Analyzing California Wildfire

October 2, 2022

Goal

Rising temperatures, drought conditions, and dry vegetation are all contributing causes of wildfires in California, but the worsening impact of climate change is further exacerbating and prolonging the fire season in the state.[1]

As part of this assignment, I analyze the wildfire in California to understand the the trends of wildfire from 2013-2020.[2] Using a open-dataset from Kaggle, I have analyzed the distributions of wildfires, the impact wildfires have and the amount of resources that are used during a wildfire.

Github: https://github.com/maitreyi-kv/cmpe255

Dataset Details

The collection includes information about the counties where wildfires have occurred, their latitude and longitude values, and the times the flames started. With this dataset the goal is to obtain insights on the areas of California that are most at risk for wildfires, as well as when and how frequently they occur.

Dataset Link: https://www.kaggle.com/datasets/ananthu017/california-wildfire-incidents-20132020

This dataset is uploaded into big query in dataset *california* under the table *cali_fire*. The result dataframe after processing is stored in table table *fire_post_processing*.

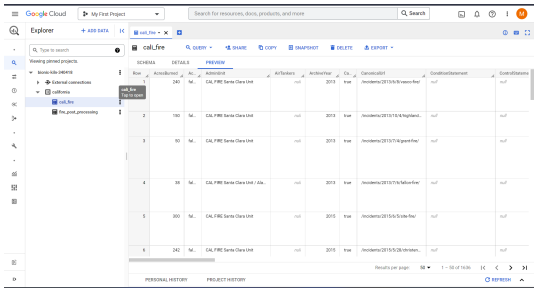

Figure 1: Dataset before and post processing on Big Query

Loading Data, Analysis and Observations

0.1 Loading Data

Pulling data from BQ to analyze

```
[ ]  query1 = """SELECT
       *
     FROM
       california.cali_fire
           """
     dfquery = client.query(query1)
```

```
[ ]  df = dfquery.to_dataframe()
     df.head(10)
```

Warning: Total number of columns (40) exceeds max_columns (20). Falling back to pandas display.

	AcresBurned	Active	AdminUnit	AirTankers	ArchiveYear	CalFireIncident	CanonicalUrl	ConditionStatement	ControlStatement	Counties	...	SearchKeywords	Start
0	240.0	False	CAL FIRE Santa Clara Unit	NaN	2013	True	/incidents/2013/6/8/vasco-fire/	None	None	Alameda	...	Vasco Fire, June 2013, Vasco Road, North Vasco...	2013-06-10:16:00+00:
1	150.0	False	CAL FIRE Santa Clara Unit	NaN	2013	True	/incidents/2013/10/4/highland-fire/	None	None	Alameda	...	Highland FIre, October 2013, Alameda County, H...	2013-10-07:35:00+00:
2	50.0	False	CAL FIRE Santa Clara Unit	NaN	2013	True	/incidents/2013/7/4/grant-fire/	None	None	Alameda	...	Grant FIre, July 2013, Grant Line Road, I-580,...	2013-07-10:00:00+00:
3	38.0	False	CAL FIRE Santa Clara Unit / Alameda County	NaN	2013	True	/incidents/2013/7/6/fallon-fire/	None	None	Alameda	...	Fallon Fire, July 2013, Fallon Road, Camino Ta...	2013-07-13:09:00+00:

Figure 2: Dataset loaded in Colab from BQ

0.2 Basic Analysis

The dataset has 1636 wildfire incidents over the course of 2013 to 2019. There are 40 features for the incident. All the wildfires in the dataset are inactive.

```
df['Active'].value_counts()
```

```
False    1609
Name: Active, dtype: int64
```

Figure 3: All wildfires are inactive

0.3 Preprocessing

I am performing preprocessing to clean the dataset. Broadly they belong to the following buckets.

1. Duration - In some of the records, the end-time is lesser than the start time. In some other cases, the end-time is so big indicating a super long wildfire (verified that these are inaccurate). I am dropping these values as it is a mistake in the data entry.

