**Study of Wildfires across US**

**GROUP No.: 22**

**Final Report**

Anusha Kondle, Nabeel Khan, Mohit Mathur, Charan Kumar

**Problem Statement** – Wildfires has been one of the most prominent disasters lately in the US- some of these occurred naturally and the others were man-made. According to Wikipedia, more than 10.2million acres of land was burned in wildfires in the Western US alone. While this affects the vegetation of the forests and their inhabitants, there are more casualties and concerns. According to WHO, Wildfires is not just affecting the wildlife but the release of carbon dioxide, carbon monoxide and other toxic substances is increasing the air pollution which in effect is causing health issues- both physiological and psychological. Climate change being one of the serious concerns, we are interested to research more about the wildfire spread across the US. In this project we aim to look at the following aspects:

1. Visualizing the trends of wildfires over the 7 years period, from 2009 to 2015
2. Highlighting some common causes of the wildfires

**Methodology** –

1. **Pre-Processing:**

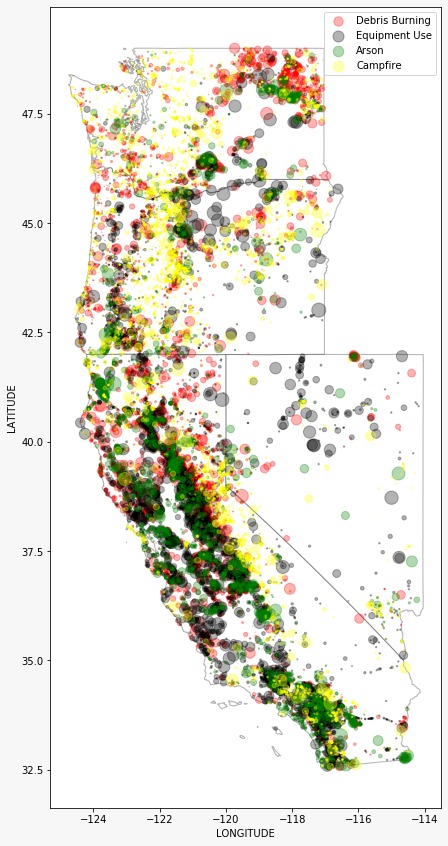
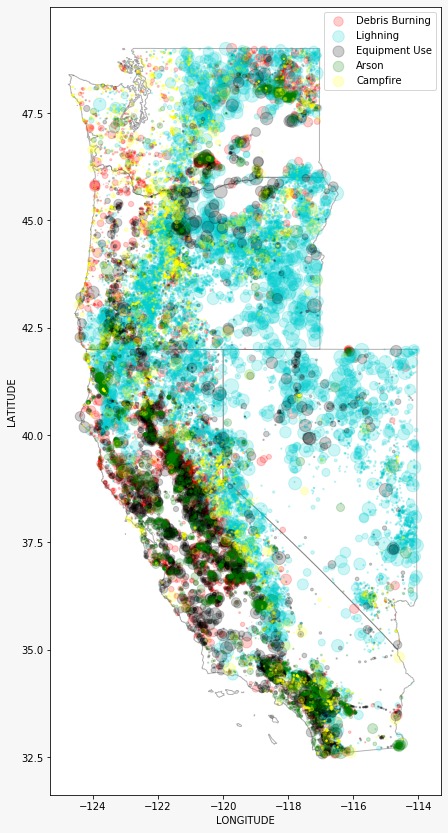
We loaded the sqlite database from Kaggle website, US map from census website and an excel list of state codes(for only mainland US) and their timezones that we created locally to map them to the original data.

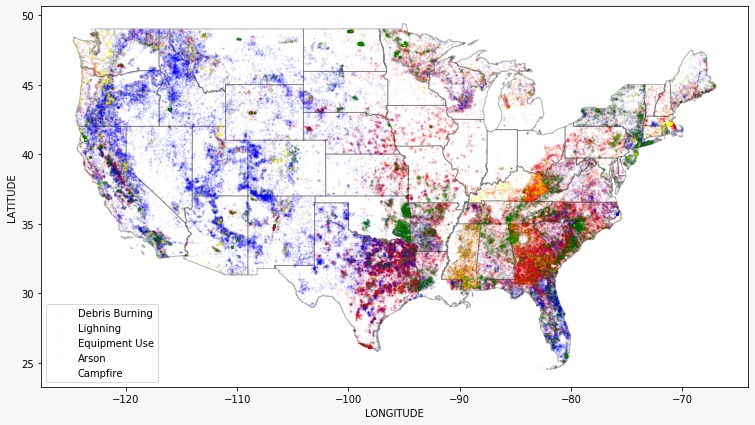
We are interested in “FIRES” table from all the tables in the database. We queried to select the following features- “FIRE\_YEAR”, “STAT\_STAT\_DESCR”, “LATITUDE”, “LONGITUDE”, “STATE”, “DISCOVERY\_DATE”, “FIRE\_SIZE” from 2009 and converted dates from Julian to Python datetime.

