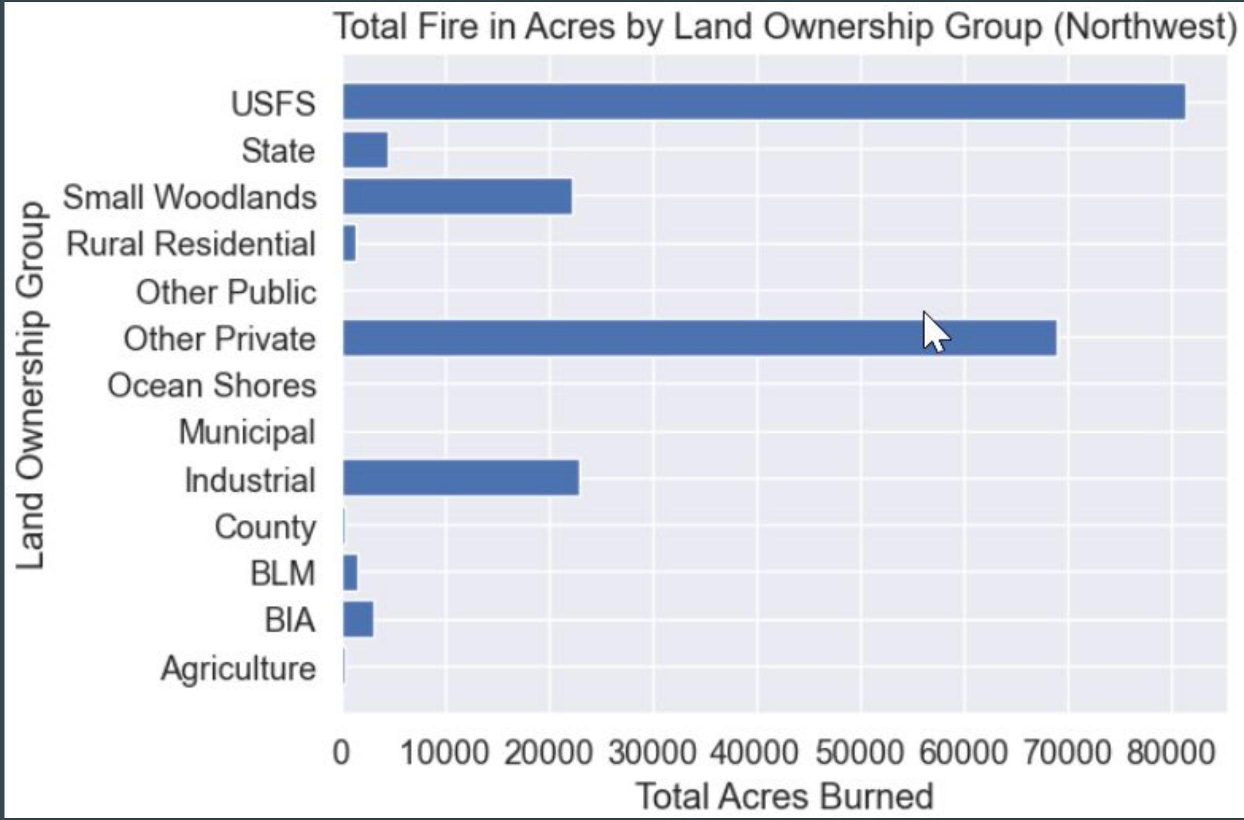
Cluster 2 Southwest Line Graph



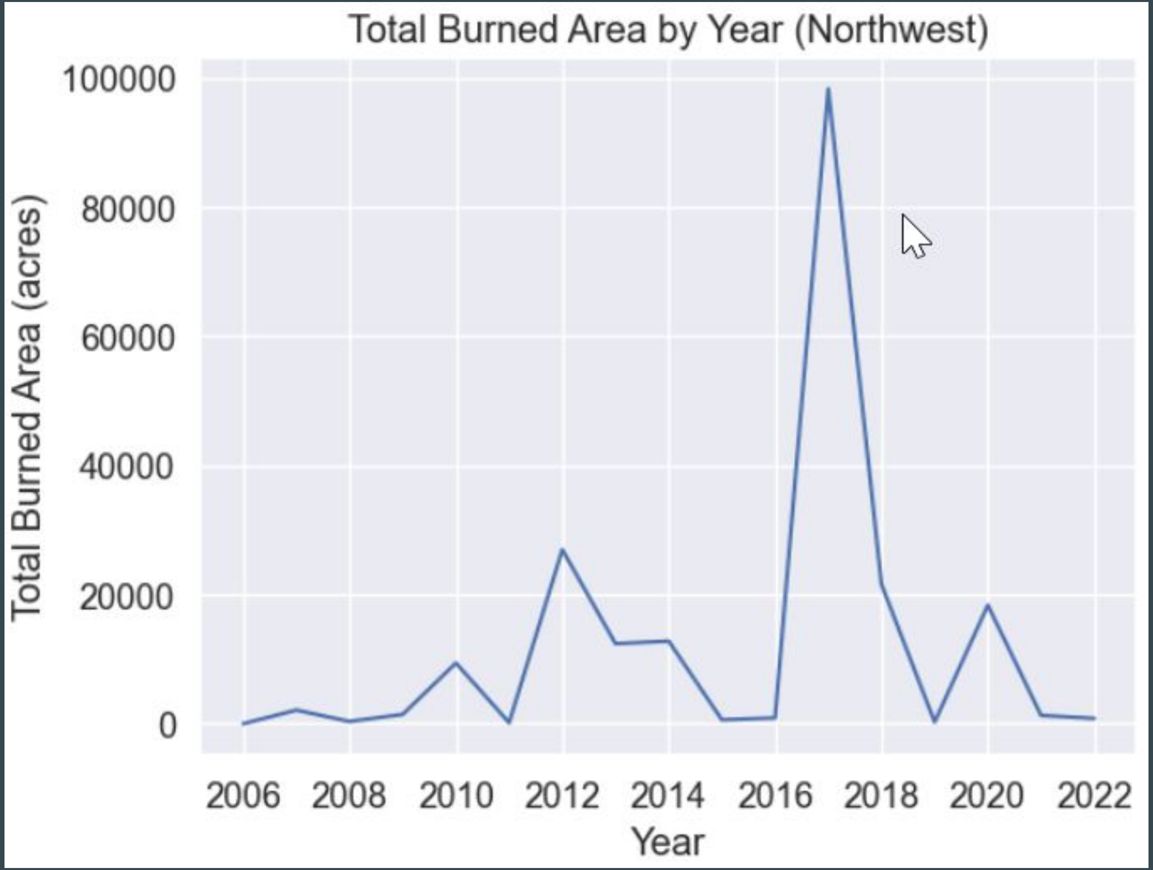
Cluster 2 Southwest Bar Graph



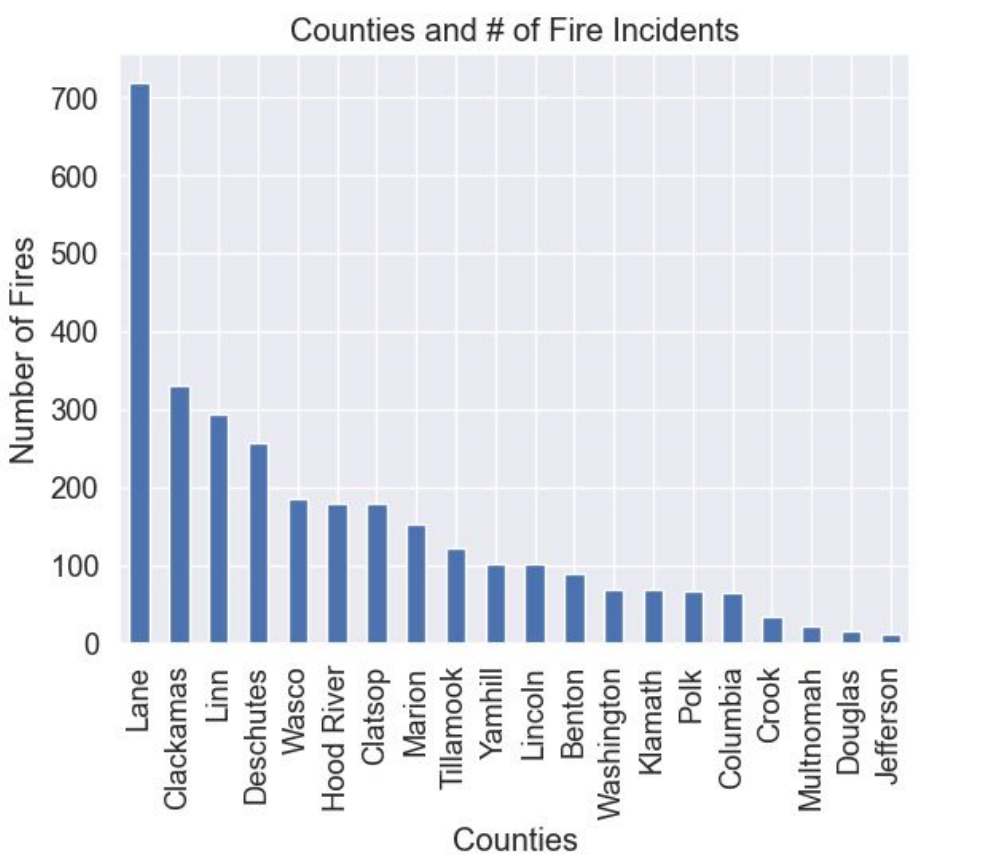
Cluster 3 Northwest Bar Graph



Cluster 3 Northwest Line Graph