Figure 4: Negative duration - incorrect data entry

2. Lat/Long - In some of the records, I can observe that the latitudes and longitudes are invalid.

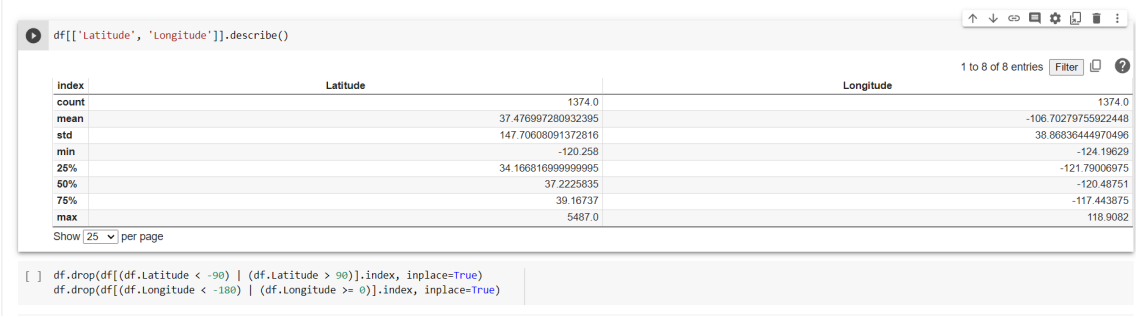

Figure 5: Negative duration - incorrect data entry

3. Duplicates - There are duplicate records in the data-set.

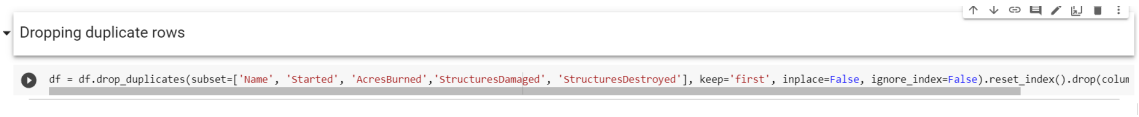

Figure 6: Removing duplicate entries

Records in the data-set which belong to either of the above buckets are dropped.

0.4 Observations and Visualization

1. How long do wildfires last for? Analyzing the duration of the wildfires to understand length.

As we can see most fires last less then a day. Long lasting fires are more more less frequently occurring than shorter fires.

