Godfrey Lozada, Rovin Soriano, Howie Nguyen, Tri Tran, Ian Lopez

Data Analysis Methods

CS105: Group 18 Project Report

Introduction:

Our dataset was found through data.gov. This dataset contains information on fires in the state of Oregon from 2000 to 2022. It includes 38 bits of information on these fires. However we only used 14 features (Serial, FireYear, Area, Size_class, EstTotalAcres, HumanOrLightning, CauseBy, General Cause, Lat_DD, Long_DD, FO_LandOwnType, County, Ign_DateTime, Discover_DateTime). Some of these features are categorical and some are numerical.

Legend:

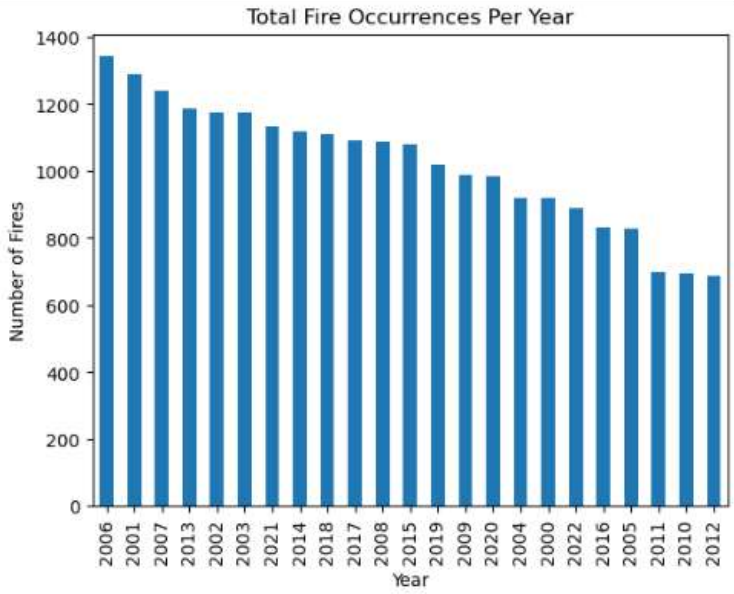
- **Serial:** It is the serial number of the fire, it is use for its unique identification
- **FireYear:** The year that the fire started
- **Area:** The area of Oregon that the fire started (NOA, SOA, EOA)
- **Size_class:** A classification of the fire based on how many acres were burnt (Numerical key provided further in the notebook)
- **EstTotalAcres:** The total amount of acres burned due to the fire
- **HumanOrLightning:** If the fire was started by humans or by lightning
- **CauseBy:** If human, this is the group that that person belonged to. If the cause was lightning, then it's just lightning.
- **GeneralCause:** The general cause of how the fire started.
- **Lat_DD:** Latitude coordinate of where the fire started.
- **Long_DD:** Longitude coordinate of where the fire started.
- **FO_LandOwnType:** The type of property where the fire started.
- **County**: The county in Oregon where the fire started.
- **Ign_DateTime:** The time at which a fire was ignited.
- **Discover_DateTime**: The time at which a fire was discovered by ODF (Oregon Department of Forestry).

Description:

The purpose of this project is to examine the relationship between fires and their aspects (like county, cause, amount burned, location) to determine correlation. Then, we will predict the class size of a fire by the features: CauseBy, FO_LandOwnType, and County. We will predict this using machine techniques discussed during lecture such as K-means clustering and K-fold validation. Additionally, we will use logistic regression to predict the chance of a major fire event

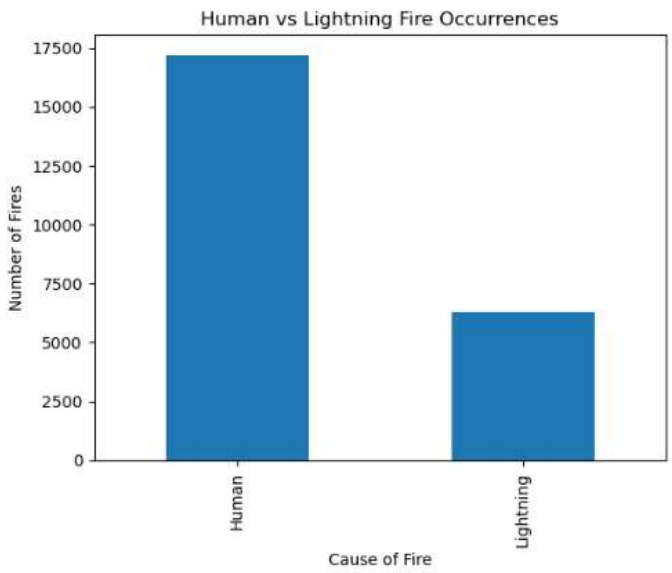
by these features: General Cause, County, and FO_LandOwnType (major fire event is classified as any fire that burns more than 500 acres of land before containment).

For this one we are trying to figure out which year have the most total fire occurrences per year:



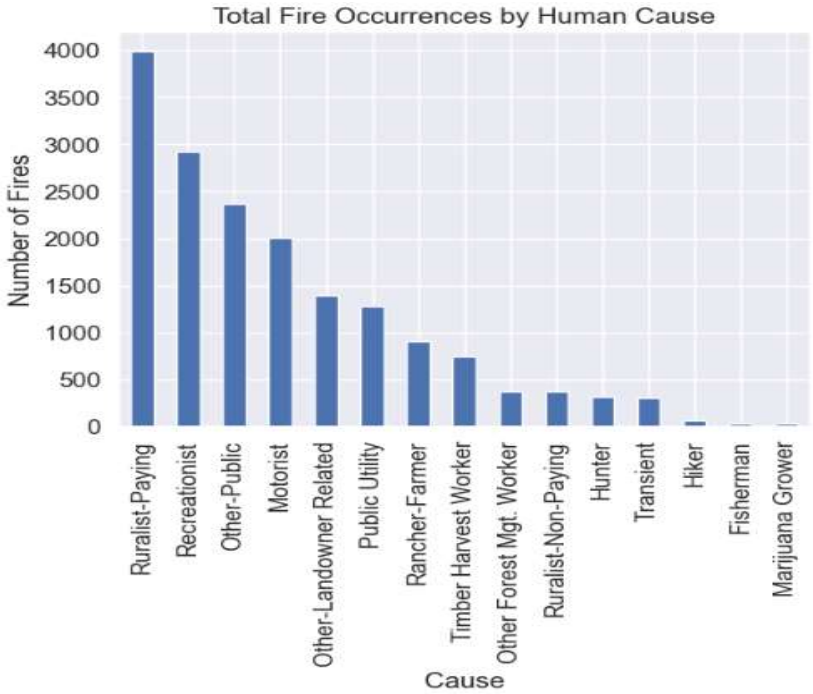
This bar chart depicts the frequency of forest fire incidents that occurred annually from 2000 to 2022, aiming to find a trend overtime. But here we can see that 2006 had the highest total fire occurrence per year and 2012 having the lowest fire occurrence total year.