Cluster 3 Northwest Bar Graph

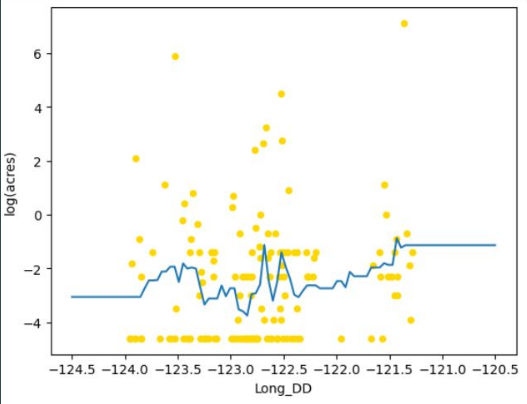


APPLYING A MODEL TO THE CLUSTERS

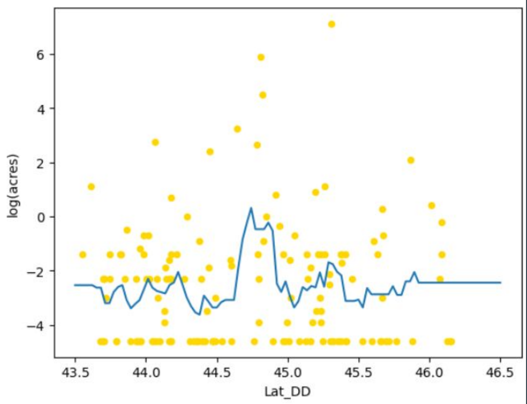
How We are analyzing each cluster:

- K-Nearest Neighbors as the regression model
- Split training and test sets
- Longitudinal/Latitudinal coordinates as inputs (individually)
- Burn area in Log(acres) as output
  - Due to large variations in fire size
- Plot line on scatterplot using training data
- Examine performance on test-set