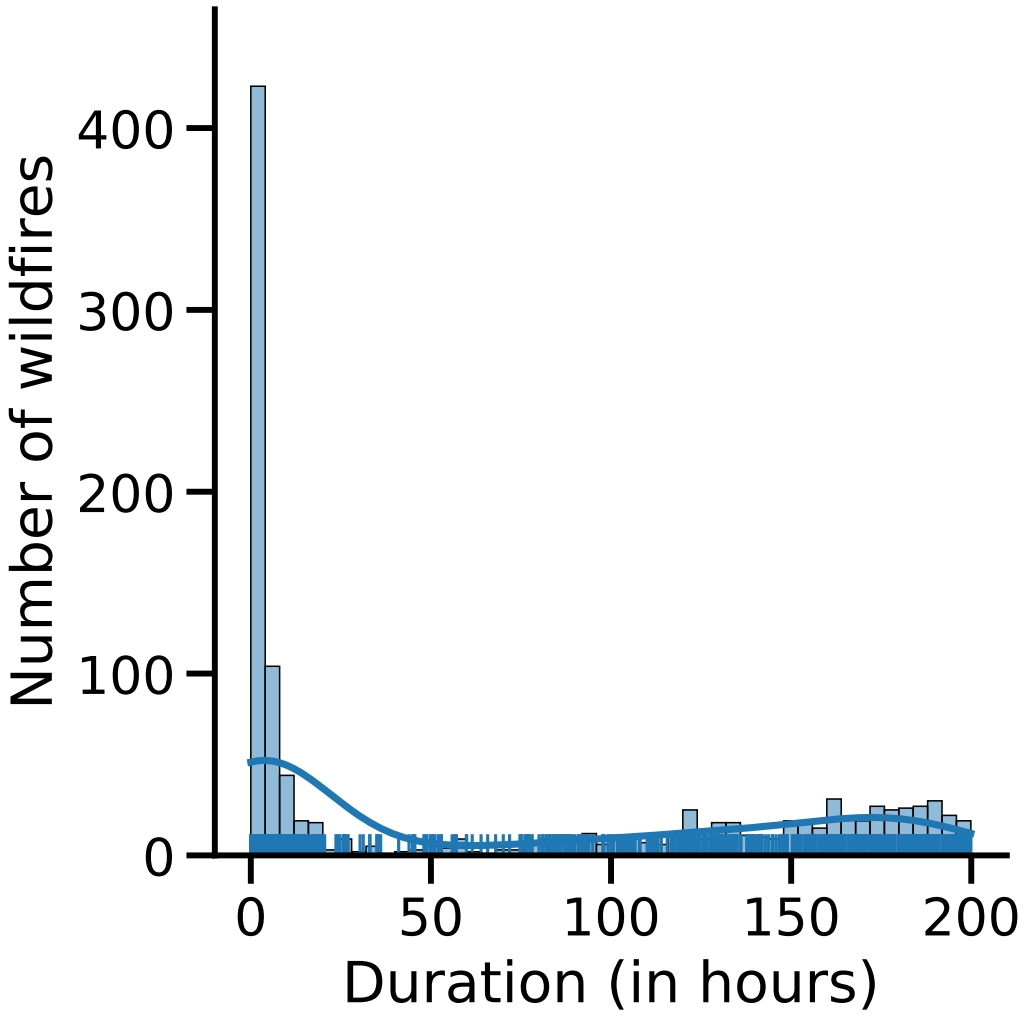

Figure 7: Length of wildfire

2. Where do wildfires occur? Analyzing the locations of the wildfires to understand if any area is more prone to wildfire.

Riverside and San Diego are places that are most affected since the last 6 years by wildfire.

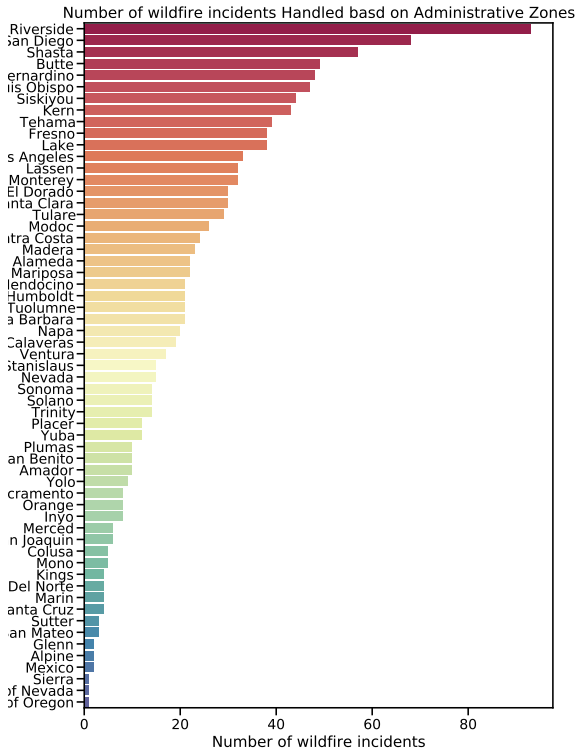

Figure 8: Most affected counties

3. Analyzing Administrative unit performing relief

There are 417 admin units in the dataset.

Looking at the administrative zones for the California Fire Dept. (shown below) we observe that the Sanoma Lake, Santa Clara, and Shasta-Trinity Unit are the top three administrative zones dealing with fire.

