Wildfires in the United States

1. Introduction

Wildfire is a phenomenon mostly caused by humans and exacerbated by nature. The goal of this project is to analyze over 20 years data for wildfires in the United States of America. The primary data used for this project is Kaggle's 1.88 Million US Wildfires. In addition, we will also use GHCN weather data curated by instructor Greg Baker, available on the course cluster gateway node at: `/home/ggbaker/public/ghcn-subset`. Our code is available on GitLab.

2. Goals

- Data processing and analysis.
    - Are there more or less wildfires overtime for the US and for a specific state? Are there bigger fires overtime?
    - What can we see about the correlations between the cause and wildfire occurrences and its scale?
- Machine Learning for prediction.
    - Using the data from previous months, train models to predict the number of fires and fire class for the subsequent months.

3. Data Preparation

The data from the Kaggle dataset was stored as a SQLite. To get the data prepared for our purposes, we used a SQLite cursor to get all the data from the database and stored them in a Pandas dataframe. This part was easy, but turned out there were some inconsistencies in the data, for example the description of the data mentioned that the dates are in YYYY-MM-DD format, but turned out they were in Georgian format, which we had to convert to standard Python datetime. There was also quite a lot of redundant or unrelated data which we had to drop and do conversions.

Once the data was prepared for processing, we used a GHCN stations dataset to find the closest GHCN to each fire. To do so, we had to apply the haversine formula for each row in our data, and calculate the distance between each fire and all the GHCN stations (around 60,000 stations). Obviously regular Pandas was not up to par for this task, so we used Dask to distribute the task between 16 threads. Next, with Spark, we used a combination of join and filter to extract the weather data on the cluster.

From the Kaggle wildfire data we were able to remove unnecessary columns and parsed necessary ones. Using the GHCN data, we located the nearest weather station and were able to extract the temperature and precipitation of the day the fire

happened as well as days leading up to the fire. The tools used were a combination of Pandas, Dask, and Spark.

By the end of data cleaning, we end up with the following data for each fire:

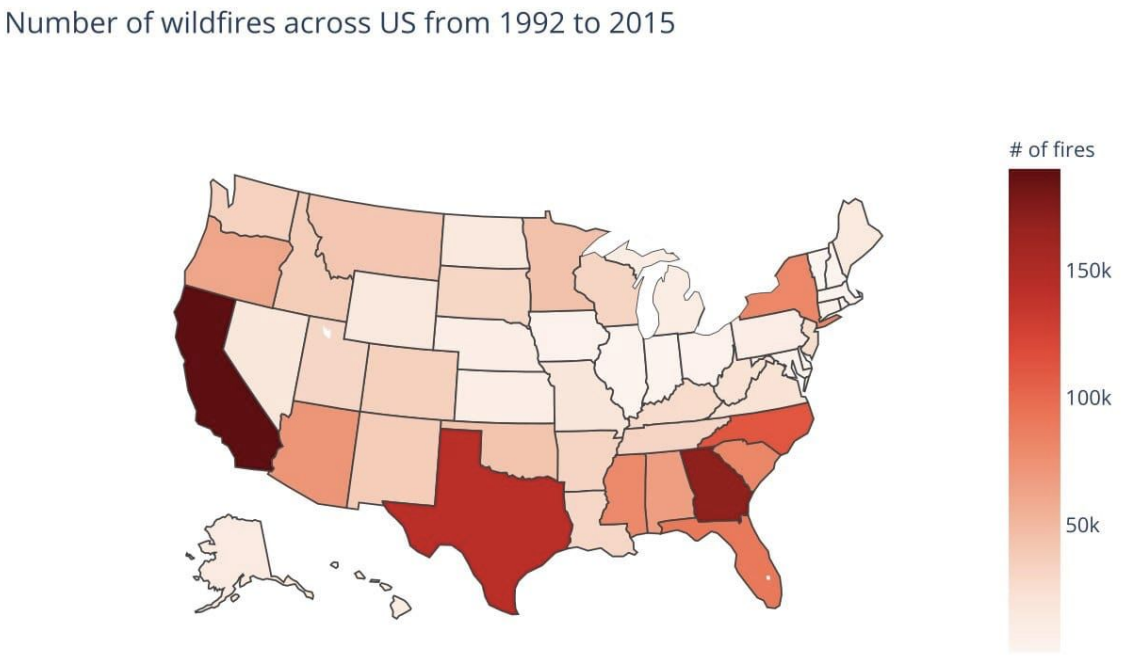
- Identifiers:
    - 'fpa_id'
    - 'nwcg_reporting_agency': unit preparing the fire report
    - 'nwcg_reporting_unit_name'
    - 'fire_name'
- Fire properties:
    - 'discovery_date'
    - 'discovery_time'
    - 'stat_cause_descr': cause of the fire
    - 'cont_date': containment date
    - 'cont_time': containment time
    - 'fire_size'
    - 'fire_size_class': A=greater than 0 but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, and G=5000+ acres). So A is the smallest fire and G is the biggest fire.
- Geography properties:
    - 'latitude'
    - 'longitude'
    - 'state'
    - 'county'
- Weather properties:
    - 'closest_weather_station'
    - 'collect_weather_date'
    - 'tmax'
    - 'prcp'