After finding the year with the highest total fire occurrence, we then wanted to find out if what caused more fire occurrence between Human and Lightning fires:



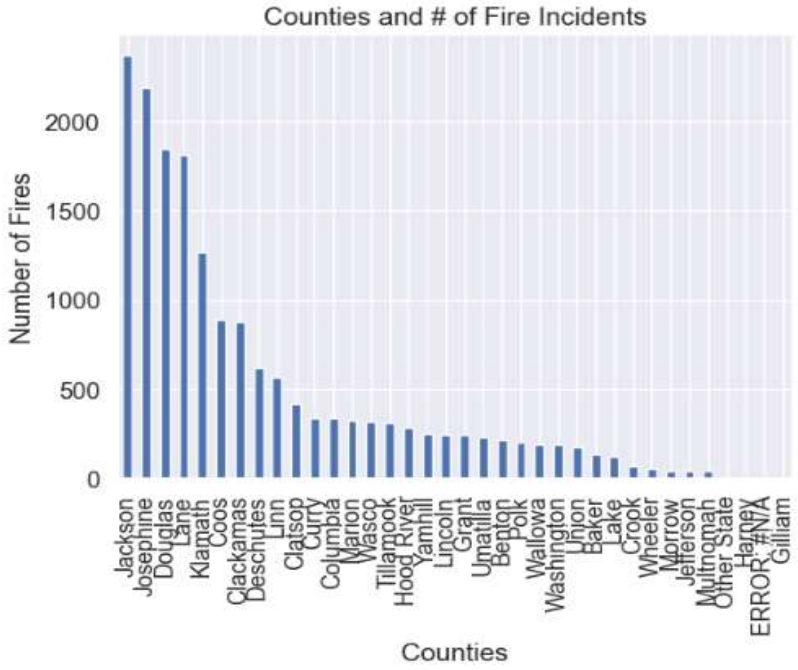
This visualization aims to compare the incidence of forest fires caused by human activities versus natural causes, specifically, lightning. From the chart, it is evident that the number of forest fires caused by human activities is over three times higher than those caused by lightning strikes.

Since we found out that Humans caused more fire than lighting we then wanted to know what are the human causes that resulted to fire incidents.



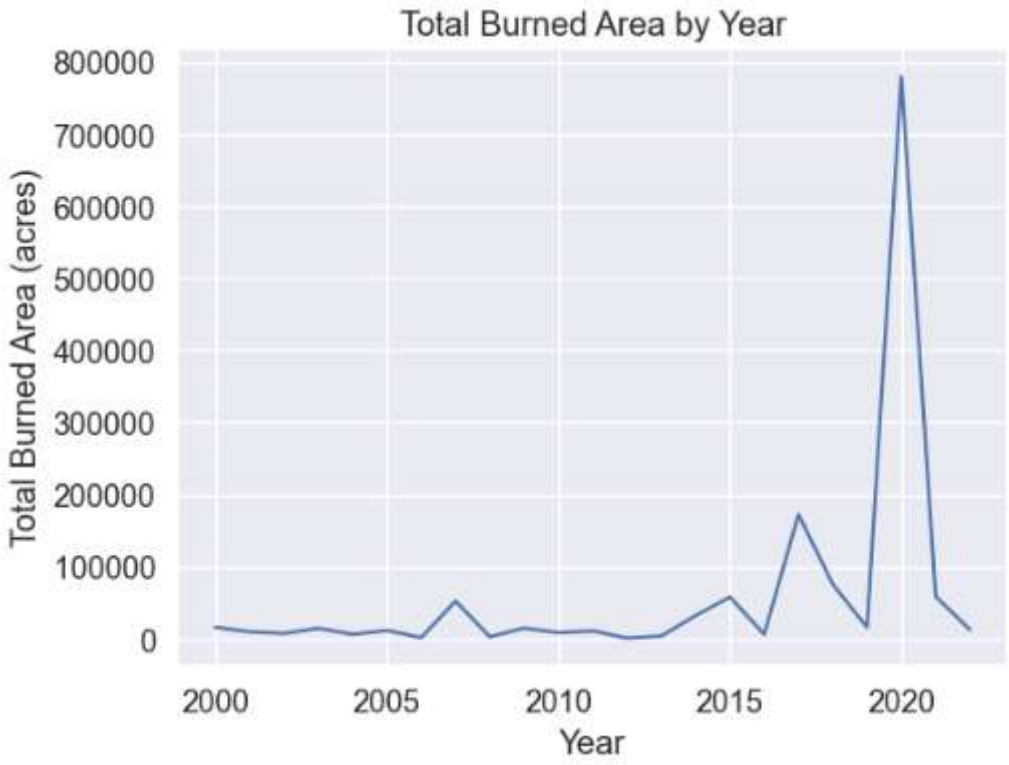
This visualization aims to analyze forest fires caused by all human activities. From the chart, it is evident that the category of Ruralist-Paying is responsible for the highest number of forest fires, with approximately 4000 incidents. On the other hand, the category of Marijuana Grower has the lowest number of forest fires, with less than 100 incidents recorded.

We then wanted to find out what counties in Oregon have the most fire incidents in total:



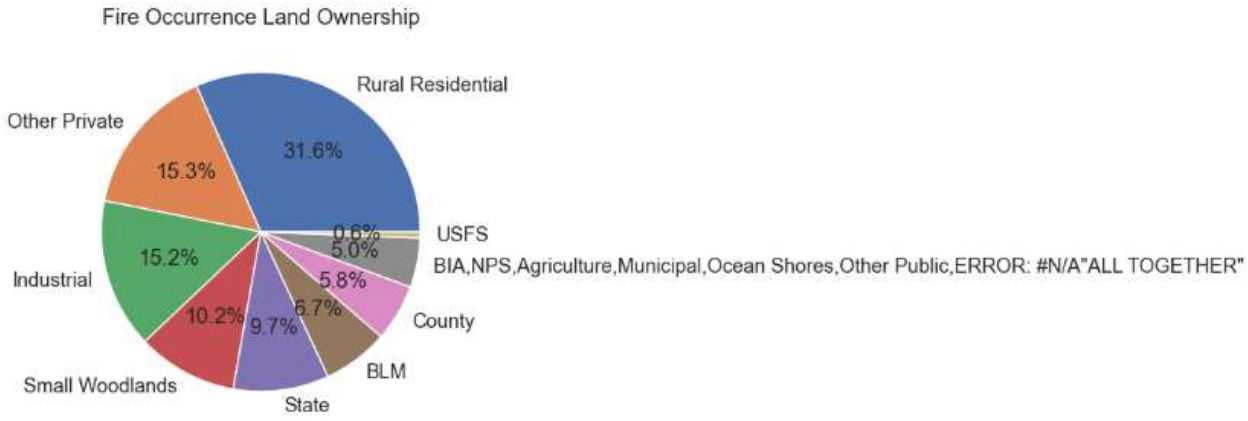
This visualization aims to investigate the number of forest fire incidents in each county of Oregon. From the chart, it is apparent that Jackson county has the highest number of fire incidents, with approximately 2400 incidents. In contrast, Harney county and Gilliam county, along with some other counties, have the lowest number of forest fires, all with less than ten recorded incidents.