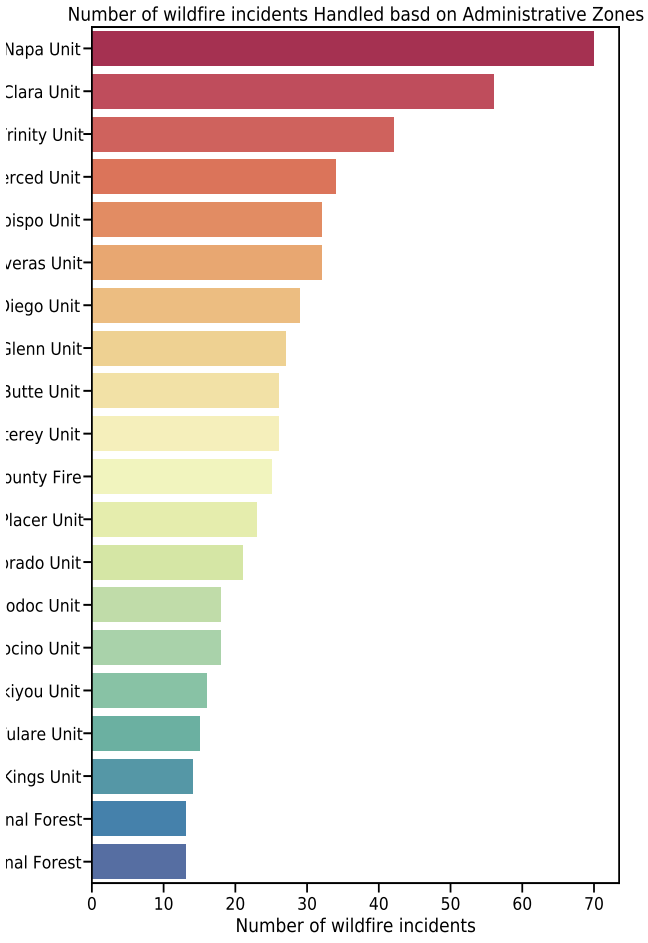

Figure 9: Top administrative unit performing relief

4. Analyzing geolocation of areas burned in fire incidents

Major wildfire incidents have mostly occured near regions of large population density and mostly concentrated towards South Eastern regions (San Diego and Riverside) and above San Francisco.

Wildfires since 2017 have been mostly concentrated in the same regions in California.

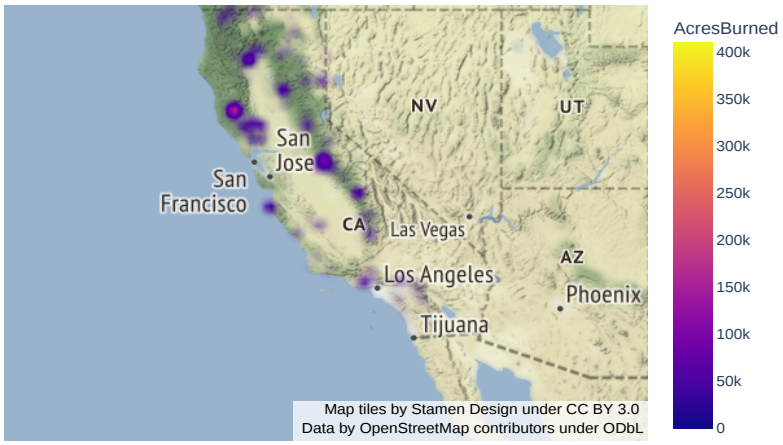

Figure 10: Density (in acres) of area burned by Wildfire

5. How many personnel's are involved in fighting Wildfires. How much fatality occurs across locations. Both personnel's and fatalities data correlates with the locations which have been heavily affected by wildfires.

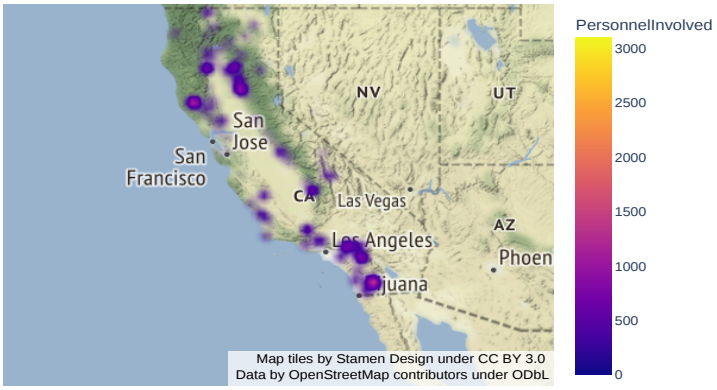

Figure 11: Personall's putting out fire in locations

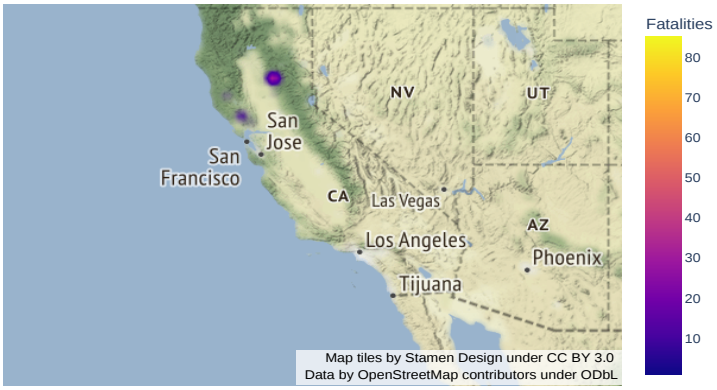

Figure 12: Fatalities occurring in locations

6. How many wildfires occur each year?

2017 has seen the most number of wildfires. From 2014 there is a constant increasing trend which could be due to global warming and climate change.