4. Data Analysis

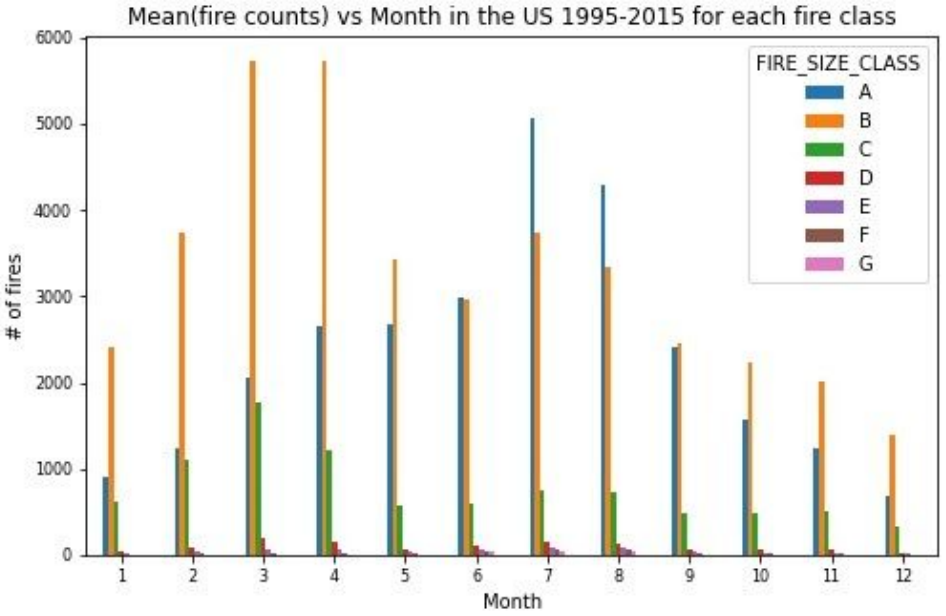
Originally, we planned to use matplotlib.pyplot for all of our visualizations, but creating a plot of number of fires for each state was pretty much impossible. Therefore, we resorted to Plotly to create a plot of the number of fires per state across the United States. Plotly provides an interactive visual of all fires, to get the visualization you can simply run `visualization.visualize_count()` and it would launch a new browser page of the plot.

For the rest of our visualizations we used the plain Pyplot (sometimes through Pandas' .plot() method). This was mostly done after grouping the data and aggregating the data.