We also wanted to find out in what year did the most area burned in total. We grouped them by the fire year and then get the sum of the burned area:



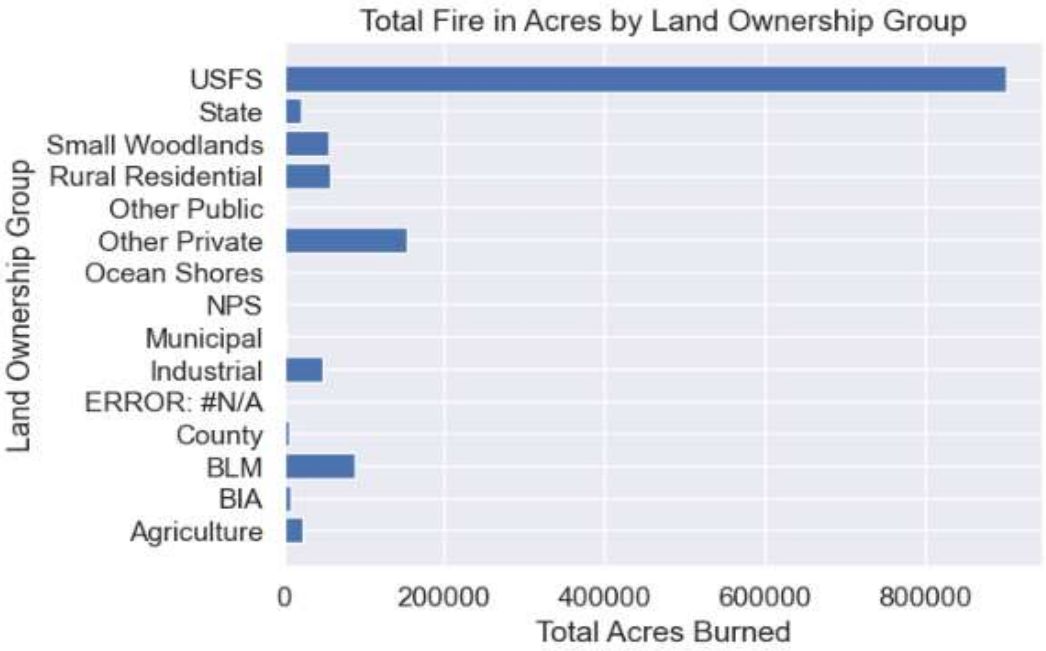
This line chart depicts the trend of the total burned area in acres resulting from fire incidents from 2000 to 2022. From the chart, it is evident that there is a consistent correlation in the total burned area from 2000 to 2015, except for a slight increase in burned area in 2006 and 2007 to about 70,000 acres, while the rest of the years recorded burned areas under 40,000 acres. However, from 2016 to 2017, there was a significant surge in burned area to approximately 180,000 acres. In 2020, there was a sharp increase in burned area to about 780,000 acres, indicating a notable deviation from the previous trend.

We then looked in the Fire occurrences within Land Ownership:



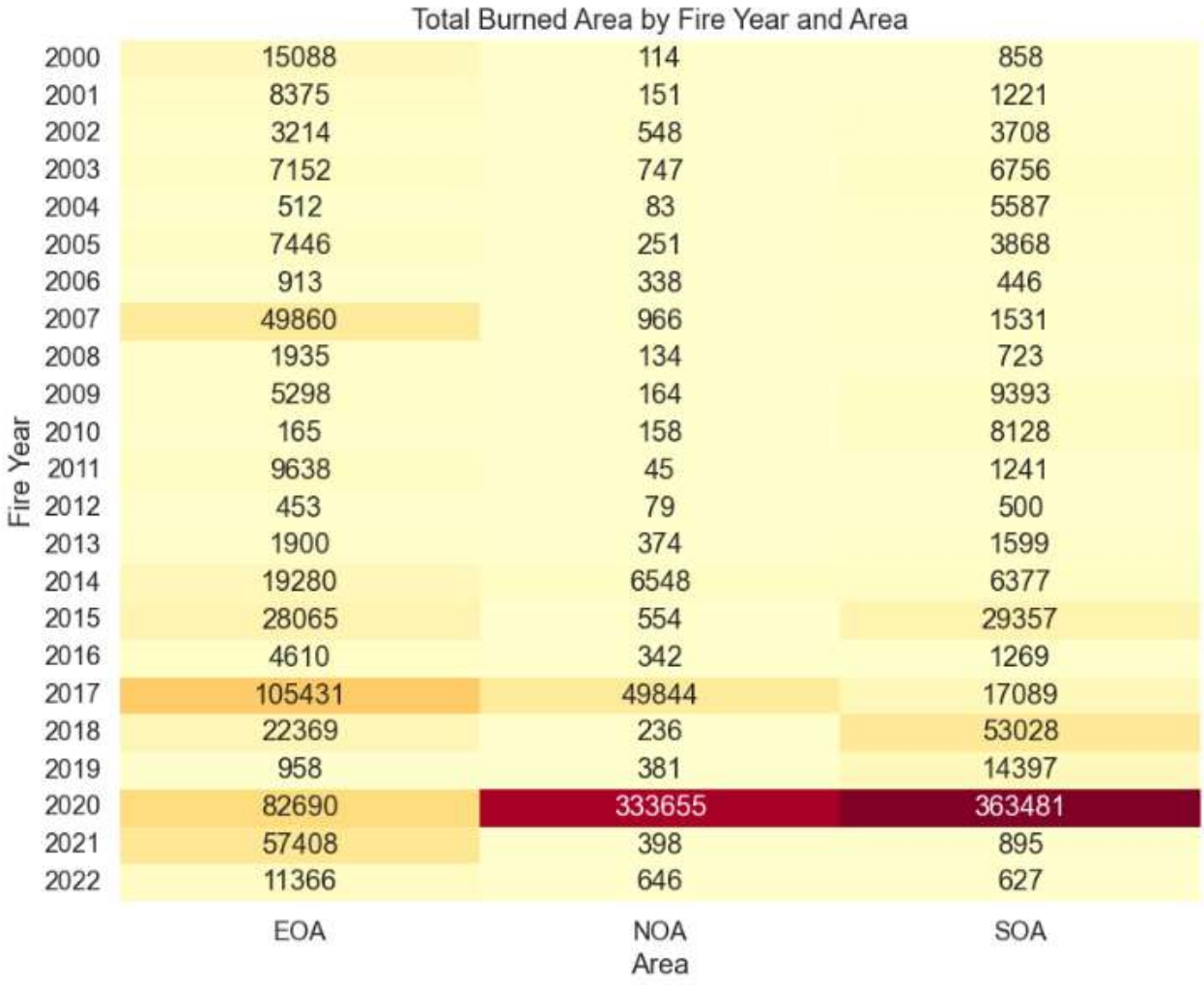
This pie chart aims to examine the land ownership distribution of all forest fire occurrences in Oregon State. From the chart, it is apparent that Rural Residential lands account for the highest percentage of fire incidents, representing 31.6% of the total. Conversely, USFS (United States Forest Service) lands have the lowest percentage of fire incidents, with only 0.6% recorded.