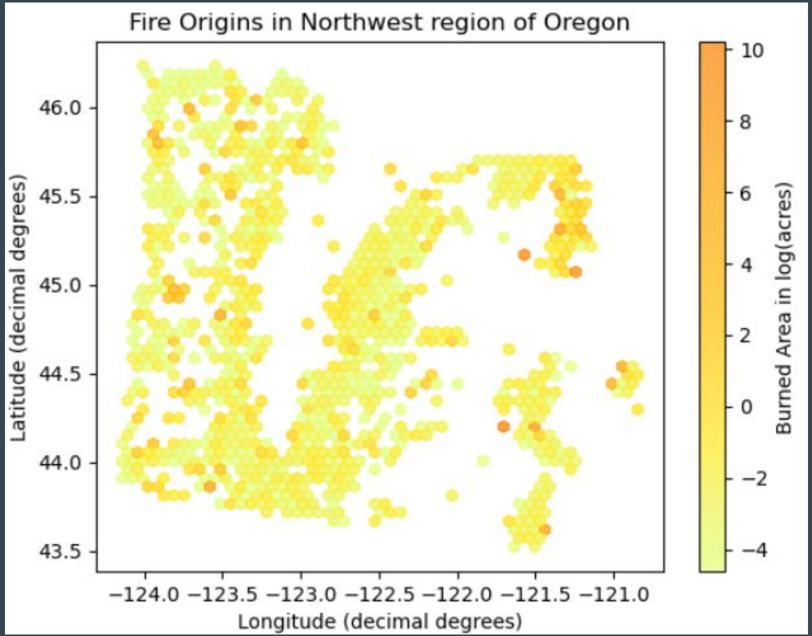
Northwest



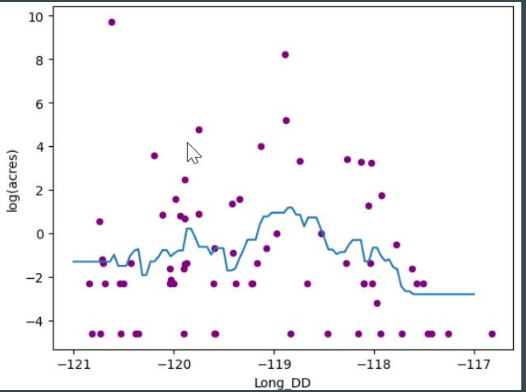
2.39 MSE



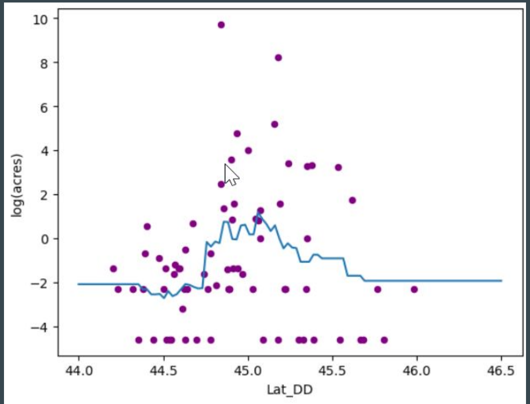
2.53 MSE



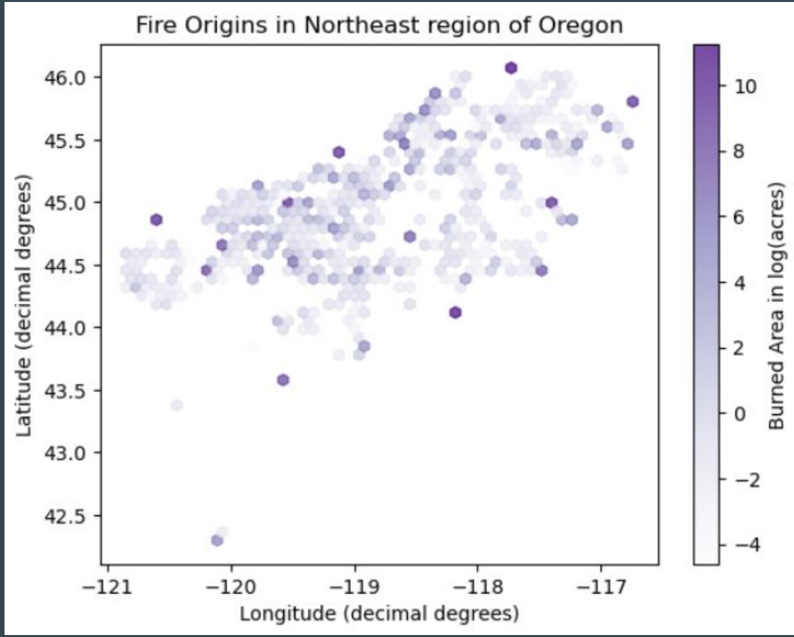
Northeast



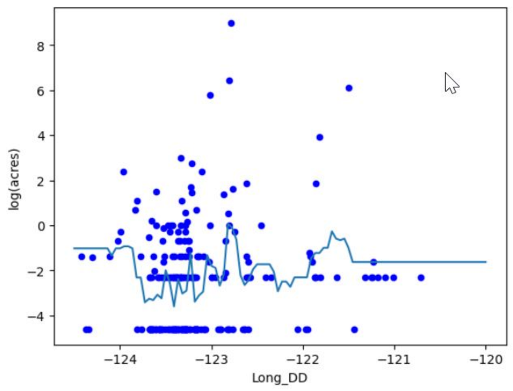
2.97 MSE



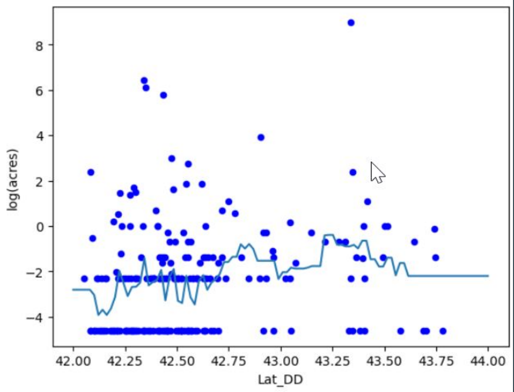
2.98 MSE



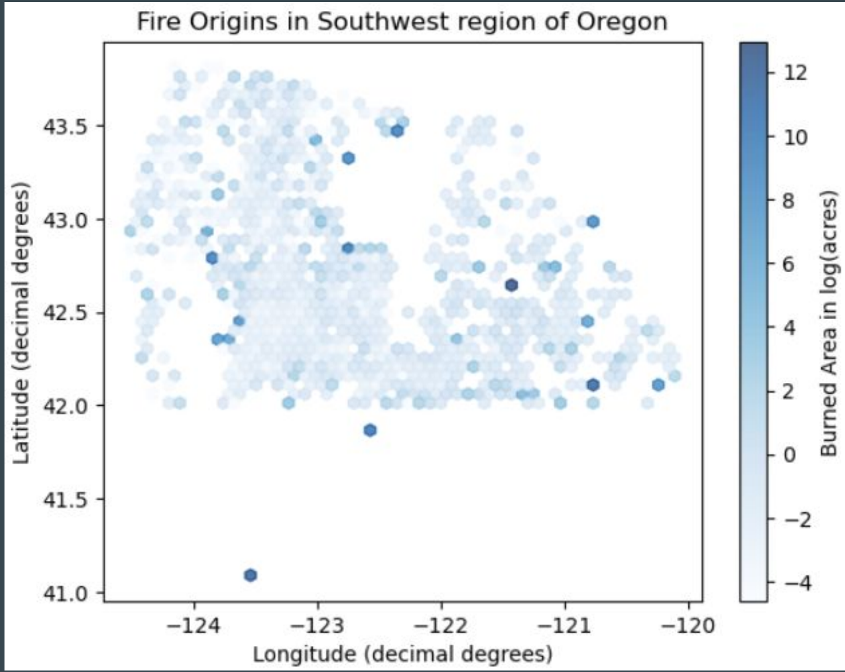
Southwest



2.48 MSE



3.13 MSE



CREATING A MODEL

Logarithmic Model: Goal

- Predict a major fire event (1000+ Acres Burnt) given following features:
  - County fire is located in
  - Cause of the fire
  - Type of land fire originated from.

Logarithmic Model: Preparing Data

- Data is full of categorical variables
  - Only Acres, Coordinates were continuous!
- Must first prepare categorical data for logarithmic model
  - “Hot-Encode” data. Have categorical variables as columns, record occurrence with 1, otherwise 0.

```
# Read in CSV file
df=pd.read_csv('FireOccurence.csv')

# Categorize/Bin major fire events (500+ acres burnt)
df.loc[df['EstTotalAcres'].between(0, 100, 'both'), 'major'] = '0'
df.loc[df['EstTotalAcres'].between(100, float("inf"), 'right'), 'major'] = '1'