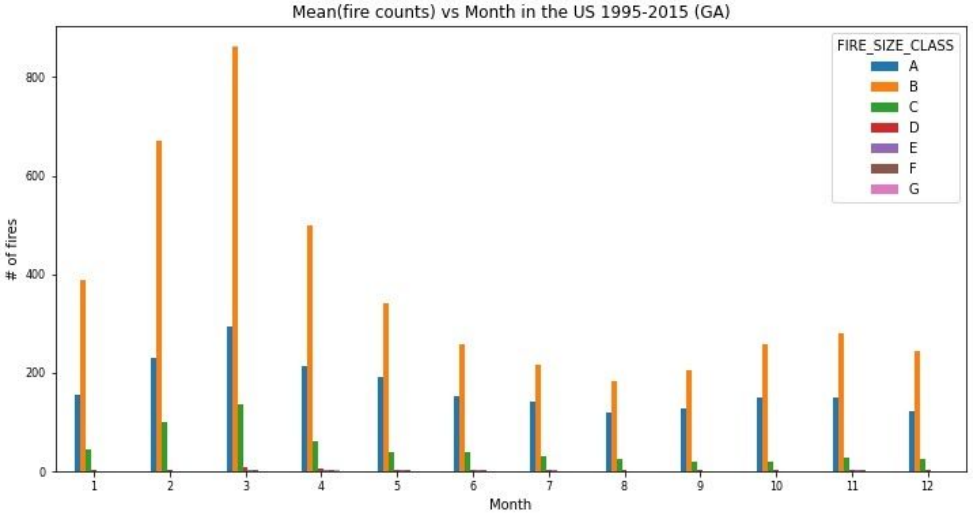
We start by looking at the overall picture. Mapping out the number of wildfires across the US below, we can see where most of the fires occur. The states in question are California, Texas and Georgia. California is notorious for its dry and hot weather and large areas of national parks that contribute to many grave fires. We will take a look at Texas and Georgia later.



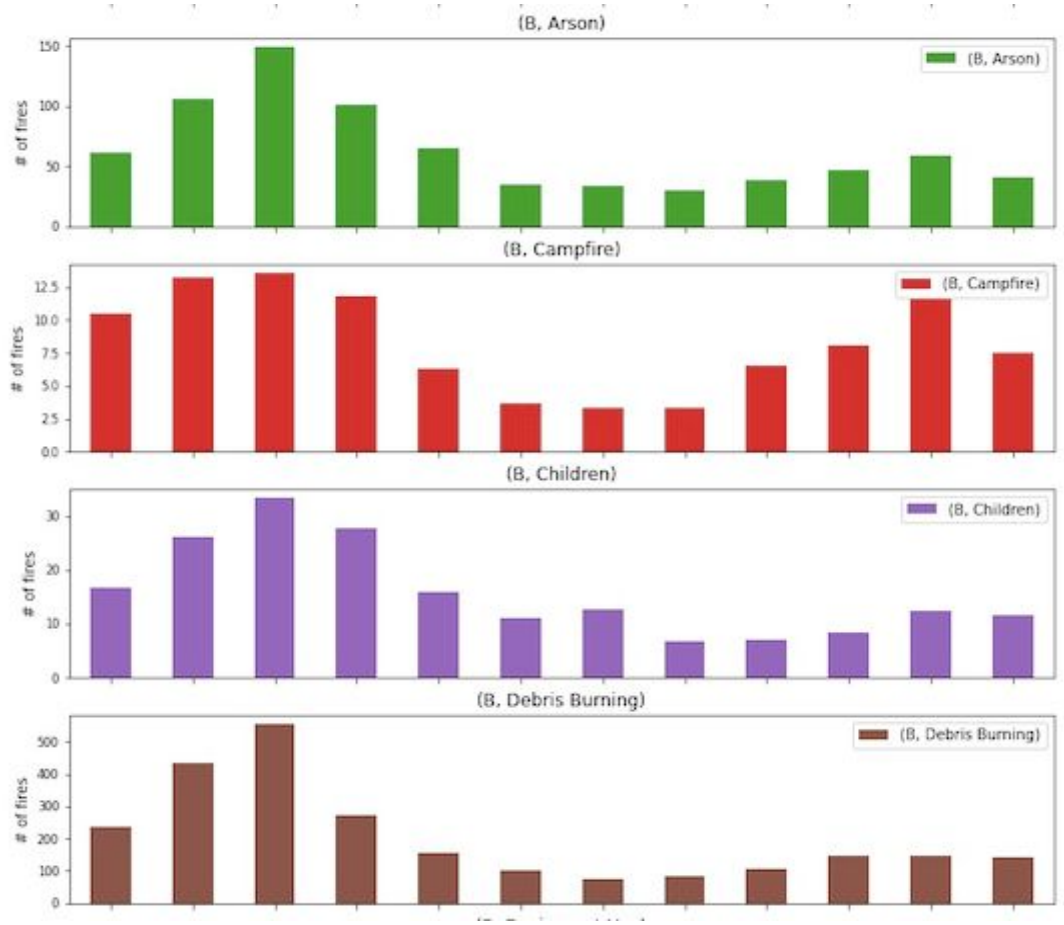
The graph below shows the mean number of wildfires in the US aggregated over the years spread across each month of the year, color coded with fire size class. From this graph we can infer that most fires are relatively small (class A and B) and occur most frequently during the months of March, April, and from June to August.