Since we saw the distribution of the fire occurrences within the Land Ownership we then check how much fire in acres each land ownership group in total:



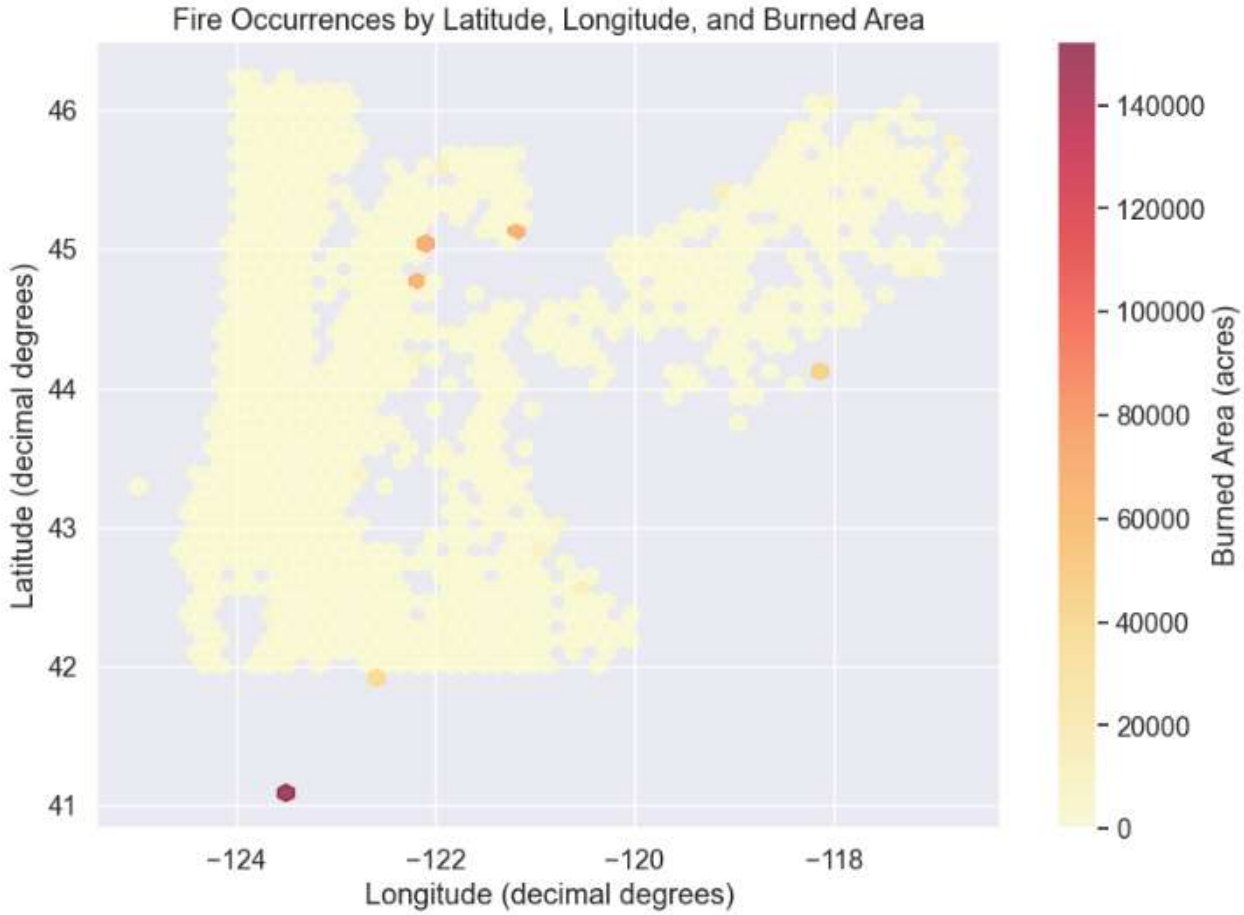
This visualization aims to analyze the total area burned by forest fires in different land ownership groups. From the chart, it is evident that USFS (United States Forest Service) lands have the highest total burned area, with approximately 860,000 acres. In contrast, Other Public lands, Ocean Shores, and NPS (National Park Service) lands recorded no burned areas in acres.

In the dataset we have the fire occurrences were divided into 3 different regions the East, North and South of Oregon and we wanted to see which area have the highest total burned area per year.



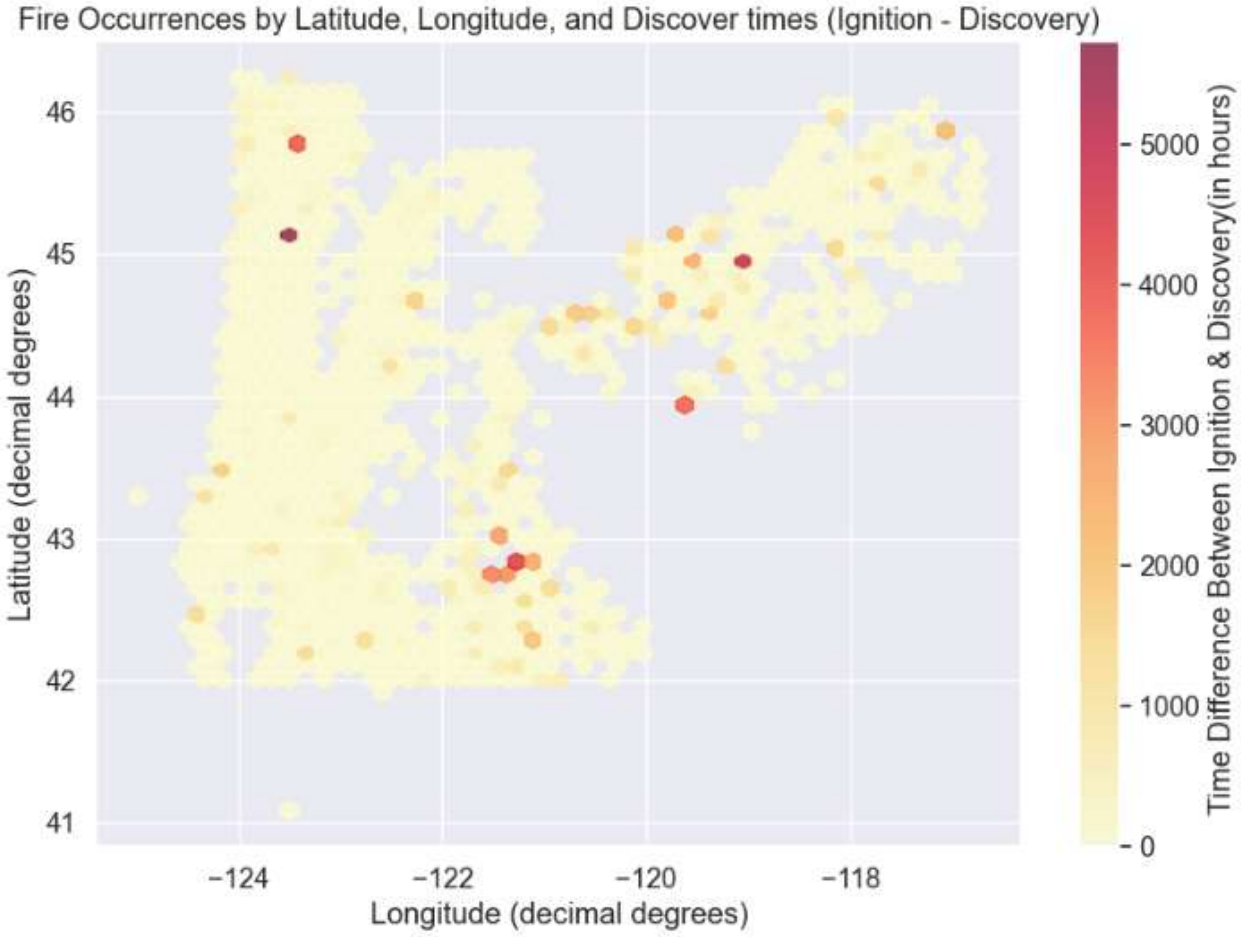
This heat map indicates the total area that was burned according to the year and ODF(Oregon Department of Forestry) Fire Protection Area. The Y-axis concerns the year. The X-axis concerns ODF(Oregon Department of Forestry) Fire Protection Area in which we have three areas: Eastern Oregon Area(EOA), Northern Oregon Area(NOA), and Southern Oregon Area(SOA). The lighter the shade the less area burned while a darker shade indicates more area burned. According to the heatmap, in 2020, each region suffered higher areas burnt, considerably more in NOA (333655) and SOA (363481) than in EOA (82690). Each area in each year suffered relatively less area burnt than 2020 with the exception of EOA in 2017. However, it is worth noting that spikes in area burnt are increasingly more common in later years (2014 and above).