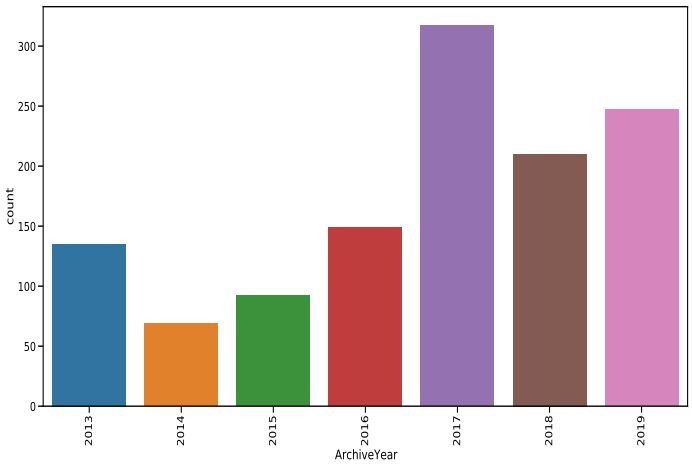

Figure 13: Wildfire by year

7. Are the acres burned directly related to no of wildfires.

Although there is a positive correlation, it is not always true as the largest wildfire is in 2018. The number of wildfires is lesser than 2017 yet the acres burned is much more.

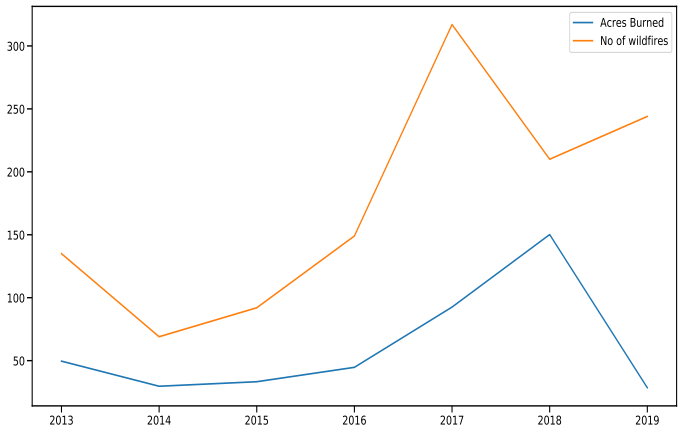

Figure 14: Correlation between wildfire and acres burned

8. What about correlation between acres burned and water tenders? And correlation between acres burned and personnel's? Are there any patterns/changes across the years?

The water tenders are is correlated. Especially during the major fire event in 2018 - most number of water tenders are used.Although, in the beginning years the number is much more. This decreases over time, and less water tender is being used. This could be due to technological advancement in mitigating wildfires. Similar patterns are seen in personnel's, they decrease over time which could be due to tech advancements.