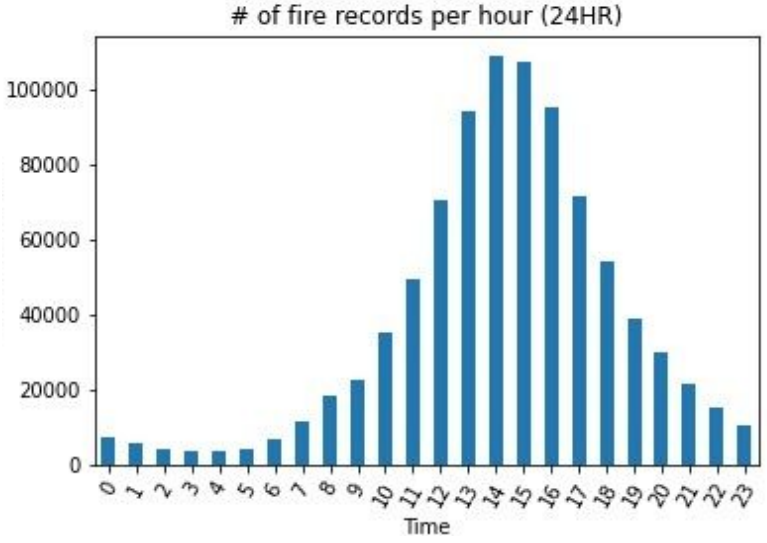
It makes sense that a large number congregates about the wildfire season in California (June to September). We dug up other states and found out that Georgia contributes to most of the B-class fires from February to April, as seen here.



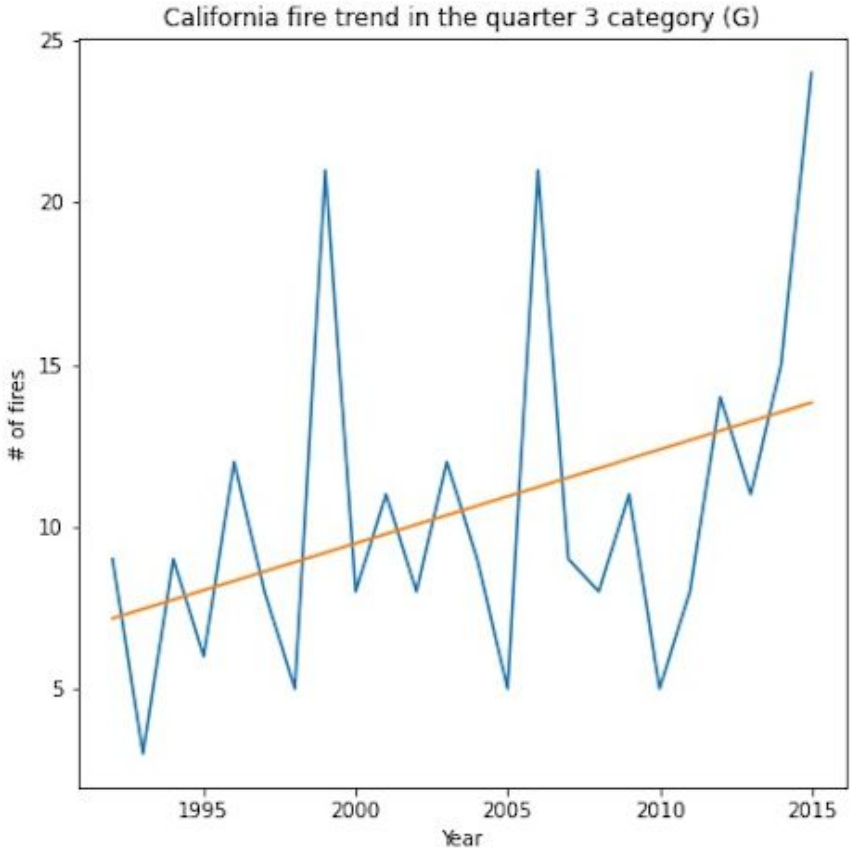
The causes of the fires are also curious. Though most of it can be attributed to debris burning, other various causes also occurred mostly within this time frame. Even though the fires are considered small, this information is still beneficial to the Georgia state fire department as they need to watch out during these months. Notably, children-caused fires can be induced although in a small number.