We wanted to see the fire occurrence in the map of Oregon and we wanted to see which part got the highest burned area.

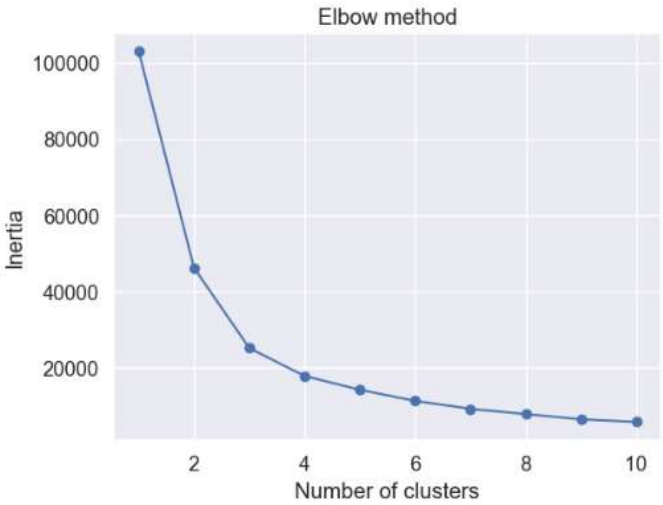


This heatmap describes the area burnt by a fire according to their location (latitude and longitude). The X-axis concerns the Longitude. The Y-axis concerns the Latitude. The heatmap colorization concerns the area burnt (in acres). A lighter shade indicates a relatively low area burnt while darker shades indicate higher area burnt. According to the graphs, 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.

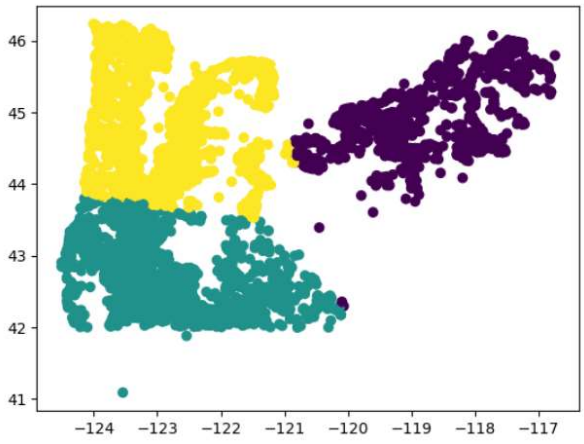
We graph the fire occurrence by its latitude and longitude and then we wanted to see the discovery times by subtracting ignition time - discovery time per fire occurrence. We wanted to see if there is a specific place where discovery time is slow.



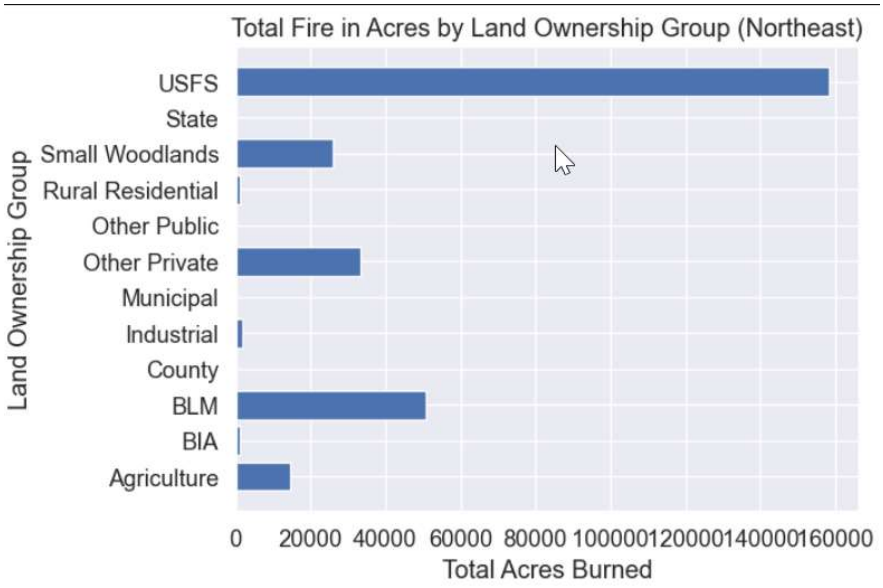
This heatmap describes the discovery times of a fire according to their location (latitude and longitude). The X-axis concerns the Longitude. The Y-axis concerns the Latitude. The heatmap colorization concerns the discovery times of a fire (the difference between the time of ignition and the time of discovery), or in other words, how long it took for a fire to be discovered after it started. A lighter shade indicates a relatively low discovery time while the darker the shade, the higher the discovery time. According to the heatmap, the fires in the Central Southern and Central part of Oregon have higher discovery times. Additionally, there are some extreme discovery times in the Northwestern part of Oregon



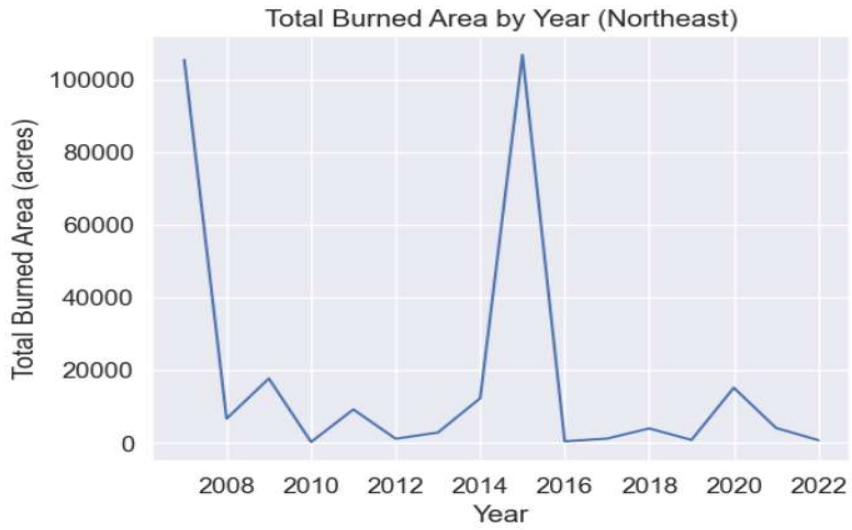
After running the code, we find that 3 clusters would be the most optimal group number to have for our testing. We have the data for a KMeans model, and have our first look at what individual clusters of coordinates we can evaluate.