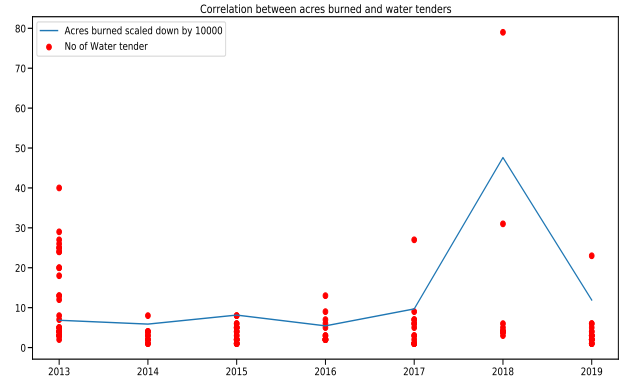

Figure 15: Correlation between wildfire and water tender

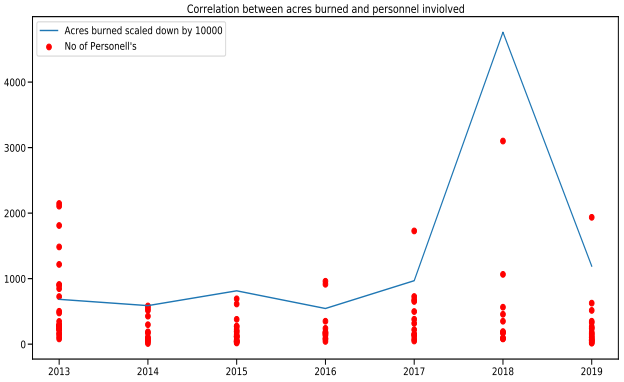

Figure 16: Correlation between wildfire and personell's

9. What is the correlation amongst all features? Is there any high correlation features in this?

The Pearson's correlation shows that there are high correlations between

1. corr of .90 between Personell's involved and Fatalities/Engines 2. corr of .99

There is Moderate correlation between Water tenders and personnel's involved. This can be visualized in the pair plot below too.

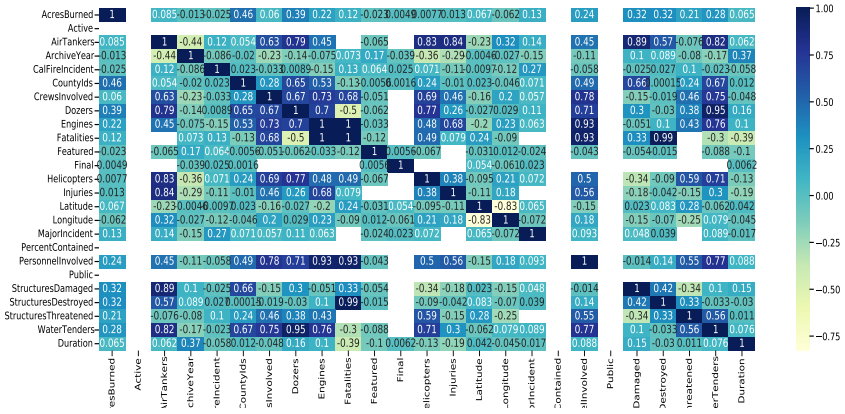

Figure 17: Pearsons correlation

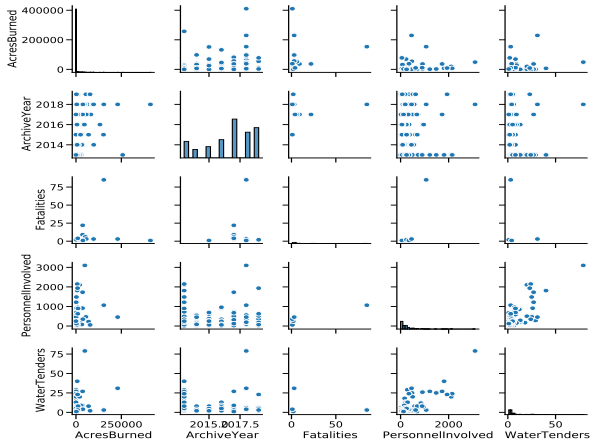

Figure 18: Pair Plots

0.5 Webapp

Using plotly and dash, I have created a app to visualize 4 different charts - Geo Location Map, Histogram, Scatter Plot, Pie Chart

1. Locations affected by wildfires using Geo Location Map
2. Counties impacted by wildfires using Histogram
3. Top 5 administrative units performing relief using Histogram
4. Distribution of wildfire by county using Scatter Plot
5. Distribution of wildfire by year using Pie Chart