Another trend we noticed is the time of the day where most of the fire happens. We found out most fires were found around 2-3pm (presumably local time zone). This could be attributed to the higher temperature in the afternoon, causing the fire to spread quicker. As we saw in some cases, some of the biggest fires were caused as trivial as people smoking but the weather could very well make it worse.



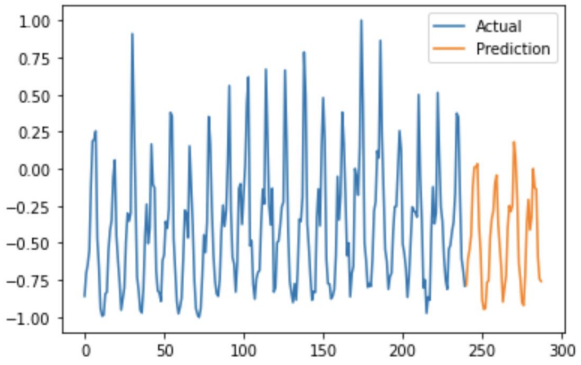
Next, using linear regression, we find a correlation between the number of large fires (G class) and its trend over the years. In general, extra large fires (5000+ acres) has an increasing trend nationally and state-wise. California has seen more and bigger fires in recent years during its wildfire season.



5. Prediction

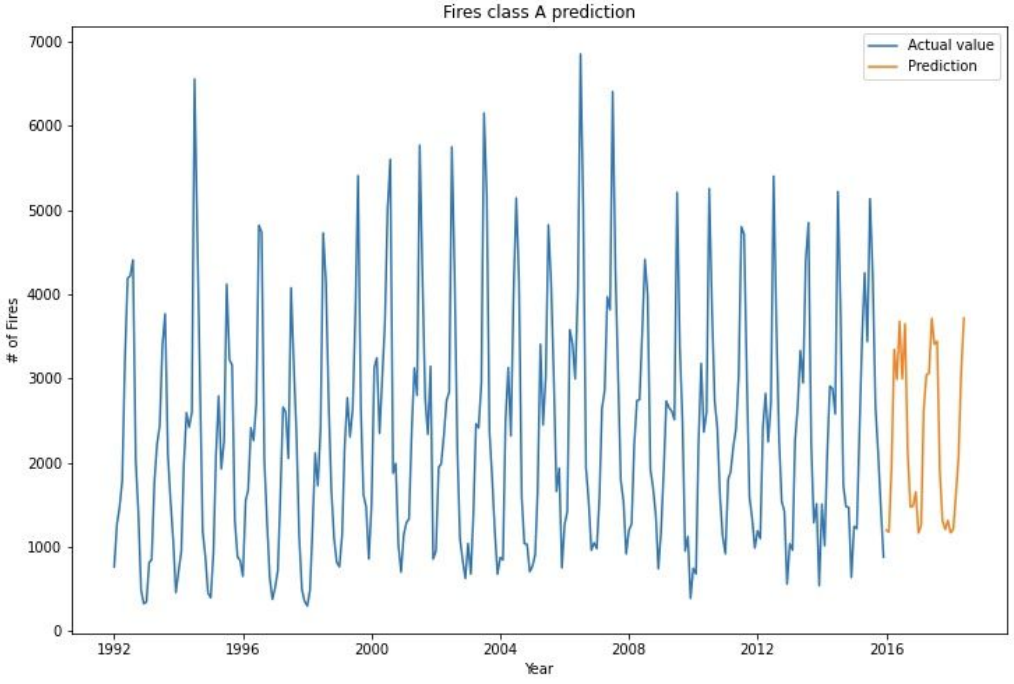
Fire Count Prediction:

For this section, we focused on predicting the number of different fire categories over the years. To do so, we tried different models; one of which was SARIMAX, which is a statistical model for predicting datasets that follow a trend and have seasonal aspects to them. While that sounds good in theory, it failed to produce any meaningful results. We then tried to use a LSTM model to create a forecasting model for our time-series data. The LSTM had some more meaningful data, but still failed to properly predict the trend for all but one fire size class.