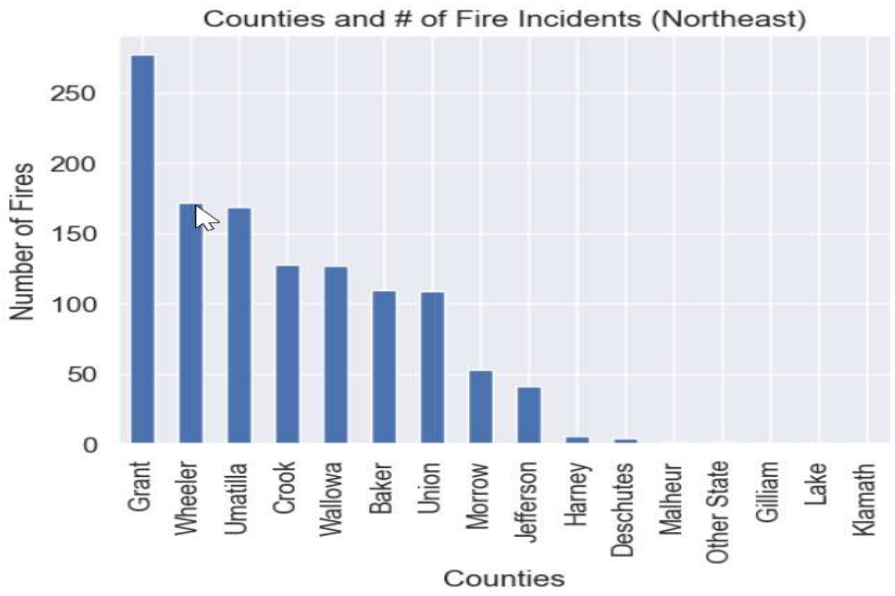
Cluster 1 Northeast



From reading the graph above, we noticed that USFS has the most fires in acres by Land Ownership by a long shot (roughly 10,000 more acres burnt than the runner up BLM). Making it the most unfortunate land to be burned up.

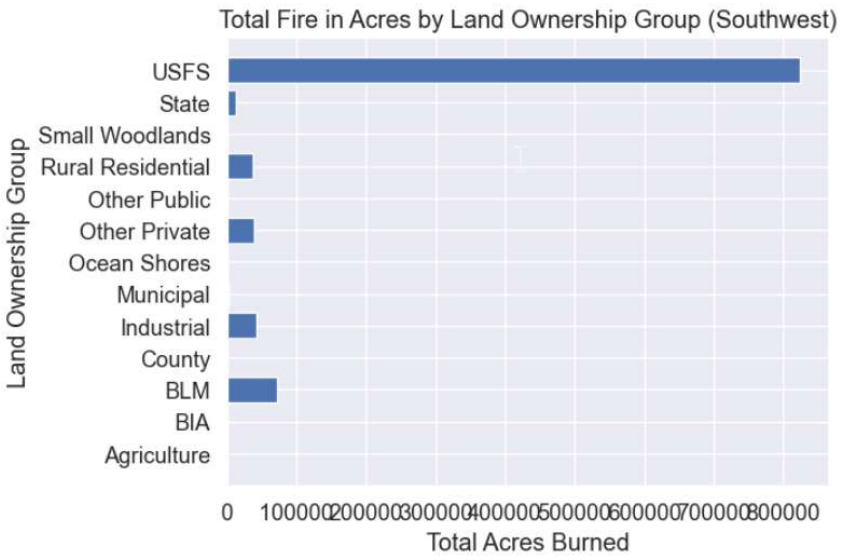


From reading the above we see that there's a big spike of more than 100000 acres of land burnt between 2014 and 2016. Burnt area then drops to a constant less than 5000 acres burnt but then bumps up to 18000 acres burnt before falling back down to less than 5000 acres in 2022.

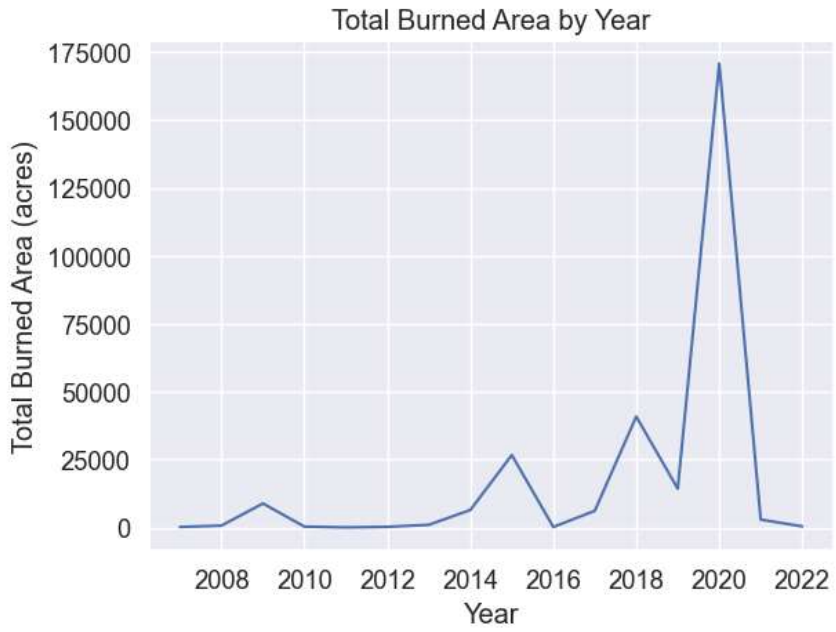


From reading the graph concerning fire incidents in the northeast, the grant county had the highest number of fire incidents overall with roughly 280 fire incidents. This is 100 more than both Wheeler and Umatilla counties.

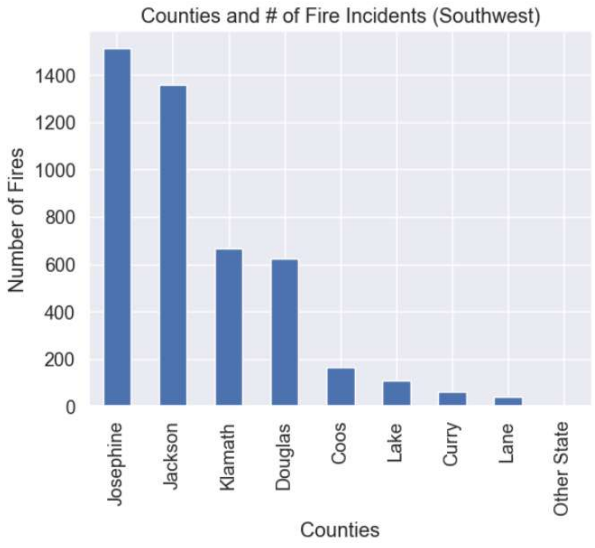
Cluster 2 Southwest Bar Chart



From reading the graph above, we have the USFS being the land ownership group with the highest total fires in acres in the southwest (around 830,000 acres). The USFS has about 760,000 more acres burnt than the runner-up BLM.