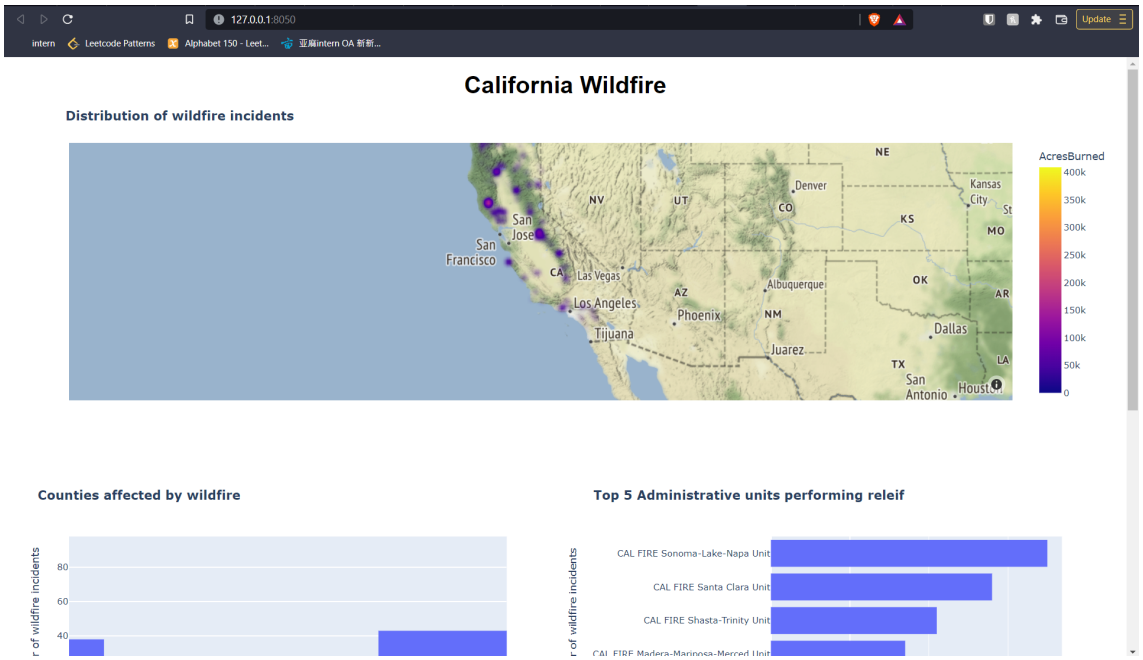

Figure 19: Webapp using plotly and Dash, Page 1

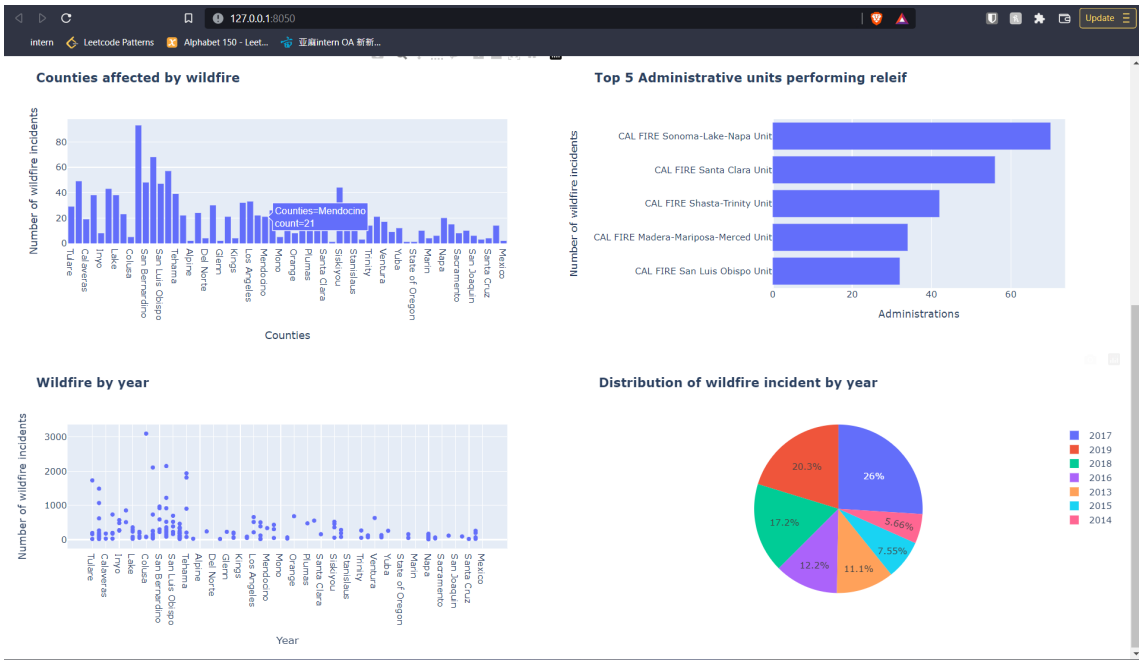

Figure 20: Webapp using plotly and Dash, Page 2

References

[1] https://earth.org/what-causes-california-wildfires/#: :text=Rising

[2] https://www.kaggle.com/datasets/ananthu017/california-wildfire-incidents-20132020

[] https://plotly.com/python/line-charts/