This is a report of the work that I’ve done over the week analysing California’s wildfires from 2013-2020. This report will be broadly split into 4 parts:

1. Data Analysis & Question Formation
2. Predicting the Data – Technical Details
3. Data Visualization
4. Pitfalls & Future Scope

**Data Analysis**

The dataset that I had been provided contained data about California’s wildfires from 2013-2020. The attributes at hand are:

* Starting date (date\_start)
* Ending date (date\_end)
* Fire duration in days (fire\_duration)
* Amount of area burned down in acres (incident\_acres\_burned)
* Geospatial data (latitude and longitude)

The task at hand is to analyze the data to make a meaningful prediction pertinent to what wildfires are going to look like in the next couple of years. There are a couple of important observations here to begin with:

* There is a substantial increase in the number of fires over the years
* The duration of these fires has also increased over the years
* Fires have become more common in “off-season” that is, in winter and spring months

After some preliminary reading on the subject, I hypothesized a few reasons for each of the above observations.

* A simple way to address the number of fires increasing over the years would be to point to the individual factors that could usually be called ‘climate change’. So factors like increasing average temperature in the area, increased droughts and the overall water shortage leaving forests dry, and the increasing elongation of the summer/fall season all attribute to the increasing number of fires in the area.
* Fire duration is another interesting metric to consider, because I noticed that some fires around 2016 and 2017 lasted over 200 days, across multiple seasons and well into the next summer. This led me to investigate further, and upon some reading, I figure that these fires are going through somewhat of a “residual” effect. When wildfires are put out by human intervention, it allows trees in that area to grow denser. These trees then get engulfed by future wildfires, spreading faster and causing longer, more dangerous wildfires.
* As a result of increasing average temperature in California as well as the increased dryness, off-season fires are starting to pop up more, somewhat blurring the definition of “wildfire season”.

For this assignment, I took up fire duration as my metric to predict and visualize, and in that process, try to verify my residual theory hypothesis. The reason I picked this metric over others was two-fold. For one, I could use the data I had on hand directly, without having to, for instance, make a time series out of the given dates. It was also easier on the side of visualization because a line graph would be a good starting point.

**Predicting the Data**

The whole process of predicting the data is broadly divided into stages.

1. Cleaning up and pre-processing the data

To start, I plotted all the latitudes and longitudes on a map of California, originally to check for “hot” zones in California year on year and ended up coming across lat-long values that were erroneous, placed far away from California. Similarly, there were minor inconsistencies in a few starting and ending date values, notably around the 12/31 day, which made me think that the issue could be algorithmic rather than a typo.

Having removed those errors, I began pre-processing the data, mostly making sure the data is sorted, and the dates are in a proper format. I made a correlation plot to understand which attributes are going to add to the prediction and which ones weren’t, thus ruling out latitude and longitude from the data. I extracted the starting month and year from the dates and dropped the date values themselves from the dataset, so that the data could be viewed from a wider perspective.

Then, I binned the months largely into ‘season’ labels, to make note of which fires were in which season.

1. Creating the machine learning model

This is a regression problem, specifically one where we’re trying to extend a graph further in time using historical data. Given the kind of data we had at hand, I decided to use the Random Forest regressor to make my model. The Random Forest algorithm takes several decision trees using each data point to make a prediction, and the results of all these decision trees is then averaged out to produce a final result. This algorithm is especially useful when the input is multivariate, like in our case. The algorithm also counteracts overfitting and deals with outliers and missing data well, due to the way it works.

Using the other data points (Month, Year, Acres, and Season label), the model predicted the duration of the fire, and performed with an accuracy of 91.1% on the test dataset. I binned the same location and area data and ran a loop to make a dummy dataset for the years and months after. Do note that the other attributes of this future data aren’t very scientific, I’d rather consider it more like a testbench for the model to train on and give us it’s prediction on fire durations for the years after.

**Pitfalls & Future Scope**

I think that given the time, a time series exploration for this data to analyze the count of fires over the years is a beneficial metric that I didn’t spend enough time on. Another factor that I wanted to address but couldn’t fully get my head around in the span of this week was handling geospatial data. While exact location prediction is both hectic and challenging because of the statistical circularity of latitude and longitude data, I think that if we could bin the data into zones and run a classifier to figure out the zonal probability of wildfires, that would be a very interesting metric to explore as well as visualize.

A problem that I faced with the data was also that there were far too little data points to consider any form of aggregation like make a quarter/month, which in some ways, tied my hands a bit when it came to selecting which values I can reasonably predict while maintaining accuracy.

As for future scope, I think that there is more analysis to be done when it comes to seasonal data and answering the question regarding whether *off season* wildfires have increased. Adding to just month based seasonal tagging, splitting summer into peak and mild summer to create more variance, or accounting for periodic climactic patterns around the Pacific and its effect on the California weather and wildfires (El Niño and La Niña).

Data relating to population and population density could also add many interesting points for both data analysis and visualization. The increase in amount of dry trash that has collected in landfills and forests could attribute to increasing amounts of area burnt down. This would also tie well into the location data that we have at hand, factoring in the populations of the counties where the wildfire started.