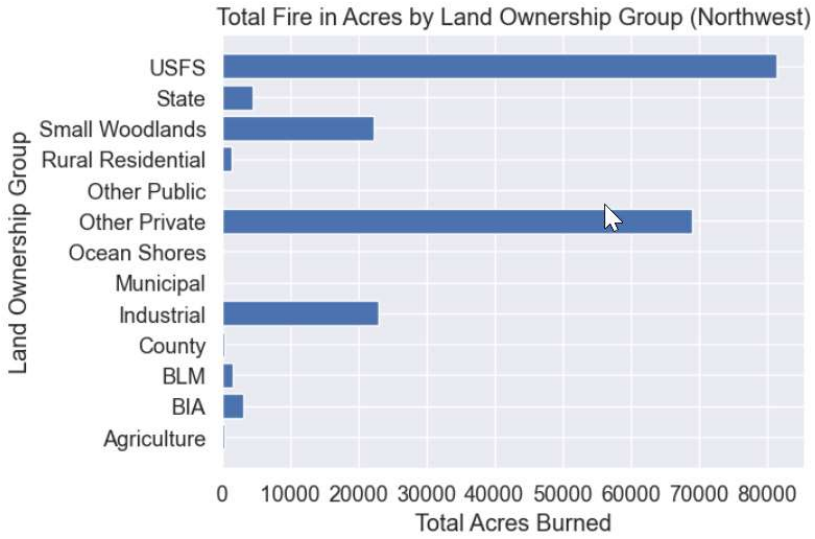


From reading this graph, in the southwest region the years with the most fire burnt is between 2020 and 2022. The peak is in 2020 with about 175000 acres of land burnt. We can see that as years continue the spikes/peaks are increasingly more common than previously.

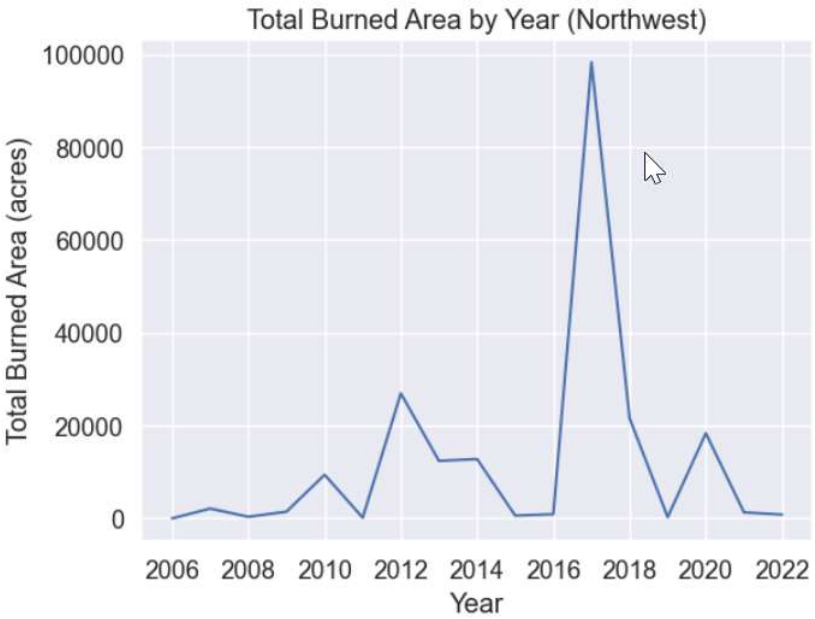


From reading this graph, we see that Josephine is the county with the highest number of fire incidents in the southwest. Josephine has around 1500 fires and the runner-up has around 1290 fires.

Cluster 3 Northwest

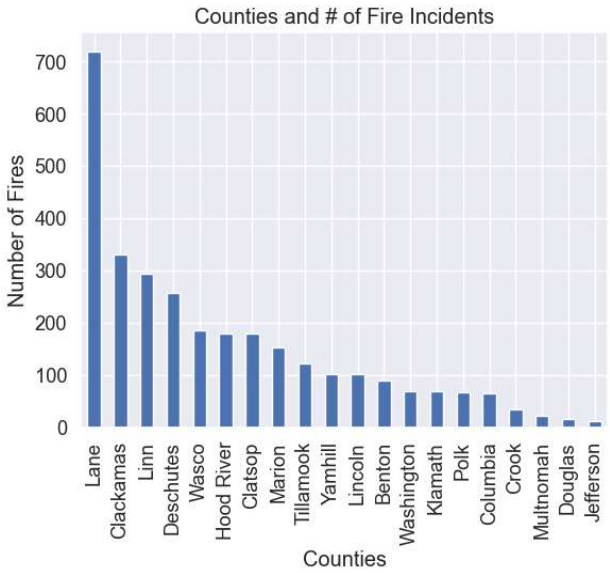


From reading the graph above for the northwest region, the land ownership group that holds the most fire burnt in acres is still USFS. USFS has a lead of around 10300 acres burnt compared to the second most group Other Private.



From reading this graph above, we can see that in the NorthWest region, the amount of area burned spikes between 2016 and 2020, with the peak occurring in 2017 (100,000 acres

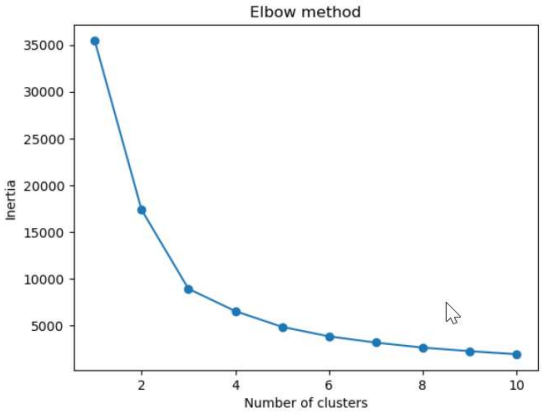
burnt). We can infer that as the years continue on, the likelihood of peaks in amounts of acres burnt will increase.

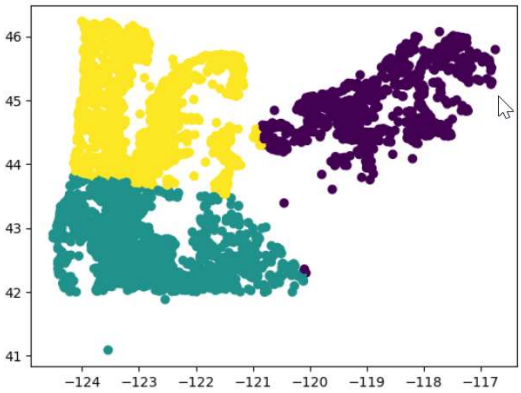


From reading the graph above, in the Northwest region, we see that Lane county is where the highest number of fire incidents occur. Lane county has around 720 fires which is 390 fires above the runner up (more than twice as much).