The plot to the left, shows the models performance over the data it was trained on. While the prediction is not perfect and the predicted numbers fall short, it resembles the pattern of the data.

Below is our attempt to predict the number of fires for the 30 months ahead of what we had. This means predictions are made from 2016-01 to 2018-06. Both plots have the date as their x axis and number of fires on the y-axis (although the one on the left has it's fire counts normalized).



Model Discussion:
As mentioned above, we used LSTM neural networks as our time-series forecasting method. The reason we chose this architecture is because as the name implies (long short-term memory) saves some date as it moves forward. This is essential since the data we have is somewhat seasonal and follows a pattern, any models that store the prior patterns would be ideal.

To avoid overfitting, we had intermediary dropout layers, to encourage the model to not memorize the data but rather keep more important data.

The model takes in the data, and would predict 30 months ahead. We will discuss this further in the model architecture.

Model Architecture:
The model has 3 LSTM layers with each connected to a dropout layer, and all connected to a dense layer. The data takes in a 3 dimensional data set, (number of batches, x_size, y_size) and outputs a value. Since our training data is normalized, the output is normalized as well. We can use the same model that we used for normalizing the data (sklean's min max normalizer) to denormalize the data. For our training we fed n data points as our x value, and for our target we chose n+30 values. This means that the model should be capable of predicting the data of 30 months ahead based on what we currently have.

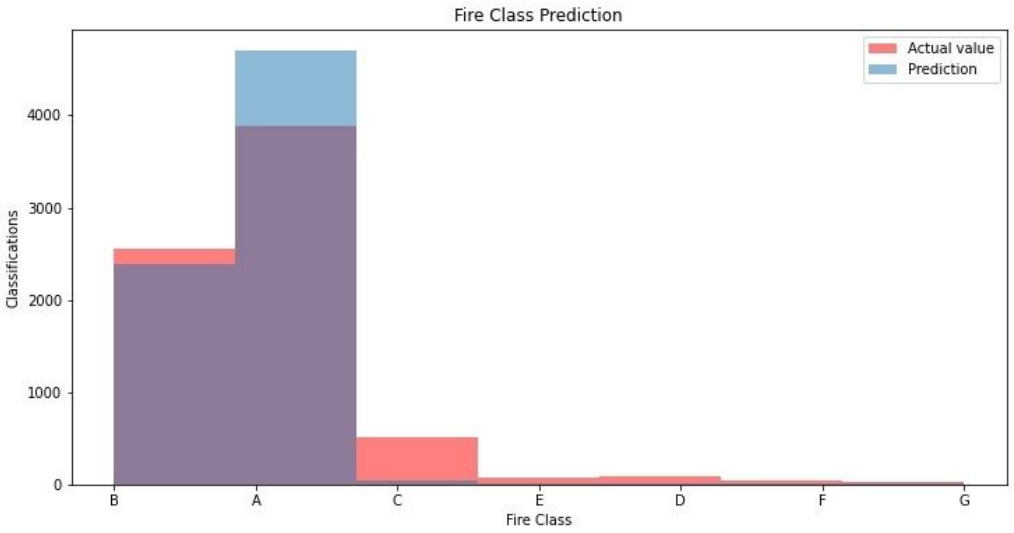
Fire Class Prediction:
Another model we tried to create was predicting the class of fire given time of the observation, month of observation, year of the observation, state, average of temperature 14 days prior to the observation and average of precipitation 14 days prior to the observation.

The reasoning for choosing these factors are as follows:

- Temperature and precipitation would give us the humidity/dryness of the environment (our thought process was the warmer and less precipitation -> dryer -> more fire).
- Latitude and Longitude would associate the fire with a climate.
- State also would give it a sense of climate but also how aggressive they would combat the fires (budget and population).
- Time of observation is hypothetically correlated with the response time.