# Dummy all variables being used, rows that have a certain event will have a "1" under respective column
major_fire = pd.get_dummies(df['major'], drop_first=True)
cause = pd.get_dummies(df['GeneralCause'], drop_first=True)
county = pd.get_dummies(df['County'], drop_first=True)
land = pd.get_dummies(df['FO_LandOwnType'], drop_first=True)

# Drop original classification
df.drop(['major', 'GeneralCause', 'County', 'FO_LandOwnType'], axis = 1, inplace = True)

# Swap in new classification
df = pd.concat([major_fire, cause, county,land])
df.head()
```

```
# For some reason fillna doesnt work for the whole thing so that sucks. Have to go with slow iterations :(
# This section takes a while so be patient please!!

for column in df:
    df[column] = df[column].fillna(0)

# Create two dataframes, one with outcomes (y) and one with input variables(x)
y = df.copy()
y = y.loc[:, ['1']]
x = df.drop(columns=['1'])

# Drop data that is not clear
x = x.drop(['ERROR: #N/A', 'Under Invest'], axis = 1)
```

Logarithmic Model: Model Evaluation

```
Cross Validation Scores:  [0.73776075 0.74457216 0.73648361 0.74414645 0.74042146 0.73914432
 0.7394636  0.74031503 0.73893146 0.73265219]
Average CV Score:  0.7393891017454236
Number of CV Scores used in Average:  10
```