Model Training and Visualization:

First we decide to split the data so we may further analyze different sections of Oregon. In order to do so, we use K-Means Clustering to find appropriate sections. We utilize the elbow method to figure out the appropriate number of clusters to be used in the model and we get the following results:

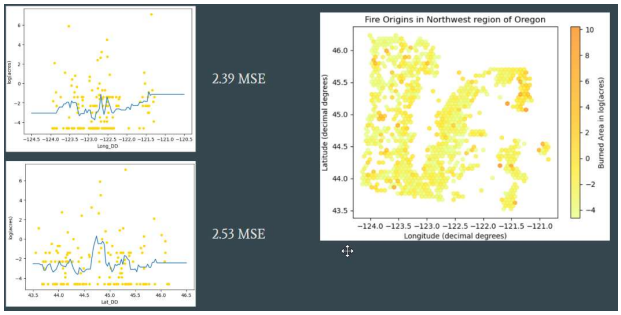




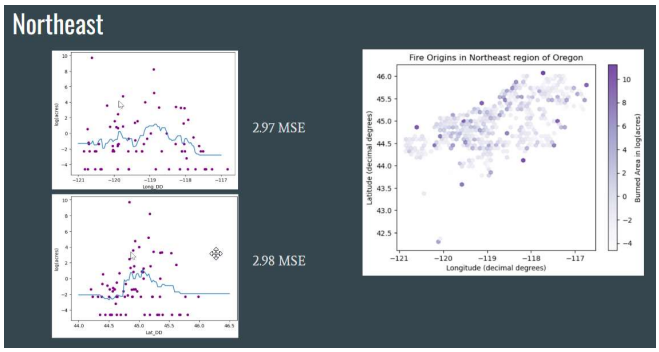
Here we can see that 3-clusters is the ideal amount, and that Oregon is to be divides into northeast, northwest, and southwest regions.

We then decided to take the clusters created by the K-Means Clustering and use the coordinate groups to analyze the data further. We took two input variables (longitude and latitude) and fed them through a 10-nearest-neighbor model to determine what size fires (in log(acres) we may see given the coordinate value. We log the values of these fires because the fire sizes vary greatly, so log() can be used to flatten the values and doesn;t undermine smaller fires. It is also worth noting in the hex-scatterplot that the dots don't necessarily represent the fire themselves, but the starting point. We run this model for all 3 different clusters and get MSE scores of anywhere between 2-3.2.

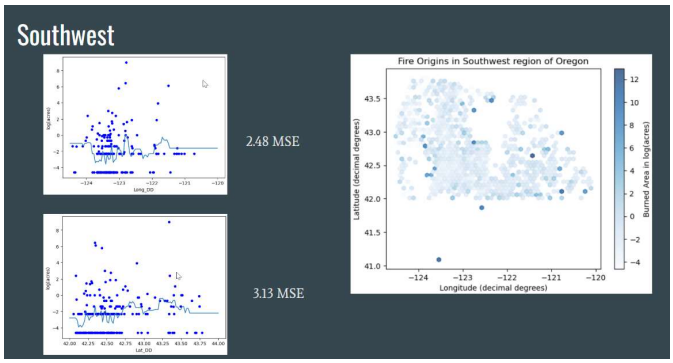
- **Northwest:** A lot of the major fires in the region seem to be split between east and west, but also seem to be in the middle when looking north-south. A lot of the major counties border the east-west areas, so it would make sense that there are the most major fires there.



- **Northeast:** This cluster has a bit more error due to the sporadic nature of the fire occurrences. However, the KNN model shows a slight increase in fire size in the middle-middle action of this region.



- **Southwest:** The fires in this section are very scattered, and the data shows it. In the KNN graphs we see a much flatter line than the other two regions. This could be because the southwest has a much higher urban and suburban population than the other two clusters. Much higher risk of fire, but lesser risk of a large wildfire.



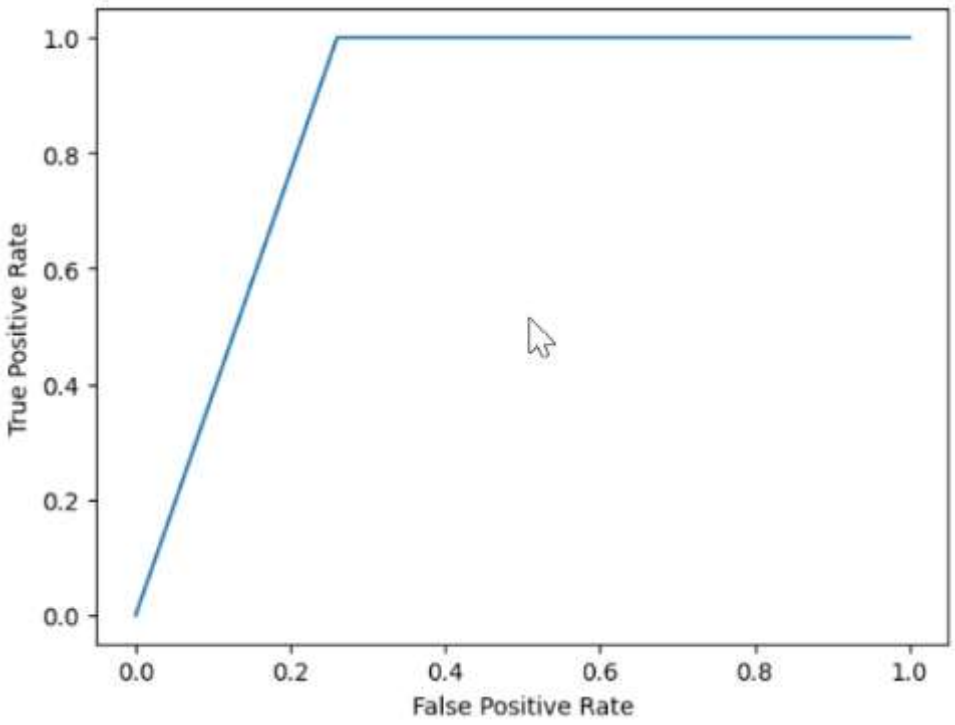
We then used the logarithmic model to predict a major fire event (burn area of 1000+ acres) given features such as county origin, cause, and land-type. This data consisted of many categorical variables, so we used “One-Hot encoding” to signify the occurrence of a feature in each fire instance. After processing, there were about 77+ columns of data! We then created a model and trained it using the fire data, and ran a 10-fold cross validation to determine the accuracy of our model, along with a confusion matrix of one of the test/training splits.

```
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

	Predicted_Minor	Predicted_Major
Minor	13812	4864
Major	0	116

What we notice here is that the model has an acceptable CV score, and that it has very very high recall. This means that the model does a good job at classifying the rows between the two outcomes. We can further prove this with an Area-Under-Curve graph, which comes out to ~0.86.



When we look at the correlation coefficients, it is important to note that because this is a logarithmic model, the values seen are the values *after running an exp()* function on the original coefficients. This is why the value is so small. We examine the correlation coefficients below, with the intercept values at ~1.3.

Cause Coefficient: County Coefficient: Land-Type Coefficient:

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

Using these coefficients if they were not ran through exp(), we get an overall formula of:

Chance of burning = -exp(Cause) - exp(County) - exp(Land-Type) + 1.3

- 0.5: 50/50 chance on major fire
- < 0.5: Lower chance to be a major fire.
- > 0.5: Higher chance to be a major fire.

Looking at this data, it is easy to see why our model often identified non-major fires and being major. Many of these coefficients are small, and don't have a big impact on lessening the chance of a major fire, besides small counties like Gilliam or land-types like NPS.

After analyzing the model, we can conclude that it does a sufficient job at protecting against/predicting major fires. Though its accuracy is way off for minor fires, the sample we used had a 100% major-fire recall score. It does a very good job at identifying what could become a major fire given the county, land, and cause.