Trying out multiple models Random Forest performed the best, but was not satisfactory. Out of all Random Forest algorithms, Gradient Boosting Random Forest performed the best but still was not good enough. After trying out different configurations the following configuration performed the best:

n_estimators = 400, max_depth=3 and min_samples_split=2. After fitting, the model had an accuracy of 66% on test data (80,20 split). Below is how it performed compared to the actual values.



We see that it failed to properly categorize C fire class and mostly predicted them as class A fires.

6. Limitations

Given more time we would have liked to collect more data such as the geographic purpose of the region where the fire happened, such as whether it was a campground, national park or a barn. We would also like to make other correlations to the economy of a state during large fire occurrences. For instance, California is often put in an emergency state during the fire season and it might be interesting to find out whether the consumption is reduced during those times.

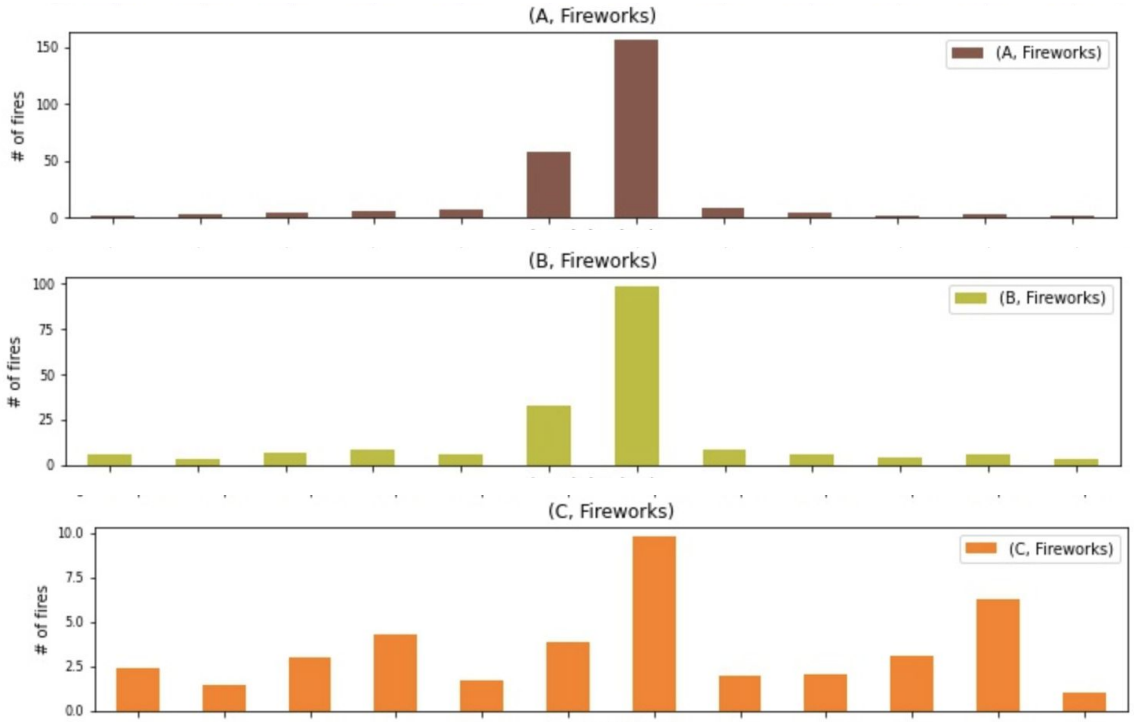
7. Project Experience

Phuong Tran: During this project, I was able to utilize my data collection and manipulation skills. Using Pandas and Spark, I cleaned, extracted, and joined fire and weather data on a distributed cluster. As a result, our group was able to use the data to create a machine learning model and perform predictions.

Parsa Alamzadeh: I focused mostly on data processing and visuals. I used Dask and Pandas for preprocessing, and scikit learn and Keras for the ML aspect of the project. I also used plotly and pyplot for the visuals.

8. Fun Fact!!!

The number one cause for small wildfires in the US from 1992 to 2015 during the month of July is Fireworks. We suspect national pride has something to do with it.
🇺🇸🇺🇸



9. Appendix

More visuals are available on the project's repository:
https://csil-git1.cs.surrey.sfu.ca/pmtran/353-forest-fires/-/tree/master/figures