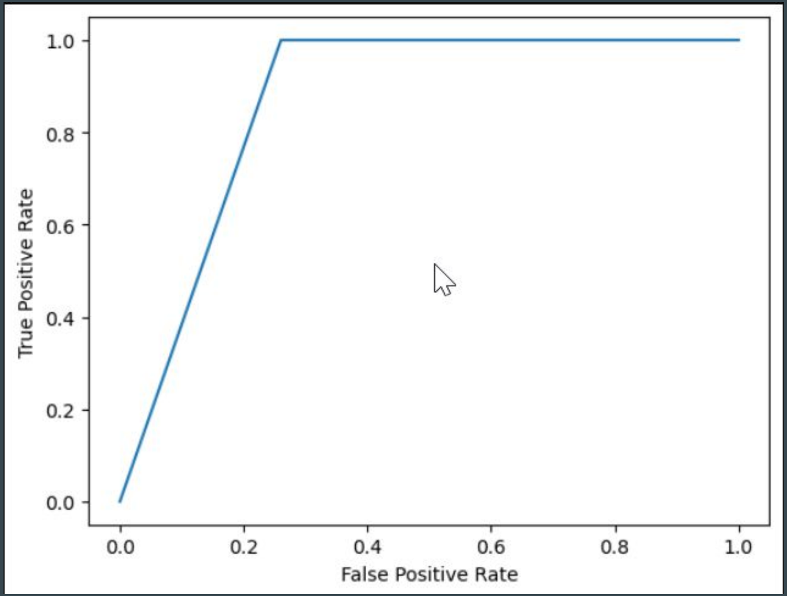
	precision	recall	f1-score	support
0.0	1.00	0.74	0.85	18676
1.0	0.02	1.00	0.05	116
accuracy			0.74	18792
macro avg	0.51	0.87	0.45	18792
weighted avg	0.99	0.74	0.85	18792

Logarithmic Model: Model Evaluation

	Predicted_Minor	Predicted_Major
Minor	13812	4864
Major	0	116

Logarithmic Model: Model Evaluation

- Log Model Intercept: ~1.3
- Area-Under-Curve: ~0.87



Logarithmic Model: Model Evaluation

Intercept: ~1.3

0	Debris Burning	0.001051
1	Equipment Use	0.001001
2	Juveniles	0.006223
3	Lightning	0.000782
4	Miscellaneous	0.001870
5	Railroad	0.024276
6	Recreation	0.001500
7	Smoking	0.003707

8	Benton	0.013302
9	Clackamas	0.004414
10	Clatsop	0.007987
11	Columbia	0.009388
12	Coos	0.004097
13	Crook	0.011819
14	Curry	0.008734
15	Deschutes	0.004623
16	Douglas	0.001857
17	Gilliam	0.519756
18	Grant	0.004264
19	Harney	0.087916
20	Hood River	0.010616
21	Jackson	0.001372
22	Jefferson	0.015180
23	Josephine	0.001666
24	Klamath	0.001997
25	Lake	0.009360

Lane	0.002096
Lincoln	0.012951
Linn	0.005793
Malheur	0.404513
Marion	0.009561
Morrow	0.022807
Multnomah	0.062234
Other State	0.126274
Polk	0.013424
Tillamook	0.009897
Umatilla	0.007156
Union	0.008807
Wallowa	0.005915
Wasco	0.008710
Washington	0.014208
Wheeler	0.009141
Yamhill	0.011960

BIA	0.027059
BLM	0.001711
County	0.003753
Industrial	0.001074
Municipal	0.015261
NPS	0.702017
Ocean Shores	0.023621
Other Private	0.001185
Other Public	0.008431
Rural Residential	0.000833
Small Woodlands	0.001633
State	0.002269
USFS	0.008662

Logarithmic Regression: Conclusion

- Definite “Hotspots” in Oregon when it comes to fires
- Model is somewhat accurate (~0.74)
    - Good at identifying/predicting major fire events
    - Less-so when it comes to minor fires
- Recall is high
    - The model can be used to predict fires that could become a major fire event if not handled appropriately.
    - Doesn’t mean it absolutely will.

QUESTIONS

1. Why are the values for the values for the AcresBurnt Scatterplot have "Log()" applied to them?

2. Despite the somewhat inaccurate Logarithmic model (~0.74 accuracy), why is the model effective at predicting the severity of fires?

3. Why are the correlation coefficients for the logarithmic model so small when first looking at them?

THANK YOU!