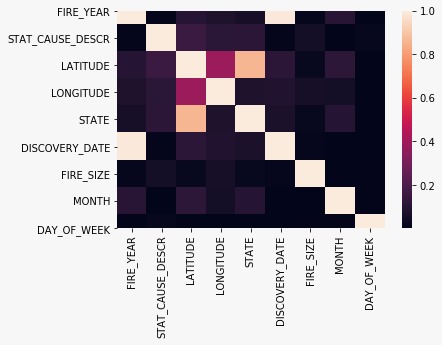
1. **Exploratory Data Analysis:**

We first examined the data to investigate the its structure and uncovered some interesting trends, patterns and relationships. Since we are only interested in west coast, we filtered our data to PST timezone. So, we did Exploratory Data Analysis and discovered some useful insights as mentioned below:

* A bar plot of different states and the number of wildfires showed that Texas and California are among the states with highest occurrences of wildfires.
* There is no correlation between the day of the week and the number of wildfires in case of lightning but we noticed that incase of Arson, it is more likely to occur during weekends.
* A bar plot of number of wildfires due to different causes showed that Lightning was the top reason among the naturally caused wildfires and Arson was the top reason among the man-made wildfire. We are specifically interested in Arson caused wildfire, since we cannot control natural wildfires but we can control manmade wildfires and Arson is intentional and malicious.
* Our analysis is focused on western US. A quick look at the trends of the western states showed that California is the most affected among the all the others.
* A summary of the data on fire\_size shows that Structure fires has caused biggest fires on an average, followed by Lightning among natural fires and Arson among man-made fires.
* The fires are more prominent during summer. Regardless of the cause of fires, we noticed that the fires peak during summers and taper down during rest of the months. We scaled the Arson plot as well to confirm this.
* We plotted a scatter plot of different kinds of fires across western US, by color coding each cause and the plot points are proportional to firesize. We observed that these plots are very clustered but not evenly spread across regions. These scatter plots also show that similar kind of fires were happening together as visualized on the US map.







1. **Random Forest Classification**:

We plotted heat map of correlation matrix between each variable and we observed that none of the variables are correlated with each other. Since there is no multicollinearity, we view this as a good sign of the data and we see no reason to further drop any other features. We believe all of these features are important for the prediction.

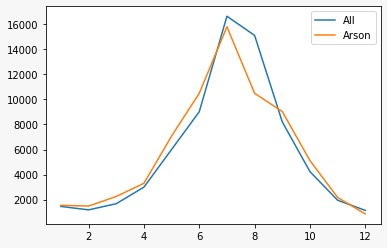
We label encoded the categorical features such as “state”, “day of the week” and the target variable (cause of fire). We trained a random forest model on this data. When we predicted using test data set, we got an accuracy of 51%. So, we merged the causes into different classes- Natural\_fire, accidental\_fire, malicious and others. This time the prediction accuracy of 66%. We plotted the confusion matrix of this classified data- prediction was best for others followed by natural and accidental but malicious showed worst performance.

Now we focused our prediction on only “Arson”. So, we classify all causes other than arson as “others” and arson is classified into “Arson” class. We again applied random forest and repeated the above procedure. This time we predicted Arson with 95% accuracy!

**Results and Conclusion** –

1. We trained a random forest model to predict if the given Wildfire is due to arson or not and we found the model to be 95.145% accurate.

Here, Latitude and Longitude were the most important features

1. It is observed that although there is no correlation of the number of Wildfires with any particular day of the week for all causes on an average except for ‘Arson’, which is dominant during weekends.
2. It can be seen that on an average, majority of the Wildfire occurrences takes place in the summer months in the US. Specifically rising from the month of April to the end of October.